

# IBM z17 (9175) Technical Guide

Ewerson Palacio

John Troy

Martin Packer

Martin Soellig

Martijn Raave

Kazuhiro Nakajima

Patrick Oughton

André Spahni

Priyal Shah

Mitchell Bride

Houda Achouri

Artem Minin

Lutz Kuehner

Markus Ertl

Octavian Lascu



**IBM Z**







**Note:** Before using this information and the product it supports, read the information in “Notices” on page 11.

**First Edition (April 2025)**

This edition applies to IBM z17 Model ME1, Machine Type 9175.

This document was created or updated on May 1, 2025.







# Contents

<b>Notices</b> .....	11
Trademarks .....	12
<b>Preface</b> .....	13
Authors .....	13
Now you can become a published author, too! .....	16
Comments welcome .....	16
Stay connected to IBM Redbooks .....	16
<b>Chapter 1. Introducing the IBM z17</b> .....	1
1.1 IBM z17 ME1 highlights .....	3
1.1.1 Supported upgrade paths .....	5
1.1.2 Capacity and performance .....	5
1.1.3 Supported operating systems .....	6
1.1.4 Supported IBM compilers .....	7
1.2 IBM z17 ME1 technical overview .....	7
1.2.1 Frames .....	7
1.2.2 CPC drawers .....	8
1.2.3 I/O subsystem and I/O drawers .....	9
1.2.4 Storage connectivity .....	9
1.2.5 Network connectivity .....	10
1.2.6 Clustering connectivity .....	12
1.2.7 Cryptography .....	12
1.2.8 Supported connectivity and crypto features .....	13
1.2.9 Special-purpose features and functions .....	14
1.3 Hardware management .....	14
1.4 Reliability, availability, and serviceability .....	15
<b>Chapter 2. Central processor complex hardware components</b> .....	19
2.1 Frames and configurations .....	20
2.1.1 IBM z17 cover (door) design .....	23
2.2 CPC drawer .....	26
2.2.1 CPC drawer interconnect topology .....	30
2.2.2 Oscillator (OSC) and Baseboard Management Controller (BMC) cards .....	31
2.2.3 System control .....	33
2.2.4 CPC drawer power .....	34
2.3 Dual chip modules .....	34
2.3.1 Processor unit chip .....	36
2.3.2 Processor unit (core) .....	37
2.3.3 PU characterization .....	38
2.3.4 Cache level structure .....	39
2.4 PCIe+ I/O drawer .....	40
2.5 Memory .....	44
2.5.1 Memory subsystem topology .....	46
2.5.2 Redundant array of independent memory (RAIM) .....	46
2.5.3 Memory configurations .....	47
2.5.4 Memory upgrades .....	52
2.5.5 Drawer replacement and memory .....	52
2.5.6 Virtual Flash Memory .....	52



2.5.7 Flexible Memory Option . . . . .	53
2.6 Reliability, availability, and serviceability. . . . .	55
2.6.1 RAS in the CPC memory subsystem . . . . .	56
2.6.2 General IBM z17 ME1 RAS features . . . . .	56
2.7 Connectivity. . . . .	58
2.7.1 Redundant I/O interconnect (RII) . . . . .	59
2.7.2 Enhanced drawer availability . . . . .	60
2.7.3 CPC drawer upgrade . . . . .	61
2.8 Model configurations . . . . .	61
2.8.1 Upgrades . . . . .	62
2.8.2 Model capacity identifier . . . . .	64
2.9 Power and cooling. . . . .	65
2.9.1 PDU-based configurations . . . . .	65
2.9.2 Power estimation tool . . . . .	67
2.9.3 Cooling . . . . .	68
2.9.4 Radiator Cooling Unit . . . . .	68
2.10 Summary. . . . .	70
<b>Chapter 3. Central processor complex design. . . . .</b>	<b>71</b>
3.1 Overview . . . . .	72
3.2 Design highlights. . . . .	73
3.2.1 Process shrink - from 7nm to 5nm . . . . .	75
3.3 CPC drawer design . . . . .	75
3.3.1 Cache levels and memory structure . . . . .	77
3.3.2 CPC drawer interconnect topology . . . . .	81
3.4 Processor unit design . . . . .	82
3.4.1 Simultaneous multithreading. . . . .	83
3.4.2 Single-instruction multiple-data . . . . .	84
3.4.3 Out-of-Order execution . . . . .	86
3.4.4 Superscalar processor . . . . .	89
3.4.5 On-chip coprocessors and accelerators . . . . .	90
3.4.6 IBM 2 <sup>nd</sup> Generation Integrated Accelerator for Artificial Intelligence . . . . .	94
3.4.7 IBM z17 DPU - Data Processing Unit . . . . .	97
3.4.8 Decimal floating point accelerator. . . . .	99
3.4.9 IEEE floating point . . . . .	100
3.4.10 Processor error detection and recovery . . . . .	100
3.4.11 Branch prediction . . . . .	101
3.4.12 Wild branch . . . . .	101
3.4.13 Translation lookaside buffer . . . . .	102
3.4.14 Instruction fetching, decoding, and grouping . . . . .	102
3.4.15 Extended Translation Facility . . . . .	103
3.4.16 Transactional Execution . . . . .	103
3.4.17 Runtime Instrumentation. . . . .	103
3.5 Processor unit functions . . . . .	103
3.5.1 Overview . . . . .	103
3.5.2 Central processors . . . . .	105
3.5.3 Integrated Facility for Linux (FC 1651) . . . . .	106
3.5.4 Internal Coupling Facility (FC 1652) . . . . .	106
3.5.5 IBM Z Integrated Information Processor (FC 1653) . . . . .	109
3.5.6 System assist processors . . . . .	114
3.5.7 Reserved processors . . . . .	114
3.5.8 Integrated firmware processors. . . . .	114
3.5.9 Processor unit assignment . . . . .	115



3.5.10	Sparing rules . . . . .	116
3.5.11	CPC drawer numbering . . . . .	116
3.6	Memory design . . . . .	117
3.6.1	Overview . . . . .	117
3.6.2	Main storage . . . . .	120
3.6.3	Hardware system area . . . . .	120
3.6.4	Virtual Flash Memory (FC 0644) . . . . .	121
3.7	Logical partitioning . . . . .	121
3.7.1	Overview . . . . .	121
3.7.2	Storage operations . . . . .	128
3.7.3	Reserved storage . . . . .	130
3.7.4	Logical partition storage granularity . . . . .	131
3.7.5	LPAR dynamic storage reconfiguration . . . . .	131
3.8	Intelligent Resource Director . . . . .	132
3.9	Clustering technology . . . . .	133
3.9.1	CF Control Code (CFCC) . . . . .	135
3.9.2	Coupling Thin Interrupts . . . . .	141
3.9.3	Dynamic CF dispatching . . . . .	142
3.10	Virtual Flash Memory . . . . .	142
3.10.1	IBM Z Virtual Flash Memory overview . . . . .	142
3.10.2	VFM feature . . . . .	143
3.10.3	VFM administration . . . . .	143
3.11	IBM Secure Service Container . . . . .	143
<b>Chapter 4.</b>	<b>Central processor complex I/O structure . . . . .</b>	<b>169</b>
4.1	Introduction to I/O infrastructure . . . . .	170
4.1.1	IBM z17 I/O infrastructure at a glance . . . . .	170
4.1.2	PCIe+ Generation 4 I/O Fanout - 2 Port . . . . .	171
4.2	I/O system overview . . . . .	172
4.2.1	Characteristics . . . . .	172
4.2.2	Supported I/O features . . . . .	174
4.3	PCIe+ I/O drawer . . . . .	175
4.3.1	PCIe+ I/O drawer offering . . . . .	177
4.4	CPC drawer fan-outs . . . . .	178
4.4.1	PCIe+ Gen4 fan-out (FC 0315) . . . . .	178
4.4.2	PCIe Gen3 and PCIe Gen4 differences . . . . .	179
4.4.3	Integrated Coupling Adapter - ICA SR2.0 (FC 0216) . . . . .	179
4.4.4	Fan-out considerations . . . . .	180
4.5	I/O features . . . . .	181
4.5.1	I/O feature card ordering information . . . . .	181
4.5.2	Physical channel ID (PCHID) report . . . . .	183
4.6	Connectivity . . . . .	184
4.6.1	I/O feature support and configuration rules . . . . .	184
4.6.2	Storage connectivity . . . . .	186
4.6.3	Network connectivity . . . . .	191
4.6.4	Parallel Sysplex connectivity . . . . .	199
4.7	Cryptographic functions . . . . .	204
4.7.1	<a href="#">CPACF functions (FC 3863) . . . . .</a>	<a href="#">204</a>
4.7.2	Crypto Express features . . . . .	204
4.7.3	IBM Fibre Channel Endpoint Security . . . . .	206
4.8	Integrated Firmware Processor . . . . .	208
<b>Chapter 5.</b>	<b>Central processor complex channel subsystem . . . . .</b>	<b>209</b>



5.1 Channel subsystem. . . . .	210
5.1.1 Multiple logical channel subsystems. . . . .	211
5.1.2 Multiple subchannel sets. . . . .	212
5.1.3 Channel path spanning. . . . .	216
5.2 I/O configuration management . . . . .	218
5.3 Channel subsystem summary. . . . .	219
5.4 IBM z17 Data Processing Unit (DPU). . . . .	219
<b>Chapter 6. Cryptographic features . . . . .</b>	<b>221</b>
6.1 Cryptography enhancements on IBM z17. . . . .	222
6.2 Cryptography overview . . . . .	223
6.2.1 Modern cryptography . . . . .	223
6.2.2 Kerckhoffs' principle . . . . .	224
6.2.3 Keys . . . . .	225
6.2.4 Algorithms. . . . .	226
6.3 Cryptography on IBM z17 . . . . .	227
6.4 CP Assist for Cryptographic Functions . . . . .	231
6.4.1 Cryptographic synchronous functions. . . . .	233
6.4.2 CPACF protected key . . . . .	234
6.5 Crypto Express8S . . . . .	236
6.5.1 Cryptographic asynchronous functions. . . . .	239
6.5.2 Crypto Express8S as a CCA coprocessor . . . . .	240
6.5.3 Crypto Express8S as an EP11 coprocessor. . . . .	247
6.5.4 Crypto Express8S as an accelerator. . . . .	248
6.5.5 Managing Crypto Express8S . . . . .	249
6.6 Trusted Key Entry workstation . . . . .	253
6.6.1 Logical partition, TKE host, and TKE target . . . . .	254
6.6.2 Optional smart card reader . . . . .	254
6.6.3 TKE hardware support and migration information. . . . .	254
6.7 Cryptographic functions comparison. . . . .	256
6.8 Cryptographic operating system support for IBM z17. . . . .	258
6.8.1 Crypto Express8S Toleration . . . . .	258
6.8.2 Crypto Express8S support of VFPE . . . . .	258
6.8.3 Crypto Express8S support of greater than 16 domains . . . . .	259
<b>Chapter 7. Operating systems support. . . . .</b>	<b>261</b>
7.1 Operating systems summary . . . . .	262
7.2 Support by operating system . . . . .	263
7.2.1 z/OS . . . . .	263
7.2.2 z/VM . . . . .	264
7.2.3 z/TPF . . . . .	266
7.2.4 VSEn. . . . .	266
7.2.5 21 <sup>st</sup> Century Software VSE <sup>n</sup> V6.3.1 . . . . .	266
7.2.6 Linux on IBM Z . . . . .	266
7.2.7 KVM hypervisor. . . . .	267
7.3 IBM z17 features and function support overview . . . . .	268
7.3.1 Supported CPC functions . . . . .	269
7.3.2 Coupling and clustering . . . . .	272
7.3.3 Storage connectivity . . . . .	272
7.3.4 Network connectivity. . . . .	275
7.3.5 Cryptographic functions . . . . .	279
7.4 Support by features and functions . . . . .	281
7.4.1 LPAR Configuration and Management. . . . .	281



7.4.2	Base CPC features and functions . . . . .	284
7.4.3	Coupling and clustering features and functions . . . . .	302
7.4.4	Storage connectivity-related features and functions . . . . .	308
7.4.5	Networking features and functions . . . . .	319
7.4.6	Cryptography Features and Functions Support . . . . .	330
7.5	z/OS migration considerations . . . . .	336
7.5.1	General guidelines . . . . .	336
7.5.2	Hardware Fix Categories . . . . .	337
7.5.3	z/OS V3.R1 . . . . .	339
7.5.4	z/OS V2.R5 . . . . .	340
7.5.5	z/OS V2.R4 . . . . .	340
7.5.6	Remote dynamic activation of I/O configurations for stand-alone Coupling Facilities, Linux on Z and z/TPF . . . . .	341
7.5.7	Coupling links . . . . .	341
7.5.8	z/OS XL C/C++ considerations . . . . .	342
7.6	z/VM migration considerations . . . . .	342
7.6.1	IBM z/VM 7.4 . . . . .	342
7.6.2	IBM z/VM 7.3 . . . . .	347
7.6.3	Capacity . . . . .	347
7.7	VSEn migration considerations . . . . .	347
7.8	Software licensing . . . . .	348
7.9	References . . . . .	350
<b>Chapter 8.</b>	<b>System upgrades . . . . .</b>	<b>353</b>
8.1	Introduction . . . . .	354
8.2	Permanent and Temporary Upgrades . . . . .	354
8.2.1	Overview . . . . .	354
8.2.2	CoD for IBM z17 systems-related terminology . . . . .	355
8.2.3	Concurrent and nondisruptive upgrades . . . . .	357
8.2.4	Permanent upgrades . . . . .	358
8.2.5	Temporary upgrades . . . . .	359
8.3	Concurrent upgrades . . . . .	360
8.3.1	PU Capacity feature upgrades . . . . .	360
8.3.2	Customer Initiated Upgrade facility . . . . .	362
8.3.3	Concurrent upgrade functions summary . . . . .	366
8.4	Miscellaneous equipment specification upgrades . . . . .	366
8.4.1	MES upgrade for processors . . . . .	367
8.4.2	MES upgrades for memory . . . . .	369
8.4.3	MES upgrades for I/O . . . . .	370
8.4.4	Feature on Demand . . . . .	371
8.4.5	Summary of plan-ahead feature . . . . .	371
8.5	Permanent upgrade by using the CIU facility . . . . .	372
8.5.1	Ordering . . . . .	373
8.5.2	Retrieval and activation . . . . .	374
8.6	On/Off Capacity on Demand . . . . .	376
8.6.1	Overview . . . . .	376
8.6.2	On/Off CoD testing . . . . .	377
8.6.3	Ordering . . . . .	377
8.6.4	Activation and deactivation . . . . .	381
8.6.5	Discontinuing and removing Capacity on Demand features . . . . .	382
8.7	z/OS Capacity Provisioning . . . . .	382
8.8	System Recovery Boost . . . . .	387
8.9	Capacity for Planned Event (CPE) . . . . .	388



8.10 Flexible Capacity for Cyber Resiliency . . . . .	388
8.11 Capacity Backup (CBU) . . . . .	390
8.11.1 Ordering . . . . .	390
8.11.2 CBU activation and deactivation . . . . .	392
8.11.3 Automatic CBU enablement for GDPS . . . . .	393
8.12 Planning for nondisruptive upgrades . . . . .	394
8.12.1 Components . . . . .	394
8.12.2 Concurrent upgrade considerations . . . . .	395
8.13 Summary of Capacity on-Demand offerings . . . . .	399
<b>Chapter 9. Reliability, availability, and serviceability . . . . .</b>	<b>401</b>
9.1 RAS strategy . . . . .	402
9.2 Technology . . . . .	402
9.2.1 Processor Unit chip . . . . .	402
9.2.2 Main memory . . . . .	405
9.2.3 I/O and service . . . . .	406
9.3 Structure . . . . .	407
9.4 Reducing complexity . . . . .	408
9.5 Reducing touches . . . . .	408
9.6 IBM z17 availability characteristics . . . . .	408
9.7 IBM z17 RAS functions . . . . .	412
9.7.1 Scheduled outages . . . . .	413
9.7.2 Unscheduled outages . . . . .	415
9.8 Enhanced drawer availability . . . . .	416
9.8.1 EDA planning considerations . . . . .	417
9.8.2 Enhanced Drawer Availability processing . . . . .	419
9.9 Concurrent Driver Maintenance . . . . .	424
9.9.1 Resource Group and native PCIe features MCLs . . . . .	425
9.10 RAS capability for the HMA and SE . . . . .	426
<b>Chapter 10. Hardware Management Console and Support Element . . . . .</b>	<b>429</b>
10.1 Introduction . . . . .	430
10.2 HMC and SE changes and new features . . . . .	430
10.2.1 Hardware Management Appliance . . . . .	443
10.2.2 HMC and SE server . . . . .	444
10.2.3 USB support for HMC and SE . . . . .	445
10.2.4 SE driver and version support with the HMC Driver 61/Version 2.17.0 . . . . .	446
10.3 HMC and SE connectivity . . . . .	446
10.3.1 Hardware Management Appliance (HMA) connectivity . . . . .	447
10.3.2 Support Element (SE) connectivity . . . . .	447
10.3.3 Network planning for the HMC and SE . . . . .	449
10.3.4 Hardware considerations . . . . .	450
10.3.5 TCP/IP Version 6 on the HMC and SE . . . . .	450
10.3.6 Assigning TCP/IP addresses to the HMC, SE and ETS . . . . .	450
10.3.7 HMC multi-factor authentication . . . . .	451
10.4 Remote Support Facility . . . . .	452
10.4.1 Security characteristics . . . . .	452
10.4.2 RSF connections to IBM and Enhanced IBM Service Support System . . . . .	453
10.5 HMC and SE capabilities . . . . .	453
10.5.1 Central processor complex management . . . . .	454
10.5.2 LPAR management . . . . .	454
10.5.3 HMC and SE remote operations . . . . .	455
10.5.4 Operating system communication . . . . .	457



10.5.5 Monitoring . . . . .	458
10.5.6 Capacity on-demand support . . . . .	460
10.5.7 Server Time Protocol support . . . . .	460
10.5.8 Security and user ID management . . . . .	462
10.5.9 Automated operations via APIs . . . . .	464
10.5.10 Cryptographic support . . . . .	464
10.5.11 Installation support for z/VM that uses the HMC . . . . .	465
10.5.12 Dynamic Partition Manager . . . . .	465
10.6 HMC,SE, and CPC microcode . . . . .	466
10.6.1 Remote Code Load (RCL) . . . . .	468
<b>Chapter 11. Environmental requirements . . . . .</b>	<b>471</b>
11.1 Introduction . . . . .	472
11.2 Power and cooling . . . . .	472
11.2.1 Intelligent Power Distribution Unit . . . . .	472
11.2.2 Cooling requirements . . . . .	476
11.3 Physical specifications . . . . .	478
11.4 Physical planning . . . . .	479
11.4.1 Top Exit Cabling without Tophat (FC 7803) . . . . .	480
11.4.2 Bottom Exit Cabling feature (FC 7804) . . . . .	481
11.4.3 Top Exit Enclosure feature (FC 5823) . . . . .	481
11.4.4 Frame Bolt-down kit . . . . .	483
11.4.5 Service clearance areas . . . . .	483
11.5 Energy management . . . . .	484
11.5.1 Environmental monitoring . . . . .	485
<b>Chapter 12. Performance and capacity planning . . . . .</b>	<b>489</b>
12.1 IBM z17 performance characteristics . . . . .	490
12.1.1 IBM z17 single-thread capacity . . . . .	490
12.1.2 IBM z17 SMT capacity . . . . .	491
12.1.3 IBM Integrated Accelerator for zEnterprise Data Compression (zEDC) . . . . .	491
12.1.4 Primary performance improvement drivers with IBM z17 . . . . .	491
12.2 IBM z17 Large System Performance Reference ratio . . . . .	492
12.2.1 LSPR workload suite . . . . .	493
12.3 Fundamental components of workload performance . . . . .	493
12.3.1 Instruction path length . . . . .	493
12.3.2 Instruction complexity . . . . .	494
12.3.3 Memory hierarchy and memory nest . . . . .	494
12.4 Relative Nest Intensity . . . . .	495
12.5 LSPR workload categories based on L1MP and RNI . . . . .	497
12.6 Relating production workloads to LSPR workloads . . . . .	497
12.7 CPU MF counter data and LSPR workload type . . . . .	498
12.8 Workload performance variation . . . . .	499
12.9 Capacity planning considerations for IBM z17 . . . . .	500
12.9.1 Collect CPU MF counter data . . . . .	500
12.9.2 Creating EDF files with CP3KEXTR . . . . .	500
12.9.3 Loading EDF files to the capacity planning tool . . . . .	501
12.9.4 IBM z17 Performance Best Practices . . . . .	502
12.9.5 IBM zPCR HiperDispatch Report . . . . .	503
12.9.6 IBM zPCR Topology Report . . . . .	506
12.9.7 IBM z17 HMC - View Partition Resource Assignments . . . . .	508
12.9.8 IBM zPCR Large Partition Support . . . . .	509
<b>Appendix A. IBM Z Integrated Accelerator for AI and IBM Spyre AI Accelerator . . . . .</b>	<b>515</b>



A.1 Overview . . . . .	516
A.2 NNPA and IBM z16 Hardware . . . . .	517
A.3 How to use IBM Z Integrated AI Accelerator in your enterprise . . . . .	518
A.4 Next Generation Artificial Intelligence Unit - AIU . . . . .	519
A.4.1 Spyre Accelerator . . . . .	520
A.4.2 Planning for the IBM Spyre Accelerator Cards . . . . .	522
<b>Appendix B. IBM Integrated Accelerator for zEnterprise Data Compression . . . . .</b>	<b>525</b>
B.1 Client value of IBM Z compression . . . . .	525
B.2 IBM z17 IBM Integrated Accelerator for zEDC . . . . .	526
B.2.1 Compression modes . . . . .	527
B.3 IBM z17 migration considerations . . . . .	527
B.3.1 All z/OS configurations stay the same . . . . .	527
B.3.2 Consider fail-over and disaster recovery sizing . . . . .	527
B.3.3 Performance metrics . . . . .	527
B.3.4 zEDC to IBM z17 zlib Program Flow for z/OS . . . . .	527
B.4 Software support . . . . .	528
B.4.1 IBM Z Batch Network Analyzer . . . . .	528
B.5 Compression acceleration and Linux on IBM Z . . . . .	529
<b>Appendix C. Tailored Fit Pricing and IBM Z Flexible Capacity for Cyber Resiliency . . . . .</b>	<b>531</b>
. . . . .	532
C.1 Tailored Fit Pricing . . . . .	532
C.2 Software Consumption Model . . . . .	533
C.2.1 International Program License Agreement in the Software Consumption Model . . . . .	534
C.3 Hardware Consumption Model . . . . .	534
C.3.1 Tailored Fit Pricing for IBM Z hardware in more detail . . . . .	535
C.3.2 Efficiency benefits of TFP-Hardware . . . . .	536
C.4 Conclusion . . . . .	537
C.5 IBM Z Flexible Capacity for Cyber Resiliency . . . . .	537
C.6 Use cases of IBM Flexible Capacity for Cyber Resiliency . . . . .	538
C.6.1 Disaster recovery and DR testing . . . . .	538
C.6.2 Frictionless compliance . . . . .	538
C.6.3 Facility maintenance . . . . .	538
C.6.4 Pro-active avoidance . . . . .	538
C.7 How does IBM Flexible Capacity for Cyber Resiliency work? . . . . .	539
C.7.1 Set up process . . . . .	539
C.7.2 Transferring workloads . . . . .	540
C.7.3 Multi-system environment . . . . .	541
C.8 Tailored fit pricing for hardware and IBM Z Flexible Capacity for Cyber Resiliency . . . . .	542
C.9 Ordering and installing IBM Z Flexible Capacity for Cyber Resiliency . . . . .	543
C.10 Terms and conditions of IBM Z Flexible Capacity for Cyber Resiliency . . . . .	544
C.11 IBM Z Flexible Capacity for Cyber Resiliency versus Capacity Back Up . . . . .	545
<b>Appendix D. Channel options . . . . .</b>	<b>547</b>
D.1 Available Channel options . . . . .	548
<b>Appendix E. Frame configurations with Power Distribution Units . . . . .</b>	<b>551</b>
. . . . .	551
E.1 Power Distribution Unit configurations . . . . .	551
E.2 New PCIe+ I/O drawers / PCHIDs location Options . . . . .	555
<b>Appendix F. Sustainability . . . . .</b>	<b>557</b>
F.1 Sustainability Improvements . . . . .	557



F.2 Improved Performance Per Kilowatt . . . . .	557
F.3 Reduced Shipping Impacts, Floor Space Savings, And Simplification . . . . .	558
F.4 Sustainability Instrumentation . . . . .	558
F.4.1 z/OS . . . . .	558
F.4.2 z/VM . . . . .	560
F.4.3 Linux on Z . . . . .	560
<b>Related publications</b> . . . . .	561
IBM Redbooks . . . . .	561
Other publications . . . . .	561
Online resources . . . . .	561
Help from IBM . . . . .	562







# Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US*

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and customer examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

## COPYRIGHT LICENSE:


This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.



## Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

CICS®	IBM Z®	System z®
Connect:Direct®	IBM z Systems®	System z10®
DB2®	IBM z13®	System z9®
Db2®	IBM z13s®	VTAM®
DS8000®	IBM z14®	WebSphere®
FICON®	IBM z16®	z Systems®
FlashCopy®	IBM z17™	z/Architecture®
GDPS®	Interconnect®	z/OS®
Guardium®	Language Environment®	z/VM®
HyperSwap®	OMEGAMON®	z/VSE®
IBM®	Parallel Sysplex®	z13®
IBM Blockchain®	Passport Advantage®	z13s®
IBM Cloud®	PIN®	z15®
IBM Security®	RACF®	z16™
IBM Spyre™	Redbooks®	z9®
IBM Sterling®	Redbooks (logo)  ®	zEnterprise®
IBM Telum®	Resource Link®	zSystems™
IBM Watson®	Sterling™	

The following terms are trademarks of other companies:

Evolution, are trademarks or registered trademarks of Kenexa, an IBM Company.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.



# Preface

This IBM® Redbooks® publication describes the features and functions of the latest member of the IBM Z® platform that was built with the IBM Telum® II processor: the IBM z17™ (machine type 9175). It includes information about the IBM z17 processor design, I/O innovations, security features, and supported operating systems.

The IBM Z platform is recognized for its security, resiliency, performance, and scale. It is relied on for mission-critical workloads and as an essential element of hybrid cloud infrastructures. The IBM z17 server adds capabilities and value with innovative technologies that are needed to accelerate the digital transformation journey.

The IBM z17 is a state-of-the-art data and transaction system that delivers advanced capabilities, which are vital to any digital transformation. The IBM z17 is designed for enhanced modularity, which is in an industry standard footprint.

This system excels at the following tasks:

- ▶ Providing AI inference with Integrated Accelerator for Artificial Intelligence
- ▶ Making use of multicloud integration services
- ▶ Securing data with pervasive encryption
- ▶ Accelerating digital transformation with agile service delivery
- ▶ Transforming a transactional platform into a data powerhouse
- ▶ Getting more out of the platform with IT Operational Analytics
- ▶ Accelerating digital transformation with agile service delivery
- ▶ Revolutionizing business processes
- ▶ Blending open source and IBM Z technologies

This book explains how this system uses new innovations and traditional IBM Z strengths to satisfy growing demand for cloud, analytics, and open source technologies. With the IBM z17 as the base, applications can run in a trusted, reliable, and secure environment that improves operations and lessens business risk.

## Authors

This book was produced by a working at IBM Redbooks, Poughkeepsie Center.

**Ewerson Palacio** is an IBM Redbooks Project Leader. He holds Bachelor's degree in Math and Computer Science. Ewerson worked for IBM Brazil for over 40 years and retired in 2017 as an IBM Distinguished Engineer. Ewerson co-authored many IBM Z Redbooks, and created and presented ITSO seminars around the globe.

**John Troy** is an IBM Z and storage hardware National Top Gun in the northeast area of the US. He has over 40 years of experience in the service field. His areas of expertise include IBM Z servers and high-end storage systems technical and customer support and services. John has also been an IBM Z hardware technical support course designer, developer, and instructor for the last eight generations of IBM high-end servers.

**Martijn Raave** is an IBM Z and LinuxONE Client Architect and Hardware Technical Specialist for IBM Northern Europe. Over a period of 27 years, his professional career has revolved around the mainframe platform, supporting several large Dutch customers in their technical



and strategic journey on IBM Z. His focus areas are hardware, resiliency, availability, architecture, but he is basically interested in any IBM Z related topic.

**Kazuhiro Nakajima** is a Senior IT Specialist at IBM Japan. He has a 35-year career at IBM Japan and has been an advanced Subject Matter Expert on IBM Z products for over 20 years. His areas of expertise include IBM Z hardware, performance, z/OS®, and IBM Z connectivity. He has co-authored several IBM Z Redbooks publications, from the IBM zEC12 to the IBM z16®.

**Martin Packer** is a mainframe performance and capacity specialist, with a penchant for SMF data analysis. He has worked with many customers around the world, having almost 40 years of mainframe experience. He has blogged, cast pods, and presented at conferences extensively. His first degree is in Mathematics and Physics, his second in Electronics and Computing.

**Priyal Sha** is a HW and Compilers Performance Analyst for IBM Z with over 20 years of experience and 14 years of experience working on compilers performance on IBM Z. Her niche area and focus is on Hardware-Software synergy and ensuring most optimal performance value of the IBM Z platform by influencing design choices in Hardware and Software. She enjoys this rare opportunity her role offers as well as the opportunity to connect and collaborate with top technical experts across the stack.

**Octavian Lascu** is a Senior IT Infrastructure Architect with Inter Computer, Romania. He has over 35 years of IT experience with IBM Z, IBM Power, and IBM Storage. Octavian was an IBM Redbooks Project Leader for over 20 years, where he has co-authored many IBM Redbooks publications covering IBM Z hardware, IBM Power, and various IBM solutions.

**Martin Soellig** is an IBM Z Technical Specialist in Germany. He has 34 years of experience working in the IBM Z field. He holds a degree in mathematics from the University of Hamburg. His areas of expertise include IBM z/OS and IBM zSystems™ hardware, specifically in Parallel Sysplex® and GDPS® environments, and also in cryptography on IBM Z.

**Pat Oughton** is an IBM Z Brand Technical Specialist in New Zealand. He joined IBM in 2015 after working as a z/OS Systems Programmer for 30 years. His areas of expertise include IBM Z installation (hardware and operating system) and IBM Parallel Sysplex Implementation. He has written three other IBM Redbooks publications.

**André Spahni** is an IBM Z Brand Technical Specialist based in Zurich, Switzerland. He has over 22 years of experience working with and supporting IBM Z clients. André has worked for EMEA 2nd level supporter and national Top Gun. His areas of expertise include IBM Z hardware, HMC/SE, and connectivity.

**Artem Minin** is currently a Technical Specialist in IBM's Washington Systems Center, a team of Subject Matter Experts that provide leading edge technical sales assistance for the design, implementation, and support of solutions that leverage IBM Z. Specifically, Artem is responsible for supporting IBM Z Data and AI solutions through PoCs, custom demos, and client workshops. He focuses on supporting client engagements which leverage open-source AI software, Machine Learning for z/OS, Cloud Pak for Data, Data Gate, DVM for z/OS, and SQL Data Insights.

**Mitchell Bride** serves as a Technical Enablement Specialist at IBM's Washington Systems Center, a premier team of subject matter experts dedicated to delivering cutting-edge technical pre-sales support and acting as consultative, trusted advisors to IBM Z clients. Since joining IBM in 2021, Mitchell has focused on advancing his expertise in IBM Z hardware, enabling organizations to maximize performance, scalability, and reliability in their mission-critical systems.



**Lutz Kuehner** is a Senior z/OS System Engineer at UBS AG in Switzerland. He has 39 years of experience in IBM z Systems®. Lutz has worked for 10 years in the IBM Z Presales support in Germany, developing and providing professional services for customers in the financial market. In addition to co-authoring several IBM Redbooks publications since 2001, he has been an official ITSO presenter at ITSO workshops.

**Markus Ertl** is a Senior IT Specialist in IBM Germany. He has more than 20 years of experience with IBM Z, working with clients from various industries. His area of expertise includes IBM Z hardware and infrastructure, performance, and Capacity on Demand topics.

**Houda Achouri** is a Senior IBM Z Technical Manager and IBM Z Technical Leader for UKI based in the UK. She has 10 years of IBM Z Hardware experience spanning 5 generations of IBM Z systems working on a diverse portfolio of clients across industries, including major financial institutions and retailers. Houda has a BSc and MSc in Mathematics and Computer Science from the University of Manchester.

**Thanks to the following people for their contributions to this project:**

John Campbell  
**IBM Z Washington Systems Center**

Robert Haimowitz  
Patrik Hysky  
**IBM Redbooks, Poughkeepsie Center**

Dave Surman, David Hollar, Michael Groetzner, Kyle Giesen, Seth Lederer, Patty Driever, Jeannie Kraus, John Torok, Brian Valentine, Patrick McKeone, Marna Walle, Leon Manten, Dalibor Kurek, Ron Geiger, Richard Gagnon, Les Geer III, Chris Filacheck, Nicole Rae, Yamil Rivera  
**IBM Poughkeepsie**



## Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

[ibm.com/redbooks/residencies.html](http://ibm.com/redbooks/residencies.html)

## Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- ▶ Use the online **Contact us** review Redbooks form found at:

[ibm.com/redbooks](http://ibm.com/redbooks)

- ▶ Send your comments in an email to:

[redbooks@us.ibm.com](mailto:redbooks@us.ibm.com)

- ▶ Mail your comments to:

IBM Corporation, IBM Redbooks  
Dept. HYTD Mail Station P099  
2455 South Road  
Poughkeepsie, NY 12601-5400

## Stay connected to IBM Redbooks

- ▶ Look for us on LinkedIn:

<http://www.linkedin.com/groups?home=&gid=2130806>

- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>

- ▶ Stay current on recent Redbooks publications with RSS Feeds:

<http://www.redbooks.ibm.com/rss.html>





# Introducing the IBM z17

The IBM Z platform is recognized for its long-standing commitment to delivering best-in-industry security, resiliency, performance, and scalability. IBM Z is relied on for mission-critical workloads and as an essential element of hybrid cloud infrastructures.

The new member of the IBM Z family, IBM z17, continues that commitment and adds value with innovative technologies that can help accelerate the digital transformation journey.

The IBM z17 system is built with the IBM Telum II processor<sup>1</sup>, which was introduced at the 2024 Hotchips Conference in August 2024. Hotchips is one of the semiconductor industry's leading conferences on high-performance microprocessors and related integrated circuits.

Along side the IBM Telum II processor IBM also announced the IBM Spyre™ Accelerator<sup>2</sup>. A purpose-built enterprise-grade accelerator offering scalable capabilities for complex AI models and generative AI use cases is being showcased. It features up to 1TB of memory, built to work in tandem across the eight cards of a regular IO drawer, to support AI model workloads across the mainframe while designed to consume no more than 75W per card. Each chip will have 32 compute cores supporting int4, int8, fp8, and fp16 datatypes for both low-latency and high-throughput AI applications. For more information on the IBM Spyre Accelerator see: Appendix A, "IBM Z Integrated Accelerator for AI and IBM Spyre AI Accelerator" on page 515.

The IBM z17 is designed to help businesses meet the following goals:

- ▶ Create value in every interaction and to optimize decision making, with the on-chip Artificial Intelligence (AI) accelerator. The Accelerator for AI is engineered for AI at scale and allows you to Integrate AI with your mission critical transactions to accelerate insights with near zero latency while ensuring data privacy and system availability.
- ▶ Act now to protect today's data against current and future threats with quantum-safe protection immediately through quantum-safe cryptography APIs and crypto discovery tools. Use AI for early detection of threats and simplify compliance while accelerating your Quantum-safe journey.

<sup>1</sup> IBM Telum II Processor: the next Telum generation microprocessor for IBM Z and IBM LinuxONE

<sup>2</sup> Statement of general direction: The IBM Spyre AI Accelerator is planned to be available starting in 4Q 2025, in accordance with applicable import/export guidelines.



- ▶ Enhance resiliency with flexible capacity to dynamically shift system resources across locations to proactively avoid disruptions.
- ▶ Modernize and integrate applications and data in a hybrid cloud environment with consistent and flexible deployment options to innovate with speed and agility.
- ▶ Reduce cost and keep up with changing regulations through a solution that helps simplify and streamline compliance tasks.

This chapter describes the basic characteristics of the IBM z17 platform. It includes the following topics:

- ▶ 1.1, “IBM z17 ME1 highlights” on page 3
- ▶ 1.2, “IBM z17 ME1 technical overview” on page 7
- ▶ 1.3, “Hardware management” on page 14
- ▶ 1.4, “Reliability, availability, and serviceability” on page 15



## 1.1 IBM z17 ME1 highlights

Each new IBM Z platform continues to deliver innovative technologies. The IBM z17 is no exception. It has a new processor chip design with each processor unit (PU) running at 5.5 GHz. IBM 17 provides 12% - 20% more processor capacity per CPC drawer compared to the IBM z16.

The new processor chip design has a enhanced cache hierarchy as introduced with the IBM z16, on-chip AI accelerator that is shared by the PU cores, transparent memory encryption, and increased uniprocessor capacity (single thread and SMT similar). Each PU chip will also include a new single Data Processing Unit (DPU), a dedicated core for I/O operations.

The on-chip AI scoring logic provides submicrosecond AI inferencing for deep learning and complex neural network models.

The enhanced cache structure features the following cache sizes:

- ▶ 256 KB L1 per PU core
- ▶ 36 MB semi-private L2 per PU core
- ▶ 360 MB (logical) shared victim virtual L3 per chip
- ▶ 2.88 GB (logical) shared victim virtual L4 per CPC drawer

The result is improved system performance and scalability with 12.5% larger L3, 40.6% larger vL3 & vL4 more cache capacity per core over the IBM z16 and reduced average access latency through a flatter topology.

The IBM z17 (machine type 9175) has one model: the ME1. The maximum number of characterizable processor units (PUs) with the IBM z17 is represented by feature names: Max43, Max90, Max136, Max183, and Max208.

The number of characterizable PUs, spare PUs, System Assist Processors (SAPs), and Integrated Firmware Processors (IFP) are included with each IBM z17 feature (see Table 1-1).

*Table 1-1 IBM z17 ME1 processor unit (PU) configurations*

Feature name	Number of CPC drawers	Feature Code	Characterizable PUs	Standard SAPs <sup>a</sup>	Spare PUs
<b>Max43</b>	1	0571	1 - 43	5	2
<b>Max90</b>	2	0572	1 - 90	10	2
<b>Max136</b>	3	0573	1 - 136	16	2
<b>Max183</b>	4	0574	1 - 183	21	2
<b>Max208</b>	4	0575	1 - 208	24	2

a. Optional SAPs are not supported on the MT 9175.

The IBM z17 memory subsystem uses proven redundant array of independent memory (RAIM) technology to ensure high availability. Up to 64 TB (16 TB per CPC drawer) of addressable memory per system can be ordered.

The IBM z17 also has unprecedented capacity to meet consolidation needs with innovative I/O features for transactional and hybrid cloud environments.



The IBM z17 (maximum configuration) can support up to 12 PCIe+ I/O drawers. Each I/O drawer can support up to 16 I/O or special purpose features for storage, network, clustering connectivity, and cryptography.

The following features were introduced with the IBM z17:

- ▶ Network Express
  - New Hardware for OSA, RoCE, and Coupling LR
- ▶ zHyperLink 2.0
- ▶ ICA SR 2.0
- ▶ Coupling Express3 Long Reach

The IBM z17 is more flexible and features simplified on-demand capacity to satisfy peak processing demands and quicker recovery times with built-in resiliency capabilities. The Capacity on Demand (CoD) function can dynamically change available system capacity. This function can help respond to new business requirements with flexibility and precise granularity.

The IBM Tailored Fit Pricing for IBM Z options delivers unmatched simplicity and predictability of hardware capacity and software pricing, even in the constantly evolving era of hybrid cloud. Consider the following points:

- ▶ The IBM z17 enhancements in resiliency include a capability that is called [IBM Z Flexible Capacity for Cyber Resiliency](#). With Flexible Capacity for Cyber Resiliency, you can remotely shift capacity and production workloads between IBM z17<sup>3</sup> systems at different sites on demand with no onsite personnel or IBM intervention. This capability is designed to help you proactively avoid disruptions from unplanned events and planned scenarios, such as site facility maintenance. Refer to “IBM Z Flexible Capacity for Cyber Resiliency” on page 537.
- ▶ IBM z17 provides no new System Recovery Boost (SRB) enhancements. As for IBM z16, SRB provides boosted processor capacity and parallelism for specific events. Client-selected middleware starts and restarts to expedite recovery for middleware regions and restore steady-state operations as soon as possible. z/OS SVC memory dump processing and HyperSwap® configuration load and reload are boosted to minimize the effect on running workloads. See: [Systems Recovery Boost content solution](#).
- ▶ On IBM z17, with the new Coupling Facility Control Code (CFCC) Level 26, the enhanced ICA-SR coupling link protocol provides improvements for read, lock, and write requests, compared to CF service times on IBM z16 systems. The improved CF service times for CF requests can translate into better Parallel Sysplex coupling efficiency; therefore, the software costs can be reduced for the attached z/OS images in the Parallel Sysplex.
- ▶ IBM z17 provides improved CF processor scalability, virtualization, consolidation, and density enhancements for CF images. The effective amount of CF capacity growth in the image will change as you add more processors to the CF image. You will get a lot more effective capacity on a z17 as you increase the number of processors up towards the limit of 16. Furthermore, a capacity study using one of the capacity planning tools will show that two CF images with eight processors each will still yield more effective capacity than one large CF image with 16 processors. That is a generally true statement of any multiprocessor image (and would be true for z/OS images, too).
- ▶ IBM z17 CF images continues to support a maximum of 16 processors in a CF image.

The IBM z17 also added functions to protect today's data now, and from future cyberattacks that can be initiated by quantum computers. The IBM z17 provides the following quantum-safe capabilities:

---

<sup>3</sup> IBM z16 also supports Flexible Capacity.



- ▶ Key generation
- ▶ Encryption
- ▶ Key encapsulation mechanisms
- ▶ Hybrid key exchange schemes
- ▶ Dual digital signature schemes

In addition to these quantum-safe cryptographic capabilities, tools (such as IBM Application Discovery and Delivery Intelligence [ADDI], Integrated Cryptographic Service Facility [ICSF], and IBM Crypto Analytics Monitor [CAT]) can help you discover where and what cryptography is used in applications. This knowledge can aid in developing a cryptographic inventory for migration and modernization planning.

### 1.1.1 Supported upgrade paths

The supported upgrade paths for the IBM z17 ME1 are shown in Figure 1-1.

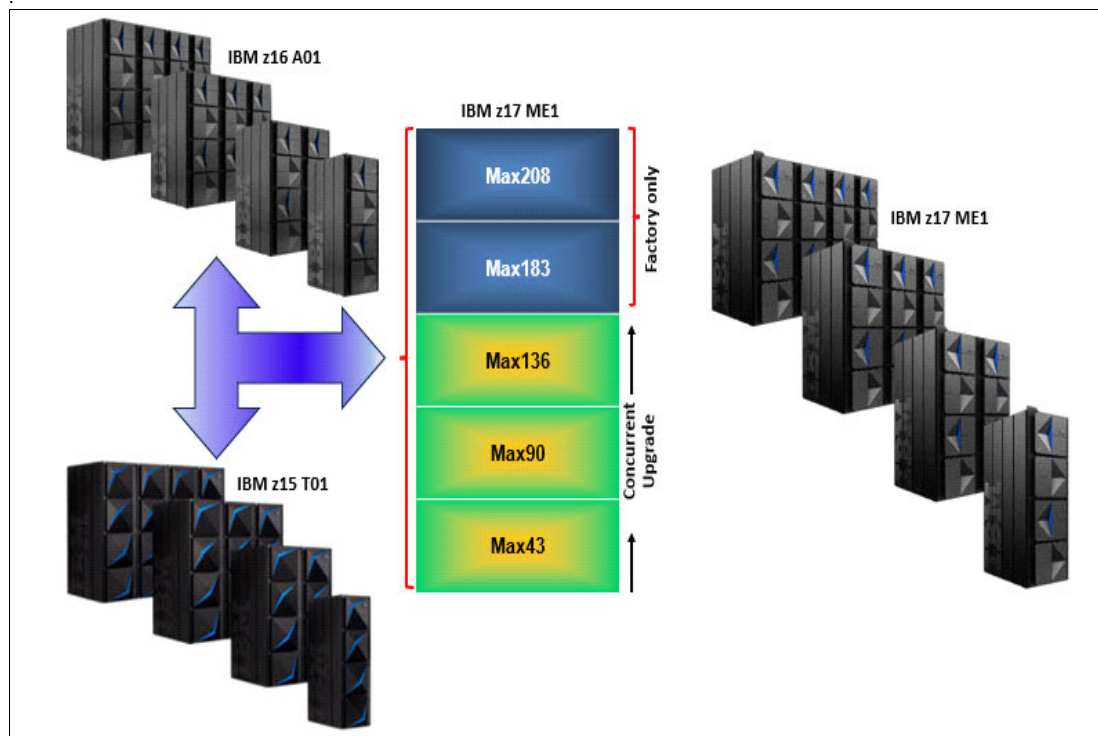


Figure 1-1 IBM z17 ME1 upgrade paths

### 1.1.2 Capacity and performance

The IBM z17 ME1 offers 337 capacity levels. In all, 208 capacity levels are based on the number of physically used CPs, plus up to 129 extra subcapacity models for the first 43 CPs.

The IBM z17 ME1 provides increased processing and enhanced I/O capabilities over its predecessor, the IBM z16 A01. This capacity is achieved by increasing the number of PUs per system, increased system cache, and introducing new I/O technologies.

The IBM z17 feature Max208 is estimated to provide up to 15% (+/- 2%) more total system capacity than the IBM z16 Model Max200, with the same amount of memory and power requirements. With up to 64 TB of main storage and enhanced SMT, the performance of the IBM z17 ME1 processors deliver considerable improvement. Uniprocessor performance also



increased significantly. An IBM z17 Model 701 offers average performance improvements of up to 11% (+/-2%) over the IBM z16 Model 701.<sup>4</sup>

The Integrated Facility for Linux (IFL) and IBM Z Integrated Information Processor (zIIP) processor units on the IBM z17 can be configured to run two simultaneous threads in a single processor (SMT). This feature increases the capacity of these processors with 25% on average<sup>4</sup> over processors that are running single thread. SMT is also enabled by default on System Assist Processors (SAPs).

Within each single drawer, IBM z17 provides 20% (+/-2%) greater capacity than IBM z16 for standard models and 40% greater capacity on the maximum configured model, which enables efficient partition scaling.

This comparison is based on the Large System Performance Reference (LSPR) mixed workload analysis. The range of performance ratings across the individual LSPR workloads is likely to feature a large spread. More performance variation of individual logical partitions (LPARs) is available when an increased number of partitions and more PUs are available. For more information, see Chapter 12, “Performance and capacity planning” on page 489.

For more information about performance, [see the LSPR website](#).

For more information about millions of service units (MSUs) ratings, see the [IBM Z Software Contracts](#) website.

### 1.1.3 Supported operating systems

The IBM z17 is supported by a large set of software products and programs, including independent software vendor (ISV) applications. Use of various features might require the latest releases.

The following operating systems are supported on the IBM z17:

- ▶ z/OS Version 3 Release 1 with PTFs
- ▶ z/OS Version 2 Release 5 with PTFs
- ▶ z/OS Version 2 Release 4<sup>5</sup> with PTFs (toleration support only)
- ▶ z/VM® Version 7 Release 4 with PTFs
- ▶ z/VM Version 7 Release 3 with PTFs
- ▶ VSE<sup>n</sup> V6.3.1 (21<sup>st</sup> Century Link Software)
- ▶ z/TPF Version 1 Release 1 with PTFs (compatibility support)

21<sup>st</sup> Century Link Software VSE<sup>n</sup> V6.3.1 is supported on IBM z17. For more information, see 7.7, “VSEn migration considerations” on page 347.

IBM plans to support the following Linux on IBM Z distributions on IBM z17:

- ▶ SUSE SLES 16.1 (Post GA)
- ▶ SUSE SLES 15.6 (GA)
- ▶ SUSE SLES 12.5 (Post GA)
- ▶ Red Hat RHEL 10.0 (Post GA)
- ▶ Red Hat RHEL 9.4
- ▶ Red Hat RHEL 8.10
- ▶ Red Hat RHEL 7.9 (Post GA)
- ▶ Canonical Ubuntu 24.04 LTS (Post GA)
- ▶ Canonical Ubuntu 22.04 LTS (Post GA)
- ▶ Canonical Ubuntu 20.04 LTS (Post GA)

<sup>4</sup> Observed performance increases vary depending on the workload types.

<sup>5</sup> z/OS 2.4 End of service support on 09/24 - requires IBM Software Support Services.



The support statements for the IBM z17 also cover the KVM hypervisor on distribution levels that have KVM support.

For more information about the features and functions that are supported on IBM z17 by operating system, see Chapter 7, “Operating systems support” on page 261.

### 1.1.4 Supported IBM compilers

The following IBM compilers for IBM Z can be used with the IBM z17:

- ▶ Enterprise COBOL for z/OS
- ▶ Enterprise PL/I for z/OS
- ▶ Automatic Binary Optimizer
- ▶ Open XL C/C++ 1.1 for z/OS
  - Available as a Web deliverable for z/OS V2.5 and V3.1
- ▶ XL C/C++ for Linux on IBM Z

The compilers increase the return on your investment in IBM Z hardware by maximizing application performance by using the compilers’ advanced optimization technology for IBM z/Architecture®.

Through their support of web services, XML, and Java, they allow for the modernization of assets in web-based applications. They also support the latest IBM middleware products (CICS®, Db2®, and IMS), which allows applications to use their latest capabilities.

To fully use the capabilities of the IBM z17, you must compile your code by using the minimum level of each compiler. To obtain the best performance, you must specify an architecture level of applicable to your environment, being mindful of potential N-1 and N-2 generations at Disaster Recovery or Secondary sites.

For more information, see 7.5.8, “z/OS XL C/C++ considerations” on page 342.

## 1.2 IBM z17 ME1 technical overview

This section briefly reviews the following main elements of the IBM z17:

- ▶ Frames
- ▶ CPC drawers
- ▶ I/O subsystem and I/O drawers
- ▶ Storage connectivity
- ▶ Network connectivity
- ▶ Clustering connectivity
- ▶ Cryptography
- ▶ Supported connectivity and crypto features
- ▶ Special-purpose features and functions

### 1.2.1 Frames

The IBM z17 ME1 uses 19-inch frames and industry-standardized power and hardware. It can be configured as a one-, two-, three-, or four-frame system. The IBM z17 ME1 packaging is compared to the two previous IBM Z platforms in Table 1-2.



Table 1-2 IBM z17 configuration options compared to IBM z15 and IBM z16 configurations

System	Number of frames	Number of CPC drawers	Number of I/O drawers	I/O and power connections	Power options	Cooling options
IBM z17 ME1	1 - 4	1 - 4	0 - 12	Rear only	<b>PDU only</b>	Radiator (air) only
IBM z16 A01	1 - 4	1 - 4	0 - 12 <sup>a</sup>	Rear only	PDU or BPA	Radiator (air) only
IBM z15® T01	1 - 4	1 - 5	0 - 12 <sup>b</sup>	Rear only	PDU or BPA	Radiator (air) or Water-Cooling Unit (WCU)

a. Maximum of 12 if ordered with a PDU or maximum of 10 if ordered with a BPA.

b. Maximum of 12 if ordered with a PDU or maximum of 11 if ordered with a BPA.

## 1.2.2 CPC drawers

The IBM z17 ME1 can be configured with up to four CPC drawers (three in the A Frame and one in the B Frame). Each CPC drawer contains the following elements:

- ▶ Processor Unit, PU chips that use Extreme Ultra-Violet (EUV) process with 5 nm silicon lithography technology. Each processor chip consists of eight PU cores and one DPU. Two processor chips are packaged in the dual-chip module (DCM).
- ▶ DCMs
 

Four interconnected DCMs that contain 64 physical PU cores plus eight DPUs per drawer (each cooled by an internal water loop). Each DCM can have 9 - 11 or 10 - 15 active PU cores (depending on the configuration that is used), and two DPU cores.
- ▶ Memory:
  - A minimum of 512 GB and a maximum of 64 TB of memory per system can be ordered, which does not include 884 GB for hardware system area (HSA).
  - Up to 48 dual inline memory modules (DIMMs) that are 32 GB, 64 GB, 128 GB, 256 GB, or 512 GB are plugged in a CPC drawer.
- ▶ Fan-outs
 

Each CPC drawer supports up to 12 PCIe+ fan-out adapters to connect to the PCIe+ I/O drawers, and Integrated Coupling Adapter Short Reach 2.0<sup>6</sup> (ICA SR 2.0) coupling links:

  - The I/O PCIe hubs support Gen5 x16 into the fanout and will drive Gen4 x16 (Bifurcated to 2 at Gen4 x8) out of the fanout to the I/O Drawers.
  - Two-port PCIe 16 gigabytes per second (GBps) I/O fan-out, each port supports one domain in the 16-slot PCIe+ I/O drawers.
  - ICA SR 2.0 PCIe fan-outs for coupling links (two PCIe links, 8 GBps each).
- ▶ Four Power Supply Units (PSUs), which provide power to the CPC drawer and are accessible from the rear.

**Note:** Loss of one PSU leaves enough power to satisfy the power requirements of the entire drawer. The PSUs can be concurrently maintained.

<sup>6</sup> ICA SR and ICA SR 1.1 adapters are not supported on the IBM z17.



- ▶ Two dual-function Base Management Cards (BMCs)/Oscillator Cards (OSCs), which provide redundant interfaces to the internal management network and provide clock synchronization to the IBM Z platform.
- ▶ Two dual-function Processor Power Cards (PPC), which control Voltage Regulation, PSU and Fan control. The PPCs are redundant and can be concurrently maintained.
- ▶ Five fans are installed at the front of the drawer to provide cooling airflow for the resources that are installed in the drawer except for the PU SCMs, which are internally water-cooled.

With the IBM z17, Virtual Flash Memory (VFM) feature is offered from the main memory capacity in 0.5 TB units (up to 6TBs maximum), to increase granularity for the feature. VFM can provide much simpler management and better performance by eliminating the I/O to the adapters in the PCIe+ I/O drawers.

### 1.2.3 I/O subsystem and I/O drawers

The IBM z17 supports a PCIe I/O infrastructure. PCIe features are installed in PCIe+ I/O drawers. Up to 12 I/O drawers per IBM z17 can be ordered, which allows for up to 192 PCIe I/O and special purpose features.

For a four CPC drawer system, up to 48 PCIe+ fan-out slots can be populated with fan-out cards for data communications between the CPC drawers and the I/O infrastructure, and for coupling. The multiple channel subsystem (CSS) architecture allows up to six CSSs, each with 256 channels.

The IBM z17 implements PCIe Generation 5 (PCIe+ Gen5), which is used to connect the PCIe Generation 4 (PCIe+ Gen4) dual port fan-out features in the CPC drawers. The I/O infrastructure is designed to reduce processor usage and I/O latency, and provide increased throughput and availability.

#### PCIe+ I/O drawer

Together with the PCIe features, the PCIe+ I/O drawer offers finer granularity and capacity over previous I/O infrastructures. It can be concurrently added and removed in the field, which eases planning. Only PCIe cards (features) are supported, in any combination.

### 1.2.4 Storage connectivity

Storage connectivity is provided on the IBM z17 by FICON® Express and the IBM zHyperLink Express features.

#### FICON Express

FICON Express features follow the established Fibre Channel (FC) standards to support data storage and access requirements, along with the latest FC technology in storage and access devices. FICON Express features support the following protocols:

- ▶ FICON  
This enhanced protocol (as compared to FC) provides for communication across channels, channel-to-channel (CTC) connectivity, and with FICON devices, such as disks, tapes, and printers. It is used in z/OS, z/VM, VSE (VSE<sup>n</sup> V6.3.1 - 21<sup>st</sup> Century Software), z/TPF (Transaction Processing Facility), and Linux on IBM Z environments.
- ▶ Fibre Channel Protocol (FCP)  
This standard protocol is used for communicating with disk and tape devices through FC switches and directors. The FCP channel can connect to FCP SAN fabrics and access



FCP/SCSI devices. FCP is used by z/VM, KVM, VSE (VSE<sup>n</sup> V6.3.1 21<sup>st</sup> Century Software), and Linux on IBM Z environments.

FICON Express32-4P features are implemented by using PCIe cards, and offers better port granularity and improved capabilities over the previous FICON Express features. FICON Express32-4P four port features support a link data rate of 32 gigabits per second (Gbps) (8, 16, or 32 Gbps auto-negotiate), and it is the preferred technology for new systems.

### zHyperLink Express

zHyperLink was created to provide fast access to data by way of low-latency connections between the IBM Z platform and storage.

The zHyperlink Express2.0 is updated to support PCIe+ Gen4 with new Gen4 retimer, Gen4 Switch with DMA, and the new CXP16 Gen4 optical transceiver. It connects to DS8K zHyperlink adapter on the other side of the link.

The zHyperLink Express2.0 feature allows you to make synchronous requests for data that is in the storage cache of the IBM DS8900F. This process is done by directly connecting the zHyperLink Express2.0 port in the IBM z17 to an I/O Bay port of the IBM DS8000®. This short distance (up to 150 m [492 feet]), direct connection is designed for low-latency reads and writes, such as with IBM DB2® for z/OS synchronous I/O reads and log writes.

Working with the FICON SAN Infrastructure, zHyperLink can improve application response time, which cuts I/O-sensitive workload response time in half without requiring application changes.<sup>7</sup>

**Note:** The zHyperLink channels complement FICON channels, but they do *not* replace FICON channels. FICON channels remain the main data driver and are mandatory for zHyperLink usage.

## 1.2.5 Network connectivity

The IBM z17 is a fully virtualized platform that can support many system images at once. Therefore, network connectivity covers not only the connections between the platform and external networks with new build and carry forward Open Systems Adapter-Express (OSA-Express) and Network Express features, but also within the IBM z17 using specialized internal connections for intra-system communication through IBM HiperSockets and Internal Shared Memory (ISM).

### Network Express

The Network Express adapter is the IBM z17 common hardware for OSA, RoCE and Coupling Express3 Long Reach (CE-LR). It converges the legacy OSA-Express and RoCE Express and CE-LR into one hardware platform offering. The adapter currently supports LR and SR, and optics will be supported for 10G and 25G.

The Network Express card is considered an OSA card even though it can perform RoCE I/O because it is not associated with a Resource Group (PSP) unlike RoCE adapters.

This new adapter supports all legacy functions available with OSA OSD, but uses EQDIO (Enhanced-QDIO) architecture while OSD uses QDIO. The Network Express card only supports the Enhanced QDIO (EQDIO) architecture, with CHPID type OSH for OSA-style I/O.

<sup>7</sup> The performance results can vary depending on the workload. Use the zBNA tool for the zHyperLink planning.



The card has 2 ports, each of which is a PCHID, regardless of variant. The PCHIDs on the card are managed by the new on-chip DPU processor, unless they are defined as NETD in the IOCDS for Physical Function (PF) Access Mode, where the PF is assigned to a customer partition instead of DPU.

Network Express supports both PCIe (e.g. RoCE) and OSA functionality on the *same port* of a card. This eliminates the need for some users of RoCE to buy a dedicated OSA card just to enable RoCE I/O. Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) is a network protocol which allows data to be transferred directly between the memory of two computers in the same Ethernet broadcast domain, significantly reducing latency, CPU load, and increasing bandwidth.

These features help reduce the use of CPU resources for applications that use the TCP/IP stack (such as IBM WebSphere® that accesses an IBM Db2 database). They also can help reduce network latency with memory-to-memory transfers by using Shared Memory Communications over RDMA (SMC-R).

With SMC-R, you can transfer huge amounts of data quickly and at low latency. SMC-R is transparent to the application and requires no code changes, which enables rapid time to value.

The RoCE function can also provide LAN connectivity for Linux on IBM Z, and complies with IEEE standards. In addition, RoCE assumes several functions of the TCP/IP stack that normally are performed by the PU, which allows significant performance benefits by offloading processing from the operating system.

Customers who require the legacy QDIO architecture (CHPID type OSD) must use OSA-Express7S 1.2 cards. The OSA CHPID types used on Network Express cards must be the same; thus, for example, if one PCHID is TYPE=OSH, then the other must also be TYPE=OSH. However, both NETH FIDs and an OSH CHPID can coexist on the same PCHID. Likewise, if one PCHID is defined as NETD for PF Access Mode, then the other PCHID can only be NETD.

## OSA-Express

The OSA-Express features provide local area network (LAN) connectivity and comply with IEEE standards. In addition, OSA-Express features assume several functions of the TCP/IP stack that normally are performed by the PU, which allows significant performance benefits by offloading processing from the operating system.

OSA-Express7S 1.2 features continue to support copper and fiber optic (single-mode and multimode) environments.

## HiperSockets

IBM HiperSockets is an integrated function of the IBM Z platforms that supplies attachments to up to 32 high-speed virtual LANs, with minimal system and network overhead.

HiperSockets is a function of the Licensed Internal Code (LIC). It provides LAN connectivity across multiple system images on the same IBM Z platform by performing memory-to-memory data transfers in a secure way.

The HiperSockets function eliminates the use of I/O subsystem operations. It also eliminates having to traverse an external network connection to communicate between LPARs in the same IBM Z platform. In this way, HiperSockets can help with server consolidation by connecting virtual servers and simplifying the enterprise network.



## Internal Shared Memory

ISM is a virtual Peripheral Component Express (PCI) network adapter that enables direct access to shared virtual memory. It provides a highly optimized network interconnect for IBM Z platform intra-communications.

Shared Memory Communications-Direct Memory Access (SMC-D) uses ISM. SMC-D optimizes operating systems communications in a way that is transparent to socket applications. It also reduces the CPU cost of TCP/IP processing in the data path, which enables highly efficient and application-transparent communications.

SMC-D requires no extra physical resources (such as RoCE, PCIe bandwidth, ports, I/O slots, network resources, or Ethernet switches). Instead, SMC-D uses LPAR-to-LPAR communication through HiperSockets or an OSA-Express feature for establishing the initial connection.

z/OS and Linux on IBM Z support SMC-R and SMC-D. Now, data can be shared by way of memory-to-memory transfer between z/OS and Linux on IBM Z.

## 1.2.6 Clustering connectivity

A Parallel Sysplex is an IBM Z clustering technology that is used to make applications that are running on logical and physical IBM Z platforms highly reliable and available. The IBM Z platforms in a Parallel Sysplex are interconnected by way of coupling links.

Coupling connectivity on the IBM z17 uses Coupling Express3 Long Reach (CE3 LR) and Integrated Coupling Adapter Short Reach (ICA SR2.0) features. The ICA SR feature supports distances up to 150 meters (492 feet); the CE3 LR feature supports unrepeated distances of up to 10 km (6.21 miles) between IBM Z platforms. ICA SR features provide sysplex and timing connectivity direct to the CPC drawer, while Coupling Express3 LR features connect into the PCIe+ I/O Drawer.

Coupling links can also carry timing information such as Server Time Protocol (STP) for synchronizing time across multiple IBM Z CPCs in a Coordinated Time Network (CTN).

For more information about coupling and clustering features, see 4.5, “I/O features” on page 181.

## 1.2.7 Cryptography

IBM z17 provides two main cryptographic functions: CP Assist for Cryptographic Functions (CPACF) and Crypto-Express8S.

### CPACF

CPACF is a high-performance, low-latency coprocessor that resides in every Telum II z17 PU chip, performs symmetric key encryption operations, and calculates message digests (hashes) in hardware. The following algorithms are supported:

- ▶ Encryption (DES, TDES, AES)
- ▶ Hashing (SHA-1, SHA-2, SHA-3, SHAKE)
- ▶ Random Number Generation (PRNG, DRNG, TRNG)

CPACF supports Elliptic Curve Cryptography (ECC) clear key, improving the performance of Elliptic Curve algorithms. The following algorithms are supported:

- ▶ ECDH[E]
- ▶ P-256, P-384, and P-521



- ▶ X25519, and X448
- ▶ ECDSA
- ▶ Keygen, sign, verify
- ▶ P-256, P-384, P521,
- ▶ EdDSA
- ▶ KeyGen, sign, verify
- ▶ Ed25519, Ed448
- ▶ Support for protected key signature creation

### Crypto-Express8S

The tamper-sensing and tamper-responding Crypto-Express8S features provide acceleration for high-performance cryptographic operations and support up to 85 domains with the IBM z17 ME1. This specialized hardware performs AES, DES and TDES, RSA, Elliptic Curve (ECC), SHA-1, and SHA-2, and other cryptographic operations.

It supports specialized high-level cryptographic APIs and functions, including those functions that are required with quantum-safe cryptography and in the banking industry. Crypto-Express8S features are designed to meet the Federal Information Processing Standards (FIPS) 140-2 Level 4 and PCI HSM security requirements for hardware security modules.

IBM z17 is an industry quantum-safe system<sup>8</sup>. Consider the following points:

- ▶ IBM z17 quantum-safe secure boot technology helps to protect IBM Z firmware from quantum attacks by using a build-in dual signature scheme with no changes required.
- ▶ IBM z17 quantum-safe technology and key management services were developed to help you protect data and keys against a potential future quantum attack, such as harvest now, decrypt later.
- ▶ IBM z17 positions customers to use quantum-safe cryptography along with classic cryptography as they begin modernizing existing applications and building new applications.

For more information about cryptographic features and functions, see Chapter 6, “Cryptographic features” on page 221.

## 1.2.8 Supported connectivity and crypto features

The IBM z17 provides a PCIe-based infrastructure for the PCIe+ I/O drawers to support the following features:

- ▶ Storage connectivity:
  - zHyperLink Express2.0 (new build)
  - FICON Express32-4P - 4 ports/adaptor FCs 0387/0388 (new build only)
  - FICON Express32S - 2 ports/adaptor FCs 0461/0462 (carry forward)
  - FICON Express16SA (carry forward only)
- ▶ Network connectivity:

<sup>8</sup> DISCLAIMER: IBM z17 with the Crypto Express 8S card provides quantum-safe APIs and access to quantum-safe algorithms that were selected as finalists during the PQC standardization process that was conducted by NIST (<https://csrc.nist.gov/Projects/post-quantum-cryptography/round-3-submissions>). *Quantum-safe cryptography* refers to efforts to identify algorithms that are resistant to attacks by classic and quantum computers to keep information assets secure, even after a large-scale quantum computer is built. (<https://www.etsi.org/technologies/quantum-safe-cryptography>). These algorithms are used to help ensure the integrity of several firmware and boot processes. IBM z16 was the Industry-first system that is protected by quantum-safe technology across multiple layers of firmware.



- OSA-Express7S 1.2 (new build and carry forward)
- OSA-Express7S GbE SX and LX (carry forward from IBM z15 only)
- OSA-Express7S 10GbE SR and LR (carry forward from IBM z15 only)
- OSA Express 7S 1000Base-T (carry forward only)
- Network Express (new build)
  - (Common Hardware for: OSA, RoCE, and Coupling LR)
- ▶ Cryptographic features:
  - Crypto Express8S, one or two HSMs<sup>9</sup> (new build and carry forward)
  - Crypto Express7S, 1- or 2-port<sup>10</sup> (carry forward only)
- ▶ Clustering connectivity<sup>11</sup>:
  - ICA SR2.0 (new build)
  - Coupling Express3 Long Reach (new build)

### 1.2.9 Special-purpose features and functions

When it comes to designing and developing the IBM Z platform, IBM takes a total systems view. The IBM Z stack is built around digital services, agile application development, connectivity, and system management. This design approach creates an integrated, diverse platform with specialized hardware and dedicated computing capabilities.

The IBM z17 delivers a range of features and functions that allows PUs to concentrate on computational tasks, while distinct, specialized features take care of the rest. For more information about these features and other IBM z17 features, see in 3.5, “Processor unit functions” on page 103.

## 1.3 Hardware management

The Hardware Management Consoles (HMCs) and Support Elements (SEs) are appliances that together provide platform management for IBM Z.

The HMC is an appliance that provides a single point of control for managing local or remote hardware elements.

For IBM z17 new built systems, IBM Z Hardware Management Appliance (FC 0129) is the only available HMC. The HMC Appliance and SE Appliance run virtualized on the SE hardware.

Standalone HMCs are no longer supported.

Existing HMA features of the IBM z15 and IBM z16 can be upgraded to driver 61 HMC code to support IBM z17, but older stand-alone HMCs (rack-mounted or tower) cannot be carried forward during an MES upgrade to IBM z17.

For more information, see Chapter 10, “Hardware Management Console and Support Element” on page 429.

<sup>9</sup> The Crypto Express8S is available with either one or two hardware security modules (HSM). The HSM is the IBM 4770 PCIe Cryptographic Coprocessor (PCleCC).

<sup>10</sup> The Crypto Express7S comes with either one (1-port) or two (2-port) hardware security modules (HSM). The HSM is the IBM 4769 PCIe Cryptographic Coprocessor (PCleCC).

<sup>11</sup> ICA SR2.0 features connect directly into the CPC (processor) Drawer, while Coupling Express3 Long Reach connects into the PCIe+ I/O Drawer.



## 1.4 Reliability, availability, and serviceability

System reliability, availability, and serviceability (RAS) is an area of continuous IBM focus and a defining IBM Z platform characteristic. The RAS objective is to reduce (or eliminate, if possible) all sources of planned and unplanned outages while providing adequate service information if an issue occurs. Adequate service information is required to determine the cause of an issue without the need to reproduce the context of an event.

IBM Z platforms are designed to enable highest availability and lowest downtime. These facts are recognized by various IT analysts, such as ITIC<sup>12</sup> and IDC<sup>13</sup>. A comprehensive, multi-layered strategy includes the following features:

- ▶ Error Prevention
- ▶ Error Detection and Correction
- ▶ Error Recovery
- ▶ System Recovery Boost

With a suitably configured IBM z17, further reduction of outages can be attained through First Failure Data Capture (FFDC), which is designed to reduce service times and avoid subsequent errors. It also improves nondisruptive replace, repair, and upgrade functions for memory, drawers, and I/O adapters. IBM z17 supports the nondisruptive download and installation of LIC updates.

IBM z17 RAS features provide unique high-availability and nondisruptive operational capabilities that differentiate IBM Z in the marketplace. IBM z17 RAS enhancements are made on many components of the CPC (processor chip, memory subsystem, I/O, and service) in areas, such as error checking, error protection, failure handling, error checking, faster repair capabilities, sparing, and cooling.

The ability to cluster multiple systems in a Parallel Sysplex takes the commercial strengths of the z/OS platform to higher levels of system management, scalable growth, and continuous availability.

The IBM z17 builds on the RAS of the IBM z16 family with the following RAS improvements:

- ▶ System Recovery Boost
  - System Recovery Boost was introduced with IBM z15. It offers customers more Central Processor (CP) capacity during system recovery operations to accelerate the startup (IPL), shutdown, or stand-alone memory dump operations (at image level - LPAR<sup>14</sup>). System Recovery Boost requires operating system support. No other IBM software changes are required to be made during the boost period.

System Recovery Boost can be used during LPAR shutdown or startup to make the running operating system and services available in a shorter period.

The System Recovery Boost provides the following options for the capacity increase:

- Subcapacity CP speed boost: During the boost period, subcapacity engines that are allocated to the boosted LPAR are transparently activated at their full capacity (CP engines).
- zIIP Capacity Boost: During the boost period, all active zIIPs that are assigned to an LPAR are used to extend the CP capacity (CP workload is dispatched to zIIP processors during the boost period).

<sup>12</sup> For more information, see [ITIC Global Server Hardware, Server OS Reliability Report](#).

<sup>13</sup> For more information, see [Quantifying the Business Value of IBM Z](#).

<sup>14</sup> LPAR that is running an Operating System image.



- System Recovery Boost enhancements that were delivered with the IBM z16 maximized service availability by using tailored short-duration boosts to mitigate the effect of the following recovery processes:
  - z/OS SVC memory dump boost boosts the system on which the SVC memory dump is taken to reduce system affect and expedite diagnostic capture. It is possible to enable, disable, or set thresholds for this option.
  - Middleware restart/recycle boost boosts the system on which a middleware instance is being restarted to expedite resource recovery processing, release retained locks, and so on. It is applicable to planned restarts, or restarts after failure, automated, or ARM-driven restarts. System Recovery Boost do not boost any system address spaces by default and must be configured by the WLM policy specification.
  - IBM HyperSwap configuration load boost boosts the system in which the HyperSwap configuration and policy information are being loaded or reloaded. This boost applies to Copy Services Manager (CSM) and GDPS. HyperSwap Configuration Load boost is enabled by default. No thresholds or criteria are applied to the boost request based on the size or number of devices that are present in the HyperSwap configuration.

Through System Recovery Boost, the IBM z17 continues to offer more CP capacity during specific system recovery operations to accelerate system (operating system and services) services when the system is being started or shutdown. System Recovery Boost is operating system-dependent. No other hardware, software, or maintenance charges are incurred during the boost period for the base functions of System Recovery Boost.

- At the time of this writing, the main System Recovery Boost users are z/OS (running in an LPAR), z/VM, VSE<sup>n</sup> V6.3.1 21<sup>st</sup> Century Software, and z/TPF) and stand-alone memory dump (SADMP).

z/VM uses the System Recovery Boost if it runs on subcapacity CP processors only (IFLs are always at their full clock speed). Second-level z/VM guest operating systems<sup>15</sup> can inherit the boost if they are running on CPs.

For more information about RAS and System Recovery Boost, see *Introducing IBM Z System Recovery Boost*, [REDP-5563](#).

- Level 2 (physical), Level 3, and Level 4 (virtual) cache enhancements include the use symbol ECC to extend the reach of older IBM Z generations cache and memory improvements for augmented availability.

The L2, L3, and L4 cache powerful symbol ECC makes it resistant to more failure mechanisms. Preemptive DRAM marking is added to the main memory to isolate and recover failures more quickly.

---

<sup>15</sup> z/OS that is configured as a guest system under z/VM does not use the boost.











## 2



# Central processor complex hardware components

This chapter provides information about the new IBM z17™ and its hardware building blocks, and how these components physically interconnect. This information is useful for planning purposes and can help in defining configurations that fit your requirements.

**Naming:** The IBM z17 Model ME1, Machine Type (M/T) 9175, is further identified in this document as IBM z17, unless otherwise specified.

This chapter includes the following topics:

- ▶ 2.1, “Frames and configurations” on page 20
- ▶ 2.2, “CPC drawer” on page 26
- ▶ 2.3, “Dual chip modules” on page 34
- ▶ 2.4, “PCIe+ I/O drawer” on page 40
- ▶ 2.5, “Memory” on page 44
- ▶ 2.6, “Reliability, availability, and serviceability” on page 55
- ▶ 2.7, “Connectivity” on page 58
- ▶ 2.8, “Model configurations” on page 61dn
- ▶ 2.9, “Power and cooling” on page 65
- ▶ 2.10, “Summary” on page 70



## 2.1 Frames and configurations

The IBM z17 Model ME1 system is designed in a 19-inch form factor with configuration of 1 - 4 frames that can be easily installed in any data center. The IBM z17 ME1 (M/T 9175) can include from one to four 42U EIA (19-inch) frames, which are bolted together. The configurations can include up to four central processor complex (CPC) drawers and up to 12 Peripheral Component Interconnect® Express+ (PCIe+) I/O drawers.

The redesigned CPC drawer and I/O infrastructure also lower power consumption, reduces the footprint, and allows installation in virtually any data center. The IBM z17 server is rated for ASHRAE class A3<sup>1</sup> data center operating environment.

The IBM z17 server is similar to IBM z16 and IBM z15, but differentiates itself from previous IBM Z server generations through the following significant changes to the modular hardware:

- ▶ All external cabling (power, I/O, and management) is performed at the rear of the system
- ▶ Flexible configurations: Frame quantity is determined by the system configuration (1 - 4 frames)
- ▶ Feature codes that reserve slots for plan-ahead CPC drawers and a new Spyre adapter
- ▶ Internal water cooling plumbing for systems with up to four CPC drawers (Frames A and B)
- ▶ PCIe+ Gen4 I/O drawers (19-inch format) supporting 16 PCIe adapters

The only power option for the IBM z17 is PDU-based power. The IBM z17 ME1 system is designed as a radiator (air) cooled system.

The IBM z17 ME1 includes the following basic hardware building blocks:

- ▶ 19-inch 42u frame (1 - 4)
- ▶ CPC (Processor) drawers (1 - 4)
- ▶ PCIe+ Gen4 I/O drawers (up to 12)
- ▶ Radiator cooling assembly (RCA) for CPC drawers cooling the Dual Chip Modules (DCM)
- ▶ Power, with Intelligent Power Distribution Units (iPDU) pairs (2 - 4 per frame, maximum 8, depending on the configuration).
- ▶ Support Elements combined with optional Hardware Management Appliance (two):
  - Single Keyboard, Mouse, Monitor (KMM) device (USB-C connection)
  - Optional IBM Hardware Management Appliance feature
- ▶ 24-port 1 GbE Switches (two or four, depending on the system configuration)
- ▶ Hardware for cable management at the rear of the system

---

<sup>1</sup> For more information, see Chapter 2, Environmental specifications in *IBM 9175 Installation Manual for Physical Planning*, GC28-7049.



An example of a fully configured system with PDU-based power, four CPC drawers, and up to 12 PCIe+ I/O drawers is shown in Figure 2-1.

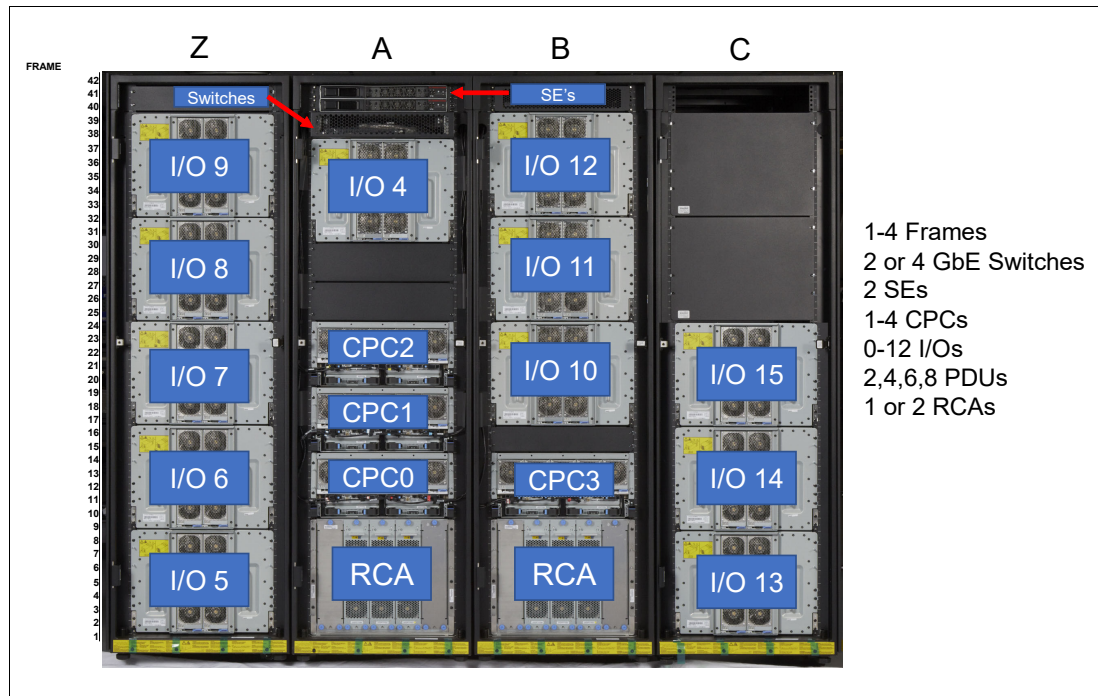


Figure 2-1 IBM z17 ME1 Maximum configuration (front view)

The key features that are used to build the system are listed in Table 2-1. For more information about the various configurations, see Appendix E, “Frame configurations with Power Distribution Units” on page 551.

Table 2-1 Key features that influence the system configurations

Feature Code	Description	Comments
0209	Model ME1	Supports CPs and specialty engines
0571	One CPC Drawer	Feature Max43
0572	Two CPC Drawers	Feature Max90
0573	Three CPC Drawers	Feature Max136
0574	Four CPC Drawers	Feature Max183
0575	Four CPC Drawers	Feature Max208
2933	CPC1 reserve	Reserve A15 location for future add CPC1 (Max39 to Max82 upgrade)
2934	CPC2 reserve	Reserve A20 location for future add CPC2 (Max82 to Max125 upgrade)
<b>Frames and cooling</b>		
4045	A Frame	Radiator (air cooled) with up to 4 PDUs
4046	B Frame	Radiator (air cooled) with 2 PDUs
4047	Z Frame	I/O drawers only without PDUs



4048	C Frame	I/O drawers only with 2 PDUs
<b>PDU power</b>		
0563	200 - 208 V 30/60A 3-Phase (Delta) PDU	North America and Japan
0564	380 - 415 V 32A 3 Phase (Wye) PDU	Worldwide (except North America and Japan)
<b>I/O</b>		
4011	PCIe+ I/O drawer	Max. 12 (PDU)
0352	Z Frame First I/O Placement	Starts IO drawer plugging sequence at Z frame bottom
5823	Top Exit Enclosure	Includes new cable management
7804	Bottom Exit Cabling	Includes rear tailgate hardware at bottom of frame
7803	Top Exit Cabling without Top Exit Enclosure	Uses rear slide plates at top of frame

## Power Options

The 9175 (z17 ME1) has the power options shown in Table 2-2:

Table 2-2 IBM z17 ME1 Power Options

Feature Description	iPDU
Number of line cords	2,4,6 or 8
3-Phase Line cords	Yes
200-240VAC (4 wire), 60A (Low Voltage)	Yes
380-415VAC (5 wire) 30/32A (High Voltage)	Yes
480VAC (5 wire Wye) <sup>a</sup>	No
PCIe+ I/O drawer max	12
Water Cooling	No
Radiator Cooling	Yes
Internal Battery Feature	No
Phase loss immunity	No
Balanced Power	No
DC Power available	No

a. Wye cords require five wires, three for Power phases, one for Neutral and one for Ground.

**Caution:** Installation of a *low voltage system* to a *high voltage facility* will cause significant damage to the system's power components.



## Considerations

Consider the following points:

- ▶ A-Frame is always present in every configuration
- ▶ 1u Support Elements (x2) are always in A-Frame at locations A41 and A42
- ▶ 1u 24-port internal Ethernet switches (x2) are always at locations A39 and A40  
Additional Ethernet switches (x2) are available when necessary in Frames C or B

### 2.1.1 IBM z17 cover (door) design

The standard cover set for IBM z17 model ME1 is shown in Figure 2-2



Figure 2-2 IBM z17 (4 frames) with covers

- ▶ A new shipping container has been developed for z17 to minimize carbon footprint impact
  - Container will include both the system frame as well as its front and rear covers (i.e., will no longer have separate package for the cover-set to be ship in and/or dispose of).

For the IBM z17 Model ME1 server, the top exit of all cables for I/O or power is always an option with no feature codes required. Adjustable cover plates are available for the openings at the top rear of each frame. See Figure on page 24.

All external I/O cabling enters the system at the rear of the frames for all adapters, management LAN, and power connections.

The Top Exit feature code (FC 5823) provides an optional new Top Exit Enclosure. The optional Top Exit Enclosure provides new fiber cable organizers within the enclosure to optimize the fiber cable storage and strain relief. It also provides mounting locations to secure Fiber Quick Connector (FQC) MPO<sup>2</sup> brackets (FC 5827) on the top of the frames.

---

<sup>2</sup> MPO - Multi-fiber Push On connector



Consider the following points:

- ▶ FC 7804 provides bottom exit cabling and power support. When ordered with FC 5827, it provides harness brackets for the bottom tailgate and fiber bundle organizer hardware.
- ▶ FC 5823 and FC 7803 are for top exit cabling and power, a bottom seal plate is installed.
- ▶ FC 5827 provides the MPO mounting brackets that cable harnesses connect to in the top and or bottom exit tailgates.

**Note:** Overhead I/O cabling is contained within the frames. Extension “chimneys” that were featured with systems before IBM z15 systems are no longer used.

For additional details about cabling feature codes, refer to 11.4, “Physical planning” on page 479.

A view of the top rear of the frame and the openings for top exit cables and power is shown in Figure . When FC 7803 is installed, the plated adjustable shields are removed, and the top exit enclosure is installed.

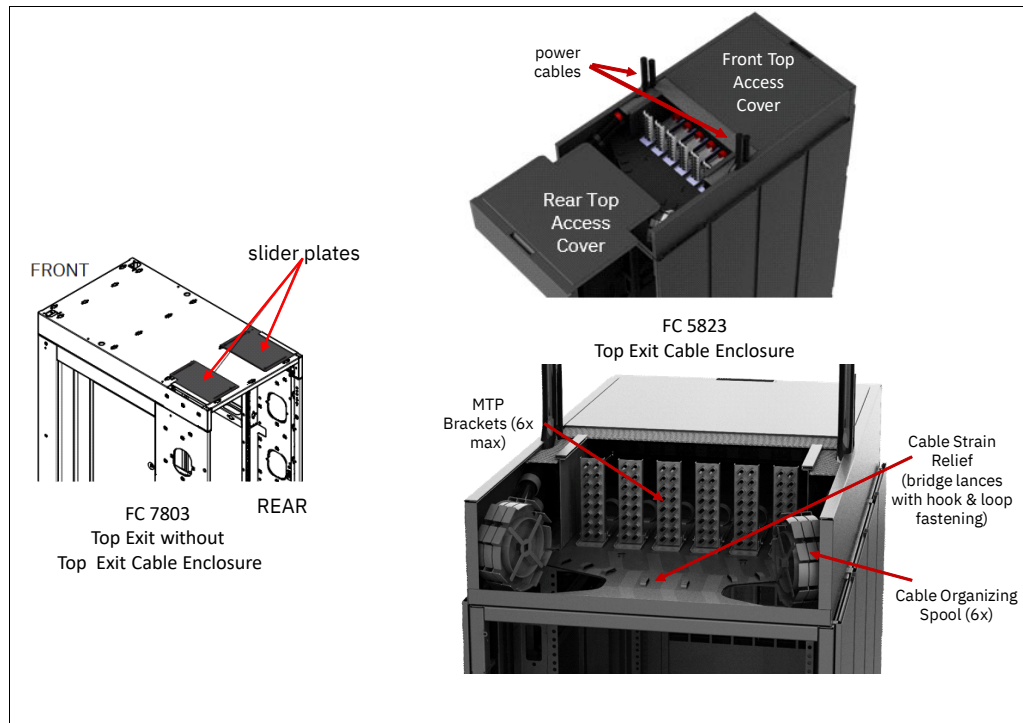


Figure 2-3 Top exit FC 7803 (top exit) without and with FC 5823 (top hat)

Care must be taken when ordering the Feature Codes (Table 2-3 on page 25) for cables that are entering the frames from above the floor, below the floor, or both, and if the Top Exit feature is wanted, which provides the Top Exit Enclosure.



Table 2-3 IBM z17 Model ME1 cabling Feature Code combinations

Environment	Bottom exit	Top exit	Feature Code	Comments
Raised Floor	Yes	No	7804 only	Ships with Bottom Exit Tailgate & supports Bottom FQC FC 5827
Raised Floor	Yes	Yes (no Top Exit Enclosure)	7803 & 7804	Ships with Bottom Exit Tailgate & supports Bottom FQC FC 5827
Raised Floor	Yes	Yes (with Top Exit Enclosure)	7804 and 5823	Top (5827) and Bottom (5824) FQC support
Raised Floor	No	Yes (no Top Exit Enclosure)	7803	Ships with Bottom Seal Plate and does not support FQC FCs
Raised Floor	No	Yes (with Top Exit Enclosure)	5823 and 7803	Ships with Bottom Seal Plate and only supports FQC FC 5824 & 5826
Non-Raised Floor	No (not supported)	Yes (no Top Exit Enclosure)	7998 <sup>a</sup> and 7803	Ships with Bottom Seal Plate and does not support FQC FCs
Non-Raised Floor	No (not supported)	Yes (with Top Exit Enclosure)	7998 <sup>a</sup> and 5823	Ships with Bottom Seal Plate and does not support FQC FCs

a. FC 7998: Non-Raised floor support (flag)

A vertical cable management guide (“spine”) can assist with proper cable management for fiber, copper, and coupling cables. A top to bottom spine is present from manufacturing with cable organizer clips that are installed for frames Z and C when present. Frames A and B contain mini-spines that serve the same purpose.

The cable retention clips can be relocated for best usage. All external cabling to the system (from top or bottom) can use the spines to minimize interference with the PDUs that are mounted on the sides of the rack. See Figure 2-4 on page 26.



The rack with the spine mounted and the optional fiber cable organizer hoops is shown in Figure 2-4. If necessary, the spine, and organizer hoops can be easily relocated for service procedures.

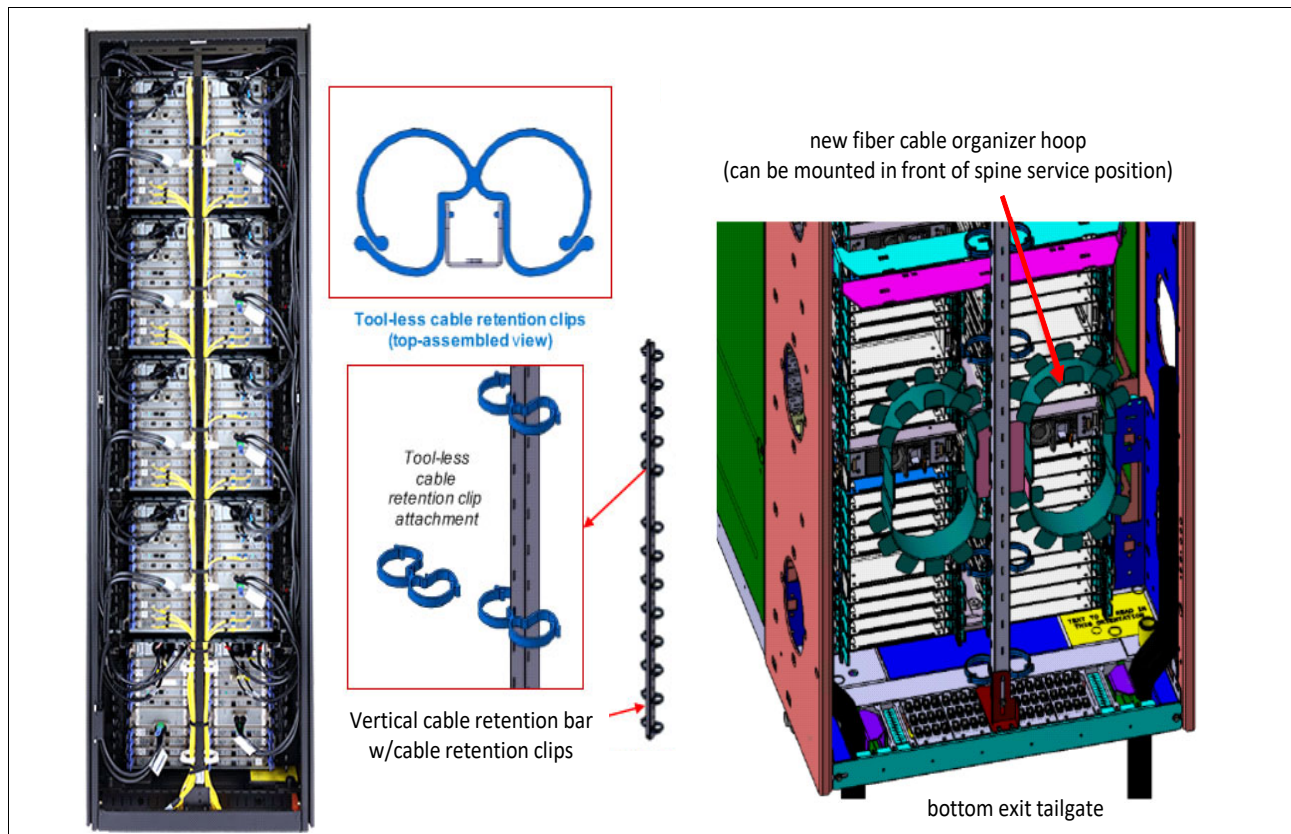


Figure 2-4 I/O cable management spine and fiber cable organizer (rear view)

## 2.2 CPC drawer

The IBM z17 Model ME1 (machine type 9175) features some design changes (compared to IBM z16), primarily with the processor modules and memory packaging in the drawers. An IBM z17 ME1 CPC drawer includes the following features:

- ▶ Four Dual chip modules (DCMs)
- ▶ Up to 48 Memory DIMMs
- ▶ Symmetric multiprocessor (SMP-10) connectivity
  - The IBM z17 SMP assembly consists of a chassis with two SMP-10 jacks and two Y-cables that route signals from the SMP-10 jacks and connect to the DCMs. There are two connectors per Y-cable to each DCM (four DCM connections total)
- ▶ Connectors to support PCIe+ Gen4 fan-out cards for PCIe+ I/O drawers or coupling fan-outs for coupling links to other CPCs



The IBM z17 ME1 can be configured with 1- 4 CPC drawers (three in the A frame and one in the B frame). A CPC drawer and its components are shown in Figure 2-5.

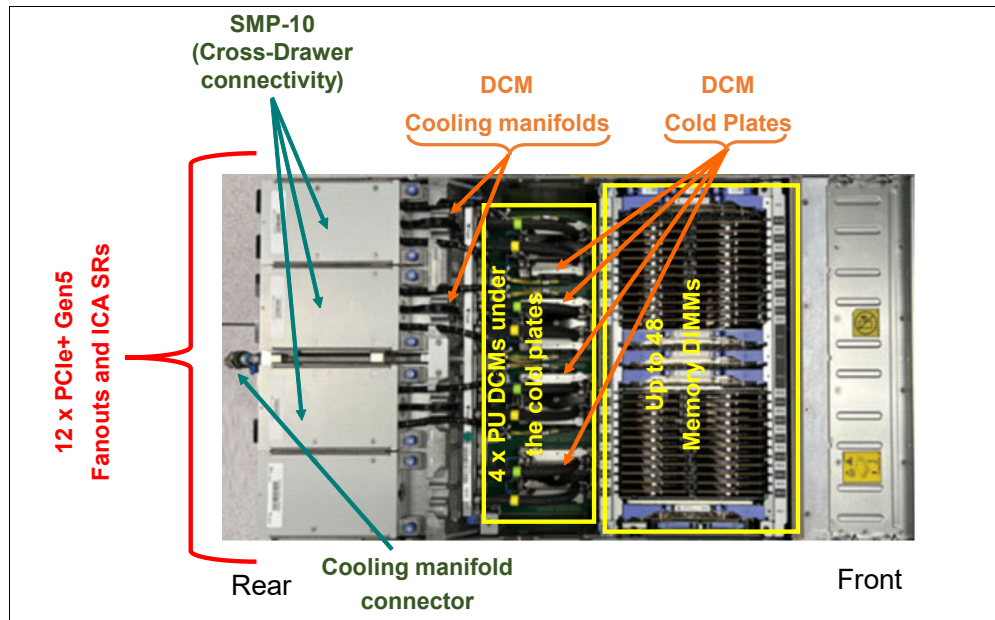


Figure 2-5 CPC drawer components (top view)

The IBM z17 Model ME1 5u CPC drawer always contains four Processor Unit (PU) DCMs, and up to 48 memory DIMM slots.

Depending on the feature, the IBM z17 ME1 contains the following CPC components:

- ▶ The number of CPC drawers installed is driven by the following feature codes:
  - FC 0571: One CPC drawer, Max43, up to 43 characterizable PUs
  - FC 0572: Two CPC drawers, Max90, up to 90 characterizable PUs
  - FC 0573: Three CPC drawers, Max136, up to 136 characterizable PUs
  - FC 0574: Four CPC drawers, Max183, up to 183 characterizable PUs
  - FC 0575: Four CPC drawers, Max208, up to 208 characterizable PUs
- ▶ The following Processor Unit DCM is used:
 

PU DCM contains two PU chips (Telum II) on one single module that use 5 nm silicon wafer technology, 42 Billion transistors, and a core that is running at 5.5 GHz, designed with nine cores per chip (eight PUs and one Data Processing Unit (DPU,) 18 cores per PU DCM - (16 PUs and two DPUs).
- ▶ Memory plugging:
  - Six memory controllers per drawer (two each on DCM2 / DCM1; one each on DCM3 / DCM0)
  - Each memory controller supports eight DIMM slots
  - All eight DIMMs on one memory controller are the same size
  - Four / Six memory controllers per drawer are populated (up to 48 DIMMs)
  - Different memory controllers can have different DIMM sizes



- ▶ Up to 12 PCIe+ Gen4<sup>3</sup> slots that can host:
  - 2-port PCIe+ Gen4 I/O fan-out for PCIe+ I/O drawers (ordered in pairs and connected for availability)
  - ICA SR 2.0 PCIe fan-out for coupling (two ports per feature)
- ▶ Management elements:
  - Two dual function base management controllers (BMC) and oscillator cards (OSC) for system control and to provide system clock (N+1 redundancy).
- ▶ CPC drawer power infrastructure consists of the following components:
  - Four Power Supply Units (PSUs) that provide power to the CPC drawer. The loss of one power supply leaves enough power to satisfy the drawer's power requirements (N+1 redundancy). The power supplies can be concurrently removed and replaced individually
  - Five 12v distribution point-of-load (POL) that plug in slots that divide the memory banks.
  - Three Voltage Regulator Modules that plug outside of the memory DIMMs
  - Two Processor Power Control cards to control the five CPC fans at the front of the CPC drawer, and contain the ambient temperature sensors
- ▶ Six SMP-10 connectors (three pairs for redundancy) that provide the CPC drawer to CPC drawer SMP/NUMA communication

The front view of the CPC drawer, which includes the cooling fans, BMC/OSC and processor power cards (PPC), is shown in Figure 2-6. The rear view of a fully populated CPC Drawer is shown in Figure 2-7 on page 29.

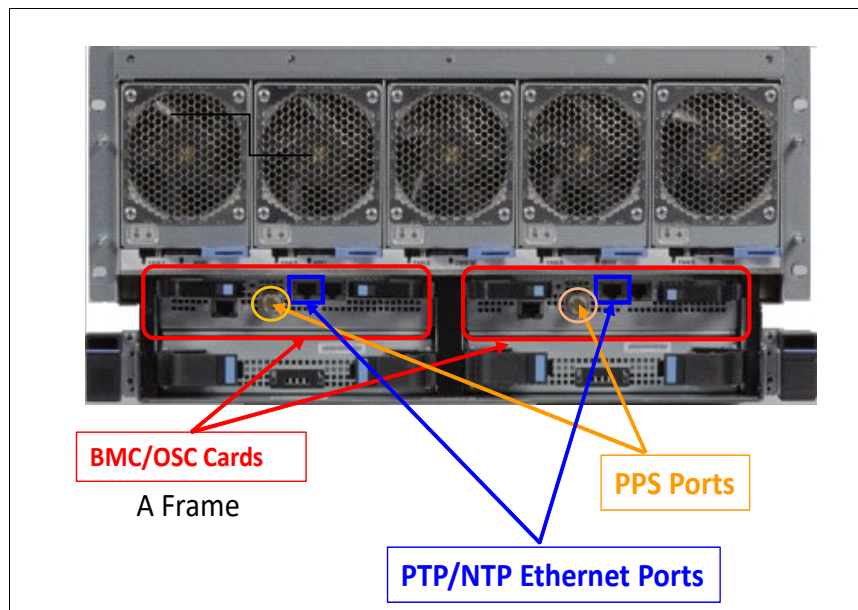


Figure 2-6 Front view of the CPC drawer

<sup>3</sup> Fan-out interfaces will run at PCIe Gen4 speeds.



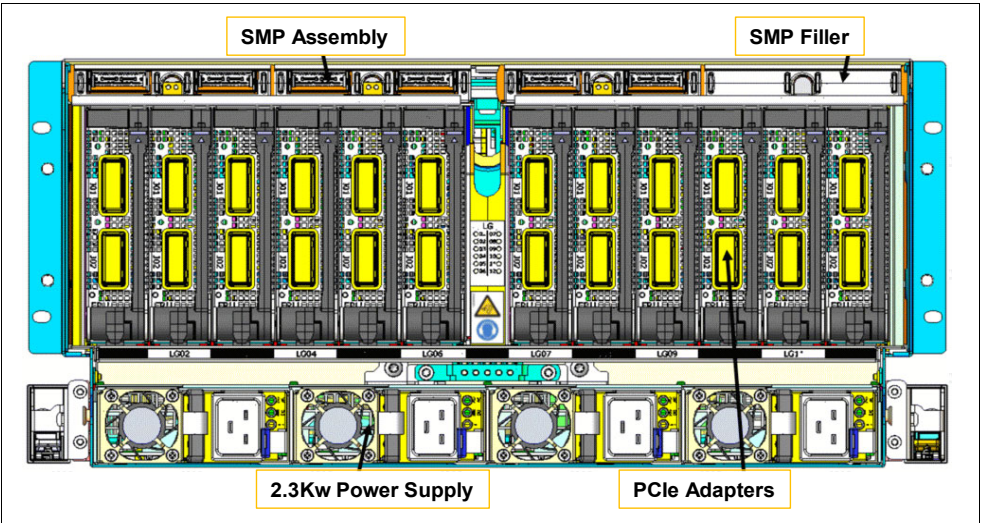


Figure 2-7 Rear view of the CPC drawer

Dual port I/O fan-outs and ICA SR adapters are plugged in specific slots for best performance and availability. Redundant power supplies and six SMP10 ports also are shown. Each pair of SMP10 ports is redundant. If a single cable fails, the repair can be performed concurrently.

The CPC drawer logical structure and component connections are shown in Figure 2-8.

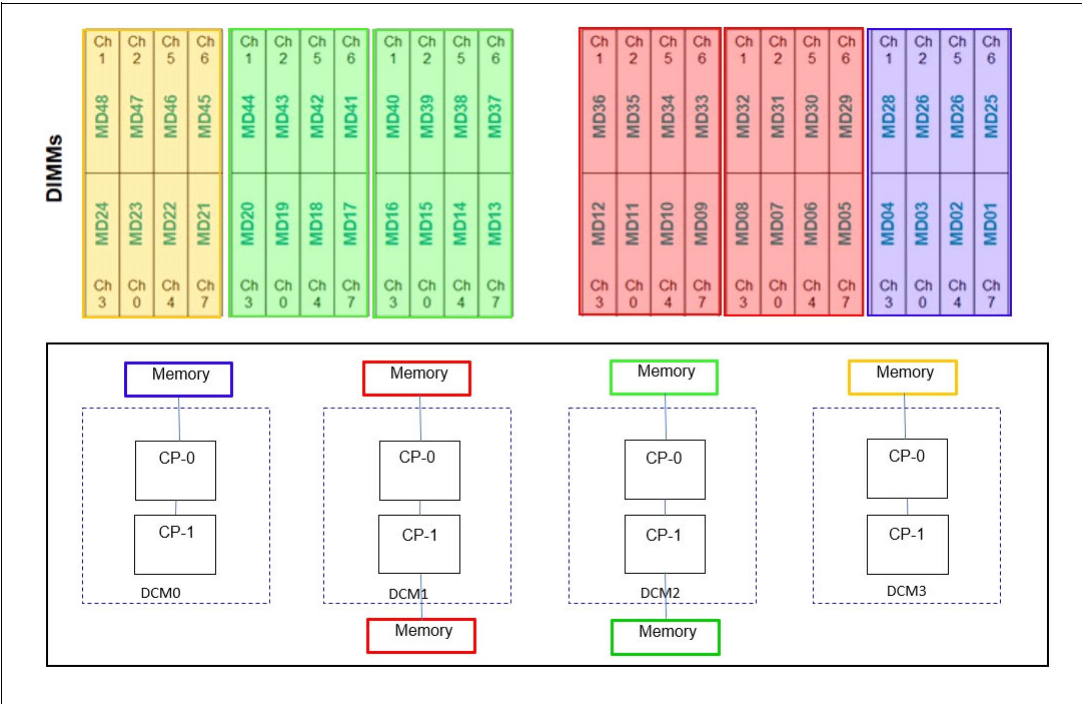


Figure 2-8 CPC Memory - Drawer logical structure

Memory is connected to the DCMs through memory control units (MCUs). Up to six MCUs are available in a CPC drawer (one or two per DCM) and provide the interface to the DIMM controller. A memory controller uses eight DIMM slots.

The buses are organized in the following configurations:



- ▶ The M-bus provides interconnects between PU chips in the same DCM
- ▶ The X-bus provides interconnects between PU chips to each other, in the same drawer
- ▶ The A-bus provides interconnects between different drawers by using SMP-10 cables

## 2.2.1 CPC drawer interconnect topology

The point-to-point SMP10 connection topology for CPC drawers is shown in Figure . Each CPC drawer communicates directly to all of the other CPC drawers DCMs by using point-to-point links.

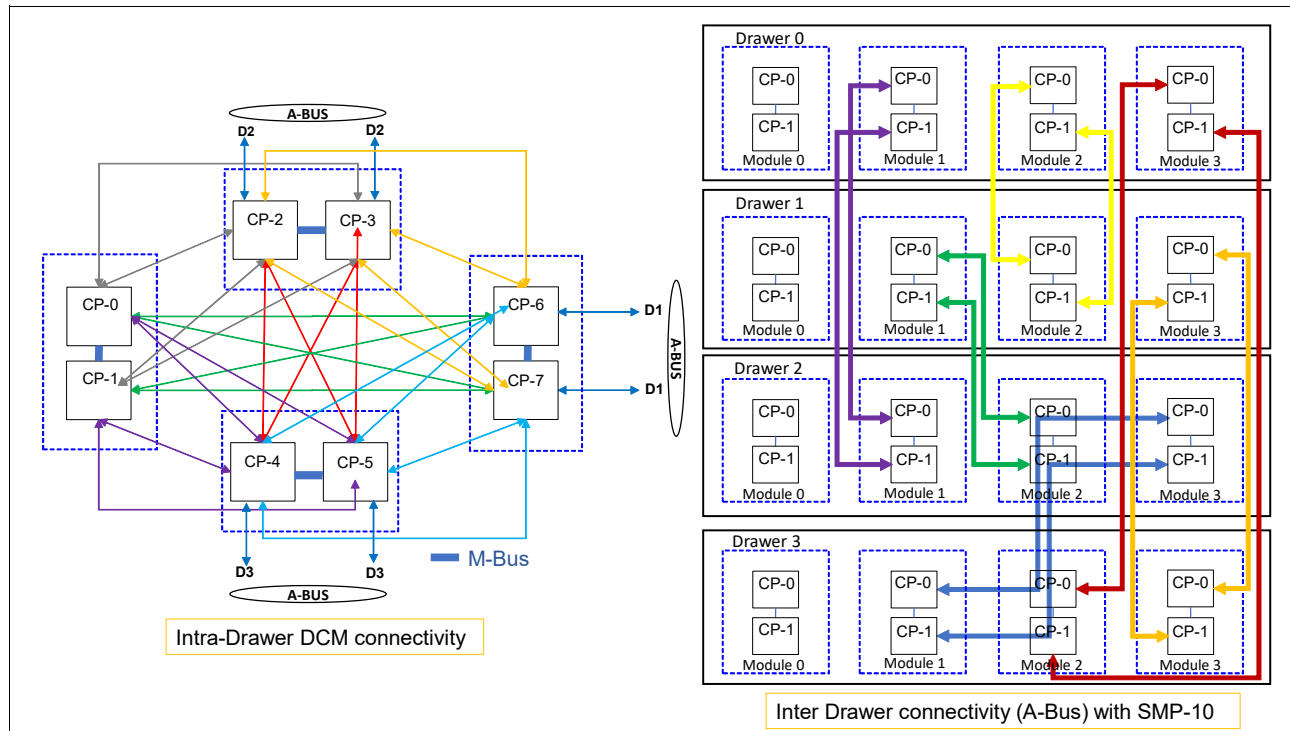


Figure 2-9 Maximum CPC drawer and SMP-10 connection

The CPC drawers that are installed in Frame A and Frame B are populated from bottom to top. The order of CPC drawer installation is listed in Table 2-4.

Table 2-4 CPC drawer installation order and position

CPC drawer <sup>a</sup>	CPC0	CPC1	CPC2	CPC3
Installation order	First	Second	Third	Fourth
Position in Frame A	A10B	A15B	A20B	B10B

a. CPC3 is factory-installed only (no field MES available).

CPC drawer installation in the A frame is concurrent. Nondisruptive addition of CPC1 or CPC2 drawers is possible in the field (MES upgrade) if the reserve features (FC 2933 or FC 2934) are included with the initial system order.

Concurrent drawer repair requires a minimum of two CPC drawers.



## 2.2.2 Oscillator (OSC) and Baseboard Management Controller (BMC) cards

With IBM z17 Model ME1, the oscillator card design and signal distribution scheme are improved; however, the RAS strategy for redundant clock signal and dynamic switchover is unchanged. One primary OSC card and one backup are used. If the primary OSC card fails, the secondary detects the failure, takes over transparently, and continues to provide the clock signal to the CPC.

### Manage System Time

HMC level 2.17.0 (Driver 61) is required to manage system time for IBM z17 Model ME1.

#### *Network Time Protocol*

The SEs provide the Simple Network Time Protocol (SNTP) client function. When Server Time Protocol (STP) is used, the time of an STP-only Coordinated Timing Network (CTN) can be synchronized to the time that is provided by a Network Time Protocol (NTP) server. This configuration allows time-of-day (TOD) synchronization in a heterogeneous platform environment and throughout the LPARs running on the CPC.

#### *Precision Time Protocol*

The time direct to CEC approach introduced with z16, while significantly improving time synchronization accuracy and thus regulatory compliance, creates some potential security vulnerabilities.

The time synchronization security introduced for IBM z17 address the potential security vulnerabilities, and will also serve as a foundation for future IBM Z time synchronization security improvements (such as quantum safe techniques) that will make IBM Z the most secure platform for time synchronization in the industry:

- ▶ Include separation of current IBM Carpo z16 ETS container into individual containers (ETS, NTP/Chrony, PTP4I<sup>4</sup>), restricting and minimizing root access
- ▶ Include full implementation of Chrony (NTP) to enable:
  - authentication and use of NTP algorithms to mitigate MiTM<sup>5</sup> attacks
- ▶ Include implementation of NTS for NTP and PTP

For IBM z17 Model ME1, Precision Time Protocol (PTP, IEEE 1588) can be used as an external time source for IBM Z Server Time Protocol (STP) for an IBM Z Coordinated Timing Network (CTN). The initial implementation for PTP connectivity was provided by using the IBM Z Support Element (SE).

As with IBM z16, on IBM z17, the external time source (PTP or NTP) is connected directly to the CPC and bypasses the SE connection. This IBM z17 feature allows more accurate sync with the external time source.

The accuracy of an STP-only CTN is improved by using an NTP or PTP server with the PPS output signal as the External Time Source (ETS). Devices with PPS output are available from several vendors that offer network timing solutions.

Consider the following points:

- ▶ A redesigned card combines the BMC and OSC that are implemented with IBM z17 Model ME1. The internal physical cards (BMC and OSC) are separate, but combined as a single FRU because of a packaging design.

<sup>4</sup> PTP4I - its the main program that implements PTP according to IEEE standard 1588 for Linux.

<sup>5</sup> MiTM - Man in the middle: a cyberattack where a hacker intercepts and alter communications between two parties to steal sensitive information.



- ▶ Two local redundant oscillator cards are available per CPC drawer, each with one PPS port and one ETS port (RJ45 Ethernet, for both PTP and NTP).
- ▶ Current design requires Pulse Per Second signal use for providing maximum time accuracy for both NTP and PTP.
- ▶ An augmented precision oscillator (20 Parts Per Million [PPM] versus 50 PPM on previous systems) is used.
- ▶ The following PPS plugging rules apply (see Figure on page 33):
  - Single CPC drawer plugs left and right OSC PPS coaxial connectors.
  - Multi-drawer plug CPC0 left OSC PPS and CPC1 left OSC PPS coaxial connectors.
  - Multi-drawer plug CPC0 left OSC ETS1 J03 ethernet and CPC1 right OSC ETS2 J03 ethernet connectors
  - Cables are routed from rear to front by using a pass-through hole in the frame, and under the CPC bezel by using a right-angle Bayonet Neill-Concelman (BNC) connector that provides the pulse per second (PPS) input for synchronization to an external time source with PPS output.
  - Cables are supplied by the customer.
  - Connected PPS ports must be assigned in the Manage System Time menus on the HMC.

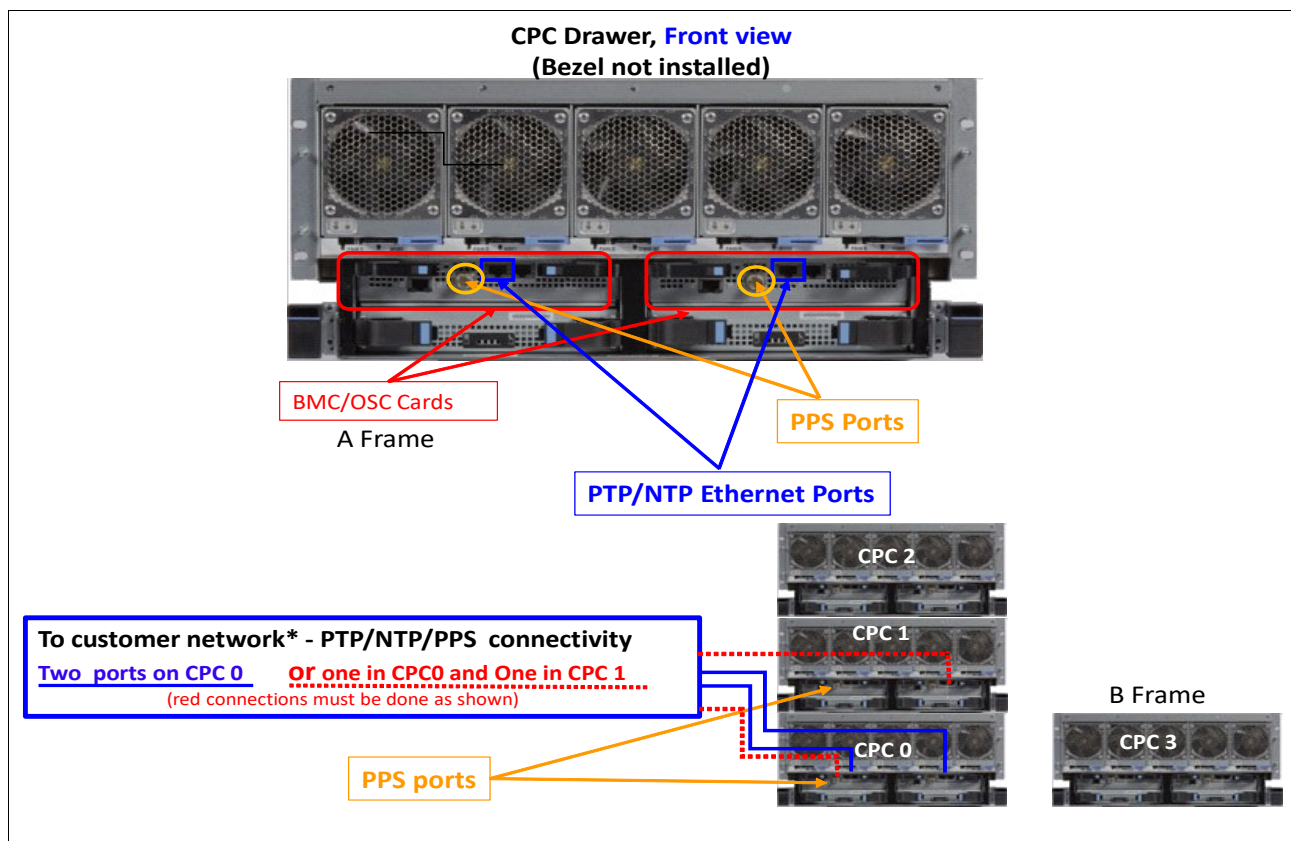


Figure 2-10 Recommended PPS cabling



**Tip:** STP is available as FC 1021. It is implemented in the Licensed Internal Code (LIC), and allows servers to maintain time synchronization with each other and synchronization to an ETS. In a multi-server STP Coordinated Timing Network (CTN) coupling/timing links are required for STP communication

For more information, see *IBM Z Server Time Protocol Guide*, [SG24-8480](#).

## 2.2.3 System control

The various system elements are managed through the base management cards (BMCs). The BMC is the replacement for the previous Flexible Support Processors (FSPs) that were used in previous systems.

With IBM z17 Model ME1, the CPC drawer BMC is combined with the Oscillator card in a single Field Replaceable Unit (FRU). Two combined BMC/OSC cards are used per CPC drawer.

Also, the PCIe+ I/O drawer has an improved BMC. Each BMC card has one Ethernet port that connects to the internal Ethernet LANs through the internal network switches (SW1, SW2, and SW3, SW4, if configured). The BMCs communicate with the SEs and provide a subsystem interface (SSI) for controlling components.

An overview of the system control design is shown in Figure 2-11.

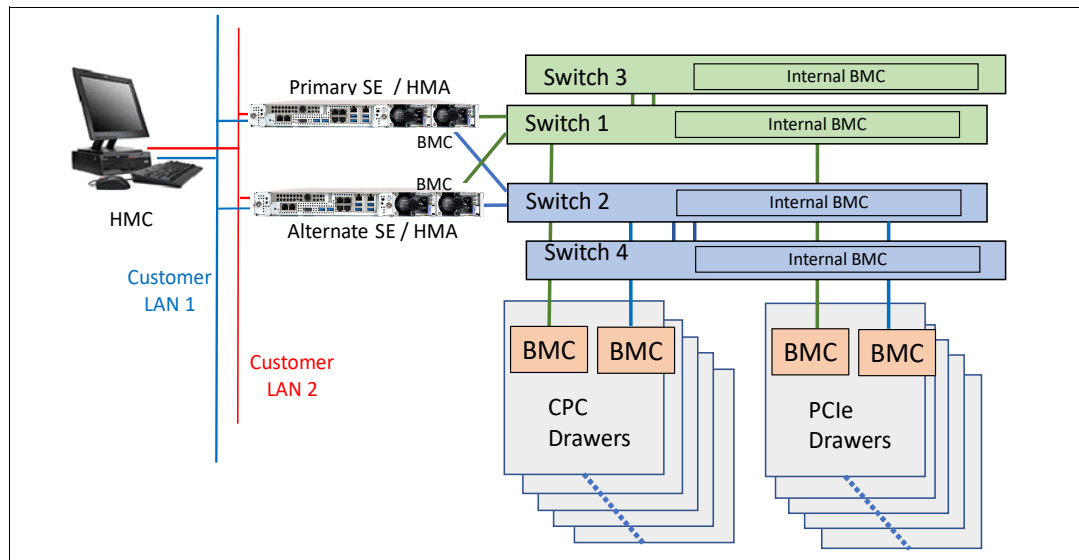


Figure 2-11 Conceptual overview of system control element network

**Note:** The maximum IBM z17 ME1 system configuration features four GbE switches, four CPC drawers, and up to 12 PCIe I/O drawers.



A typical BMC operation is to control a power supply. An SE sends a command to the BMC to start the power supply. The BMC cycles the various components of the power supply, monitors the success of each step and the resulting voltages, and reports this status to the SE.

SEs are duplexed ( $N+1$ ), and each element has at least one BMC. Two internal Ethernet LAN switches and two SEs, for redundancy, and crossover connectivity between the LANs, are configured so that both SEs can operate on both LANs.

The Hardware Management Appliances (HMAs) and SEs are connected directly to one or two Ethernet Customer LANs. One or more HMCs can be used.

With IBM z17 current ordered systems, the Hardware Management Appliance (HMA, FC 0355) is the only HMC-orderable feature. The HMA is running on the same hardware (virtual appliance) with the Support Elements (the SE code runs as a guest of the HMA).

## 2.2.4 CPC drawer power

The IBM z17 ME1 power for the CPC drawer is similar to IBM z16 A01. It uses the following combinations of Power Supply Units (PSUs), Points of Load (POLs), Voltage Regulator Modules (VRMs), Fans, and Processor Power Control (PPC) cards:

- ▶ PSUs: Four new 2.3 KW PSUs provide 12 VDC bulk and 12 VDC standby power and are installed at the rear of the CPC.
- ▶ POLs: Five Point of Load N+2 Redundant cards are installed next to the Memory DIMMs.
- ▶ VRMs: Three Voltage Regulator Sticks (N+2 redundancy)
- ▶ Fans: Five fan assemblies are installed in the front of the CPC drawer
- ▶ PPCs: Two redundant processor power and control cards connect to the CPC trail board. The control function is powered from 12 V standby that is provided by the PSU. The PPC card also includes pressure, temperature, and humidity sensors.

## 2.3 Dual chip modules

The DCM is a multi-layer metal substrate module that holds two PU chips. The PU chip size is  $565.6 \text{ mm}^2$  (23.75mm x 23.82mm) and has 43 billion transistors. Each CPC drawer has four PU DCMs.

The PU DCM is shown in Figure 2-12 on page 35. The DCM features a thermal cap that is placed over the chips. Each DCM is water-cooled by way of a cold plate manifold assembly.



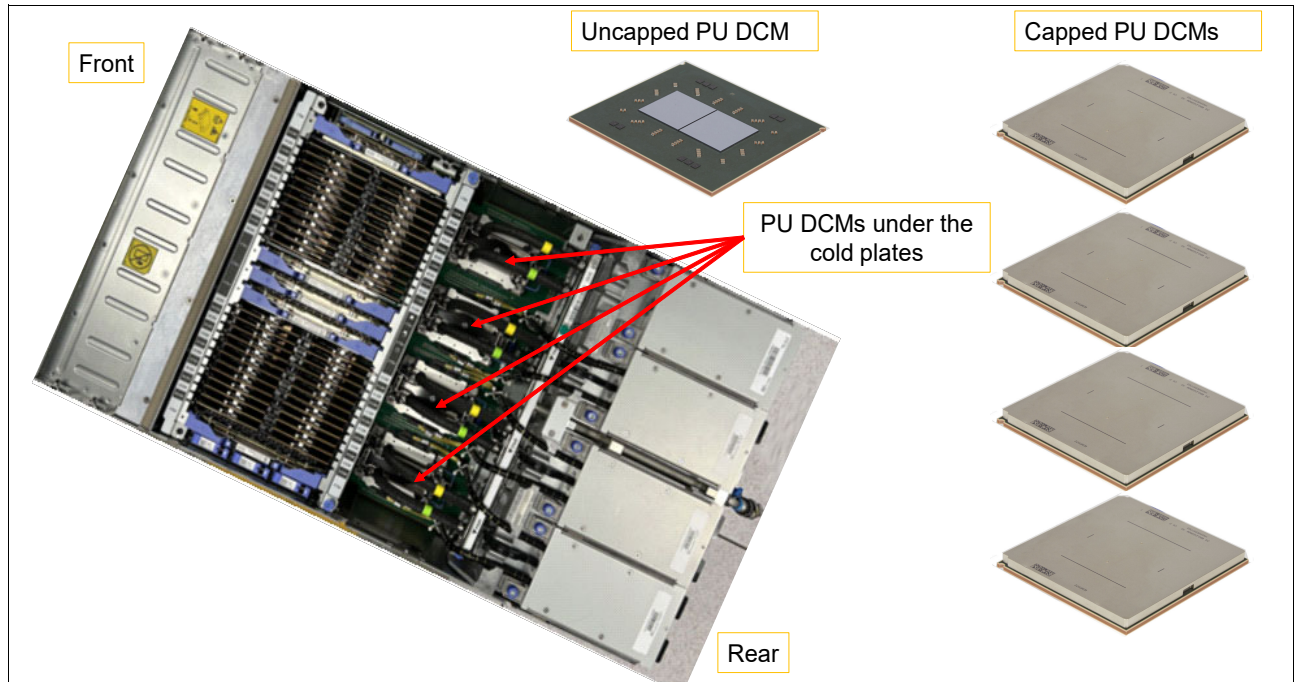


Figure 2-12 Dual chip modules (PU DCM)

Each DCM socket size is 71.5 mm X 79 mm and holds two 5nm FinFET PU chips measuring 23.75 mm x 23.82 mm (565 mm<sup>2</sup>).

The DCMs are each plugged into a socket that is part of the CPC drawer packaging. Each DCM is cooled by water flowing through a manifold assembly where cold plates are attached and used to help remove the heat from the processor chips.



### 2.3.1 Processor unit chip

A floor plan representation of the IBM z17 Telum II PU chip is shown in Figure 2-13.

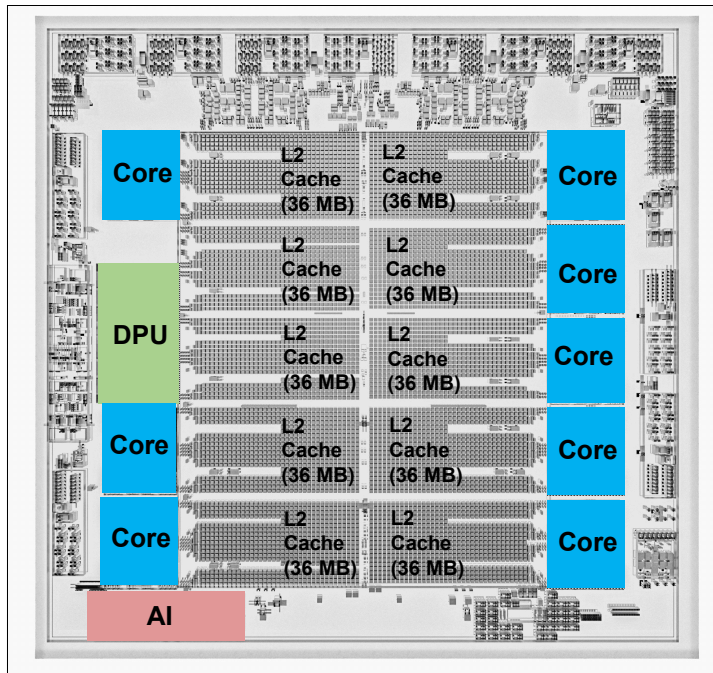


Figure 2-13 Single PU chip floor plan

The IBM z17 Model ME1 PU chip (two PUs packaged on each DCM) includes the following features and improvements:

- ▶ 5 nm silicon lithography wafer technology
- ▶ 565.6 mm<sup>2</sup> chip size
- ▶ 43 Billion transistors
- ▶ 5.5 GHz base clock frequency
- ▶ Nine core design with increased on-chip cache sizes
- ▶ Two PCIe Gen5 interfaces
- ▶ DDR5 memory controller
- ▶ AES256 Memory Encryption
- ▶ Two M-Bus to support DCM internal chip to chip connectivity
- ▶ Six X-Bus (12 per DCM) to support DCM to DCM connectivity in the CPC drawer
- ▶ One A-Bus to support drawer to drawer connectivity
- ▶ New cache structure design compared to IBM z16 PU:
  - L1D (data) and L1I (instruction) caches - ON-core (128 KB each)
  - 10 x L2 - 36 MB dense SRAM - outside the core, semi-private to the core
  - L3 (virtual) - up to 360 MB
  - L4 (virtual) - up to 2.88 GB
- ▶ 40% more cache per chip
- ▶ DPU - Data Processing Unit
  - 32 small processor cores



- Protocol Accelerators
- Interface to one of the L2 caches
- PCIe Gen5 x16 (bifurcated as Gen4 x8 used for I/O)
- ▶ Improved branch prediction design by using SRAM
- ▶ Improved Gzip Compression
- ▶ 2nd Generation AI Unit
  - On chip Artificial Intelligence Unit AIU: Deep learning focus for AI inference
  - AIU Intelligent Routing
  - Hardware acceleration for NLP and transformer models
  - Pounces AI ecosystem support
- ▶ Significant architecture changes: COBOL compiler and more
- ▶ Speeds and Feeds
  - 10% increase in performance per thread
  - 20% increase in standard models drawer capacity
  - 14% increase in max config system capacity
  - Increased Memory capacity to 64 TB (40% increase over IBM z16)

### 2.3.2 Processor unit (core)

Each processor unit, or core, is a superscalar and out-of-order processor that supports 10 concurrent issues to execution units in a single CPU cycle. Figure 2-14 shows the core floor plan, which contains the following units:

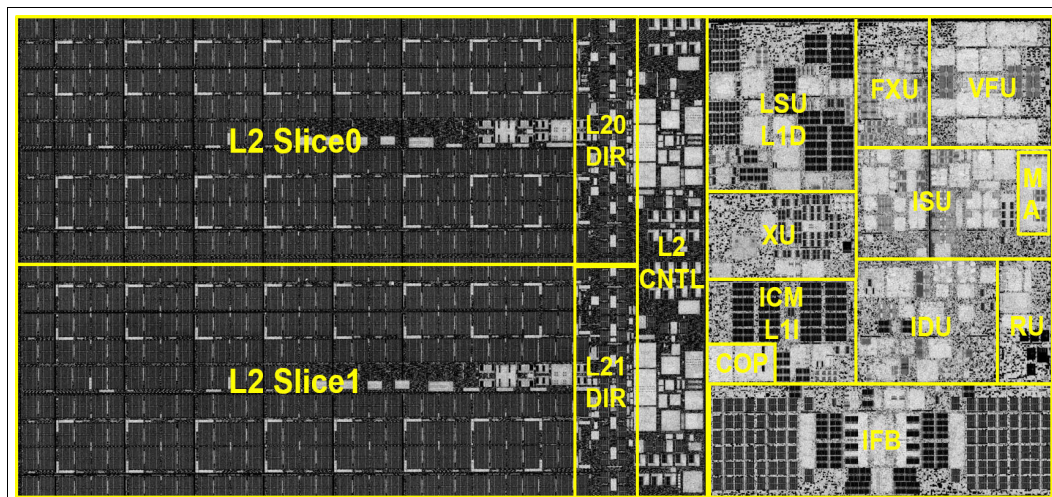


Figure 2-14 Processor core floor plan

- ▶ Fixed-point unit (FXU): The FXU handles fixed-point arithmetic.
- ▶ Load-store unit (LSU): The LSU contains the data cache. It is responsible for handling all types of operand accesses of all lengths, modes, and formats as defined in the z/Architecture.
- ▶ Instruction fetch and branch (IFB) (prediction) and Instruction cache and merge (ICM). These two sub units (IFB and ICM) contain the instruction cache, branch prediction logic, instruction fetching controls, and buffers. Its relative size is the result of the elaborate branch prediction.
- ▶ L1D (data) and L1I (instructions) are incorporated into the LSU and ICM, respectively.



- ▶ Instruction decode unit (IDU): The IDU is fed from the IFU buffers and is responsible for parsing and decoding of all z/Architecture operation codes.
- ▶ Translation unit (XU): The XU has a large translation lookaside buffer (TLB) and the Dynamic Address Translation (DAT) function that handles the dynamic translation of logical to physical addresses.
- ▶ Instruction sequence unit (ISU): This unit enables the out-of-order (OoO) pipeline. It tracks register names, OoO instruction dependency, and handling of instruction resource dispatch.
- ▶ Instruction fetching unit (IFU) (prediction): These units contain the instruction cache, branch prediction logic, instruction fetching controls, and buffers. Its relative size is the result of the elaborate branch prediction design.
- ▶ Recovery unit (RU): The RU keeps a copy of the complete state of the system that includes all registers, collects hardware fault signals, and manages the hardware recovery actions.
- ▶ Dedicated Co-Processor (CoP): The dedicated coprocessor is responsible for data compression and encryption functions for each core.
- ▶ Core pervasive unit (PC) for instrumentation and error collection.
- ▶ Modulo arithmetic (MA) unit: Support for Elliptic Curve Cryptography

#### Vector and Floating point Units (VFU):

- BFU: Binary floating point unit
  - DFU: Decimal floating point unit
  - DFx: Decimal fixed-point unit
  - FPD: Floating point divide unit
  - VXx: Vector fixed-point unit
  - VXs: Vector string unit
  - VXP: Vector permute unit
  - Vxm: Vector multiply unit
- ▶ L2 – Level 2 cache

### 2.3.3 PU characterization

The PUs are characterized for client use. The characterized PUs can be used in general to run supported operating systems, such as z/OS, z/VM, and Linux on IBM Z. They also can run specific workloads, such as Java, XML services, IPsec, and some Db2 workloads, or clustering functions, such as the Coupling Facility Control Code (CFCC).

The maximum number of characterizable PUs depends on the IBM z17 Model ME1 CPC drawer feature code. Some PUs are characterized for system use; some are characterized for client workload use.

By default, two spare PUs are available to assume the function of failed PUs. The maximum number of PUs that can be characterized for client use are listed in Table 2-5 on page 39.



Table 2-5 PU characterization

Feature	CPs	IFLs	Unassigned IFLs	zIIPs	ICFs	IFPs	Std SAP <sup>a</sup> s	Spare PUs
Max43	0 - 43	0 - 43	0 - 42	0 - 42	0 - 43	2	5	2
Max90	0 - 90	0 - 90	0 - 89	0 - 89	0 - 90		10	
Max136	0 - 136	0 - 136	0 - 135	0 - 135	0 - 136		16	
Max183	0 - 183	0 - 183	0 - 182	0 - 182	0 - 183		21	
Max208	0 - 208	0 - 208	0 - 207	0 - 207	0 - 208		24	

a. Additional SAPs option is not available for the IBM z17

The rule for the CP to zIIP purchase ratio, that for every CP purchased, up to two zIIPs can be purchased has been removed starting with IBM z16.

Converting a PU from one type to any other type is possible by using the Dynamic Processor Unit Reassignment process. These conversions occur concurrently with the system operation.

**Note:** The addition of ICFs, IFLs, and zIIPs to the IBM z17 Model ME1 does not change the system capacity setting or its million service units (MSU) rating.

## 2.3.4 Cache level structure

The cache structure for the IBM z16 A01 is shown in Figure 2-15. Figure 2-16 shows the IBM z17 ME1 cache structure.

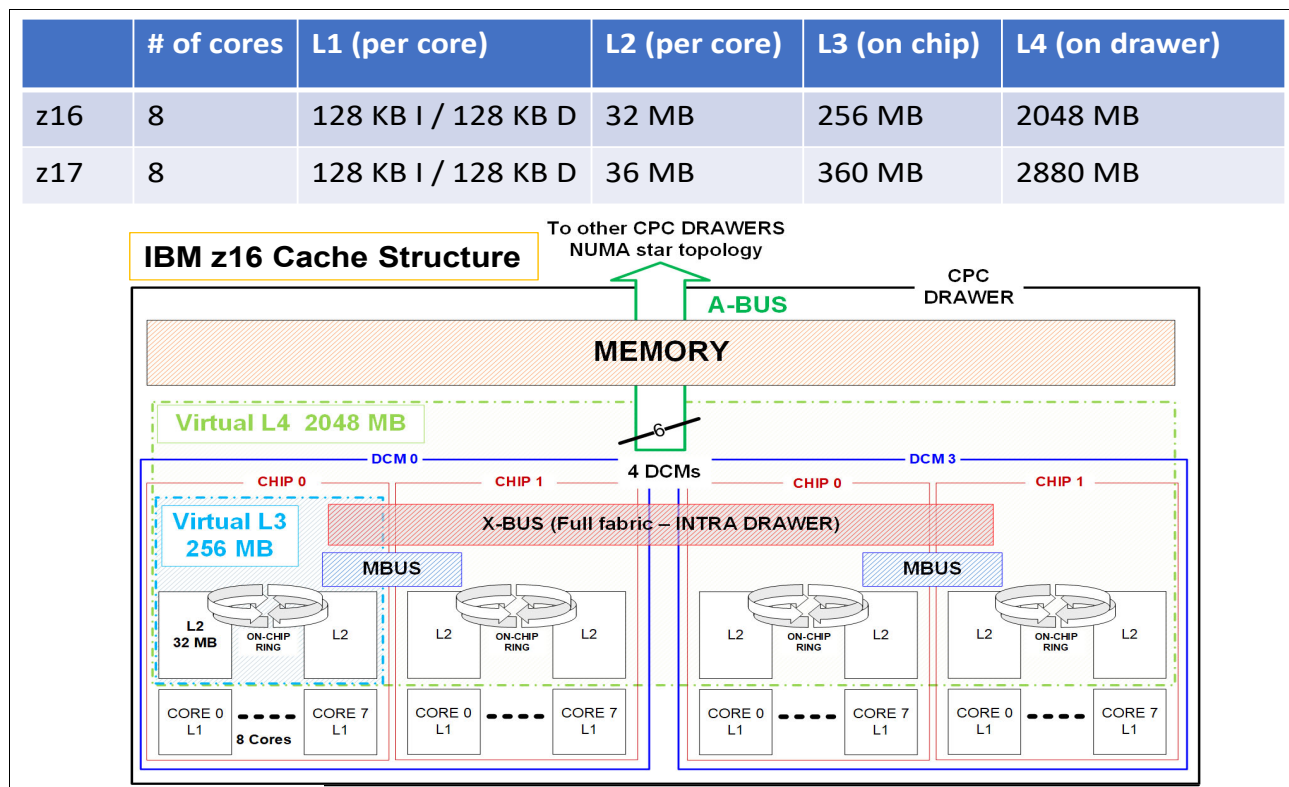


Figure 2-15 Cache structure IBM z16



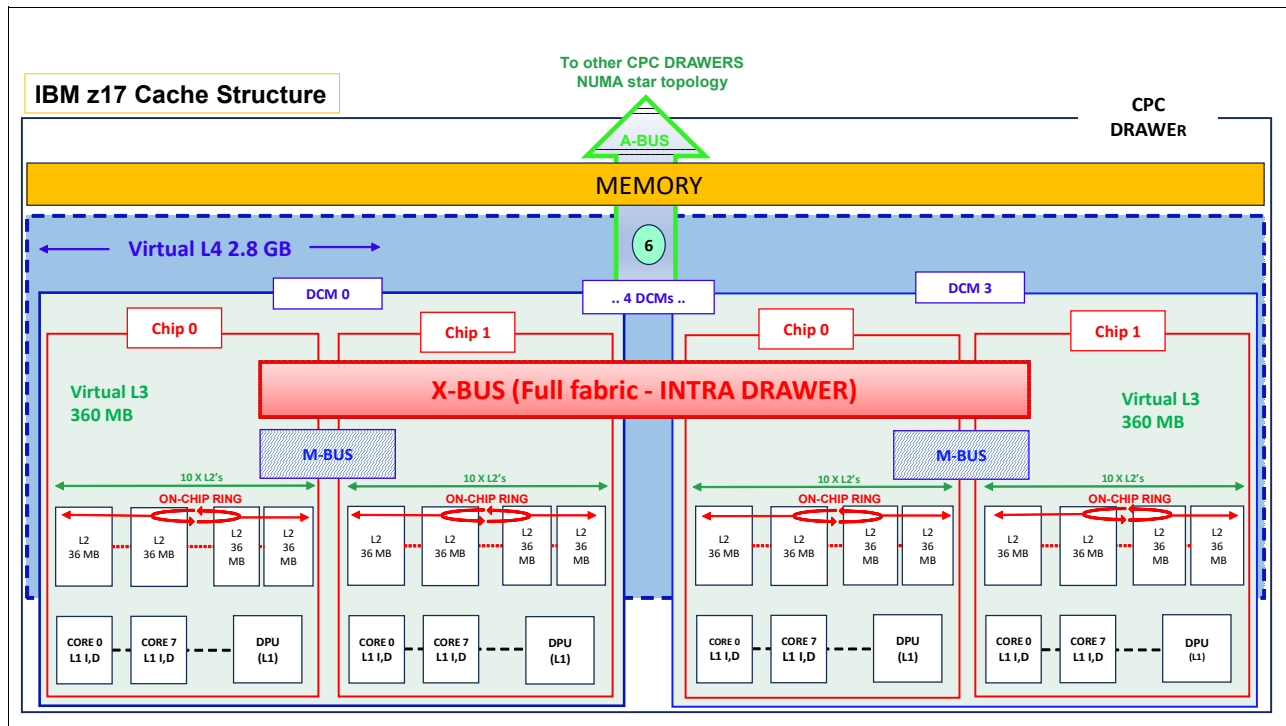


Figure 2-16 Cache structure on IBM z17

## 2.4 PCIe+ I/O drawer

As shown in Figure 2-17, each PCIe+ I/O drawer includes 16 slots to support the PCIe I/O infrastructure with a bandwidth of 16 GBps and includes the following features:

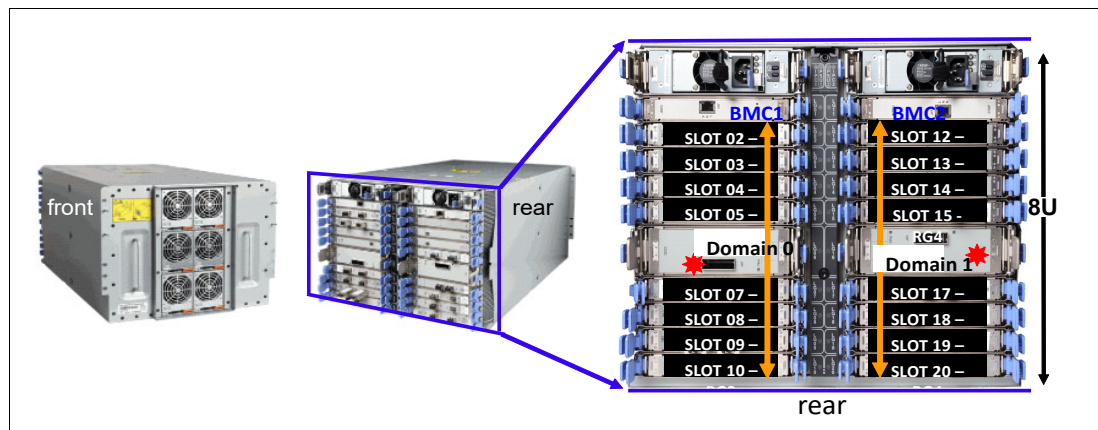


Figure 2-17 PCIe I/O drawer (front and rear view)

- ▶ A total of 16 I/O cards are spread over two I/O domains (0 and 1):
  - Each I/O slot reserves up to four PCHIDs.
  - Left-side slots are numbered LG01 - LG10 and right side slots are numbered LG11 - LG20 from the rear of the rack. A location and LED identifier panel is at the center of the drawer.



- With IBM z17 Model ME1, the numbering of the PCHIDs by location for the I/O drawers has changed. I/O drawers in the Z frame will start with PCHID 100 and continue incrementing assigning PCHIDs in a specific sequence depending on the number of drawers and location.

For more information about examples of the various configurations, see Appendix E, “Frame configurations with Power Distribution Units” on page 551.

- ▶ Two PCIe+ Gen4 Switch Cards provide connectivity to the PCIe+ Gen4 Fanouts that are installed in the CPC drawers.
- ▶ Each I/O drawer domain has four dedicated support partitions (two per domain) to manage the native PCIe cards.
- ▶ Two Baseboard Management Controllers (BMC) cards are used to control the drawer function.
- ▶ Redundant N+1 power supplies (two) are mounted on the rear and redundant blowers (six) are mounted on the front.

The following example shows a maximum configuration of CPC drawers and I/O features ordered, the layout of the PCIe+ I/O drawers, and order of install, see Figure 2-18.

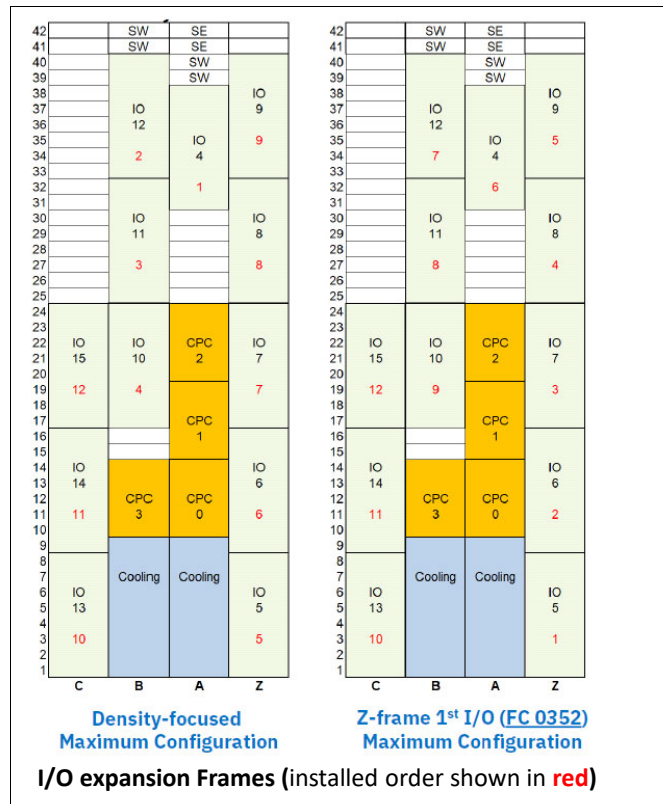


Figure 2-18 I/O drawer installation order for an IBM z17 with four frames

- ▶ The IBM z17 introduces a new Feature Code (FC 0352 - Z-frame 1st I/O). This FC will allow clients to prepare to carry their Z-Frame forward into their future IBM z System.
- ▶ Feature codes used for reserving CPC drawer slots for future growth is possible by ordering FC 2933 (CPC1) and FC2934 (CPC2). Care must be taken to reserve CPC drawers. If not ordered, the open space may be populated with I/O drawers.



The I/O drawer plugging sequence for z17 *without* Z-frame 1st I/O Placement FC 0352 is shown in Table 2-6. and with Z-frame 1st I/O Placement FC 0352 shown in

Table 2-6

I/O Drawer Plug Sequence	I/O Drawer Location						
	Max43 1-CPC	Max43 1-CPC CPC1 Reserved	Max43 1-CPC CPC1/CPC2 Reserved	Max90 2-CPC	Max90 2-CPC CPC2 Reserved	Max136 3-CPC	Max183 Max208 4-CPC
1	A31B	A31B	A31B	A31B	A31B	A31B	A31B
2	A23B	A23B	Z01B	A23B	Z01B	Z01B	B33B
3	A15B	Z01B	Z09B	Z01B	Z09B	Z09B	B25B
4	Z01B	Z09B	Z17B	Z09B	Z17B	Z17B	B17B
5	Z09B	Z17B	Z25B	Z17B	Z25B	Z25B	Z01B
6	Z17B	Z25B	Z33B	Z25B	Z33B	Z33B	Z09B
7				Z33B	C01B	C01B	Z17B
8				C01B	C09B	C09B	Z25B
9				C09B	C17B	C17B	Z33B
10				C17B	C25B	C25B	C01B
11				C25B	C33B	C33B	C09B
12				C33B			C17B

The I/O drawer plugging sequence for z17 *with* Z-frame 1st I/O Placement FC 0352 is shown in Table 2-7.

Table 2-7 I/O Drawer plugging locations *with* first FC0352

I/O Drawer Plug Sequence	I/O Drawer Location						
	Max43 1-CPC	Max43 1-CPC & CPC1 Reserved	Max43 1-CPC & CPC1/CPC2 Reserved	Max90 2-CPC	Max90 2-CPC & CPC2 Reserved	Max136 3-CPC	Max183 Max208 4-CPC
1	Z01B	Z01B	Z01B	Z01B	Z01B	Z01B	Z01B
2	Z09B	Z09B	Z09B	Z09B	Z09B	Z09B	Z09B
3	Z17B	Z17B	Z17B	Z17B	Z17B	Z17B	Z17B
4	Z25B	Z25B	Z25B	Z25B	Z25B	Z25B	Z25B
5	Z33B	Z33B	Z33B	Z33B	Z33B	Z33B	Z33B
6	A31B	A31B	A31B	A31B	A31B	A31B	A31B
7				A23B	C01B	C01B	B33B
8				C01B	C09B	C09B	B25B
9				C09B	C17B	C17B	B17B
10				C17B	C25B	C25B	C01B
11				C25B	C33B	C33B	C09B
12				C33B			C17B



## Consideration for PCHID numbering:

New for IBM z17, the PCHID range assignment will be fixed for specific I/O drawer locations, and other I/O drawers will be different dependant on the configuration. Table 2-8 demonstrates the PCHID range for all I/O Drawers by configuration with the number of CPC and I/O Drawers, and without or with FC 0352 Z-frame 1st.

Table 2-8

PCHID Range	I/O Drawer Location				
	Max43 1-CPC w/o Z-FIRST	Max43 1-CPC w/ Z-FIRST	Max90 2-CPC	Max136 3-CPC	Max183 Max208 4-CPC
100-13F	Z01B	Z01B	Z01B	Z01B	Z01B
140-17F	Z09B	Z09B	Z09B	Z09B	Z09B
180-1BF	Z17B	Z17B	Z17B	Z17B	Z17B
1C0-1FF		Z25B	Z25B	Z25B	Z25B
200-23F		Z33B	Z33B	Z33B	Z33B
240-27F	A31B	A31B	A31B	A31B	A31B
280-2BF	A23B		A23B		B33B
2C0-2FF	A15B		C01B	C01B	C01B
300-33F			C09B	C09B	C09B
340-37F			C17B	C17B	C17B
380-3BF			C25B	C25B	B25B
3C0-3FF			C33B	C33B	B17B

- Max43 configurations are limited to maximum six I/O Drawers
- PCHID assignment will be fixed for all the I/O drawers in the Z frame, A31B, C09B and C17B. All other drawers may inherit a different PCHID assignment depending on the configuration.

## Consideration for PCHID identification:

For IBM z17 Model ME1, the orientation of the PCIe features in a I/O drawer are horizontal. The top of the adapter is closest to the center of the drawer for the left and right side of the drawer.

The port and PCHID layout where the top of the adapter (port D1) is closest to the location panel on both sides of the drawer are shown in Figure 2-19 on page 44.



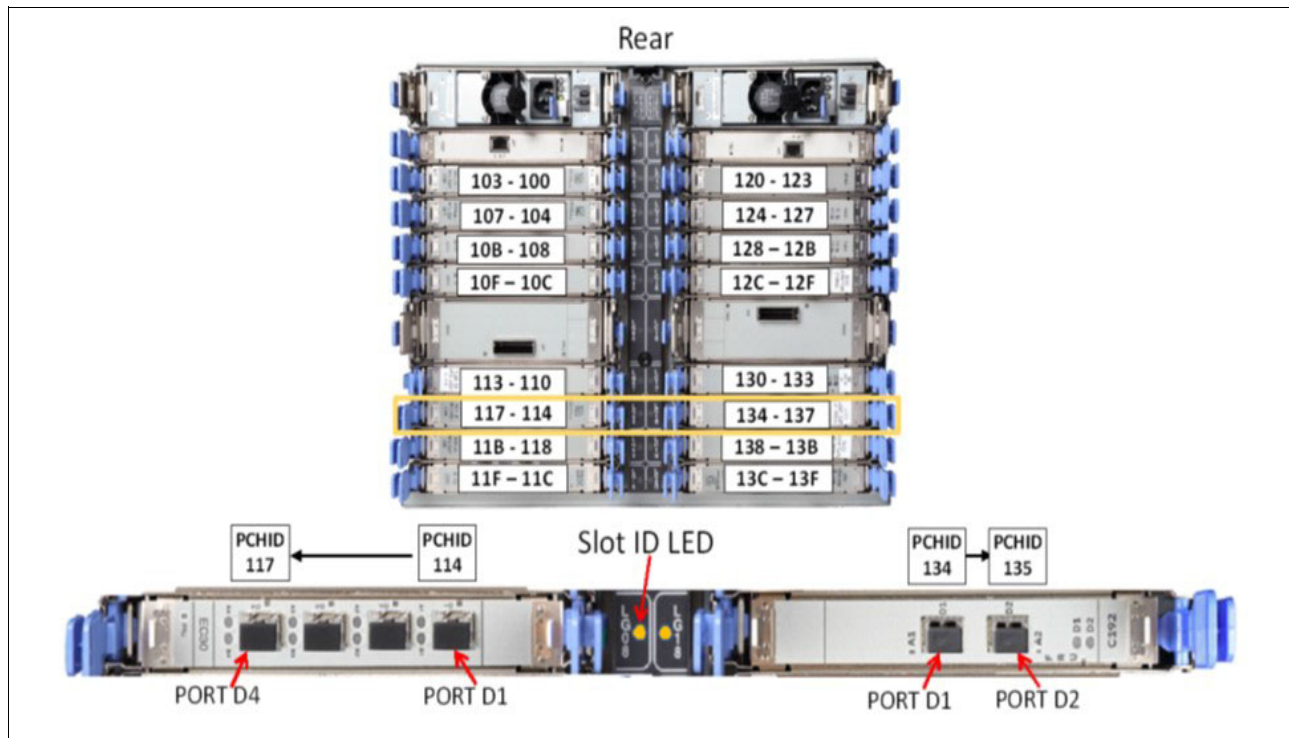


Figure 2-19 I/O feature orientation in PCIe I/O drawer (rear view)

**Note:** The Configuration Mapping Tool (available on ResourceLink) can be used to print a CHPID Report that displays the drawer and PCHID/CHPID layout.

## 2.5 Memory

The maximum physical memory size is directly related to the number of CPC drawers in the system. Each CPC drawer can contain up to 14 TB of customer memory, for a total of 64 TB of memory per system.

The minimum and maximum memory sizes that you can order for each IBM z17 Model ME1 feature are listed in Table 2-9.

Table 2-9 Purchased memory that is available for assignment to LPARs

Feature	Number of CPC drawers	Customer memory GB	Flexible memory GB
Max43	1	512 - 15616	NA
Max90	2	512- 31488	512 - 15616
Max136	3	512 - 47872	512 - 31488
Max183	4	512 - 64256	512- 47872
Max208	4	512 - 64256	512 - 47872

The following memory types are available:



- ▶ Purchased: Memory that is available for assignment to LPARs.
- ▶ Hardware System Area (HSA): Standard 884 GB of addressable memory for system use outside of customer memory.
- ▶ Standard: Provides minimum physical memory that is required to hold customer purchase memory plus 884 GB HSA.
- ▶ Flexible: Provides more physical memory that is needed to support that activation of base customer memory and HSA on a multiple CPC drawer IBM z17 with one drawer out of service (concurrent drawer replacement; not available on Max43 feature).

The memory granularity, which is based on the installed customer memory, is listed in Table 2-10.

*Table 2-10 Customer offering memory increments*

Memory increment (GB)	Offered memory sizes (GB)
64	512 - 768
128	896 - 2048
256	2304 - 3840
512	4352 - 17152
1024	18176 - 33536
2048	35584 - 64256



## 2.5.1 Memory subsystem topology

The IBM z17 memory subsystem uses high-speed, differential-ended communications memory channels to link a host memory to the main memory storage devices.

The CPC drawer memory topology of an IBM z17 is shown in Figure 2-20.

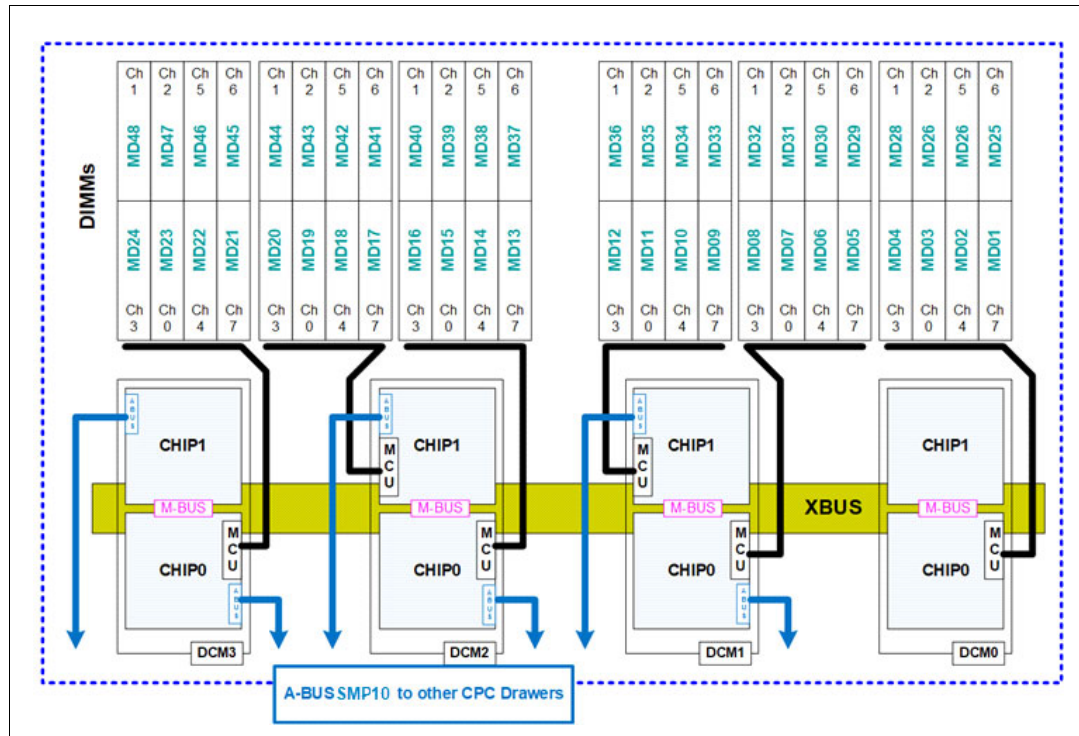


Figure 2-20 CPC drawer memory topology at maximum configuration

Consider the following points regarding the topology:

- ▶ Six memory controllers (MCUs) populated per drawer, one per PU chip (two chips do not have their MCUs populated)
- ▶ Each memory controller supports eight DIMM slots
- ▶ The eight DIMMs on a controller are all the same size
- ▶ Four, five, or six memory controllers per drawer are populated (32, 40, or 48 DIMMs)
- ▶ Different memory controllers might have different size DIMMs
- ▶ Features with different DIMMs sizes can be mixed in the same drawer
- ▶ The eight DIMMs per MCU must be the same size
- ▶ Addressable memory is required for partitions and HSA

## 2.5.2 Redundant array of independent memory (RAIM)

The IBM z17 Model ME1 server uses the enhanced RAIM design. Installed Physical Memory (DIMM capacity) in configuration reports is the addressable memory size. Memory is protected by RAIM. DIMM size includes RAIM overhead.

Servers are configured with the most efficient configuration of memory DIMMs that can support Addressable Memory that is required for Customer Ordered Memory plus HSA.



In some cases, Available Addressable Memory might be available to support one or more concurrent LIC CC Customer Memory upgrades with no DIMM changes.

IBM z17 implements enhanced RAIM design that includes the following features:

- ▶ Eight Channel Reed-Solomon<sup>6</sup> RAIM
- ▶ 90 → 80 DRAMs accessed across memory channels (11% reduction, excluding unused spare)
- ▶ Staggered Memory Refresh → Leverage RAIM to hide memory refresh penalty
- ▶ z17 Memory Buffer Chip Interface:
  - Open-top Memory Buffer (OCMB); Fully (meso)synchronous OCMB
  - Lane sparing replaced with lane degrade
  - 256 B fetch support, remove 128 B store support
- ▶ Up 16TB / Drawer (with six MCUs populated with 512GB DIMMs)
- ▶ Host-side Memory Encryption
  - Memory encryption is performed by the memory controllers (MCUs) using an encryption mechanism which combines an encryption key with part of the memory address to encrypt and protect the data. Encryption and decryption occur between RAIM error correction and memory. Encrypting data is done post-RAIM encoding during store, and decryption is done pre-RAIM decoding during fetch operations. IBM z17 implements 256 bit AES keys while IBM z16 uses 128 bit keys. ECC protected Memory Encryption keys are auto generated once per IML, and their values are not exposed.

### 2.5.3 Memory configurations

Memory sizes in each CPC drawer do not have to be similar. RAIM is now built into the bank of eight DIMMs (included in the reported memory; no longer be part of the RAIM memory equation). Consider the following points:

- ▶ A total of 17 configurations that support memory is available per drawer (numbered 22 - 46).
- ▶ Different CPC drawers can contain different amounts of memory.
- ▶ A drawer can have a mix of DIMM sizes (see Table 2-11).
- ▶ Total memory includes HSA. Customer memory is remaining memory that is available to customer after HSA is subtracted.
- ▶ Each drawer contains minimum 1024 GB of memory.

Supported drawer memory configurations are listed in Table 2-11. Each CPC drawer is included from manufacturing with one of these memory configurations.

Table 2-11 Drawer memory plugging configurations (all values in GB)

CFG number	M0CP0 MD01-04 and MD25-28	M1CP0 MD05-08 and MD29-32	M1CP1 MD09-12 and MD33-36	M2CP0 MD13-16 and MD37-40	M2CP1 MD17-20 and MD41-44	M3CP0 MD21-24 and MD45-48	Physical	INC	- HSA 884 GB
22	0	32	32	32	32	0	1024		140
23	0	64	32	64	32	0	1536	512	652

<sup>6</sup> Reed-Solomon error correction



CFG number	M0CP0 MD01-04 and MD25-28	M1CP0 MD05-08 and MD29-32	M1CP1 MD09-12 and MD33-36	M2CP0 MD13-16 and MD37-40	M2CP1 MD17-20 and MD41-44	M3CP0 MD21-24 and MD45-48	Physical	INC	- HSA 884 GB
24	0	64	64	64	64	0	2048	512	1164
25	64	64	64	64	64	0	2560	512	1676
26	64	64	64	64	64	64	3072	512	2188
27	128	64	64	64	64	64	3584	512	2700
28	128	64	64	64	64	128	4096	512	3212
29	128	128	64	64	64	128	4608	512	3724
30	128	128	64	64	128	128	5120	512	4236
31	128	128	128	128	128	128	6144	1024	5260
32	256	128	128	128	128	128	7168	1024	6284
36	256	128	128	128	128	256	8192	1024	7308
37	256	256	128	128	128	256	9216	1024	8332
38	256	256	128	128	256	256	10240	1024	9356
39	256	256	256	256	256	256	12288	1024	11404
45	256	512	512	256	256	0	14336	1024	13452
46	256	512	512	512	256	0	16384	1024	15500


Consider the following points:

- ▶ A CPC drawer contains a minimum of (32, 4x8) 32 GB DIMMs as listed in drawer configuration number 22 in Table 2-11 on page 47.
- ▶ A CPC drawer can have more memory that is installed than what is enabled for customer use. The amount of memory that can be enabled by the customer is the total physically installed memory minus the 884 GB of HSA memory.
- ▶ A CPC drawer can have available unused memory, which can be ordered as a memory upgrade and enabled by LIC-CC concurrently without DIMM changes.
- ▶ DIMM changes require a disruptive power-on reset (POR) on IBM z17 ME1 with a single CPC drawer. DIMM changes can be done concurrently on IBM z17 models with multiple CPC drawers using Enhanced Drawer Availability (EDA).

DIMM plugging for the configurations in each CPC drawer do not have to be similar. Each memory 8-slot DIMM bank must have the same DIMM size; however, a drawer can have a mix of DIMM banks.

The support element View Hardware Configuration task can be used to determine the size and quantity of the memory plugged in each drawer. Figure 2-21 on page 49 shows an example of configuration number 27 from the previous tables, and displays the location and description of the installed memory modules.



 **View Hardware Configuration - P00USXE8**

Machine Type - Model: 9175 - ME1  
Machine serial number: 0000000USXE8  
Processor location: ASYS

Select	Location	Identifier	Description
<input checked="" type="radio"/>	A10BMD01	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD02	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD03	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD04	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD05	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD06	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD07	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD08	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD09	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD10	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD11	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD12	32AE	Memory DIMM 256 GB DDR4 (16 Gb DRAM)
<input type="radio"/>	A10BMD13	3299	Memory DIMM 256 GB DDR5 (16 Gb DRAM)
<input type="radio"/>	A10BMD14	3299	Memory DIMM 256 GB DDR5 (16 Gb DRAM)
<input type="radio"/>	A10BMD15	3299	Memory DIMM 256 GB DDR5 (16 Gb DRAM)
<input type="radio"/>	A10BMD16	3299	Memory DIMM 256 GB DDR5 (16 Gb DRAM)
<input type="radio"/>	A10BMD17	3299	Memory DIMM 256 GB DDR5 (16 Gb DRAM)
<input type="radio"/>	A10BMD18	3299	Memory DIMM 256 GB DDR5 (16 Gb DRAM)
<input type="radio"/>	A10BMD19	3299	Memory DIMM 256 GB DDR5 (16 Gb DRAM)
<input type="radio"/>	A10BMD20	3299	Memory DIMM 256 GB DDR5 (16 Gb DRAM)
<input type="radio"/>	A10BMD21	3299	Memory DIMM 256 GB DDR5 (16 Gb DRAM)

Details... Cancel Help

Figure 2-21 View Hardware Configuration task on the Support Element

Table 2-12 on page 50 lists the physical memory plugging configurations by feature code from manufacturing when the system is ordered. Consider the following points:

- ▶ The CPC drawer columns for the specific feature contain the Memory Plug Drawer Configuration number that is referenced in Table 2-11 on page 47 and the Population by DIMM Bank that is listed in Table 2-12 on page 50.
- ▶ Dial Max indicates the maximum memory that can be enabled by way of the LICC concurrent upgrade.

If more storage is ordered by using other feature codes, such as Virtual Flash Memory or Flexible Memory, the extra storage is installed and plugged as necessary.

For example, a customer orders FC 3943 that features 9472 GB customer memory and FC 0573 Max136 (3 CPC drawers). The drawer configurations include the following components: CPC0 (3584 GB), CPC1 (3584 GB), CPC2 (3584 GB) - Configuration #27 (3584 GB)  
Total 3584 + 3584 + 3584 - 884 HSA= 9868 GB (Dial Max).



Table 2-12 Memory features and physical plugging - Standard Offering

Feature Code	Customer Increment Growth	Customer Increment GB	ME1 Max43		ME1 Max90			ME1 Max136				ME1 Max183 / Max208				
			CPC0 -884G HSA	Dial Max (GB)	CPC0 -884G HSA	CPC1	Dial Max (GB)	CPC0 -884G HSA	CPC1	CPC2	Dial Max (GB)	CPC0 -884G HSA	CPC1	CPC2	CPC3	Dial Max (GB)
3911		512	23	652	22	22	1164	22	22	22	2188	22	22	22	22	3212
3912		576	23	652	22	22	1164	22	22	22	2188	22	22	22	22	3212
3913	64	640	23	652	22	22	1164	22	22	22	2188	22	22	22	22	3212
3914		704	24	1164	22	22	1164	22	22	22	2188	22	22	22	22	3212
3915		768	24	1164	22	22	1164	22	22	22	2188	22	22	22	22	3212
3916		896	24	1164	22	22	1164	22	22	22	2188	22	22	22	22	3212
3917		1024	24	1164	22	22	1164	22	22	22	2188	22	22	22	22	3212
3918		1152	24	1164	22	22	1164	22	22	22	2188	22	22	22	22	3212
3919		1280	25	1676	23	23	2188	22	22	22	2188	22	22	22	22	3212
3920	128	1408	25	1676	23	23	2188	22	22	22	2188	22	22	22	22	3212
3921		1536	25	1676	23	23	2188	22	22	22	2188	22	22	22	22	3212
3922		1664	25	1676	23	23	2188	22	22	22	2188	22	22	22	22	3212
3923		1792	26	2188	23	23	2188	22	22	22	2188	22	22	22	22	3212
3924		1920	26	2188	23	23	2188	22	22	22	2188	22	22	22	22	3212
3925		2048	26	2188	23	23	2188	22	22	22	2188	22	22	22	22	3212
3926		2304	27	2700	24	24	3212	23	23	23	3724	22	22	22	22	3212
3927		2560	27	2700	24	24	3212	23	23	23	3724	22	22	22	22	3212
3928		2816	28	3212	24	24	3212	23	23	23	3724	22	22	22	22	3212
3929	256	3072	28	3212	24	24	3212	23	23	23	3724	22	22	22	22	3212
3930		3328	30	4236	25	25	4236	23	23	23	3724	23	22	22	22	3724
3931		3584	30	4236	25	25	4236	23	23	23	3724	23	22	22	22	3724
3932		3840	30	4236	25	25	4236	24	24	24	5260	23	23	22	22	4236
3933		4352	31	5260	26	26	5260	24	24	24	5260	23	23	23	23	5260
3934		4864	31	5260	26	26	5260	24	24	24	5260	23	23	23	23	5260
3935		5376	32	6284	27	27	6284	25	25	25	6796	24	23	23	23	5772
3936		5888	32	6284	27	27	6284	25	25	25	6796	24	24	24	23	6796
3937		6400	36	7308	28	28	7308	25	25	25	6796	24	24	24	24	7308
3938		6912	36	7308	28	28	7308	26	26	26	8332	24	24	24	24	7308
3939		7424	37	8332	29	29	8332	26	26	26	8332	25	24	24	24	7820
3940		7936	37	8332	29	29	8332	26	26	26	8332	25	25	25	24	8844
3941		8448	38	9356	30	30	9356	27	27	27	9868	25	25	25	24	8844
3942		8960	38	9356	30	30	9356	27	27	27	9868	25	25	25	25	9356
3943		9472	39	11404	31	31	11404	27	27	27	9868	26	26	25	25	10380
3944		9984	39	11404	31	31	11404	28	28	28	11404	26	26	25	25	10380
3945		10496	39	11404	31	31	11404	28	28	28	11404	26	26	26	26	11404
3946	512	11008	39	11404	31	31	11404	28	28	28	11404	26	26	26	26	11404
3947		11520	45	13452	32	32	13452	29	29	29	12940	27	26	26	26	11916
3948		12032	45	13452	32	32	13452	29	29	29	12940	27	27	27	26	12940
3949		12544	45	13452	32	32	13452	29	29	29	12940	27	27	27	26	12940
3950		13056	45	13452	32	32	13452	30	30	30	14476	27	27	27	27	13452



3951		13568	46	15500	36	36	15500	30	30	30	14476	28	28	27	27	14476
3952		14080	46	15500	36	36	15500	30	30	30	14476	28	28	27	27	14476
3953	512	14592	46	15500	36	36	15500	31	31	31	17548	28	28	28	28	15500
3954		15104	46	15500	36	36	15500	31	31	31	17548	29	28	28	28	16012
3955		15616			37	37	17548	31	31	31	17548	29	28	28	28	16012
3956		16128			37	37	17548	31	31	31	17548	29	29	29	28	17036
3957		16640			37	37	17548	31	31	31	17548	29	29	29	28	17036
3958		17152			37	37	17548	31	31	31	17548	29	29	29	29	17548
3959		18176			38	38	19596	32	32	32	20620	30	30	30	30	19596
3960		19200			38	38	19596	32	32	32	20620	30	30	30	30	19596
3961		20224			39	39	23692	32	32	32	20620	31	30	30	30	20620
3962		21248			39	39	23692	36	36	32	22668	31	31	31	30	22668
3963		22272			39	39	23692	36	36	32	22668	31	31	31	30	22668
3964		23296			39	39	23692	37	36	36	24716	31	31	31	31	23692
3965	1024	24320			45	45	27788	37	36	36	24716	32	32	31	31	25740
3966		25344			45	45	27788	37	37	37	26764	32	32	31	31	25740
3967		26368			45	45	27788	37	37	37	26764	32	32	32	31	26764
3968		27392			45	45	27788	38	38	37	28812	36	32	32	32	28812
3969		28416			46	46	31884	38	38	37	28812	36	32	32	32	28812
3970		29440			46	46	31884	38	38	38	29836	36	36	36	32	30860
3971		30464			46	46	31884	39	38	38	31884	36	36	36	36	31884
3972		31488			46	46	31884	39	38	38	31884	37	36	36	36	32908
3973		32512						39	39	38	33932	37	36	36	36	32908
3974		33536						39	39	38	33932	37	37	36	36	33932
3975		35584						45	39	39	38028	37	37	37	37	35980
3976		37632						45	39	39	38028	38	38	38	38	40076
3977		39680						45	45	45	42124	38	38	38	38	40076
3978		41728						45	45	45	42124	39	38	38	38	42124
3979		43776						46	46	45	46220	39	39	39	38	46220
3980	2048	45824						46	46	46	48268	39	39	39	38	46220
3981		47872						46	46	46	48268	39	39	39	45	50316
3982		49920										39	39	39	45	50316
3983		51968										39	45	45	45	54412
3984		54016										45	45	45	45	56460
3985		56064										45	45	45	46	58508
3986		58112										45	45	46	46	60556
3987		60160										45	46	46	46	62604
3988		62208										45	46	46	46	62604
3989		64256										46	46	46	46	64652



## 2.5.4 Memory upgrades

Memory upgrades can be ordered and enabled by LIC, upgrading the DIMM cards, adding DIMM cards, or adding a CPC drawer.

For a model upgrade that results in the addition of a CPC drawer, the minimum memory increment is added to the system. Each CPC drawer has a minimum physical memory size of 1024 GB.

### Memory allowed upgrades for Max208 models

Once a model Max208 is shipped, then the memory upgrades via MES are limited dependent on what configuration exists in the installed CPC drawers.

- ▶ If a CPC drawer is plugged with CFG numbers 22 to 38 (see Table 2-11 on page 47) then the max memory configuration for that CPC drawer is 38
- ▶ If a CPC drawer is plugged with CFG number 39, then the max memory configuration for that CPC drawer is 39
- ▶ If a CPC drawer is plugged with CFG number 46, then it's maxed out anyway

### Model Upgrades

During a model upgrade, adding a CPC drawer is a concurrent operation<sup>7</sup>. Adding physical memory to the added drawer is also concurrent. If all or part of the added memory is enabled for use, it might become available to an active LPAR if the partition includes defined reserved storage. (For more information, see 3.7.3, “Reserved storage” on page 130.) Alternatively, the added memory can be used by a defined LPAR that is activated after the memory is added.

**Note:** Memory downgrades within an IBM z17 are not supported. Feature downgrades (removal of a CPC quantity feature) are not supported.

## 2.5.5 Drawer replacement and memory

With Enhanced Drawer Availability, supported for IBM z17 ME1, there must be sufficient resources available to accommodate the resources that are rendered unavailable when a CPC drawer is removed for upgrade or repair.

For more information, see 2.7.1, “Redundant I/O interconnect (RII)” on page 59.

Removing a CPC drawer often results in removing active memory. With the flexible memory option, removing the affected memory and reallocating its use elsewhere in the system is possible. (For more information, see 2.5.7, “Flexible Memory Option” on page 53.) This process requires more available memory to compensate for the memory that becomes unavailable with the removal of the drawer.

## 2.5.6 Virtual Flash Memory

IBM Virtual Flash Memory (VFM) FC 0566 offers up to 6.0 TB of virtual flash memory in 512 GB (0.5 TB) increments (a maximum of 12 features for IBM z17 ME1) for improved application availability and to handle paging workload spikes.

<sup>7</sup> CPC 1 and CPC 2 can be added concurrently in the field if FC 2933 and FC 2934 are ordered with the initial configuration. The addition of a fourth CPC Drawer (CPC 3) is *not* supported. Four CPC drawer systems are factory-built only.



No application changes are required to change from IBM Flash Express to VFM. Consider the following points:

- ▶ Dialed memory + VFM = total hardware plugged
- ▶ Dialed memory + VFM + Flex memory option = total hardware plugged

VFM helps improve availability and handling of paging workload spikes when V2.3, V2.4, V2.5, or V3.1 is run. With this support, z/OS helps improve system availability and responsiveness by using VFM across transitional workload events, such as market openings and diagnostic data collection. z/OS also helps improve processor performance by supporting middleware use of pageable large (1 MB) pages and to help eliminate delays that can occur when collecting diagnostic data.

VFM also can be used by coupling facility images running on IBM z15 or IBM z16, to provide extended capacity and availability for workloads that use IBM WebSphere MQ Shared Queues structures.

VFM can help organizations meet their most demanding service level agreements and compete more effectively. VFM is easy to configure in the LPAR Image Profile and provides rapid time to value.

**Note:** Beginning with IBM z17, a coupling facility (CF) partition can no longer use VFM.

## 2.5.7 Flexible Memory Option

With the Flexible Memory Option, more physical memory is supplied to support the activation of the purchased memory entitlement in a single CPC drawer that is out of service during activation (POR), or in a scheduled concurrent drawer upgrade (memory add) or drawer maintenance (N+1 repair) with the use of enhanced drawer availability.

When you order memory, you can request extra flexible memory. The extra physical memory, (if required) is calculated by the configuration and priced accordingly.

The required memory DIMMs are installed before shipping and are based on a target capacity that the customer specifies. These memory DIMMs are enabled by using a Licensed Internal Code Configuration Control (LICCC) order that the client places when they determine more memory capacity is needed.

The flexible memory sizes that are available for the IBM z17 ME1 are listed in Table 2-13 on page 54.



Table 2-13 Flexible Memory Options

Feature Code	Customer Increment Growth	Customer Increment GB	ME1 Max90			ME1 Max136				ME1 Max183/Max208				
			CPC0 -884G HSA	CPC1	Dial Max (GB)	CPC0 -884G HSA	CPC1	CPC2	Dial Max (GB)	CPC0 -684G HSA	CPC1	CPC2	CPC3	Dial Max (GB)
3911		512	23	23	652	22	22	22	1164	22	22	22	22	2188
3912		576	23	23	652	22	22	22	1164	22	22	22	22	2188
3913	64	640	23	23	652	22	22	22	1164	22	22	22	22	2188
3914		704	24	24	1164	22	22	22	1164	22	22	22	22	2188
3915		768	24	24	1164	22	22	22	1164	22	22	22	22	2188
3916		896	24	24	1164	22	22	22	1164	22	22	22	22	2188
3917		1024	24	24	1164	22	22	22	1164	22	22	22	22	2188
3918		1152	24	24	1164	22	22	22	1164	22	22	22	22	2188
3919		1280	25	25	1676	23	23	22	1676	22	22	22	22	2188
3920	128	1408	25	25	1676	23	23	22	1676	22	22	22	22	2188
3921		1536	25	25	1676	23	23	22	1676	22	22	22	22	2188
3922		1664	25	25	1676	23	23	22	1676	22	22	22	22	2188
3923		1792	26	26	2188	23	23	23	2188	22	22	22	22	2188
3924		1920	26	26	2188	23	23	23	2188	22	22	22	22	2188
3925		2048	26	26	2188	23	23	23	2188	22	22	22	22	2188
3926		2304	27	27	2700	24	24	23	2700	23	23	23	23	3724
3927		2560	27	27	2700	24	24	23	2700	23	23	23	23	3724
3928		2816	28	28	3212	24	24	24	3212	23	23	23	23	3724
3929	256	3072	28	28	3212	24	24	24	3212	23	23	23	23	3724
3930		3328	30	30	4236	25	25	24	3724	23	23	23	23	3724
3931		3584	30	30	4236	25	25	24	3724	23	23	23	23	3724
3932		3840	30	30	4236	25	25	25	4236	24	24	24	24	5260
3933		4352	31	31	5260	26	26	26	5260	24	24	24	24	5260
3934		4864	31	31	5260	26	26	26	5260	24	24	24	24	5260
3935		5376	32	32	6284	27	27	27	6284	25	25	25	25	6796
3936		5888	32	32	6284	27	27	27	6284	25	25	25	25	6796
3937		6400	36	36	7308	28	28	28	7308	25	25	25	25	6796
3938		6912	36	36	7308	28	28	28	7308	26	26	26	26	8332
3939		7424	37	37	8332	29	29	29	8332	26	26	26	26	8332
3940		7936	37	37	8332	29	29	29	8332	26	26	26	26	8332
3941		8448	38	38	9356	30	30	30	9356	27	27	27	27	9868
3942		8960	38	38	9356	30	30	30	9356	27	27	27	27	9868
3943		9472	39	39	11404	31	31	30	10380	27	27	27	27	9868
3944		9984	39	39	11404	31	31	30	10380	28	28	28	28	11404
3945		10496	39	39	11404	31	31	31	11404	28	28	28	28	11404
3946	512	11008	39	39	11404	31	31	31	11404	28	28	28	28	11404
3947		11520	45	45	13452	32	32	31	12428	29	29	29	29	12940
3948		12032	45	45	13452	32	32	31	12428	29	29	29	29	12940
3949		12544	45	45	13452	32	32	32	13452	29	29	29	29	12940



3950		13056	45	45	13452	32	32	32	13452	30	30	29	29	13452
3951		13568	46	46	15500	36	36	32	14476	30	30	30	30	14476
3952		14080	46	46	15500	36	36	32	14476	30	30	30	30	14476
3953	512	14592	46	46	15500	36	36	36	15500	31	31	30	30	15500
3954		15104	46	46	15500	36	36	36	15500	31	31	30	30	15500
3955		15616				37	37	36	16524	31	31	31	30	16524
3956		16128				37	37	36	16524	31	31	31	30	16524
3957		16640				37	37	37	17548	31	31	31	31	17548
3958		17152				37	37	37	17548	31	31	31	31	17548
3959		18176				38	38	38	19596	32	32	31	31	18572
3960		19200				38	38	38	19596	32	32	32	32	20620
3961		20224				39	39	38	21644	32	32	32	32	20620
3962		21248				39	39	38	21644	36	36	32	32	21644
3963		22272				39	39	39	23692	36	36	36	36	23692
3964		23296				39	39	39	23692	36	36	36	36	23692
3965	1024	24320				45	45	39	25740	37	37	36	36	24716
3966		25344				45	45	39	25740	37	37	37	37	26764
3967		26368				45	45	45	27788	37	37	37	37	26764
3968		27392				45	45	45	27788	38	38	38	37	28812
3969		28416				46	46	45	29836	38	38	38	37	28812
3970		29440				46	46	45	29836	38	38	38	38	29836
3971		30464				46	46	46	31884	39	39	38	38	31884
3972		31488				46	46	46	31884	39	39	38	38	31884
3973		32512								39	39	39	38	33932
3974		33536								39	39	39	38	33932
3975		35584								45	45	39	39	38028
3976		37632								45	45	39	39	38028
3977	2048	39680								45	45	45	45	42124
3978		41728								45	45	45	45	42124
3979		43776								46	46	45	45	44172
3980		45824								46	46	46	46	48268
3981		47872								46	46	46	46	48268

## 2.6 Reliability, availability, and serviceability

IBM Z servers continue to deliver enterprise class RAS with IBM z17. The main philosophy behind RAS is about preventing or tolerating (masking) outages. It is also about providing the necessary instrumentation (in hardware, LIC and microcode, and software) to capture or collect the relevant failure information to help identify an issue without requiring a re-creation of the event. These outages can be planned or unplanned, as shown in the following examples:

- ▶ An unplanned outage because of loss of cache coherency
- ▶ An unplanned outage because of a firmware problem
- ▶ A planned outage for a disruptive firmware upgrade
- ▶ A planned outage to add physical memory to a single CPC drawer configuration



The IBM Z hardware has decades of intense engineering behind it, which results in a robust and reliable platform. The hardware has many RAS features that are built into it.

For more information, see Chapter 9, “Reliability, availability, and serviceability” on page 401.

### 2.6.1 RAS in the CPC memory subsystem

The IBM z17 ME1 server includes the following RAS features:

- ▶ Up to 48 DIMM per drawer in two rows (64 TB system maximum):
  - Organized in eight-card RAIM groups:
    - 50% reduced RAIM overhead (compared to IBM z15)
    - RAIM protection (similar to IBM z16)
    - Up to three chip marks + one channel mark
  - DDR5 DRAM with on chip power regulation: N+1 new voltage regulators
  - Standard Open Memory Interface (OMI); up to six OMI per drawer:
    - CRC/Retry for soft errors
    - Degraded bus lanes 4 → 2 on hard error
    - No waiting for all eight cards (use first seven to respond)
  - Time Domain Reflectometry (TDR)
    - DDR4: CP end only [Memory ASIC does not have TDR circuitry]
    - DDR5: TDR supported on both ends [uses new IBM Memory ASIC]
- ▶ Concurrent service and upgrade by way of Concurrent Drawer Repair (CDR)

### 2.6.2 General IBM z17 ME1 RAS features

The IBM z17 ME1 server includes the following RAS features:

- ▶ The server provides a true N+1 pumps and N+1 fans for the cooling function
- ▶ The Power/Thermal Subsystem is continued from IBM z16 A01 to IBM z17 ME1.
  - It uses switchable, intelligent Power Distribution Units (PDUs)
  - Redundant (N+1), number of PDUs (up to four pairs, configuration dependent)
- ▶ CPC drawer is packaged to fit in the 19-inch frame. CPC drawer power and cooling includes the following components:
  - PSUs: AC to 12 V bulk/standby (N+1 redundant). The PSU FRU includes the fan. CPC drawers have four PSUs for PDU-based configurations
  - N+2 Voltage Regulation Module (VRM): On-point of load cards (POL) or voltage regulation sticks (VRS)
  - Power Processor Control Card: Power control card to control CPC fans (N+1 redundant) and are available in quantities of two
  - Fans: Drawer has five fans and are N+1 redundant
  - BMC/OSCs: Redundant (N+1)
- ▶ PU DCMs are all water-cooled
- ▶ The PCIe+ I/O drawer power supplies for the IBM z17 also are based on the N+1 design
  - The second power supply can maintain operations and avoid an unplanned outage of the system



- ▶ N+2 redundant environmental sensors (ambient temperature, relative humidity, and air density<sup>8</sup>)

The internal intelligent Power Distribution Unit (iPDU) provides the following capabilities:

- ▶ Individual outlet control by way of Ethernet:
  - Provide a System Reset capability
  - Power cycle an SE if a hang occurs
  - Verify power cables at installation
- ▶ System Reset Function:
  - No physical EPO switch is available on the IBM z17, which provides a means for the service technician<sup>9</sup> to put a server into a known state
  - This function does not provide the option to power down and keep the power down to the system. The power must be unplugged or the customer-supplied power turned off at the panel.
- ▶ Other characteristics:
  - PDU Firmware can be concurrently updated
  - Concurrently repairable
  - Power redundancy check
- ▶ Cable verification test by way of PDU:
  - By power cycling individual iPDU outlets, the system can verify proper cable connectivity
  - Power cable test runs during system Power On
  - Runs at system installation and at every system Power On until the test passes
- ▶ PCIe service enhancements:
  - Mandatory end-to-end cyclic redundancy check (ECRC)
  - Customer operation code is separate from maintenance code
  - Native PCIe firmware stack that is running on the integrated firmware processor (IFP) to manage isolation and recovery

The power service and control network (PSCN) is used to control and monitor the elements in the system and include the following components:

- ▶ Ethernet Top of Rack (TOR) switches provide the internal PSCN connectivity:
  - Switches are redundant (N+1)
  - Concurrently maintainable
  - Each switch has one integrated power supply
  - BMCs are cross wired to the Ethernet switches
- ▶ Redundant SEs
 

Each SE has two power supplies (N+1) and input power is cross-coupled from the PDUs
- ▶ Concurrent CPC upgrades
 

CPC1 to (CPC1 + CPC2) and (CPC1 + CPC2) to (CPC1+CPC2+CPC3) if CPC1 Reserve or CPC2 Reserve features are part of the initial system order (FC 2933 or FC 2934)
- ▶ All PCIe+ I/O drawer MESs and rebalance are concurrent
- ▶ All LICC model changes are concurrent

<sup>8</sup> The air density sensor measures air pressure and is used to control blower speed.

<sup>9</sup> This function is available to IBM System Service Representatives (SSRs) only.



## System Recovery Boost

With IBM z17, System Recovery Boost feature (first introduced with IBM z15) enables the restoration of service from, and catch up after, planned and unplanned outages faster than on any previous IBM Z with no extra software cost. It also provides faster site switching changeover between systems active in different sites (in a GDPS configured environment).

Service restoration involves speeding up IPL and shutdown operations of an image (LPAR), and short-duration recovery process boosts for specific sysplex and operating system events. For more information see *Introducing IBM Z System Recovery Boost*, [REDP-5563](#).

**Important:** The base System Recovery Boost capability is built into IBM z17 firmware and does not require ordering of extra features.

IBM z17 servers continue to deliver robust server designs through new technologies, hardening innovative and classic redundancy. For more information, see Chapter 9, “Reliability, availability, and serviceability” on page 401.

## 2.7 Connectivity

Connections to PCIe+ I/O drawers and Integrated Coupling Adapters are driven from the CPC drawer fan-out cards. These fan-outs are installed in the rear of the CPC drawer.

Figure 2-22 shows two locations of the fan-out slots. Each slot is identified with a location code (label) of LGxx.

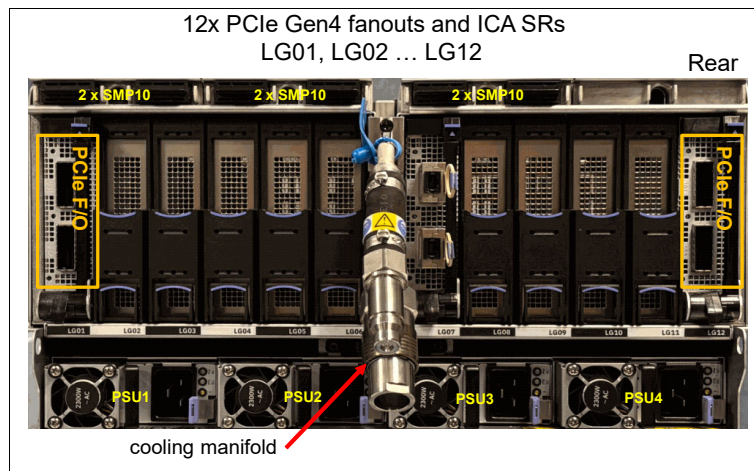


Figure 2-22 Fan out locations in the CPC drawer

Up to 12 PCIe fan-outs (LG01 - LG12) can be installed in each CPC drawer.

A fan-out can be repaired concurrently with the use of redundant I/O interconnect. For more information, see 2.7.1, “Redundant I/O interconnect (RII)” on page 59.

The following types of fan-outs are available:

- ▶ PCIe+ Generation 4 dual port fan-out card:
  - This fan-out provides connectivity to the PCIe switch cards in the PCIe+ I/O drawer.
- ▶ Integrated Coupling Adapter (ICA SR2.0):



- This adapters provides coupling connectivity to IBM z17, IBM z16, and IBM z15 servers.
- One, two, or three pairs of redundant SMP-10<sup>10</sup> connectors provide connectivity to the other one, two, or three CPC drawers in the configuration.

When configured for availability, the channels and coupling links are balanced across CPC drawers. In a system that is configured for maximum availability, alternative paths maintain access to critical I/O devices, such as disks and networks. The CHPID Mapping Tool can be used to assist with configuring a system for high availability.

Enhanced Drawer Availability allows a single CPC drawer in a multidrawer CPC to be removed and reinstalled (serviced) concurrently for an upgrade or a repair. Removing a CPC drawer means that the connectivity to the I/O devices that are connected to that CPC drawer is lost. To prevent connectivity loss, the redundant I/O interconnect feature allows you to maintain connection to critical I/O devices installed in PCIe+ I/O Drawers when a CPC drawer is removed. ICA SR2.0 can also be installed (configured) for coupling and timing links redundancy.

### 2.7.1 Redundant I/O interconnect (RII)

Redundancy is provided for PCIe I/O interconnects.

The PCIe+ I/O drawer supports up to 16 PCIe features, which are organized in two hardware domains (for each drawer). The infrastructure for the fan-out to I/O drawers and external coupling is shown in Figure 2-23.

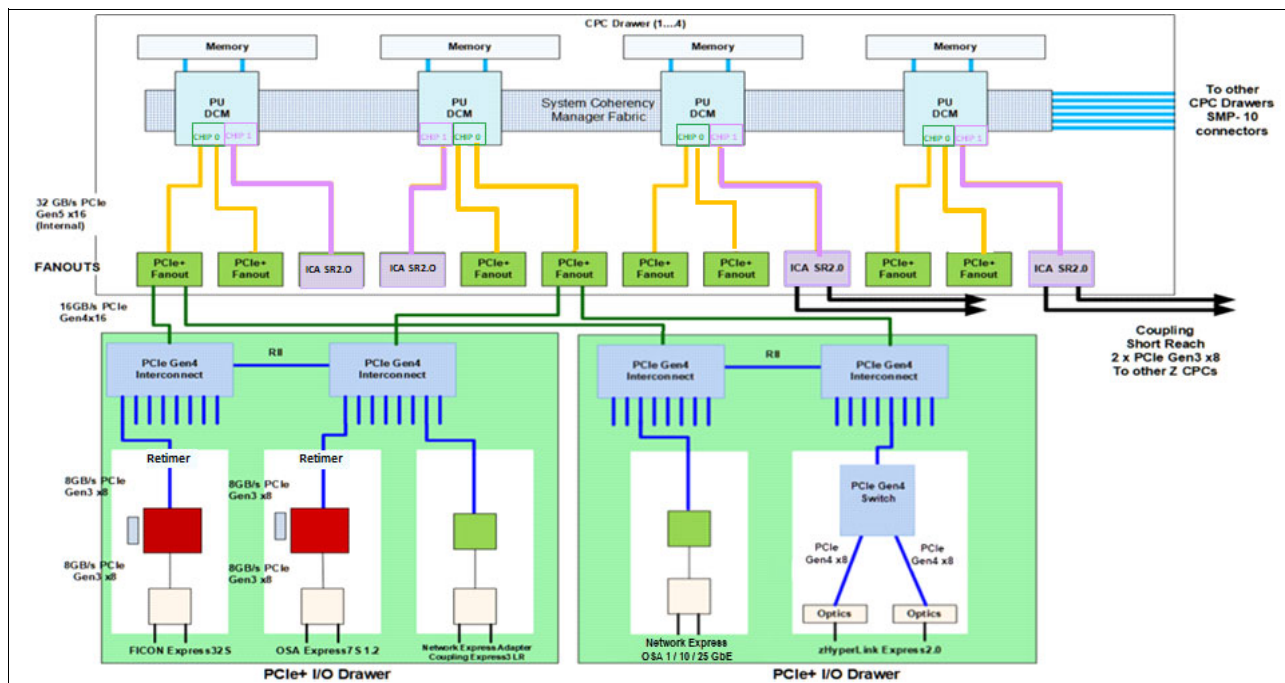


Figure 2-23 Infrastructure for PCIe+ I/O drawers (IBM z17 Max43 system with two PCIe+ I/O drawers)

The PCIe+ Gen4 fan-out cards are used to provide the connection from the PU DCM PCIe Bridge Unit (PBU), which splits the PCIe Gen5 (@32GBps) processor busses into two PCIe Gen4 x16 (@16 GBps) interfaces to the PCIe switch card in the PCIe+ I/O drawer.

<sup>10</sup> SMP-10 is new with the IBM z17 (IBM z16 uses SMP-9 for the same functions)



The PCIe switch card spreads the x16 PCIe bus to the PCIe I/O slots in the domain.

In the PCIe+ I/O drawer, the two PCIe switch cards (LG06 and LG16) provide a backup path (Redundant I/O Interconnect [RII]) for each other through the passive connection in the PCIe+ I/O drawer backplane.

During a CPC Drawer PCIe+ Gen4 fan-out or cable failure, all 16 PCIe cards in the two domains can be driven through a single PCIe switch card (see Figure 2-24).

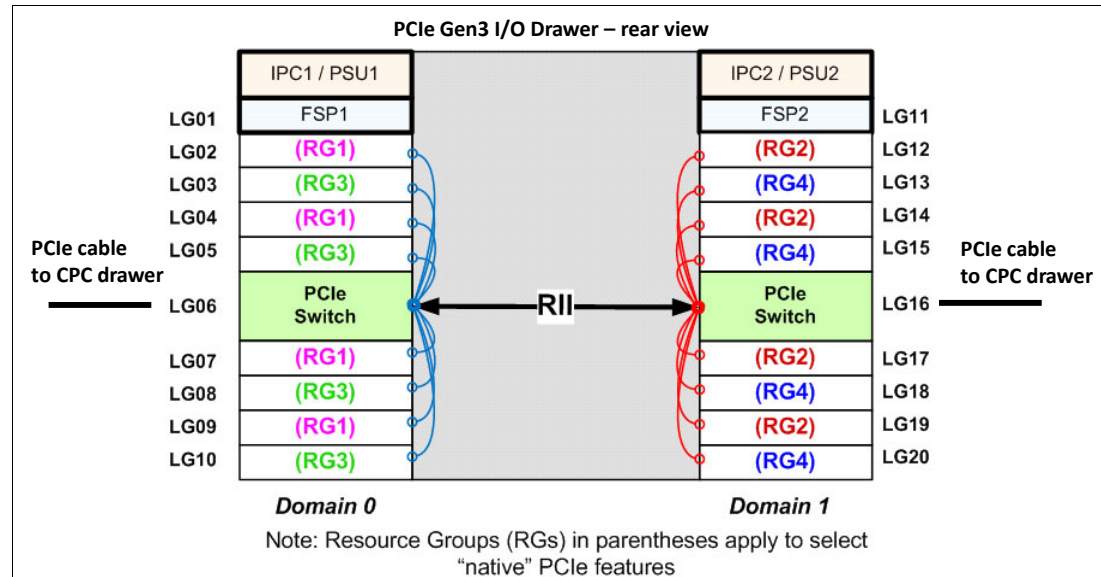


Figure 2-24 Redundant I/O Interconnect

To support RII between domain pair 0 and 1, the two interconnects to each pair must be driven from two different PCIe fan-outs. Normally, each PCIe interconnect in a pair supports the eight features in its domain. In backup operation mode, one PCIe interconnect supports all 16 features in the domain pair. Refer to 4.3, “PCIe+ I/O drawer” on page 175.

**Note:** The PCIe Interconnect (switch) adapter *must* be installed in the PCIe+ I/O drawer to maintain the interconnect across I/O domains. If the adapter is removed (for a service operation), the I/O cards in that domain (up to eight) become unavailable.

## 2.7.2 Enhanced drawer availability

With EDA, the effect of CPC drawer replacement is minimized. In a multiple CPC drawer system, a single CPC drawer can be concurrently removed and reinstalled for an upgrade or repair. Removing a CPC drawer without affecting the workload requires sufficient resources in the remaining CPC drawer.

Before removing the CPC drawer, the contents of the PUs and memory of the drawer must be relocated. PUs must be available on the remaining CPC drawers to replace the deactivated drawer. Also, sufficient redundant memory must be available if no degradation of applications is allowed.

To ensure that the CPC configuration supports removal of a CPC drawer with minimal effect on the workload, consider the flexible memory option. Any CPC drawer can be replaced, including the first CPC drawer that initially contains the HSA.



Removal of a CPC drawer also removes the CPC drawer connectivity to the PCIe I/O drawers, and coupling links. The effect of the removal of the CPC drawer on the system is limited by the use of redundant I/O interconnect. (For more information, see 2.7.1, “Redundant I/O interconnect (RII)” on page 59.) However, all ICA SR2.0 links that are installed in the removed CPC drawer must be configured offline.

If the enhanced drawer availability and flexible memory options are *not* used when a CPC drawer must be replaced, the memory in the failing drawer also is removed. This process might be necessary during an upgrade or a repair action.

Until the removed CPC drawer is replaced, a power-on reset of the system with the remaining CPC drawers is supported. The CPC drawer then can be replaced and added back into the configuration concurrently.

### 2.7.3 CPC drawer upgrade

All fan-outs that are used for I/O and coupling links are rebalanced concurrently as part of a CPC drawer addition to support better RAS characteristics.

## 2.8 Model configurations

When an IBM z17 is ordered, the PUs are characterized according to their intended usage. The PUs can be ordered as any of the following items:

<b>CP</b>	The processor is purchased and activated. PU supports running the z/OS, VSE <sup>n</sup> V6.3.1 (21 <sup>st</sup> Century Software), z/VM, z/TPF, and Linux on IBM Z <sup>11</sup> operating systems. It also can run Coupling Facility Control Code.
<b>IFL</b>	The Integrated Facility for Linux (IFL) is a processor that is purchased and activated for use by z/VM for Linux guests and Linux on IBM Z <sup>12</sup> operating systems.
<b>Unassigned IFL</b>	A processor that is purchased for future use as an IFL. It is offline and cannot be used until an upgrade for the IFL is installed. It does not affect software licenses or maintenance charges.
<b>ICF</b>	An internal coupling facility (ICF) processor that is purchased and activated for use by the Coupling Facility Control Code.
<b>Unassigned ICF</b>	A processor that is purchased for future use as an ICF. It is offline and cannot be used until an upgrade for the ICF is installed. It does not affect software licenses or maintenance charges.
<b>zIIP</b>	An “Off Load Processor” for workload that supports applications such as DB2 and z/OS Container Extensions. It also can be used for “System Recovery Boost” on page 58.
<b>Unassigned zIIP</b>	A processor that is purchased for future use as zIIP. It is offline and cannot be used until an upgrade for the zIIP is installed. It does not affect software licenses or maintenance charges.
<b>Additional SAP</b>	Additional SAP(s) (System Assist Processors) are <i>not supported</i> on the IBM z17 ME1.

<sup>11</sup> The KVM hypervisor is part of supported Linux on IBM Z distributions.

<sup>12</sup> See 1.1.3, “Supported operating systems” on page 6



A minimum of one PU that is characterized as a CP, IFL, or ICF is required per system. The maximum number of characterizable PUs is 208. A zIIP requires at least one CP to be present in the configuration. The maximum number of zIIPs is up to one minus the total capacity of a MaxXX model. For instance, a Max43 can have up to 42 zIIPs. Refer to Table 2-5 on page 39.

The following components are present in the IBM z17, but they are not part of the PUs that clients purchase and require no characterization:

- ▶ SAPs are used by the channel subsystem. The number of predefined SAPs depends on the IBM z17 model. See Table 2-14
- ▶ Two IFPs, which are used in the support of designated features and functions, such as RoCE (NETD and NETH CHPD types - all features), Coupling Express3 LR, zHyperLink Express 2.0, Internal Shared Memory (ISM) SMC-D, and other management functions.
- ▶ Two spare PUs, which can transparently assume any characterization during a permanent failure of another PU.

The IBM z17 uses features to define the number of PUs that are available for client use in each configuration. The models are listed in Table 2-14.

Table 2-14 IBM z17 processor configurations

Feature	CPC Drawers	PUs per drawer	Active PUs				zIIP	IFP	Opt SAPs	Base SAPs	Spares
			CPs	IFLs	ICFs	uIFLs					
Max43	1	48	0 - 43	0 - 43	0 - 43	0 - 42	0 - 42	2	no	5	2
Max90	2	48	0 - 90	0 - 90	0 - 90	0 - 89	0 - 89	2		10	2
Max136	3	48	0 - 136	0 - 136	0 - 136	0 - 135	0 - 135	2		16	2
Max183	4	48	0 - 183	0 - 183	0 - 183	0 - 182	0 - 182	2		21	2
Max208	4	57	0 - 208	0 - 208	0 - 208	0 - 207	0 - 207	2		24	2

- ▶ Not all PUs available on a model are required to be characterized with a feature code. Only the PUs purchased by a customer are identified with a feature code
- ▶ zIIP maximum quantity for new build systems depends on the Model (MaxXX)
- ▶ All PU conversions can be performed concurrently

A *capacity marker* identifies the number of CPs that were purchased. This number of purchased CPs is higher than or equal to the number of CPs that is actively used. The capacity marker marks the availability of purchased but unused capacity that is intended to be used as CPs in the future. This status often is present for software-charging reasons.

Unused CPs are not a factor when establishing the millions of service units (MSU) value that is used for charging monthly license charge (MLC) software, or when charged on a per-processor basis.

## 2.8.1 Upgrades

Concurrent upgrades of CPs, IFLs, ICFs, and zIIPs, are available for the IBM z17. However, concurrent PU upgrades require that more PUs are installed but not activated.



Spare PUs are used to replace defective PUs. Two spare PUs always are on an IBM z17 ME1 server. In the rare event of a PU failure, a spare PU is activated concurrently and transparently, and is assigned the characteristics of the failing PU.

If an upgrade request cannot be accomplished within the existing configuration, a hardware upgrade is required.

The following upgrade paths for the IBM z17 are shown in Figure 2-25:

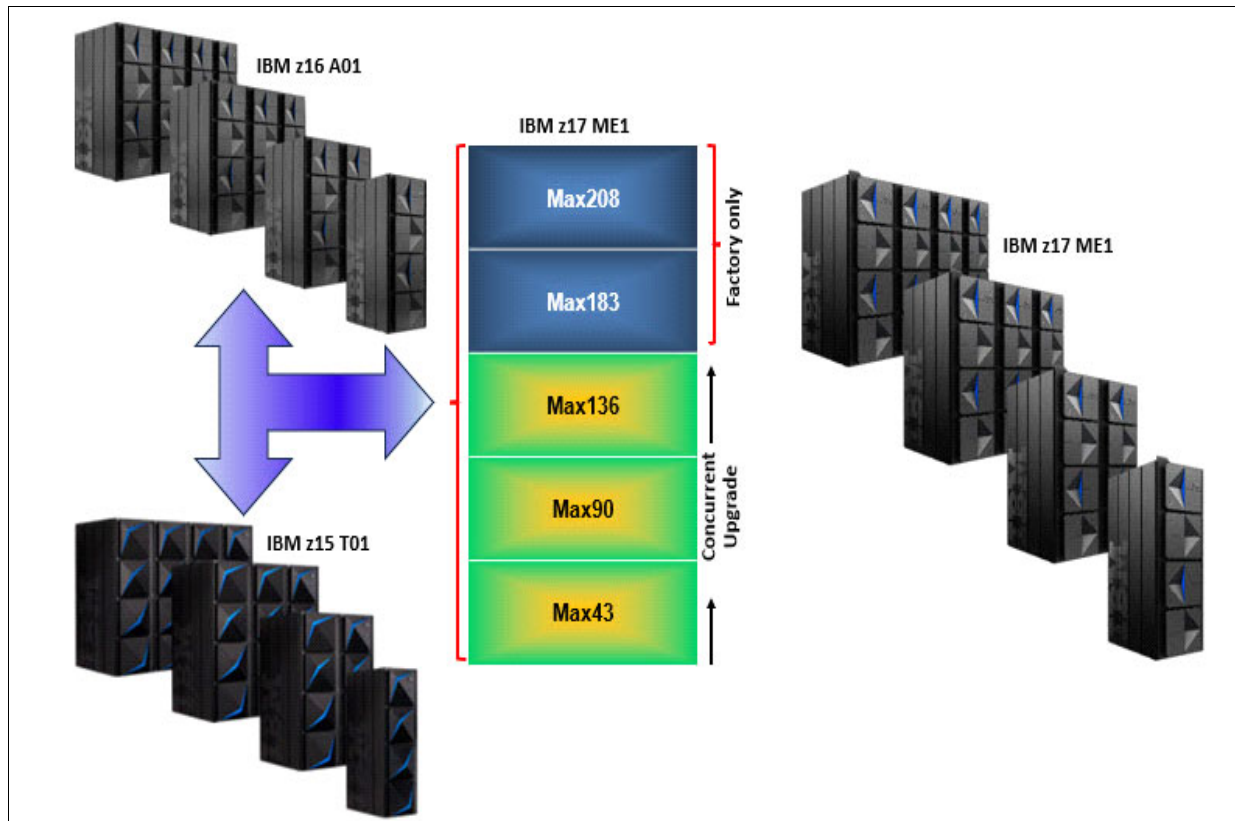


Figure 2-25 IBM z17 ME1 system upgrade paths

- ▶ IBM z17 ME1 to IBM z17 ME1 upgrades:
  - Max43 - Max90, Max136 are concurrent (if FC 2933 or 2934<sup>13</sup> initially ordered)
  - No upgrade to Max183 or Max208 (these features are factory-built only)
  - More I/O drawers can be added based on available space in current frames or I/O expansion frames
- ▶ IBM z15 (M/T 8561) to IBM z17 ME1
- ▶ IBM z16 (M/T 3931) to IBM z17 ME1:
  - Model to model MAXxx upgrades cannot be performed if I/O drawers are populating the CPC drawer positions.
  - Supported model upgrade paths are listed in Table 2-15 on page 64.

<sup>13</sup> FCs 2933 and 2934 are CPC1 (A15B) and CPC2 (A20B) CPC reserves



Table 2-15 Supported model MAXxx upgrade paths

	Max43	Max90	Max136	Max183	Max208
Max43	--	Y	Y	N	N
Max90	N	--	Y	N	N
Max136	N	N	--	N	N
Max183	N	N	N	--	N
Max208	N	N	N	N	--

## 2.8.2 Model capacity identifier

To recognize how many PUs are characterized as CPs, the Store System Information (STSI) instruction returns a Model Capacity Identifier (MCI). The MCI determines the number and speed of characterized CPs. Refer to [z/Architecture Principles of Operation](#).

Characterization of a PU as an IFL, ICF, or zIIP is not reflected in the output of the STSI instruction because characterization has no effect on software charging. For more information about STSI output, see “Processor identification” on page 395.

The following distinct model capacity identifier ranges are recognized (one for full capacity and three for granular capacity):

- ▶ For full-capacity engines, model capacity identifiers 701 - 7K8 are used. They express capacity settings for 1 - 208 characterized CPs.
- ▶ Three model subcapacity identifier ranges offer a unique level of granular subcapacity engines at the low end. They are available for up to 43 characterized CPs. These three subcapacity settings are applied to up to 43 CPs, which combined offer 129 more capacity settings, as described next.

### Granular capacity

The IBM z17 ME1 server offers 129 capacity settings (granular capacity) for up to 43 CPs. When subcapacity settings are used, PUs beyond 43 can be characterized only as specialty engines. For models with more than 43 CPs, all CPs are running at full capacity (7xx).

The three defined ranges of subcapacity settings include model capacity identifiers numbered 401- 443, 501 - 543, and 601 - 643.

**Consideration:** All CPs have the same capacity identifier. Specialty engines (IFLs, zIIPs, and ICFs) operate at full speed.

### List of model capacity identifiers

Regardless of the number of CPC drawers, a configuration with one characterized CP is possible, as listed in Table 2-16.

Table 2-16 Model capacity identifiers

Feature	Model capacity identifier
Max43 - (FC0571)	701 - 743, 601 - 643, 501 - 543, and 401 - 443
Max90 - (FC0572)	701 - 790, 601 - 643, 501 - 543, and 401 - 443
Max136 - (FC0573)	701 - 7D6, 601 - 643, 501 - 543, and 401 - 443



Max183 - (FC0574)	701 - 7I3, 601 - 643, 501 - 543, and 401 - 443
Max208 - (FC0575)	701 - 7K8, 601 - 643, 501 - 543, and 401 - 443

For more information about temporary capacity increases, see Chapter 8, “System upgrades” on page 353.

## 2.9 Power and cooling

The IBM z17 ME1 power and cooling system is similar to IBM z16 A01 systems, which is packaged in an industry-standard, 19-inch form factor frame for all the internal system elements. The configuration can consist of 1 - 4 frames.

Consider the following points:

- ▶ The power subsystem is based on the Power Distribution Units (PDUs) that are mounted at the rear of the system in pairs
- ▶ The system uses 3-phase power:
  - Low voltage 4 wire “Delta”
  - High voltage 5 wire “Wye”
- ▶ No EPO (emergency power off) switch.  
IBM z17 has a support element task to simulate the EPO function (used only when necessary to perform a System Reset Function).
- ▶ No DC input feature is available.
- ▶ The IBM z17 ME1 system has a radiator unit (RU) with N+1 design for the pumps and blowers.
- ▶ The 19-inch frame is capable of top or bottom exit of power. All power cords are 4.26 meters (14 feet). Combined with the Top Exit I/O Cabling feature, more options are available when you are planning your computer room cabling.
- ▶ The new PSCN structure uses industry standard Ethernet switches (up to four).

### 2.9.1 PDU-based configurations

The IBM Z operate with redundant power infrastructure. The IBM z17 ME1 is designed with a new power infrastructure that is based on intelligent PDUs that are mounted vertically on the rear side of the 19-inch racks and Power Supply Units for the internal components. The PDU configuration supports radiator-cooled systems only.

The PDUs are controlled by using an Ethernet port and support the following input:

- ▶ 3-phase 200 - 240 V AC (4-wire “Delta”)
- ▶ 3-phase 380 - 415 V AC (5-wire “Wye”)

The power supply units convert the AC power to DC power that is used as input for the Point of Loads (POLs) in the CPC drawer and the PCIe+ I/O drawers.

The power requirements depend on the number of CPC drawers (1 - 4), number of PCIe I/O drawers (0 - 12) and I/O features that are installed in the PCIe I/O drawers.

PDUs are installed and serviced from the rear of the frame. Unused power ports should *never* be used by any external device.



A view of a maximum configured system with PDU-based power is shown in Figure 2-26.

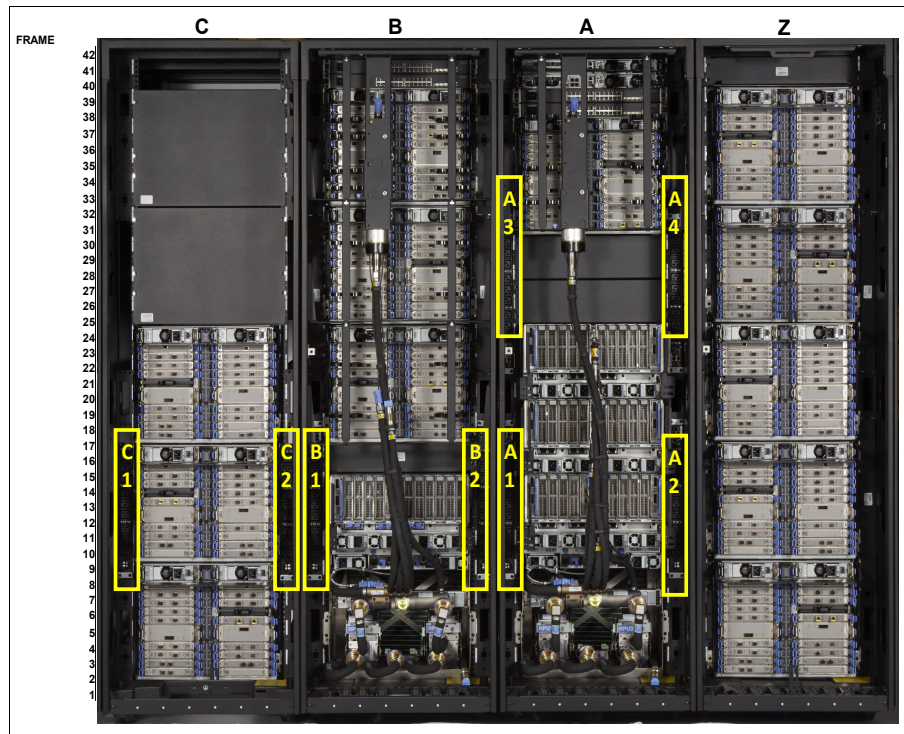


Figure 2-26 Rear view of maximum configured system PDU power

Each PDU installed requires a customer-supplied power feed. The number of power cords that are required depends on the system configuration.

**Note:** For initial installation, all power sources are required to run the system checkout diagnostics successfully.

PDUs are installed in pairs. A system can have 2, 4, 6, or 8 PDUs, depending on the configuration. Consider the following points:

- ▶ Paired PDUs are A1/A2, A3/A4, B1/B2, and C1/C2.
- ▶ From the rear of the system, the odd-numbered PDUs are on the left side of the rack; the even-numbered PDUs are on the right side of the rack.
- ▶ The total loss of one PDU in a pair does not affect the system operation.

Components that plug into the PDUs for redundancy (by using two power cords) include the following features:

- ▶ CPC Drawers, PCIe+ I/O drawers, Radiators, and Support Elements
- ▶ The redundancy for each component is achieved by plugging the power cables into the paired PDUs.

For example, the top Support Element (1), has one power supply plugged into PDU A1 and the second power supply plugged into the paired PDU A2 for redundancy.



**Note:** Customer power sources should always maintain redundancy across PDU pairs; that is, one power source or distribution panel supplies power for PDU A1 and the separate power source or distribution panel supplies power for PDU A2.

As a best practice, connect the odd-numbered PDUs (A1, B1, C1, and D1) to one power source or distribution panel, and the even-numbered PDUs (A2, B2, C2, and D2) to a separate power source or distribution panel.

The frame count rules (number of frames) for IBM z17 ME1 are listed in Table 2-17.

Table 2-17 Frame count rules for IBM z17 ME1

Frame Count	I/O drawers												
CPC drawers	0	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	1	2	2	2	2	2	3	3	3	3
2	1	1	1	2	2	2	2	2	3	3	3	3	3
3	1	1	2	2	2	2	2	3	3	3	3	3	N/A
4	2	2	2	2	2	3	3	3	3	3	4	4	4

The number of CPC drawers and I/O drawers determines the number of racks in the system and the number of PDUs in the system.

The PDU/line cord rules (number of PDU/Cord pairs) for IBM z17 ME are listed in Table 2-18.

Table 2-18 PDU/line cord rules (# PDU/Cord pairs) for IBM z17 ME1

PDU/Linecord	I/O drawers												
CPC drawers	0	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	1	2	2	2	2	2	3	3	3	3
2	2	2	2	2	2	2	2	2	3	3	3	3	3
3	2	2	2	2	2	2	2	3	3	3	3	3	N/A
4	3	3	3	3	3	3	3	3	3	4	4	4	4

## 2.9.2 Power estimation tool

The power estimation tool for the IBM z17 allows you to enter your precise server configuration to obtain an *estimate* of power consumption by using the following process:

1. Log in to the [Resource link](#) with your user ID.
2. Click **Planning** → **Tools** → **Power and weight estimation**
3. Specify the quantity for the features that are installed in your system

This tool estimates the power consumption for the specified configuration. The tool does *not* verify that the specified configuration can be physically built.

**Tip:** The exact power consumption for your system varies. The object of the tool is to estimate the power requirements to aid you in planning for your system installation. Actual power consumption after installation can be confirmed by using the HMC Monitors Dashboard task.



## 2.9.3 Cooling

The PU DCMs for IBM z17 ME1 are cooled by a cold plate that is connected to the internal water-cooling loop. In an air-cooled system, the radiator unit dissipates the heat from the internal water loop with air. The radiator unit provides improved availability with N+ 1 pumps and blowers.

For all IBM z17 ME1 servers, the CPC drawer components (except for PU DCMs) and the PCIe+ I/O drawers are air cooled by redundant fans. Airflow of the system is directed from front (cool air) to the back of the system (hot air).

### **Radiator-cooled (air-cooled) models (FC 4045, FC 4046)**

The IBM z17 ME1 PU DCMs in the CPC drawers are cooled by closed loop water. The internal closed water loop removes heat from PU DCMs by circulating water between the radiator heat exchanger and the cold plate that is mounted on the PU DCMs.

For more information, see 2.9.4, “Radiator Cooling Unit” on page 68.

Although the PU DCMs are cooled by water, the heat is exhausted into the room from the radiator heat exchanger by forced air with blowers. At the system level, these IBM z17 ME1 are still air-cooled systems.

## 2.9.4 Radiator Cooling Unit

A total of 1 - 2 Radiator Cooling Units (RCUs) are used in the system: One in Frame A and one in Frame B in support of water cooling of the PU DCMs within the CPC drawers. The unit includes N+1 pumps and N+1 fans. Water loops to each drawer are directly delivered by way of hoses to each drawer from manifolds.

The RCU discharges heat from the internal frame water loop to the customer's data center.

The RCU contains two independent pump FRUs. Because the cooling capability is a redundant N+1 design, a single working pump and blower can support the entire load. The replacement of one pump or blower can be done concurrently and does not affect performance.

Each RCU provides cooling to PU DCMs with closed loop water within the respective frame. No connection to an external chilled water supply is required. The IBM z17 ME1 server will use a new coolant which consists of 40% propylene glycol and 60% DI (deionized) water. This new coolant has many advantages:

- ▶ Elimination of the IBM Z fill and drain tool (used with IBM z16 and z15)
- ▶ Elimination of SSR fluid handling in the field
- ▶ Reduced install time
- ▶ Elimination of customer requirement to store the fill and drain tool and BTA canisters
- ▶ Elimination of customer requirement to discard coolant at the end of life Fill and Drain Tool



Each RCU contains up to four independent fan assemblies that can be concurrently serviced. The number of fans present depends on the number of CPC drawers that are installed in the frame.

The water pumps, manifold assembly, radiator assembly (which includes the heat exchanger), and fan assemblies are the main components of the IBM z17 RCU, as shown in Figure 2-27.

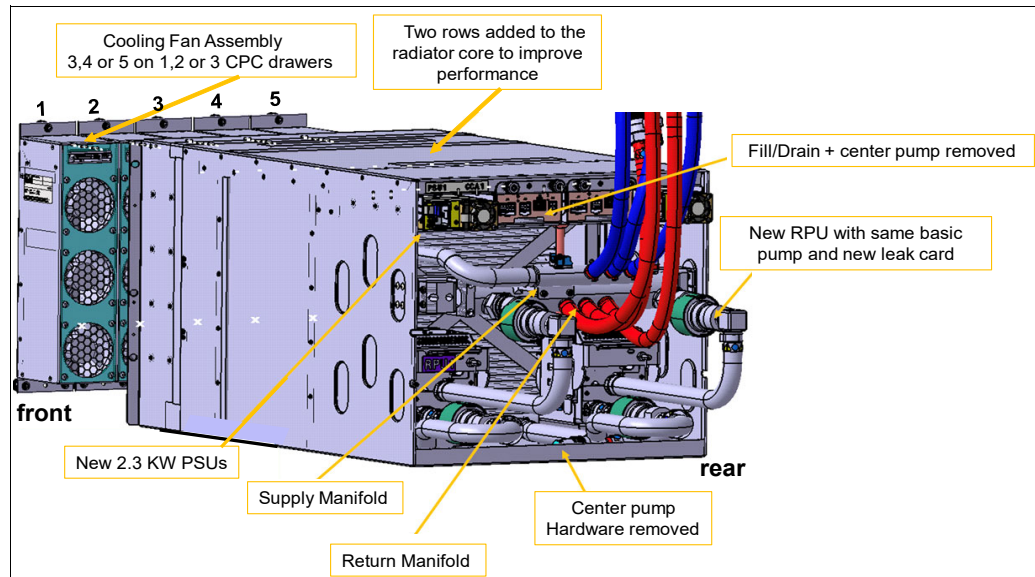


Figure 2-27 Radiator Cooling Unit details

The closed water loop in the radiator unit is shown in Figure 2-28. The warm water that is exiting from the PU DCMs cold plates enters pumps through a common manifold and is pumped through a heat exchanger where heat is extracted by the air flowing across the heat exchanger fins. The cooled water is then recirculated back into the PU DCMs cold plates.

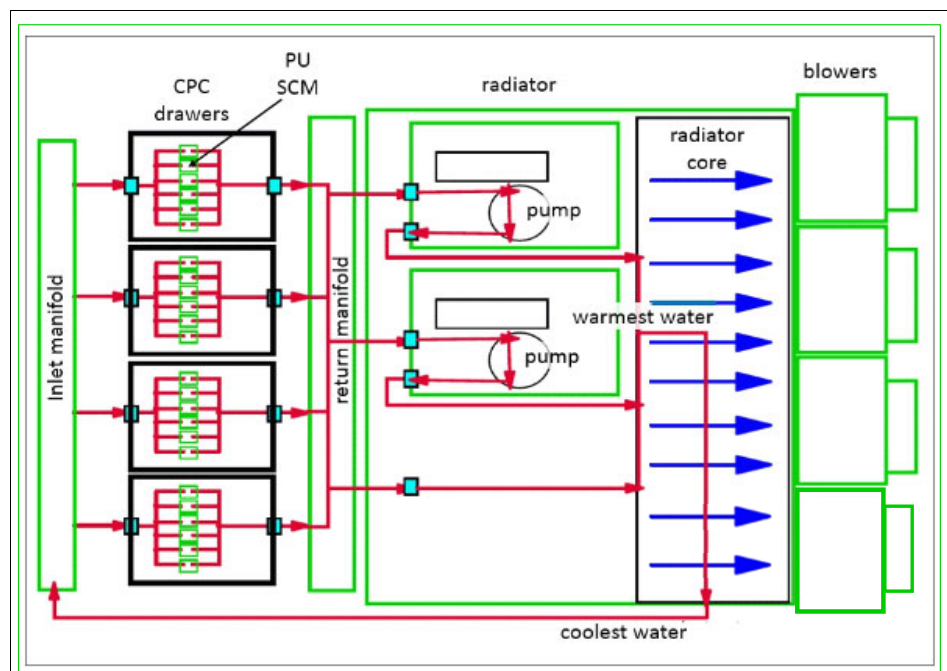


Figure 2-28 Radiator cooling system



## 2.10 Summary

All aspects of the IBM z17 ME1 structure are listed in Table 2-19

Table 2-19 System structure summary

Description	Max43	Max90	Max136	Max183	Max208
Maximum number of characterized PUs	43	90	136	183	208
Number of CPs	0 - 43	0 - 90	0 - 136	0 - 183	0 - 208
Number of IFLs	0 - 43	0 - 90	0 - 136	0 - 183	0 - 208
Number of unassigned IFLs	0 - 42	0 - 89	0 - 135	0 - 182	0 - 207
Number of ICFs	0 - 43	0 - 90	0 - 136	0 - 183	0 - 208
Number of unassigned ICFs	0 - 42	0 - 89	0 - 135	0 - 182	0 - 207
Number of zIIPs	0 - 42	0 - 89	0 - 135	0 - 182	0 - 207
Number of unassigned zIIPs	0 - 42	0 - 89	0 - 135	0 - 182	0 - 207
Standard SAPs	5	10	16	21	24
Number of IFP	2	2	2	2	2
Standard spare PUs	2	2	2	2	2
Enabled Memory sizes GB	512 - 15616	512 - 31488	512 - 47872	512 - 64256	512 - 64256
Flexible memory sizes GB	N/A	512 - 13652	512 - 27988	512 - 42324	512 - 47872
L1 cache per PU (I/D)	128/128 KB	128/128 KB	128/128 KB	128/128 KB	128/128 KB
L2 private unified cache per core	36 MB	36 MB	36 MB	36 MB	36 MB
L3 virtual cache on PU chip	36 x 10 360 MB	36 x 10 360 MB	36 x 10 360 MB	36 x 10 360 MB	36 x 10 360 MB
L4 virtual cache on chips in drawer	36 x 10 x 8 = = 2.88 GB	36 x 10 x 8 = 2.88GB	36 x 10 x 8 = 2.88GB	36 x 10 x 8 = 2.88GB	36 x 10 x 8 = 2.88 GB
Cycle time (ns)	0.181	0.181	0.181	0.181	0.181
Clock frequency	5.5 GHz	5.5 GHz	5.5 GHz	5.5 GHz	5.5 GHz
Maximum number of PCIe fan-outs	12	24	36	48	48
PCIe bandwidth	32 GBps	32 GBps	32 GBps	32 GBps	32 GBps
Number of support elements	2	2	2	2	2
External AC power	3-phase	3-phase	3-phase	3-phase	3-phase



## 3



# Central processor complex design

This chapter describes the design of the IBM z17 processor unit. By understanding this design, users become familiar with the functions that make the IBM z17 a system that accommodates a broad mix of workloads for large enterprises.

**Note:** The IBM z17 Model ME1, Machine Type (M/T) 9175 (M/T 91751), is further identified in this document as IBM z17, unless otherwise specified.

For more information about the processor unit, see *z/Architecture Principles of Operation*, SA22-7832.

This chapter includes the following topics:

- ▶ 3.1, “Overview” on page 72
- ▶ 3.2, “Design highlights” on page 73
- ▶ 3.3, “CPC drawer design” on page 75
- ▶ 3.4, “Processor unit design” on page 82
- ▶ 3.5, “Processor unit functions” on page 103
- ▶ 3.6, “Memory design” on page 117
- ▶ 3.7, “Logical partitioning” on page 121
- ▶ 3.8, “Intelligent Resource Director” on page 132
- ▶ 3.9, “Clustering technology” on page 133
- ▶ 3.10, “Virtual Flash Memory” on page 142
- ▶ Chapter 3.3.11, “IBM Secure Service Container” on page 143



## 3.1 Overview

The IBM z17 symmetric multiprocessor (SMP) system is the next step in an evolutionary trajectory that began with the introduction of the IBM System/360 in 1964. Over time, the design was adapted to the changing requirements that were dictated by the shift toward new types of applications on which clients depend.

IBM Z servers offer high levels of reliability, availability, serviceability (RAS), resilience, and security. The IBM z17 fits into the IBM strategy in which mainframes play a central role in creating an infrastructure for cloud, artificial intelligence, and analytics, which is underpinned by security. The IBM z17 server is designed so that everything around it, such as operating systems, middleware, storage, security, and network technologies that support open standards, helps you achieve your business goals.

The IBM z17 extends the platform's capabilities and adds value with breakthrough technologies, such as the following examples:

- ▶ An industry-first system that uses quantum-safe technologies, cryptographic discovery tools, and end-to-end data encryption to protect against future attacks now.
- ▶ A continuous compliance solution to help keep up with changing regulations, which reduces cost and risk.
- ▶ A consistent cloud experience to enable accelerated modernization, rapid delivery of new services, and end-to-end automation.
- ▶ New options in flexible and responsible consumption to manage system resources across geographical locations, with sustainability that is built in across its lifecycle.

The modular CPC drawer design aims to reduce (or in some cases even eliminate) planned and unplanned outages. The design does so by offering concurrent repair, replace, and upgrade functions for processors, memory, and I/O.

For more information about the IBM z17 RAS features, see Chapter 9, "Reliability, availability, and serviceability" on page 401.

IBM z17 ME1 servers include the following features:

- ▶ Ultra-high frequency, large, high-speed buffers (caches) and memory
- ▶ Superscalar processor design
- ▶ Improved out-of-order core execution
- ▶ Simultaneous multithreading (SMT)
- ▶ Single-instruction multiple-data (SIMD)
- ▶ On-core integrated accelerator for Z SORT, one per PU core
- ▶ IBM integrated accelerator for zEnterprise® Data Compression (zEDC) (on-chip compression accelerator), one per PU chip
- ▶ New on-chip Data Processing Unit (DPU), one per PU chip
- ▶ On-chip Second Generation Artificial Intelligence (AI unit or AIU), one per PU chip, at speed and scale that is designed to leave no transaction behind
- ▶ Quantum-safe cryptography support
- ▶ Flexible configuration options

It is the next implementation of IBM Z servers to address the ever-changing IT environment.



For more information about frames and configurations, see Chapter 2, “Central processor complex hardware components” on page 19, and Appendix E, “Frame configurations with Power Distribution Units” on page 551.

## 3.2 Design highlights

The physical packaging of the IBM z17 server's CPC drawer is a continuation and evolution of the previous generations of IBM z Systems. Its modular CPC drawer and dual chip module (DCM) design address the increasing costs that are related to building systems with evergreater capacities.

The modular CPC drawer design is flexible and expandable. It offers unprecedented capacity and security features to meet consolidation needs.

IBM z17 servers CPC continues the line of mainframe processors that are compatible with an earlier version. The IBM z17 brings the following processor design enhancements:

- ▶ 5 nm silicon lithography
- ▶ Eight cores per PU chip design with 43 billion transistors per PU chip, compared to 22.5 billion for z16
- ▶ Level 2 cache increase from 30MB to 36MB per core
- ▶ Four PU Dual Chip Modules per CPC Drawer
- ▶ Each PU chip features two PCIe Generation 5 ports (x16 @ 32 GBps)
- ▶ Optimized pipeline
- ▶ Improved SMT and SIMD
- ▶ Improved branch prediction
- ▶ Improved co-processor functions (CPACF)
- ▶ IBM integrated accelerator for Z Sort (on-core sort accelerator)
- ▶ IBM Second Generation integrated accelerator for AI (on-chip AI accelerator)
- ▶ IBM integrated Data Processing Unit (DPU)
- ▶ Improved transparent memory encryption with 256 bit AES

It uses 24-, 31-, and 64-bit addressing modes, multiple arithmetic formats, and multiple address spaces for robust interprocess security.

The IBM z17 system design features the following main objectives:

- ▶ Offer a data-centric approach to information (data) security that is simple, transparent, and consumable (extensive data encryption from inception to archive, in-flight, and at-rest).
- ▶ Offer a flexible infrastructure to concurrently accommodate a wide range of operating systems and applications, from the traditional systems (for example, z/OS and z/VM) to the world of Linux, cloud, analytics, Artificial Intelligence, and mobile computing.
- ▶ Offer state-of-the-art integration capability for server consolidation by using virtualization capabilities in a highly secure environment:
  - Logical partitioning, which allows up to 85 independent logical servers
  - z/VM, which can virtualize hundreds to thousands of servers as independently running virtual machines (guests)



- HyperSockets, which implement virtual LANs between logical partitions (LPARs) within the system
- Efficient data transfer that uses direct memory access (SMC-D), Remote Direct Memory Access (SMC-R), and reduced storage access latency for transactional environments
- The IBM Z Processor Resource/System Manager (PR/SM) is designed for Common Criteria Evaluation Assurance Level 5+ (EAL 5+) certification for security; therefore, an application that is running on one partition (LPAR) cannot access another application on a different partition, which provides essentially the same security as an air-gapped system.
- The Secure Execution feature, which securely separates second-level guest operating systems that are running under KVM for Z from each other and securely separates access to second-level guests from the hypervisor.

This configuration allows for a logical and virtual server coexistence and maximizes system utilization and efficiency by sharing hardware resources.

- ▶ Offer high-performance computing to achieve the outstanding response times that are required by new workload-type applications. This performance is achieved by high-frequency, enhanced superscalar processor technology, out-of-order core execution, large high-speed buffers (cache) and memory, an architecture with multiple complex instructions, and high-bandwidth channels.
- ▶ Offer the high capacity and scalability that are required by the most demanding applications, from the single-system and clustered-systems points of view.
- ▶ Offer the capability of concurrent upgrades for processors, memory, and I/O connectivity, which prevents system outages in planned situations.
- ▶ Implement a system with high availability and reliability. These goals are achieved with redundancy of critical elements and sparing components of a single system, and the clustering technology of the Parallel Sysplex environment.
- ▶ Have internal and external connectivity offerings, supporting open standards, such as Gigabit Ethernet (GbE) and Fibre Channel Protocol (FCP).
- ▶ Provide leading cryptographic performance. Every processor unit (PU) includes a dedicated and optimized CP Assist for Cryptographic Function (CPACF). Optional Crypto Express features with cryptographic coprocessors provide the highest standardized security certification.<sup>1</sup> These optional features also can be configured as Cryptographic Accelerators to enhance the performance of Secure Sockets Layer/Transport Layer Security (SSL/TLS) transactions.
- ▶ Provide on-chip compression. Every PU chip design incorporates a compression unit, which is the IBM Integrated Accelerator for z Enterprise Data Compression (zEDC). This configuration is different from the CMPSC (Compression Coprocessor) that is implemented in each core.
- ▶ Provide a second generation dedicated on-chip integrated AI Accelerator for high-speed inference to enable real-time AI embedded directly in transactional workloads, and improvements for performance, security, and availability.
- ▶ Provide an on-chip Data Processing Unit (sometimes known as an I/O Engine), moving functionality from the Application-specific integrated circuit (ASIC) on the I/O adapters and in-boarding it into 32 Assist Processors in the IBM z17 PU chip. The DPU decreases channel latencies and improves performance and power efficiency.
- ▶ Be self-managing and self-optimizing, adjusting itself when the workload changes to achieve the best system throughput. This process can be done by using the Intelligent

<sup>1</sup> Federal Information Processing Standard (FIPS) 140-2 Security Requirements for Cryptographic Modules.



Resource Director or the Workload Manager functions, which are assisted by HiperDispatch.

- ▶ Have a balanced system design with pervasive encryption, which provides large data rate bandwidths for high-performance connectivity along with processor and system capacity, while protecting every byte that enters and exits the IBM z16.

The remaining sections in this chapter describe the IBM z17 system structure. It shows a logical representation of the data flow from PUs, caches, memory cards, and various interconnect capabilities.

### 3.2.1 Process shrink - from 7nm to 5nm

z17 PU chips are manufactured using Samsung's mature 5nm process<sup>2</sup> whereas z16 is manufactured using their 7nm process<sup>3</sup>. Because of their maturity both processes are high yield.

Both "7nm" and "5nm" are industry standard terms for process nodes. These terms do not indicate that any physical feature (such as gate length, metal pitch or gate pitch) of the transistors is that size. However there is a considerable increase in density going from 7nm to 5nm.

The increase in density, has a number of benefits.

In particular the Telum 2 processor chip has 43.0 billion transistors, compared to 22.5 billion for Telum.

The chip design uses this increase in density, along with a slight chip size increase from 530mm<sup>2</sup> to 565.6mm<sup>2</sup> so that

- ▶ Power consumption and heat dissipation are improved.
- ▶ Level 2 cache sizes are larger, increasing from 8 x 32MB to 10 x 36MB.
- ▶ The Data Processing Unit (DPU) has been introduced.

## 3.3 CPC drawer design

An IBM z17 ME1 system can have up to four CPC drawers in a full configuration, with up to 208 PUs that can be characterized for customer use, and up to 64 TB of customer usable memory.

The following types of CPC drawer configurations are available for IBM z17 ME1:

- ▶ One drawer: Max43
- ▶ Two drawers: Max90
- ▶ Three drawers: Max136
- ▶ Four drawers: Max183
- ▶ Four drawers: Max208

**Note:** Max183 and Max208 are factory-build only. It is not possible to upgrade in the field to Max183 or Max208.

<sup>2</sup> See [https://en.wikipedia.org/wiki/5\\_nm\\_process](https://en.wikipedia.org/wiki/5_nm_process)

<sup>3</sup> See [https://en.wikipedia.org/wiki/7\\_nm\\_process](https://en.wikipedia.org/wiki/7_nm_process)



The IBM z17 ME1 has up to 24 memory controller units (MCUs) for a Max208 feature (two MCUs per PU chip and up to six MCUs (out of eight) populated per CPC drawer). The MCU configuration uses an eight channel Reed-Solomon (R-S) redundant array of independent memory (RAIM).

The RAIM design is an 8-channel R-S RAIM design on the IBM z17. The DIMM sizes (32, 64, 128, 256 or 512 GB) include RAIM overhead. An IBM z17 CPC drawer can have up to 48 memory DIMMs:

- ▶ DDR4 Memory DIMMS in 32, 64, 128 GB sizes
- ▶ DDR5 Memory DIMMS in 32, 64, 128, 256, 512 GB sizes

The IBM z17 microprocessor chip integrates a cache hierarchy design with only two levels of physical cache (L1 and L2). The cache hierarchy (L1, L2) is implemented with dense static random access memory (SRAM).

eDRAM is no longer used in the IBM processor. On an IBM z17, L2 cache (36 MB) is semi-private with 18 MB dedicated to the associated core, and 18 MB shared with the system (the 50/50 split is adjustable). Level 3 (L3) and Level 4 (L4) caches are now virtual caches and are allocated on L2.

Two processor chips (up to eight active cores per PU chip) are combined in a Dual Chip Module (DCM) and four DCMs are assembled in a CPC drawer (eight PU chips). An IBM z17 can have from one CPC drawer (Max43) up to four (Max183 and Max208).

Figure 3-1 shows the new IBM Telum II processor.

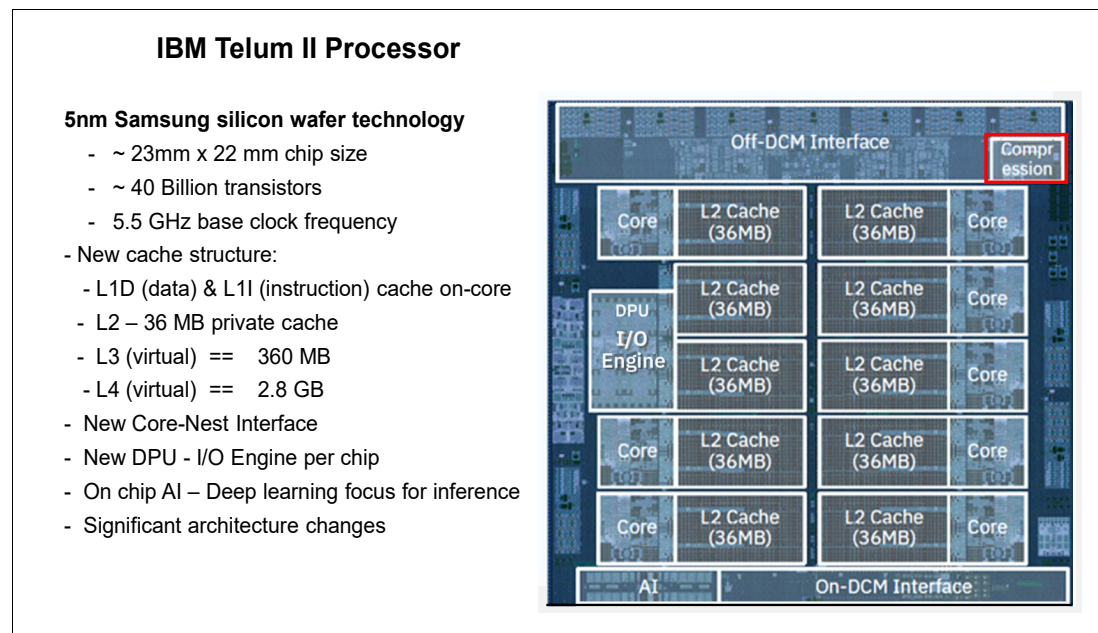


Figure 3-1 IBM Telum II processor

The new IBM z17 Dual Chip Module (DCM) is shown in Figure 3-2 on page 77.



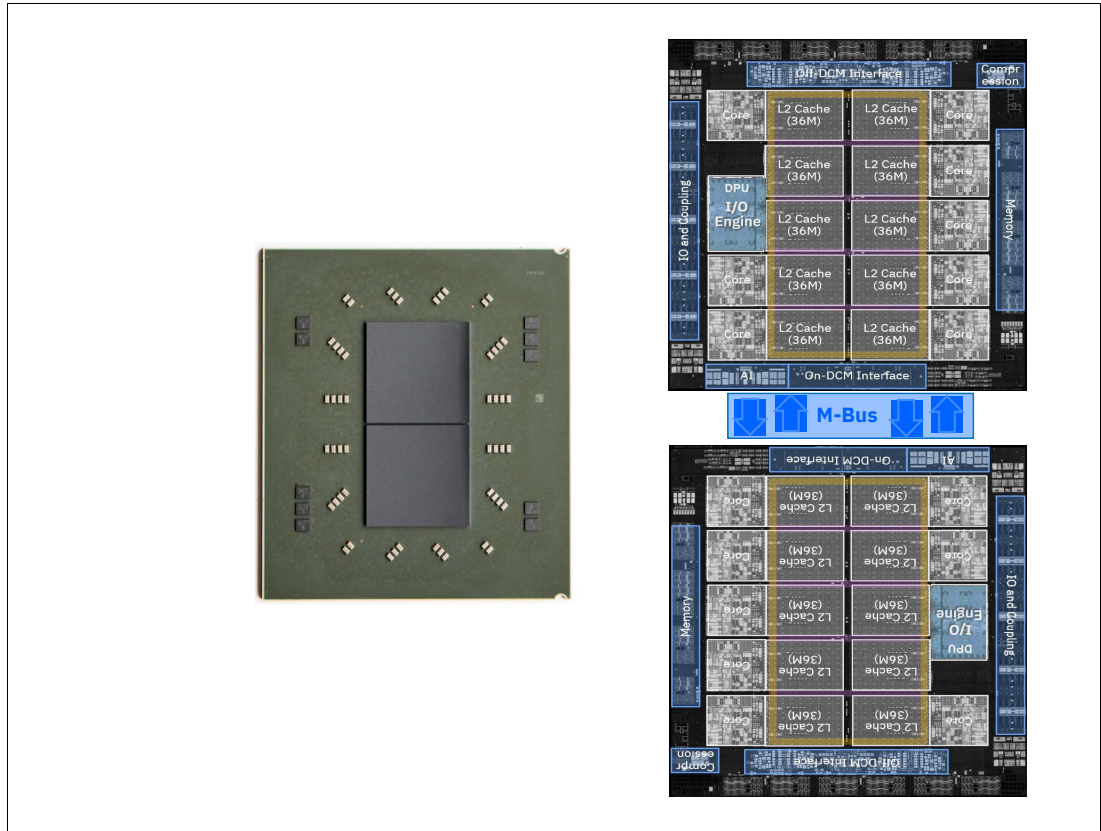


Figure 3-2 IBM z17 Dual Chip Module

Concurrent maintenance allows dynamic Central Processing Complex (CPC) drawer add and repair.<sup>4</sup>

IBM z17 processors use 5nm extreme ultraviolet (EUV) lithography chip technology with advanced low latency pipeline design, which creates high-speed yet power-efficient circuit designs. The PU DCM has a dense packaging, which allows closed water loop cooling. The heat from the closed loop is dissipated into the air by a radiator unit (RU).

The external water-cooling option is no longer available starting with IBM z16. For more information, see 2.9, “Power and cooling” on page 65.

### 3.3.1 Cache levels and memory structure

The IBM z17 includes an optimized memory subsystem design that focuses on keeping data “closer” to the PU core. The IBM z17 ME1 cache hierarchy is shown in Figure 3-3 on page 78.

<sup>4</sup> For configurations with two or more CPC drawers installed.



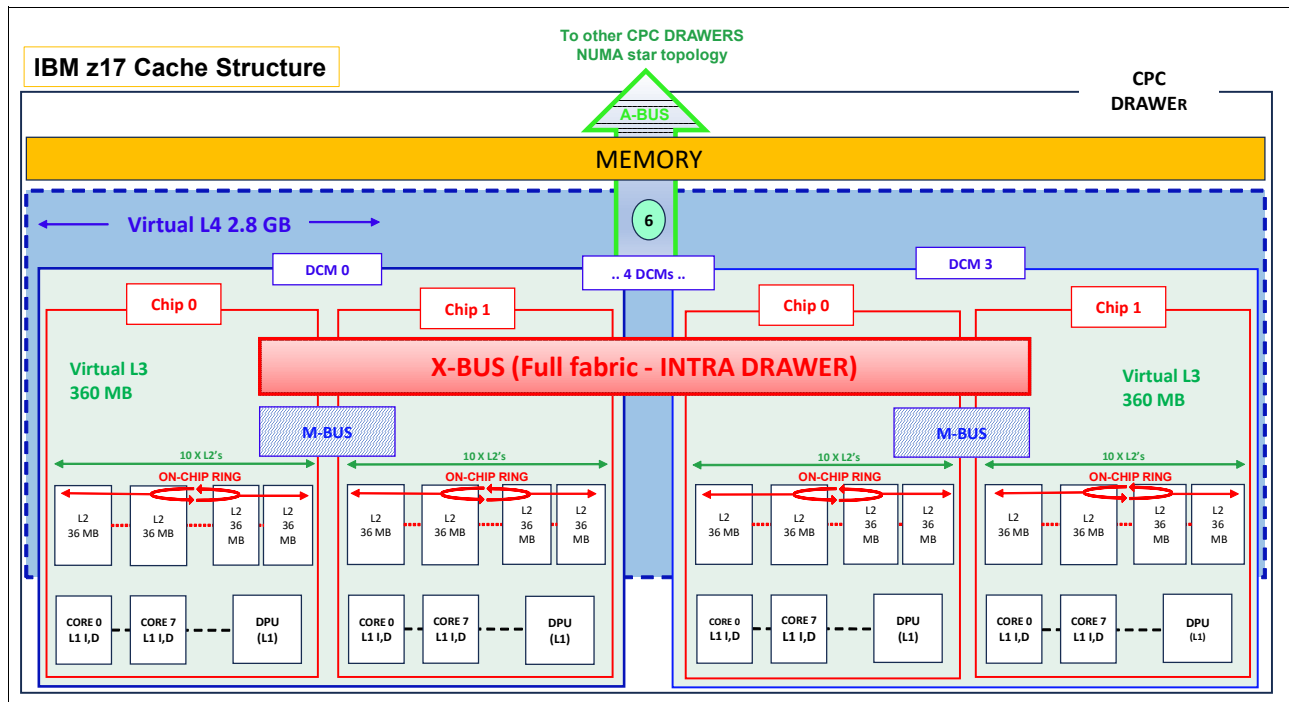


Figure 3-3 IBM z17 cache levels and memory hierarchy

On IBM z17, L1 and L2 caches are implemented on the PU chip, and L3 and L4 caches are implemented as virtual caches and dynamically allocated on the shared part of the L2 semi-private cache.

The cache structure of the IBM z17 features the following characteristics:

- Large L1, L2 caches (more data closer to the core).
- L1 cache is implemented by using SRAM technology and has the same size as on IBM z16 (128 KB for instructions and 128 KB for data).
- L2 cache (36 MB in total) uses SRAM technology, and is semi-private to each PU core with 18 MB dedicated to the associated core, and 18 MB shared with the system (the 50/50 split is adjustable).
- L3 cache (up to 360 MB) now becomes a virtual cache and can be allocated on any of the shared parts of the L2 caches. It operates at the DCM level.
- L4 cache (up to 2880 MB) is also a virtual cache and can be allocated on any of the shared parts of the L2 caches. It operates at the drawer level.



Figure 3-4 shows the cache structure that is implemented in an IBM z17 CPC drawer.

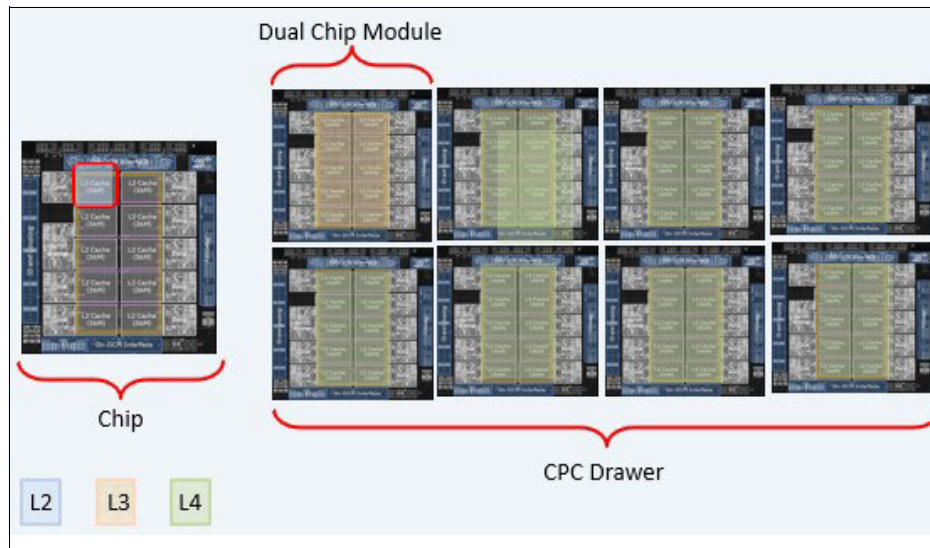


Figure 3-4 IBM z17 cache structure at CPC drawer level

Main storage has up to 16TB addressable memory per CPC drawer, which uses up to 48 DDR4 and DDR5 DIMMs. A system with four CPC drawers can have up to 64TB of main storage.

### Considerations

Cache sizes are limited by ever-diminishing cycle times because they must respond quickly without creating bottlenecks. Access to large caches costs more cycles. Instruction and data cache (L1) sizes must be limited because larger distances must be traveled to reach long cache lines. This L1 access time generally occurs in one cycle, which prevents increased latency.

Also, the distance to remote caches as seen from the microprocessor becomes a significant factor. For example, on an IBM z15 server, access to L4 physical cache (on the SC chip and which might not even be in the same CPC drawer) requires several cycles to travel the distance to the cache. (Off-drawer access can take hundreds of cycles.) On an IBM z17, having an L4 virtual, physically allocated on the shared L2 requires fewer processor cycles in many instances.

Although large caches mean increased access latency, the new technology 5nm EUV chip lithography and the lower cycle time allows IBM z17 servers to increase the size of L2 cache level within the PU chip.

To overcome the inherent delays of the SMP CPC drawer design and save cycles to access the remote virtual L4 content, the system keeps instructions and data as close to the processors as possible. This configuration can be managed by directing as much work of a specific LPAR workload to the processors in the same CPC drawer as the L4 virtual cache.

This configuration is achieved by having the IBM Processor Resource/Systems Manager (PR/SM) scheduler, the z/OS WLM and dispatcher work together. A good LPAR design can assist them in keeping as much work as possible within the boundaries of as few processors and L4 virtual cache space (which is best within a CPC drawer boundary) without affecting throughput and response times.



The cache structures of IBM z17 ME1 systems are compared with the previous generation of IBM Z (IBM z16 A01) in Figure 3-5.

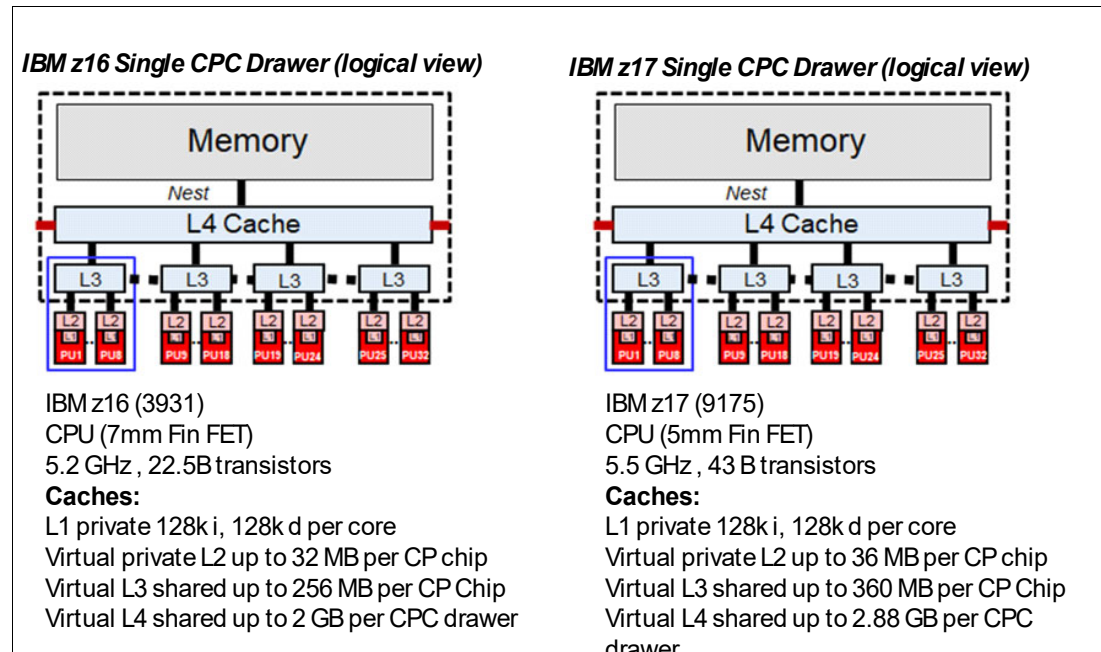


Figure 3-5 IBM z16 and IBM z17 cache level comparison

Compared to IBM z16, the IBM z17 has a larger L2 cache. Cache L3 and L4 are virtual caches in the IBM z16 and IBM z17. More affinity exists between the memory of a partition, the L4 virtual cache in a drawer, and the cores in the PU chips. As in IBM z16, the IBM z17 cache level structure is focused on keeping more data closer to the PU. This design can improve system performance on many production workloads.

## HiperDispatch

To help avoid latency in a high-frequency processor design, PR/SM and the dispatcher must schedule and dispatch a workload to run on as small a portion of the system as possible. The cooperation between z/OS and PR/SM is bundled in a function called HiperDispatch. HiperDispatch uses the IBM z17 cache topology, which features reduced cross-cluster “help” and better locality for multitasking address spaces.

PR/SM can use dynamic PU reassignment to move processors (CPs, ZIIPs, IFLs, ICFs, and spares) to a different chip and drawer to improve the reuse of shared caches by processors of the same partition. It can use dynamic memory relocation (DMR) to move a running partition’s memory to different physical memory to improve the affinity and reduce the distance between the memory of a partition and the processors of the partition. These are relatively infrequent events.

For more information about HiperDispatch, see 3.7, “Logical partitioning” on page 121.



### 3.3.2 CPC drawer interconnect topology

CPC drawers are interconnected in a point-to-point topology that allows each CPC drawer to communicate with every other CPC drawer. Data transfer does not always have to go through another CPC drawer to address the requested data or control information.

The IBM z17 ME1 inter-CPC drawer communication structure is shown in Figure 3-6.

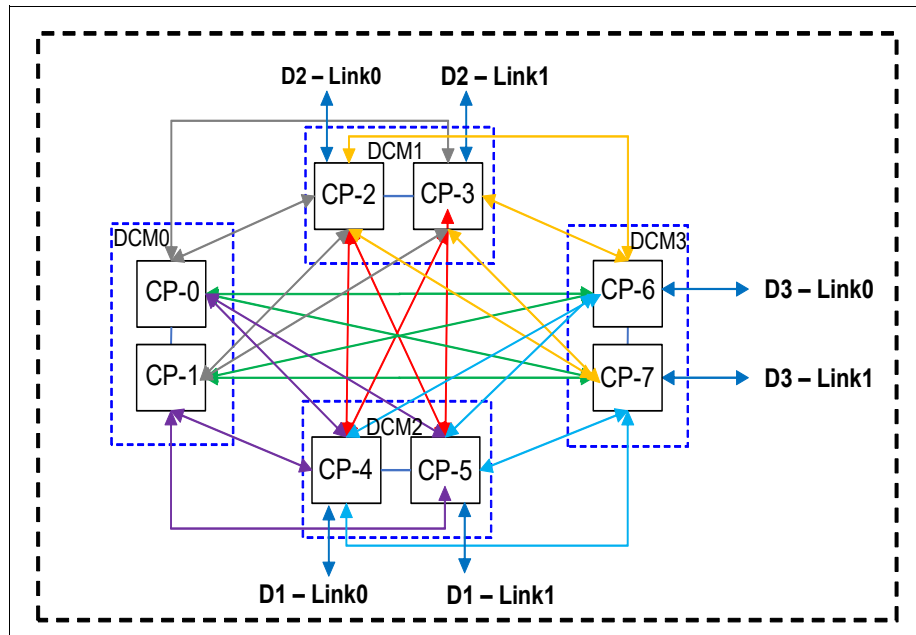


Figure 3-6 IBM z17 CPC drawer communication topology

A simplified topology of a four-CPC drawer system is shown in Figure 3-7.

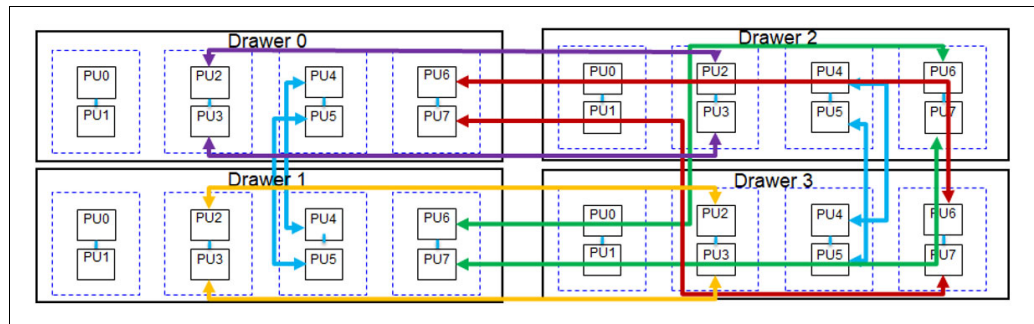


Figure 3-7 Point-to-point topology with four CPC drawers

Inter-CPC drawer communication occurs at the Level 4 virtual cache level, which is implemented on the semi-private part of one of the Level 2 caches in a chip module. The Level 4 cache function regulates coherent drawer-to-drawer traffic.

Note: PU chips 0 and 1 (the first DCM) of each drawer don't have direct connections to other drawers; Transfer is through other DCMs in the drawer.



## 3.4 Processor unit design

Processor cycle time is especially important for processor-intensive applications. Current systems design is driven by processor cycle time, although improved cycle time does not automatically mean that the performance characteristics of the system improve, or at least not as much as the cycle time improvement.

The IBM z17 ME1 core frequency is 5.5 GHz (increased from 5.2 GHz in the IBM z16 A01), with increased number of processors that share larger caches to have shorter access times and improved capacity and performance.

Through innovative processor design (significant architecture changes, new cache structure, new Core-Nest interface, new branch prediction design that uses dense SRAM, on-chip AI accelerator for inference, and new on-chip I/O processor), the IBM Z processor's performance continues to evolve.

Enhancements were made on the processor unit design, including the following examples:

- ▶ Branch prediction mechanism
- ▶ Floating point unit
- ▶ Divide engine scheduler
- ▶ Load/Store Unit and Operand Store Compare (OSC)
- ▶ Relative nest intensity (RNI) redesigns

For more information about RNI, see 12.4, “Relative Nest Intensity” on page 495.

Performance was enhanced through the following changes to the IBM z17 processor design:

- ▶ Core optimization to enable performance and capacity growth.
- ▶ A larger cache Level 2 (SRAM) and virtual Level 3 and Level 4 cache to reduce latency.
- ▶ DPU On-chip IBM I/O processor. For more information see 3.4.7, “IBM z17 DPU - Data Processing Unit” on page 97.
- ▶ Enhancement of nest-core staging.
- ▶ On-chip IBM second generation Integrated Accelerator for AI. For more information, see Chapter 2, “Central processor complex hardware components” on page 19, and Appendix A, “IBM Z Integrated Accelerator for AI and IBM Spyre AI Accelerator” on page 515.

Because of these enhancements, the IBM z17 processor full speed z/OS single-thread performance is on average 1.11 times faster than the IBM z16 at equal N-way, and an average 1.15 times faster for the max capacity (IBM z17 Max 208). For more information about performance, see Chapter 12, “Performance and capacity planning” on page 489.

IBM z13® servers introduced architectural extensions with instructions that reduce processor quiesce effects, cache misses, and pipeline disruption, and increase parallelism with instructions that process several operands in a single instruction (SIMD). The processor architecture was further developed for IBM z16 and IBM z17.

IBM z17 includes the following enhancements:

- ▶ Improved Out-of-Order core execution
- ▶ Improvements in branch prediction and handling
- ▶ Pipeline optimization

The IBM z17 enhanced Instruction Set Architecture (ISA) includes a set of instructions that are added to improve compiled code efficiency. These instructions optimize PUs to meet the



demands of various business and analytics workload types without compromising the performance characteristics of traditional workloads.

### 3.4.1 Simultaneous multithreading

Aligned with industry directions, IBM z17 servers can process up to two simultaneous threads in a single core while sharing certain resources of the processor, such as execution units, translation lookaside buffers (TLBs), and caches. When one thread in the core is waiting for other hardware resources, the second thread in the core can use the shared resources rather than remaining idle. This capability is known as *simultaneous multithreading* (SMT).

An operating system with SMT support can be configured to dispatch work to a thread on a zIIP (for eligible workloads in z/OS) or an IFL (for z/VM and Linux on IBM Z) core in single thread or SMT mode so that HiperDispatch cache optimization can be considered. For more information about operating system support, see Chapter 7, “Operating systems support” on page 261. All SAP processors except one also work in SMT mode.

SMT technology allows instructions from more than one thread to run in any pipeline stage at a time. SMT can handle up to four pending translations.

Each thread has its own unique state information, such as Program Status Word (PSW) and registers. The simultaneous threads cannot necessarily run instructions instantly and must at times compete to use certain core resources that are shared between the threads. In some cases, threads can use shared resources that are not experiencing competition.

Two threads (A and B) that are running on the same processor core on different pipeline stages and sharing the core resources is shown in Figure 3-8.

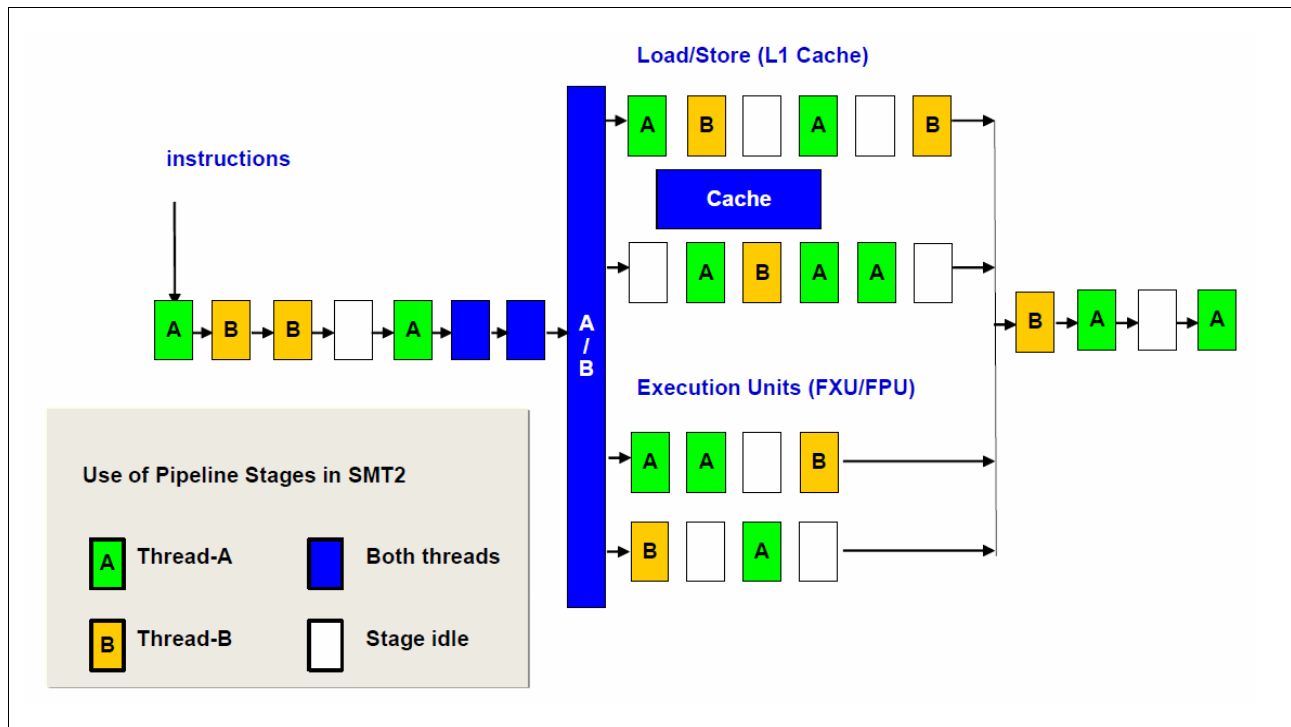


Figure 3-8 Two threads running simultaneously on the same processor core

The use of SMT provides more efficient use of the processors' resources and helps address memory latency, which results in overall throughput gains. The active thread shares core



resources in space, such as data and instruction caches, TLBs, branch history tables, and, in time, pipeline slots, execution units, and address translators.

Although SMT increases the processing capacity, the performance in some cases might be superior if a single thread is used. Enhanced hardware monitoring supports measurement through CPUMF (via SMF 113 for z/OS) and z/OS RMF for thread usage and capacity.

For workloads that need maximum thread speed, the partition's SMT mode can be turned off. For workloads that need more throughput to decrease the dispatch queue size, the partition's SMT mode can be turned on.

SMT use is functionally transparent to middleware and applications, and no changes are required to run them in an SMT-enabled partition.

### 3.4.2 Single-instruction multiple-data

The IBM z17 superscalar processor has 32 vector registers and an instruction set architecture that includes a subset of instructions (known as SIMD) that were added to improve the efficiency of complex mathematical models and vector processing. These new instructions allow a larger number of operands to be processed with a single instruction. The SIMD instructions use the superscalar core to process operands in parallel.

SIMD provides the next phase of enhancements of IBM Z analytics capability. The set of SIMD instructions is a type of data parallel computing and vector processing that can decrease the number of instructions in a program and accelerate code that handles integer, string, character, and floating point data types. The SIMD instructions improve performance of complex mathematical models and allow integration of business transactions and analytic workloads on IBM Z servers.

The 32 vector registers feature 128 bits. The instructions include string operations, vector integer, and vector floating point operations. Each register contains multiple data elements of a fixed size. The following instruction codes specify which data format to use and the size of the elements:

- ▶ Byte (16 8-bit operands)
- ▶ Halfword (eight 16-bit operands)
- ▶ Word (four 32-bit operands)
- ▶ Doubleword (two 64-bit operands)
- ▶ Quadword (one 128-bit operand)

The 128-bit collection of elements in a register is called a *vector*. A single instruction operates on all of the elements in the register. Instructions include a nondestructive operand encoding that allows the addition of the register vectors A1, A2 and A3 and registers vectors B1, B2 and B3 and stores the result in the register vector Cn ( $C_n = A_n + B_n$  with  $n = 1$  to 3).



A schematic representation of a SIMD instruction with 16-byte size elements in each vector operand is shown in Figure 3-9.

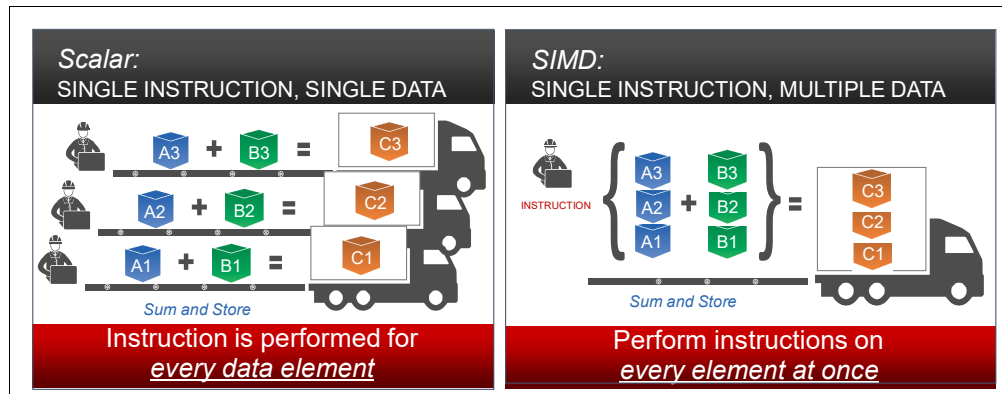


Figure 3-9 SIMD operation logic

The vector register file overlays the floating-point registers (FPRs), as shown in Figure 3-10. The FPRs use the first 64 bits of the first 16 vector registers, which saves hardware area and power, and makes it easier to mix scalar and SIMD codes. Effectively, the core gets 64 FPRs, which can further improve FP code efficiency.

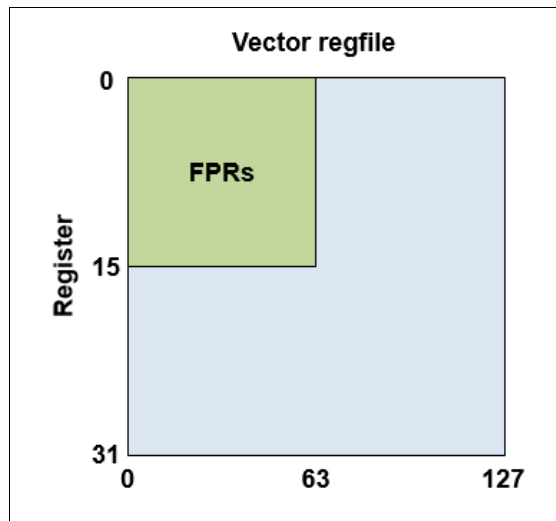


Figure 3-10 Floating point registers overlaid by vector registers

SIMD instructions include the following examples:

- ▶ Integer byte to quadword add, sub, and compare
- ▶ Integer byte to doubleword min, max, and average
- ▶ Integer byte to word multiply
- ▶ String find 8-bit, 16-bit, and 32-bit
- ▶ String range compare
- ▶ String find any equal
- ▶ String load to block boundaries and load/store with length

For most operations, the condition code is not set. A summary condition code is used only for a few instructions.

In z17 vector processing has been enhanced:



- ▶ With Vector-Enhancements Facility 3 new instructions have been added.
- ▶ With Vector-Packed-Decimal-Enhancement 3 performance improvements for COBOL programs when compiled using the NUMCHECK option to detect and convert data.
- ▶ The number of vector rename registers has been increased - to speed up arithmetic and AI workloads.
- ▶ Integer arithmetic extension to 64-bit and 128-bit elements.

### 3.4.3 Out-of-Order execution

IBM z17 servers continue with the Out-of-Order core design. This optimized Out-of-Order feature yields significant performance benefits for compute-intensive applications. It does so by reordering instruction execution, which allows later (younger) instructions to be run ahead of a stalled instruction, and reordering storage accesses and parallel storage accesses. Out-of-Order maintains good performance growth for traditional applications.

Out-of-Order execution can improve performance in the following ways:

- ▶ Reordering instruction execution  
Instructions stall in a pipeline because they are waiting for results from a previous instruction or the execution resource that they require is busy. In an in-order core, this stalled instruction stalls all later instructions in the code stream. In an out-of-order core, later instructions are allowed to run ahead of the stalled instruction.
- ▶ Reordering storage accesses  
Instructions that access storage can stall because they are waiting on results that are needed to compute the storage address. In an in-order core, later instructions are stalled. In an out-of-order core, later storage-accessing instructions that can compute their storage address are allowed to run.
- ▶ Hiding storage access latency  
Many instructions access data from storage. Storage accesses can miss the L1 and require 7 - 50 more clock cycles to retrieve the storage data. In an in-order core, later instructions in the code stream are stalled. In an out-of-order core, later instructions that are not dependent on this storage data are allowed to run.

The IBM z17 processor includes pipeline enhancements that benefit Out-of-Order execution. The processor design features advanced micro-architectural innovations that provide the following benefits:

- ▶ Maximized instruction-level parallelism (ILP) for a better cycles per instruction (CPI) design.
- ▶ Maximized performance per watt.
- ▶ Enhanced instruction dispatch and grouping efficiency.
- ▶ Increased Out-of-Order resources, such as Global Completion Table entries, physical GPR entries, and physical FPR entries.
- ▶ Improved completion rate.
- ▶ Reduced cache/TLB miss penalty.
- ▶ Improved execution of D-Cache store and reload and new Fixed-point divide.
- ▶ New Operand Store Compare (OSC) (load-hit-store conflict) avoidance scheme.
- ▶ Enhanced branch prediction structure and sequential instruction fetching.
- ▶ Load Indexed Address instruction to speed up address manipulation, especially in Linux.



Out-of-Order execution relies on a technique called Register Renaming, where program registers are mapped onto Rename Registers. The more Rename Registers the more effective Out-of-Order execution can be. In z17 the number of Rename Registers has been increased from 128 to 160.

### Program results

Out-of-Order execution does not change any program results. Execution can occur out of (program) order, but all program dependencies are accepted, and the same results are seen as in-order (program) execution. The design was optimized by increasing the Global Completion Table (GCT) from 48x3 to 60x3, which increased the issue queue size from 2x30 to 2x36 and designed a new Mapper.

This implementation requires special circuitry to make execution and memory accesses display in order to the software. The logical diagram of an IBM z17 core is shown in Figure 3-11.

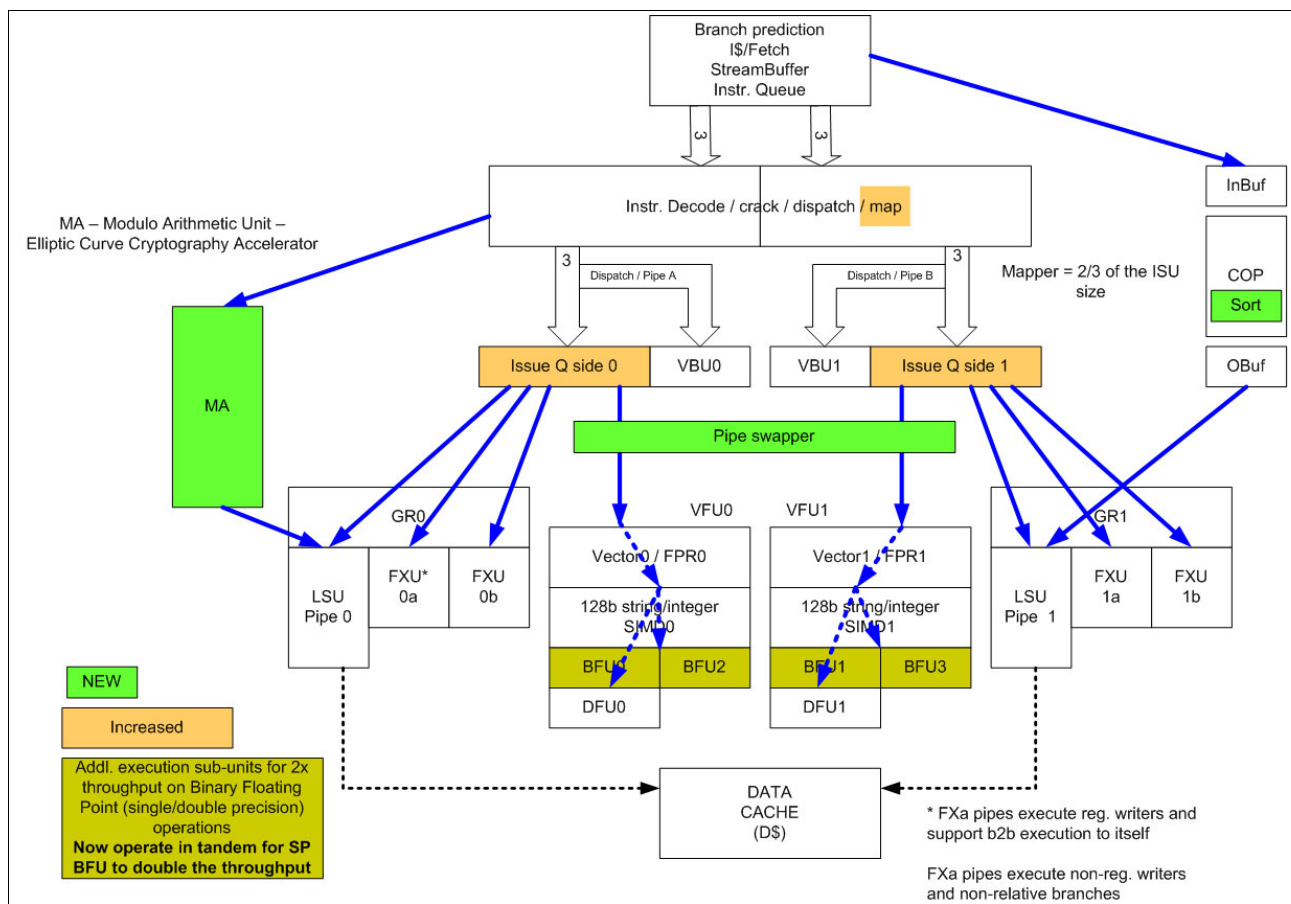


Figure 3-11 IBM z17 PU core logical diagram

Memory address generation and memory accesses can occur out of (program) order. This capability can provide a greater use of the IBM z17 superscalar core, and improve system performance.

The IBM z17 ME1 processor unit core is a superscalar, out-of-order, SMT processor with eight execution units. Up to six instructions can be decoded per cycle, and up to 12 instructions or operations can be started to run per clock cycle (0.181ns). The execution of the instructions can occur out of program order and memory address generation and memory



accesses can also occur out of program order. Each core includes special circuitry to display execution and memory accesses in order to the software.

The IBM z17 superscalar PU core can have up to 10 instructions or operations that are running per cycle. This technology results in shorter workload runtime.

### **Enhanced branch prediction**

If the branch prediction logic of the microprocessor makes the wrong prediction, all instructions in the parallel pipelines are removed. The wrong branch prediction is expensive in a high-frequency processor design. Therefore the branch prediction techniques that are used are important to prevent as many wrong branches as possible.

Equally, if the branch prediction logic fails to keep ahead of instruction execution the latter has to pause. So the logic has to be highly performant and efficient.

In general branch prediction logic has two objectives:

- ▶ Predicting whether a conditional branch will be taken. Generally this is referred to as “Direction”. Perhaps confusingly, this is sometimes also known as “Branch Prediction”. A better term would be “Branch Direction Prediction”.

You can think of this as “do we take the branch?” For conditional branches this is not known for sure until we execute the branch instruction, evaluating the condition.

- ▶ Predicting where execution will branch to. Generally this is known as the “Target”. It’s also known as “Branch Target Prediction”.

You can think of this as “where do we branch to, if we do?” Often this is not known for sure until the branch is taken, for example as subroutine returns are implemented as branches and it could be called from different places.

With successive generations of IBM Z processor, optimisations have been made - both for prediction accuracy and for efficiency.

To ensure branch prediction and branch target prediction are effective, various history-based prediction mechanisms are used, as shown in the in-order part of the IBM z17 PU core logical diagram in Figure 3-11 on page 87.

The Branch Target Buffer (BTB) runs ahead of instruction cache prefetches to prevent branch misses in an early stage. Furthermore, a branch history table (BHT) offers a high branch prediction success rate in combination with a pattern history table (PHT) and the use of tagged multi-target prediction technology branch prediction.

### ***Branch Direction Prediction***

Since z15 Branch Direction Prediction has used a two table TAGE<sup>5</sup> Pattern History Table (PHT) to predict based on history: Each of the tables uses a different history length:

- ▶ A “short history” table, using the last 9 branches
- ▶ A “long history” table, using the last 17 branches

If both tables supply a hit the longer one wins, being more specific and therefore more likely to reflect the current behaviour.

---

<sup>5</sup> Tagged GEometric predictor.



For z17 these tables are paired, with common predictions residing in a slightly faster PHT-1 and less common ones in a slightly slower PHT-2. PHT-1 and PHT-2 are both the same size. The intention of doubling up is to increase PHT capacity and to improve branch prediction's ability to keep ahead.

Branch direction prediction also uses a Perceptron-based mechanism, introduced with z14. "Perceptron" is an industry term. This is a neural network algorithm that learns to correlate branch history over time to predict the direction of branches that the other mechanisms cannot catch with sufficient accuracy.

### ***Branch Target Prediction***

Branch Target Prediction is implemented using a two level Branch Target Buffer (BTB):

- ▶ BTB1 ("small" and "fast")
- ▶ BTB2 (large, dense-SRAM)

Starting with z16, BTB1 and BTB2 feature dynamic (variable) capacity, adapting to changing conditions:

- ▶ **BTB1:** First Level Branch Target Buffer, which is smaller than IBM z15, dynamic director, variable capacity:
  - Minimum total branches in all parents (all *large* branches) = 8 K
  - Maximum total branches in all parents (all *medium* branches) = 12 K
- ▶ **BTB2:** Second Level Branch Target Buffer, also variable capacity (variable directory), up to 260 k branches

Branch Target Prediction has two further mechanisms of note:

- ▶ Two table TAGE Changing Target Buffer (CTB): A two-level table (with different history lengths). Branches are remembered that have different targets depending on history. This is typically subroutine returns and branch tables.
- ▶ Return Address Table Call/Return Stack (RAT CRS): Multi-level CRS that is implemented as a table lookup

## **3.4.4 Superscalar processor**

A *scalar processor* is a processor that is based on a single-issue architecture, which means that only a single instruction is run at a time. A *superscalar processor* allows concurrent (parallel) execution of instructions by adding resources to the microprocessor in multiple pipelines, each working on its own set of instructions to create parallelism.

A superscalar processor is based on a multi-issue architecture. However, when multiple instructions can be run during each cycle, the level of complexity is increased because an operation in one pipeline stage might depend on data in another pipeline stage. Therefore, a superscalar design demands careful consideration of which instruction sequences can successfully operate in a long pipeline environment.

IBM z17 is a superscalar processor. Each processor unit, or core, is a superscalar and out-of-order processor that supports 10 concurrent issues to execution units in a single CPU cycle:

- ▶ Fixed-point unit (FXU): The FXU handles fixed-point arithmetic.



- ▶ Load-store unit (LSU): The LSU contains the data cache. It is responsible for handling all types of operand accesses of all lengths, modes, and formats as defined in the z/Architecture.
- ▶ Instruction fetch and branch (IFB) (prediction) and Instruction cache and merge (ICM). These two sub units (IFB and ICM) contain the instruction cache, branch prediction logic, instruction fetching controls, and buffers. Its relative size is the result of the elaborate branch prediction.
- ▶ L1 data and L1 instruction are incorporated into the LSU and ICM, respectively.

### COBOL enhancements

IBM z17 core implements new instructions for the compiler to accelerate numeric formatting, and hardware support for new numeric conversion instructions (exponents and arithmetic common in financial applications).

## 3.4.5 On-chip coprocessors and accelerators

This section introduces the IBM Integrated Accelerator for zEnterprise Data Compression (zEDC) and the CPACF enhancements for IBM z17.

### IBM integrated Accelerator for zEDC (on-chip)

Introduced in IBM z15, the On-Chip data compression accelerator (Nest Accelerator Unit - NXU, see Figure 3-12 on page 91) provides real value for existing and new data compression use cases.

IBM z17 Compression/Decompression accelerator is implemented in the Nest Accelerator Unit (NXU) on each processor chip of the IBM z17 microprocessor. IBM z17 On-Chip Compression delivers industry-leading throughput and replaces the zEDC Express PCIe adapter available on the IBM z14® and earlier servers.

One Nest Accelerator Unit (NXU) is used per processor chip, which is shared by all cores on the chip and features the following benefits:

- ▶ Brand new concept of sharing and operating an accelerator function in the nest
- ▶ Supports DEFLATE compliant compression/decompression and GZIP CRC/ZLIB Adler
- ▶ Low latency
- ▶ High bandwidth
- ▶ Problem state execution
- ▶ Hardware/Firmware interlocks to ensure system responsiveness
- ▶ Designed instruction
- ▶ Run in millicode

The On-Chip Compression Accelerator removes this virtualization constraint because it is shared by all PUs on the processors chip; therefore, it is available to all LPARs and guests.

Moving the compression function from the I/O drawer to the processor chip means that compression can operate directly on L2 cache and data does not need to be passed by using I/O.

Data compression is running in one of the two execution modes available: Synchronous mode or Asynchronous mode:

- ▶ Synchronous execution occurs in problem states where the user application starts the instruction in its virtual address space.



- Asynchronous execution is optimized for Large Operations under z/OS for authorized applications (for example, BSAM/QSAM) and issues I/O by using EADMF for asynchronous execution.

Asynchronous execution maintains the current user experience and provides a transparent implementation for existing authorized users of zEDC.

The On-Chip data compression implements compression as defined by RFC1951 (DEFLATE).

Figure 3-12 shows the nest compression accelerator (NXU) for On-Chip Compression acceleration.

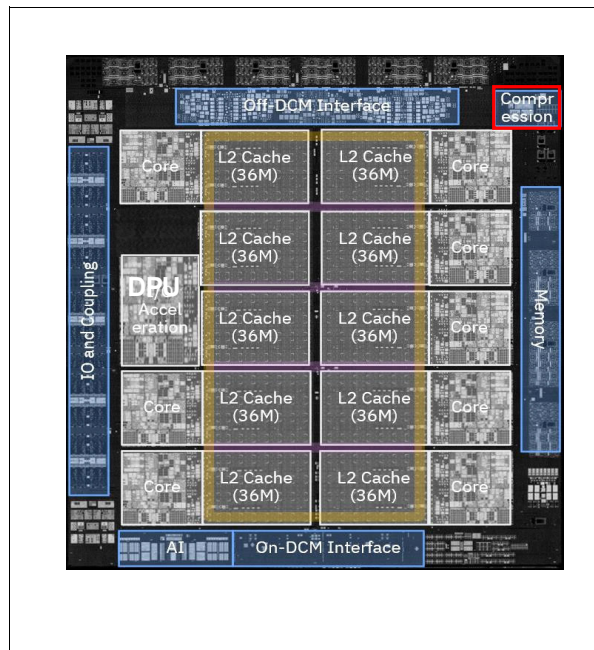


Figure 3-12 Integrated Accelerator for zEDC (NXU) on the IBM z17 PU chip

For more information about sizing, migration considerations, and software support, see Chapter B, “IBM Integrated Accelerator for zEnterprise Data Compression” on page 525.

### Coprocessor units (on-core)

A data compression coprocessor and a cryptography coprocessor unit is available on *each core* in the IBM z17 chip.

The compression engine uses static dictionary compression and expansion that is based on CMPSC<sup>6</sup> instruction. The compression dictionary uses the level 1 (L1) cache (instruction cache).

The cryptography coprocessor is used for CPACF, which offers a set of symmetric cryptographic functions for encrypting and decrypting of clear key operations.

The compression and cryptography coprocessors feature the following characteristics:

- Each core has an independent compression and cryptographic engine.
- The coprocessor was redesigned to support SMT operation and for throughput increase.
- It is available to any processor type (regardless of the processor characterization).
- The owning processor is busy when its coprocessor is busy.

<sup>6</sup> For more information about CMPSC instruction see the latest : [z/Architecture Principles of Operation, SA22-7832](#)



The location of the coprocessor on the IBM z17 chip is shown in Figure 3-13.

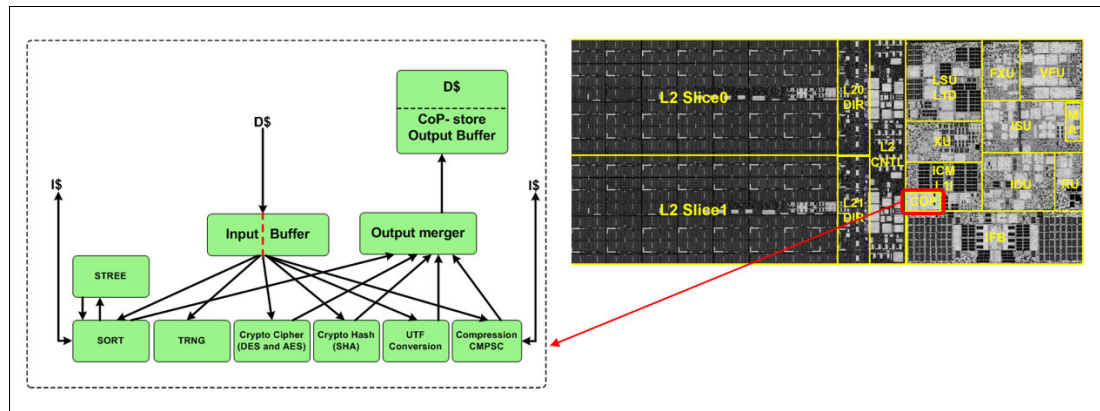


Figure 3-13 IBM z17 on Core co-processor

### On-core compression (CMPSC) on IBM z17

The compression coprocessor on IBM z17 provides the same functions that are available on IBM z16.

### On-core cryptography coprocessor (CPACF)

CPACF accelerates the encrypting and decrypting of SSL/TLS transactions, virtual private network (VPN)-encrypted data transfers, and data-storing applications that do not require FIPS 140-2 level 4 security. The assist function uses a special instruction set for symmetrical clear key cryptographic encryption and decryption, and for hash operations. This group of instructions is known as the *Message-Security Assist (MSA)*.

For more information about these instructions, see the latest version of the [z/Architecture Principles of Operation](#), SA22-7832.

## Crypto functions enhancements

The IBM z17 microprocessor structure is optimized and aligned to the new cache hierarchy. Co-processor results (data) are stored by way of level 1 (L1) cache.

The crypto/hashing/UTF-conversion/compression engines were redesigned for increased throughput.

New CPACF accelerator that is built into every core supports Pervasive Encryption by providing fast synchronous cryptographic services:

- ▶ Encryption (DES, TDES, and AES)
- ▶ Hashing (SHA-1, SHA-2, SHA-3, and SHAKE)
- ▶ Random Number Generation (PRNG, DRNG, and TRNG)
- ▶ Elliptic Curve operations (ECDH[E], ECDSA, and EdDSA)

For more information about cryptographic functions on IBM z17 servers, see Chapter 6, “Cryptographic features” on page 221.

## IBM Integrated Accelerator for Z Sort (on-core)

Sorting data is a significant part of IBM Z workloads including batch workloads, database query processing, and utility processing. The amount of data that is stored and processed on IBM Z continues to grow at a high rate, which drives an ever-increasing sort workload.



Introduced on IBM z15 was the sort accelerator that is known as the IBM Integrated Accelerator for Z Sort (see Figure 3-13 on page 92). The SORTL hardware instruction that is implemented on each core is used by DFSORT and the Db2 utilities for z/OS Suite to allow the use of a hardware-accelerated approach to sorting.

The IBM Integrated Accelerator for Z Sort feature termed as “ZSORT” helps to reduce the CPU costs and improve the elapsed time for eligible workloads. One of the primary requirements for ZSORT is providing enough virtual, real, and auxiliary storage.

Sort jobs that run in memory-constrained environments in which the amount of memory that is available to be used by DFSORT jobs is restricted might not achieve optimal performance results or might not be able to use ZSORT.

The 64-bit memory objects (above-the-bar-storage) can use the ZSORT accelerator for sort workloads for optimal results. Because ZSORT is part of the CPU and memory latency is much less than disk latency, sorting in memory is more efficient than sorting with memory and disk workspace. By allowing ZSORT to process the input completely in memory, it can achieve the best results in elapsed time and CPU time.

Because the goal of ZSORT is to reduce CPU time and elapsed time, it can require more storage than a DFSORT application that does not use ZSORT.

**Note:** Not all sorts are eligible to use ZSORT. IBM’s zBNA tool provides modeling support for identifying potential ZSORT-eligible candidate jobs and estimates the benefits of ZSORT. The tool uses information in the SMF type 16 records.

The following restrictions disable ZSORT and revert to the use of traditional sorting technique:

- ▶ SORTL facility is not enabled/unavailable on the processor
- ▶ ZSORT is not enabled
- ▶ OPTION COPY or SORT FIELDS=COPY is specified
- ▶ Use of:
  - INREC
  - JOINKEYS
  - MERGE FIELDS
  - MODS(EXIT) statements
  - OUTREC
  - OUTFIL
  - SUM FIELDS
- ▶ Program-invoked sorts
- ▶ Memory objects cannot be created
- ▶ Insufficient memory object storage available (required more than currently available)
- ▶ Unsupported sort fields specified (examples Unicode, Locale, and ALTSEQ)
- ▶ Unknown file size or file size=0.
- ▶ SIZE/FILSZ=Uxxxxxx is specified
- ▶ SORTIN/SORTOUT is a VSAM Cluster
- ▶ Sort control field positions are beyond 4092 and VLSHRT is specified
- ▶ Use of EXCP access method was requested
- ▶ Insufficient storage (for example, above or below the line)



- ▶ Sorting key greater than 4088 bytes or greater than 4080 bytes if EQUALS is specified
- ▶ For variable records, the record length (LRECL) must be greater than 24
- ▶ zHPF is unavailable for a sort that cannot be performed entirely in memory
- ▶ Insufficient amount of sort workspace

### 3.4.6 IBM 2<sup>nd</sup> Generation Integrated Accelerator for Artificial Intelligence

The IBM Z processor chip was enhanced from one generation to another. This enhancement enables various data manipulations (such as compression, sorting, cryptography) directly in hardware, on the processor chip by way of purpose-built accelerators. It also provides eligible workloads with low latency time, high performance, and high throughput.

The new IBM z17 microprocessor chip, also called the IBM Telum II processor, integrates an I/O engine and redesigned AI accelerator. These innovations bring incredible value to applications and workloads that are running on IBM Z platform.

Customers can benefit from the integrated AI accelerator by adding AI operations that are used to perform fraud prevention and fraud detection, customer behavior predictions, and supply chain operations. All of these operations are done in real time and fully integrated in transactional workloads. As a result, valuable insights are gained from their data instantly.

The integrated accelerator for AI delivers AI inference in real time, at large scale, and high throughput rate, with no transaction left behind. The AI capability applies directly to the running transaction. It shifts the traditional paradigm of applying AI to the transactions that were completed. This innovative technology also can be used for intelligent IT workloads placement algorithms, which contributes to the better overall system performance.

The Telum II processor also integrates powerful mechanisms of data prefetch, fast and high capacity level 1 (L1) and level 2 (L2) caches, enhanced branch prediction, and other improvements and innovations that streamlines the data processing by the AI accelerator. The hardware, firmware, and software are vertically integrated to deliver the new AI for inference functions seamless to the applications.

The location of the integrated accelerator for AI on the Telum II chip is shown in Figure 3-14.

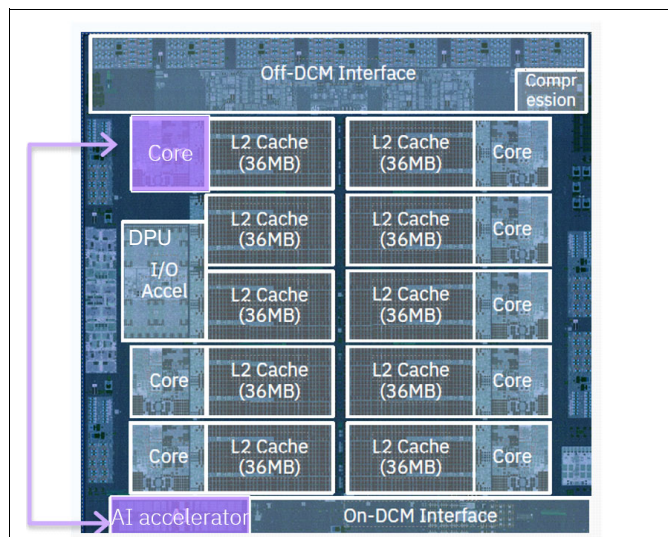


Figure 3-14 Integrated Accelerator for AI on the IBM Telum II processor



The AI accelerator is driven by the new Neural Networks Processing Assist (NNPA) instruction.

NNPA is a new nonprivileged Complex Instruction Set Computer (CISC) memory-to-memory instruction that operates on tensor objects that are in user program's memory. AI functions and macros are abstracted by NNPA.

Figure 3-15 shows the AI accelerator and its components:

- ▶ Data movers surround the compute arrays that consist of the Processor Tiles (PT)
- ▶ Processing Elements (PE)
- ▶ Special Function Processors (SFP)

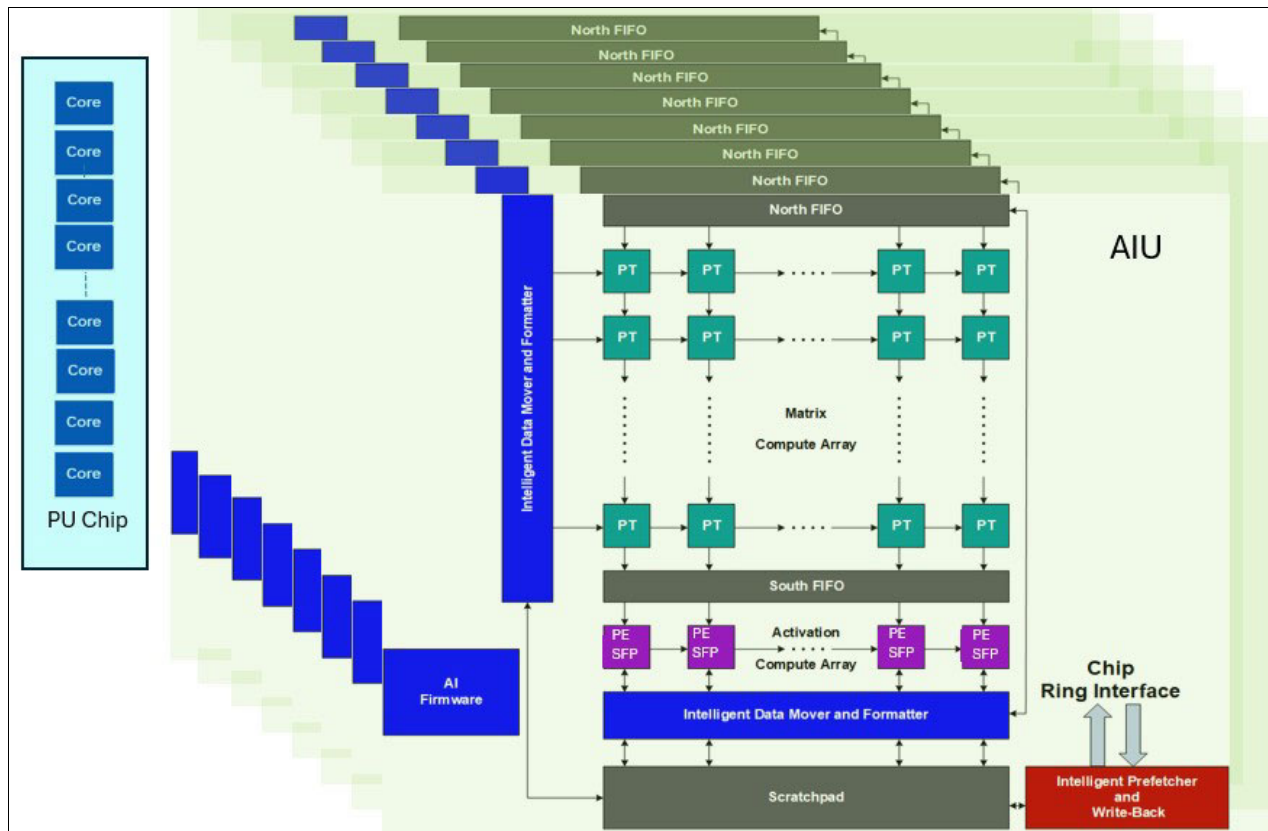


Figure 3-15 IBM z17 2<sup>nd</sup> generation Integrated Accelerator for AI logical diagram

As shown in Figure 3-15 above, all cores in a PU chip have access to its local AIU.

With IBM z17, the cores in one PU chip can transparently access the AIUs in the remote PU chips in the same CPC drawer (access to AIUs in IBM z16 was limited to the AIUs in the same PU chip).

Intelligent data movers and prefetchers are connected to the chip by way of ring interface for high-speed, low-latency, read/write cache operations at 200+ GBps read/store bandwidth, and 600+ GBps bandwidth between engines.



Compute Arrays consist of 128 processor tiles with 8-way FP-16 FMA SIMD, which are optimized for matrix multiplication and convolution, and 32 processor tiles with 8-way FP-16/FP-32 SIMD, which are optimized for activation functions and complex functions.

The integrated AI accelerator delivers more than 24 Trillions of instructions per second (TOPs) per chip and over 752 TOPs in fully configured IBM z17 system with the 32 chips. The AI accelerator is shared by all cores on the chip and by all cores on the remote chips in the same drawer. The firmware, running on the cores and accelerator, orchestrates and synchronizes the execution on the accelerator.

### ***Using IBM Z Integrated AI Accelerator in your enterprise***

Figure 3-16 shows the software ecosystem and high-level integration of the AI accelerator into enterprise AI/Machine Learning solution stack. Great flexibility and interoperability are available for training and building models.

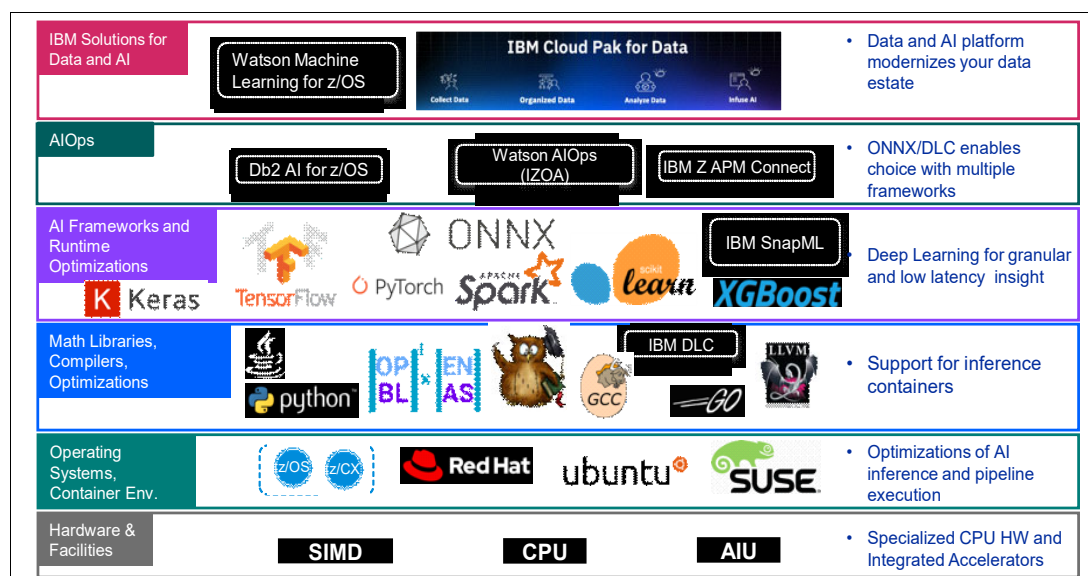


Figure 3-16 Software ecosystem for the AI accelerator

Acknowledging the diverse AI training frameworks, customers can train their models on platforms of their choice, including IBM Z (on-premises and in hybrid cloud) and then, deploy it efficiently on IBM Z in colocation with the transactional workloads. No other development effort is needed to enable this strategy.

IBM has invested into Open Neural Network Exchange (ONNX), which is a standard format for representing AI models that allows a data scientist to build and train a model in the framework of choice without worrying about the downstream inference implications.

To enable deployment of ONNX models, IBM provides an ONNX model compiler that is optimized for IBM Z. IBM also optimized key Open Source frameworks, such as TensorFlow and TensorFlow Serving, for use on IBM Z platform.

IBM open-sourced zDNN library provides common APIs for the functions that allow to convert tensor format to the accelerator required one. Customers can run zDNN under z/OS (in zCX) and Linux on IBM Z.

A Deep Learning Compiler (DLC) for z/OS and for Linux on IBM Z provides the AI functions to the applications.



### 3.4.7 IBM z17 DPU - Data Processing Unit

The IBM z17 Data Processing Unit encompasses a comprehensive refactoring of the I/O subsystem.

With the DPU, functionality from I/O Adapters' ASICs is moved into the Z CP chip. The design aims to build an I/O subsystem with better qualities of services than the existing overprovisioned I/O subsystem. It delivers value by improved performance and power efficiency. It also deliver value to our clients with decreased channel latencies.

Generally, it carries functional existing capabilities forward in the channels that it supports, with some important improvements and increased capabilities. Refer to Figure..

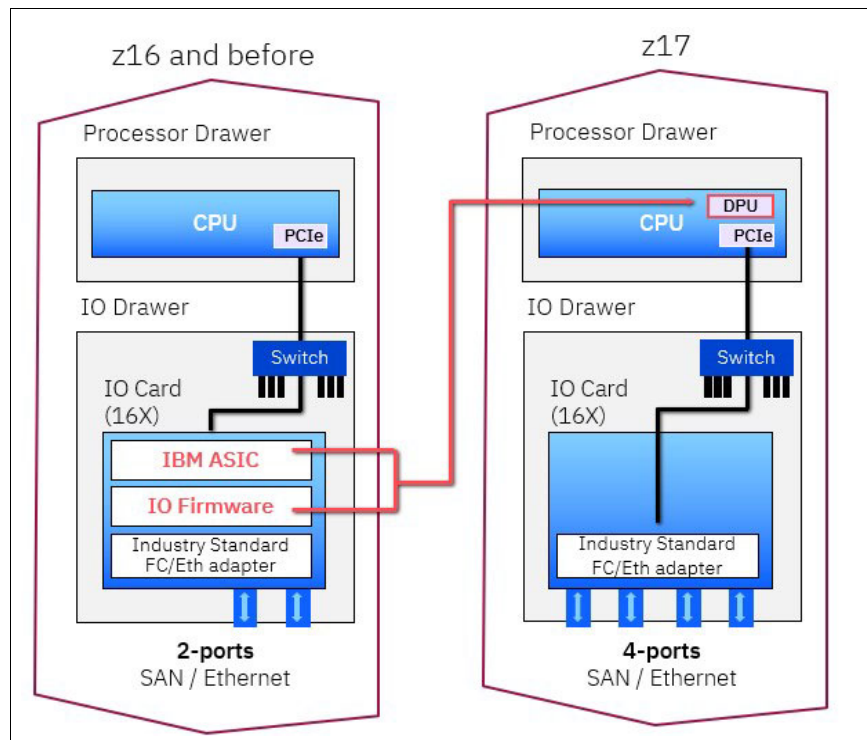


Figure 3-17 IBM z17 DPU

The DPU's goal is to deliver improved IBM Z platform efficiencies:

- ▶ to improve peak I/O start rates and reduce latencies
- ▶ to provide focused per port recovery for the most common types of failures
- ▶ to improve recurring networking costs for clients by providing integrated RoCE SMC-R and OSA support
- ▶ to provide single port serviceability for all managed I/O adapters
- ▶ to reduce dependence on the PCI support partition by providing physical function support for PCIe Native use cases

#### Supported Protocols

The following protocols will run on the DPU:

- ▶ 1. Legacy Mode FICON
- ▶ 2. HPF (High Performance FICON)
- ▶ 3. FCP (SCSI over fiber channel)
- ▶ 4. OSA (Open Systems Adapter)
- ▶ 5. OSA-ICC (Open Systems Adapter - Integrated Console Controller)



- ▶ 6. Physical function support for Native Ethernet exploitation.
  - This support also allows a port to be shared between a PCIe Native protocol and OSA

## DPU Components

The DPU engine design reduces the amount of custom hardware in the I/O cards, eliminating anything that's protocol specific (like data routers), providing strategic hardware assists where appropriate, and doing as much processing as possible in firmware.

There is one DPU complex per CP (PU Chip), where the logic physically occupies the space of 1.5 z-cores. Each PU Chip has a single associated DPU.

A DPU comprises 32 cores arranged in 4 clusters of eight. The DPU has a coprocessor interface as a microarchitectural feature. Figure 3-18 shows the IBM z17 Telum II DPU I/O Engine.

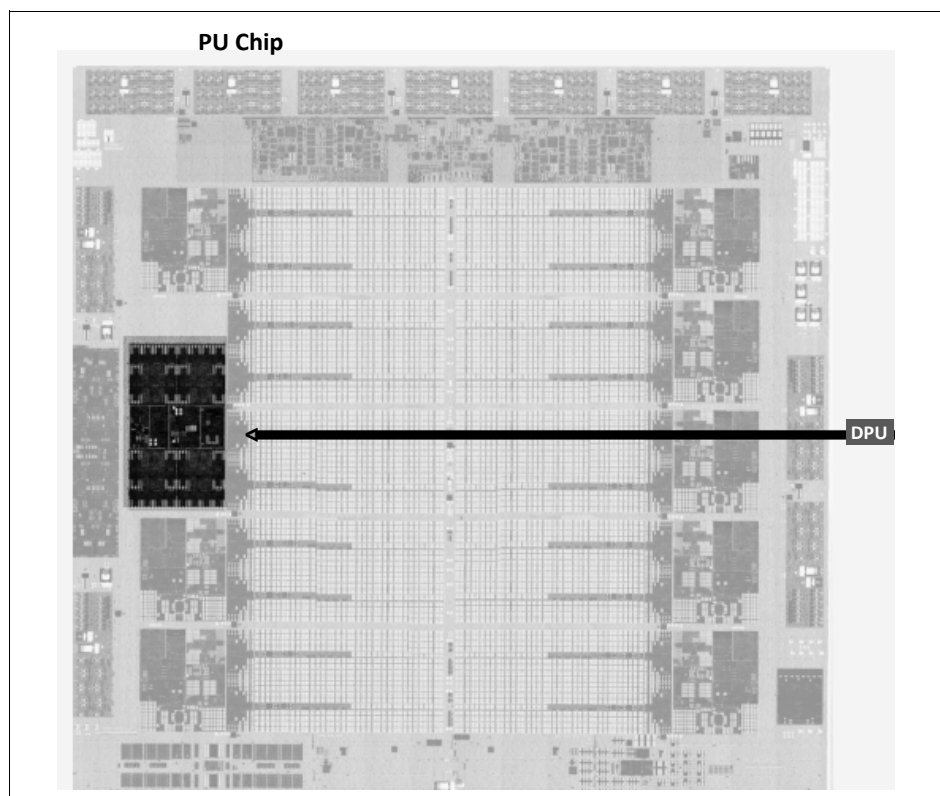


Figure 3-18 Location of the DPU I/O Engine in the Telum II chip

## DPU Sharing

Each DPU is shared by PU cores not only on the chip it resides on but other PU chips in the drawer. This is necessary because each CHPID is managed by a single DPU.



### 3.4.8 Decimal floating point accelerator

Each of the microprocessors (cores) on the 8-core chip has a Decimal Floating Point (DFP) accelerator. Its implementation meets business application requirements for better performance, precision, and function.

Base 10 arithmetic is used for most business and financial computation. Floating point computation that is used for work that is typically done in decimal arithmetic involves frequent data conversions and approximation to represent decimal numbers. This process makes floating point arithmetic complex and error-prone for programmers who use it for applications in which the data is typically decimal.

Hardware DFP computational instructions provide the following features:

- ▶ Data formats of 4, 8, and 16 bytes
- ▶ An encoded decimal (base 10) representation for data
- ▶ Instructions for running decimal floating point computations
- ▶ An instruction that runs data conversions to and from the decimal floating point representation

#### Benefits of the DFP accelerator

The DFP accelerator offers the following benefits:

- ▶ Avoids rounding issues, such as those issues that occur with binary-to-decimal conversions.
- ▶ Controls binary-coded decimal (BCD) operations better.
- ▶ Follows the standardization of the dominant decimal data and decimal operations in commercial computing, which supports the industry standardization (IEEE 745R) of decimal floating point operations. Instructions are added in support of the Draft Standard for Floating-Point Arithmetic - IEEE 754-2008, which is intended to supersede the ANSI/IEEE Standard 754-1985.
- ▶ Allows COBOL programs that use zoned-decimal operations to take advantage of the z/Architecture DFP instructions.

IBM z17 servers have two DFP accelerator units per core, which improve the decimal floating point execution bandwidth. The floating point instructions operate on newly designed vector registers (32 new 128-bit registers).

IBM z17 servers include decimal floating point packed conversion facility support with the following benefits:

- ▶ Reduces code path length because extra instructions to format conversion are no longer needed.
- ▶ Packed data is operated in memory by all decimal instructions without general-purpose registers, which were required only to prepare for decimal floating point packed conversion instruction.
- ▶ Converting from packed can now force the input packed value to positive instead of requiring a separate OI, OILL, or load positive instruction.
- ▶ Converting to packed can now force a positive zero result instead of requiring ZAP instruction.



Cobol and PL/I compilers were updated to support the new IBM z17 enhancements:

- ▶ BCD to HFP conversions
- ▶ Numeric editing operation
- ▶ Zoned decimal operations

### Software support

DFP is supported in the following programming languages and products:

- ▶ Release 4 and later of the High Level Assembler
- ▶ C/C++, which requires supported z/OS version
- ▶ Enterprise PL/I Release 3.7 and Debug Tool Release 8.1 or later
- ▶ Java Applications that use the BigDecimal Class Library
- ▶ SQL support as of Db2 Version 9 and later

## 3.4.9 IEEE floating point

Binary and hexadecimal floating-point instructions are implemented in IBM z17 servers. They incorporate IEEE standards into the system.

The IBM z17 core implements two other execution subunits for 2x throughput on BFP (single/double precision) operations (see Figure 3-11 on page 87).

The key point is that Java and C/C++ applications tend to use IEEE BFP operations more frequently than earlier applications. Therefore, the better the hardware implementation of this set of instructions, the better the performance of applications.

## 3.4.10 Processor error detection and recovery

The PU uses a process called *transient recovery* as an error recovery mechanism. When an error is detected, the instruction unit tries the instruction again, and attempts to recover the error. If the second attempt is unsuccessful (that is, a permanent fault exists), a relocation process is started that restores the full capacity by moving work to another PU.

Relocation under hardware control is possible because the R-unit has the full designed state in its buffer. PU error detection and recovery are shown in Figure 3-19.

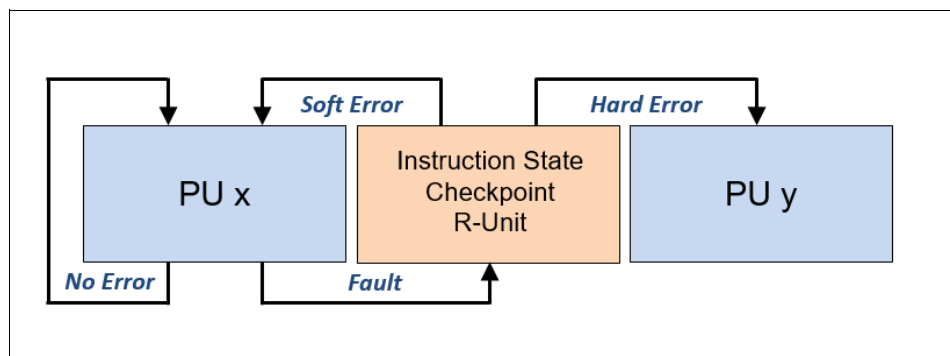


Figure 3-19 PU error detection and recovery



### 3.4.11 Branch prediction

Because of the ultra-high frequency of the PUs, the penalty for a wrongly predicted branch is high. Therefore, a multi-pronged strategy for branch prediction is implemented on each core based on gathered branch history that is combined with other prediction mechanisms.

The BHT (Branch History Table) implementation on processors provides a large performance improvement. Originally introduced on the IBM ES/9000 9021 in 1990, the BHT is continuously improved.

It offers significant branch performance benefits. The BHT allows each PU to take instruction branches that are based on a stored BHT, which improves processing times for calculation routines.

In addition to the BHT, IBM z17 servers use the following techniques to improve the prediction of the correct branch to be run:

- ▶ BTB
- ▶ PHT
- ▶ CTB

The success rate of branch prediction contributes significantly to the superscalar aspects of IBM z17 processor. This success is because the architecture rules prescribe that the correctly predicted result of the branch is essential for successful parallel execution of an instruction stream.

IBM z17 integrates a new branch prediction design that uses SRAM and supports the following:

- ▶ BTB1: 8K - 12K
- ▶ BTB2: up to 260K
- ▶ TAGE PHT: 4K x 2
- ▶ TAGE PHT2: 4K x 2
- ▶ TAGE CTB: 1K x 2

### 3.4.12 Wild branch

When a bad pointer is used or when code overlays a data area that contains a pointer to code, a random branch is the result. This process causes a 0C1 or 0C4 abend. Random branches are difficult to diagnose because clues about how the system got to that point are not evident.

With the wild branch hardware facility, the last address from which a successful branch instruction was run is kept. z/OS uses this information with debugging aids, such as the **SLIP** command, to determine from where a wild branch came.

It also can collect data from that storage location. This approach decreases the number of debugging steps that are necessary when you want to know from where the branch came.



### 3.4.13 Translation lookaside buffer

The TLB in the instruction and data L1 caches use a secondary TLB to enhance performance.

The size of the TLB is kept as small as possible because of its short access time requirements and hardware space limitations. Because memory sizes recently increased significantly as a result of the introduction of 64-bit addressing, a smaller working set is represented by the TLB.

To increase the working set representation in the TLB without enlarging the TLB, large (1 MB) page and giant page (2 GB) support is available and can be used when suitable. For more information, see “Large page support” on page 119.

With the enhanced DAT-2 (EDAT-2) improvements, the IBM Z servers support 2 GB page frames.

#### IBM z16 & z17 TLB

IBM z16 switches to a logical-tagged L1 directory and inline TLB2. Each L1 cache directory entry contains the virtual address and Address Space Control Element (ASCE) because it no longer must access TLB for L1 cache hit. TLB2 is accessed in parallel to L2, which saves significant latency compared to TLB1-miss.

The new translation engine allows up to four translations pending concurrently. Each translation step is ~2x faster, which helps second level guests.

In z17 the TLB lookup pipeline is modified to handle both Demand and Prefetches.

### 3.4.14 Instruction fetching, decoding, and grouping

The superscalar design of the microprocessor allows for the decoding of up to six instructions per cycle and the execution of up to 12 instructions per cycle. Both execution and storage accesses for instruction and operand fetching can occur out of sequence.

#### Instruction fetching

Instruction fetching normally tries to get as far ahead of instruction decoding and execution as possible because of the relatively large instruction buffers that are available. In the microprocessor, smaller instruction buffers are used. The operation code is fetched from the I-cache and put in instruction buffers that hold prefetched data that is awaiting decoding.

z17 has a new Prefetch Pipeline, driven by the branch prediction logic. It looks for lines that are not in the Level 1 instruction cache but are likely to be in the Level 2 cache.

#### Instruction decoding

The processor can decode up to six instructions per cycle. The result of the decoding process is queued and later used to form a group.

#### Instruction grouping

From the instruction queue, up to 12 instructions can be completed on every cycle. A complete description of the rules is beyond the scope of this publication.

Compilers and JVMs are responsible for selecting instructions that best fit with the superscalar microprocessor. They abide by the rules to create code that best uses the



superscalar implementation. All IBM Z compilers and JVMs are constantly updated to benefit from new instructions and advances in microprocessor designs.

### 3.4.15 Extended Translation Facility

The z/Architecture instruction set includes instructions in support of the Extended Translation Facility. They are used in data conversion operations for Unicode data, which causes applications that are enabled for Unicode or globalization to be more efficient. These data-encoding formats are used in web services, grid, and on-demand environments in which XML and SOAP technologies are used. The High Level Assembler supports the Extended Translation Facility instructions.

### 3.4.16 Transactional Execution

The Transactional Execution (TX) capability, which is known in the industry as *hardware transactional memory*, runs a group of instructions atomically; that is, all of their results are committed or no result is committed. The execution is optimistic. The instructions are run, but previous state values are saved in a transactional memory. If the transaction succeeds, the saved values are discarded; otherwise, they are used to restore the original values.

The Transaction Execution Facility provides instructions, including declaring the beginning and end of a transaction, and canceling the transaction. TX is expected to provide significant performance benefits and scalability by avoiding most locks. This benefit is especially important for heavily threaded applications, such as Java.

z17 is the last processor generation to support the Transaction Execution Facility.

### 3.4.17 Runtime Instrumentation

Runtime Instrumentation (RI) is a hardware facility for managed run times, such as the Java Runtime Environment (JRE). RI allows dynamic optimization of code generation as it is being run. It requires fewer system resources than the current software-only profiling, and provides information about hardware and program characteristics. RI also enhances JRE in making the correct decision by providing real-time feedback.

## 3.5 Processor unit functions

The PU functions are described in this section.

### 3.5.1 Overview

All PUs on an IBM z17 are physically identical. When the system is initialized, two integrated firmware processors (IFP) are allocated from the pool of PUs that is available for the entire system. The other PUs can be characterized to specific functions (CP, IFL, ICF, zIIP, or SAP).

The function that is assigned to a PU is set by the Licensed Internal Code (LIC). The LIC is loaded when the system is initialized at power-on reset (POR) and the PUs are *characterized*.

Only characterized PUs include a designated function. Non-characterized PUs are considered spares. You must order at least one CP, IFL, or ICF on an IBM z17.



This design brings outstanding flexibility to IBM z17 servers because any PU can assume any available characterization. The design also plays an essential role in system availability because PU characterization can be done dynamically, with no system outage.

For more information about software level support of functions and features, see Chapter 7, “Operating systems support” on page 261.

## Concurrent upgrades

For all IBM z17 ME1 features that have more processor units (PUs) installed (non-characterized) than activated, concurrent upgrades can be done by using LIC activation. This activation assigns a PU function to a previously non-characterized PU. No hardware changes are required.

The upgrade can be done concurrently through the following facilities:

- ▶ Customer Initiated Upgrade (CIU) for permanent upgrades
- ▶ On/Off Capacity on Demand (On/Off CoD) for temporary upgrades
- ▶ Capacity BackUp (CBU) for temporary upgrades
- ▶ Flexible Capacity for Cyber Resilience upgrades

If the PU chips in the installed CPC drawers have no available remaining PUs, an upgrade results in a feature upgrade and the installation of an extra CPC drawer. Field add (MES) of a CPC drawer is possible for IBM z17 Model ME1 features Max43 and Max90 only. These features can be upgraded to a Max136 provided initial order for the CPC Reserve features FC 2933 or FC 2934. CPC drawer installation is nondisruptive, but takes more time than a simple LIC upgrade. Features Max183 and Max208 are factory build only.

For more information about Capacity on Demand, see Chapter 8, “System upgrades” on page 353.

## PU sparing

If a PU failure occurs, the failed PU's characterization is dynamically and transparently reassigned to a spare PU. IBM z17 servers have two spare PUs. PUs that are not characterized on a CPC configuration also can be used as extra spare PUs.

For more information about PU sparing, see 3.5.10, “Sparing rules” on page 116.

## PU pools

PUs that are defined as CPs, IFLs, ICFs, and zIIPs are grouped in their own pools from where they can be managed separately. This configuration significantly simplifies capacity planning and management for LPARs. The separation also affects weight management because CP and zIIP weights can be managed separately.

For more information, see “[PU weighting](#)” on page 105.

All assigned PUs are grouped in the PU pool. These PUs are dispatched to online logical PUs. For example, consider an IBM z17 with 10 CPs, 2 IFLs, 5 zIIPs, and 1 ICF. This system has a PU pool of 18 PUs, called the *pool width*. Subdivision defines the following pools:

- ▶ A CP pool of 10 CPs
- ▶ An ICF pool of one ICF
- ▶ An IFL pool of two IFLs
- ▶ A zIIP pool of five zIIPs

PUs are placed in the pools in the following circumstances:



- ▶ When the system is POREd.
- ▶ At the time of a concurrent upgrade.
- ▶ As a result of adding PUs during a CBU.
- ▶ Following a capacity on-demand upgrade through On/Off CoD or CIU.

PUs are removed from their pools:

- ▶ When a concurrent downgrade occurs as the result of the removal of a CBU.
- ▶ Through the On/Off CoD process.
- ▶ The conversion of a PU.

When a dedicated LPAR is activated, its PUs are configured from the appropriate pools. This process also is the case when an LPAR logically configures a PU as on, if the width of the pool allows for it.

For an LPAR, logical PUs are dispatched on physical PUs in the supporting pool. The logical CPs are dispatched from the CP pool, logical zIIPs from the zIIP pool, logical IFLs from the IFL pool, and the logical ICFs from the ICF pool.

### PU weighting

Because CPs, zIIPs, IFLs, and ICFs have their own pools from where they are dispatched, shared logical processors are given pool-specific weights. For more information about PU pools and processing weights, see the *Processor Resource/Systems Manager Planning Guide*, SB10-7178.

## 3.5.2 Central processors

A central processor (CP) is a PU that uses the full z/Architecture instruction set. It can run z/Architecture-based operating systems (z/OS, z/VM, z/TPF, VSE<sup>n</sup> (V6.3.1 from 21<sup>st</sup> Century Software), and Linux on IBM Z), and the Coupling Facility Control Code (CFCC). Up to 208 PUs can be characterized as CPs, depending on the configuration.

The IBM z17 can be initialized in LPAR (PR/SM) mode or in Dynamic Partition Manager (DPM) mode.

CPs are defined as dedicated or shared. Reserved CPs can be defined to an LPAR to allow for nondisruptive image upgrades. If the operating system in the LPAR supports the logical processor add function, reserved processors are no longer needed. Regardless of the installed model, an LPAR can have up to 208 logical CPs that are defined (the sum of active and reserved logical CPs). In practice, define no more CPs than the operating system supports.

All PUs that are characterized as CPs within a configuration are grouped into the CP pool. The CP pool can be seen on the Hardware Management Console (HMC) workplace. Any z/Architecture operating systems and CFCCs can run on CPs that are assigned from the CP pool.

The IBM z17 ME1 server recognizes four distinct capacity settings for CPs. Full-capacity CPs are identified as CP7. In addition to full-capacity CPs, three subcapacity settings (CP6, CP5, and CP4), each for up to 43 PUs, are offered.

The following capacity settings appear in hardware descriptions:

- ▶ CP4 Feature Code 1647 (up to 43 PUs)
- ▶ CP5 Feature Code 1648 (up to 43 PUs)
- ▶ CP6 Feature Code 1648 (up to 43 PUs)



- CP7 Feature Code 1650 (up to 208 PUs)

Granular capacity adds 129 subcapacity settings to the 208 capacity settings that are available with full capacity CPs (CP7). Each of the 129 subcapacity settings applies to up to 43 CPs only, independent of the model installed.

**Note:** Information about CPs in the remainder of this chapter applies to all CP capacity settings, unless indicated otherwise. For more information about granular capacity, see 2.3.3, “PU characterization” on page 38.

### 3.5.3 Integrated Facility for Linux (FC 1651)

An IFL is a PU that can be used to run Linux, Linux guests on z/VM operating systems, and an IBM Secure Service Container (SSC). Up to 208 PUs can be characterized as IFLs, depending on the configuration.

**Note:** IFLs can be dedicated to a Linux, a z/VM, or an SSC LPAR, or can be shared by multiple Linux guests, z/VM LPARs, or SSC that are running on the same IBM z17 server. Only z/VM, Linux on IBM Z operating systems, SSC, and designated software products can run on IFLs. IFLs are orderable by using FC 1651.

#### IFL pool

All PUs that are characterized as IFLs within a configuration are grouped into the IFL pool. The IFL pool can be seen on the HMC workplace.

IFLs do not change the model capacity identifier of the IBM z17. Software product license charges that are based on the model capacity identifier are not affected by the addition of IFLs.

#### Unassigned IFLs

An IFL that is purchased but not activated is registered as an unassigned IFL (FC 1654). When the system is later upgraded with another IFL, the system recognizes that an IFL was purchased and is present.

The allowable number of IFLs and Unassigned IFLs per feature is listed in Table 3-1.

Table 3-1 IFLs and Unassigned IFLs per feature

Features	Max43	Max90	Max136	Max183	Max208
Maximum of IFLs FC 1651	43	90	136	183	208
Maximum of Unassigned IFLs FC 1654	42	89	135	182	207

### 3.5.4 Internal Coupling Facility (FC 1652)

An Internal Coupling Facility (ICF) is a PU that is used to run the CFCC for Parallel Sysplex environments. Within the sum of all unassigned PUs in up to four CPC drawers, up to 208 ICFs can be characterized, depending on the model. However, the maximum number of ICFs that can be defined on a coupling facility LPAR is limited to 16. ICFs are orderable by using FC 1652.



## Unassigned ICFs

An ICF that is purchased but not activated is registered as an unassigned ICF (FC 1655). When the system is later upgraded with another ICF, the system recognizes that an ICF was purchased and is present.

The allowable number of ICFs and Unassigned ICFs for each model is listed in Table 3-2.

Table 3-2 ICFs and Unassigned ICFs per feature

Features	Max43	Max90	Max136	Max183	Max208
Maximum of ICFs FC 1652	43	90	136	183	208
Maximum of Unassigned ICFs FC 1655	42	89	135	182	207

ICFs exclusively run CFCC. ICFs do not change the model capacity identifier of the IBM z17 system. Software product license charges that are based on the model capacity identifier are not affected by the addition of ICFs.

All ICFs within a configuration are grouped into the ICF pool. The ICF pool can be seen on the HMC workplace.

The ICFs can be used by coupling facility LPARs only. ICFs are dedicated or shared. ICFs can be dedicated to a CF LPAR, or shared by multiple CF LPARs that run on the same system. However, having an LPAR with dedicated and shared ICFs at the same time is not possible.

## Coupling Thin Interrupts

With the introduction of Driver 15F (zEC12 and zBC12), the IBM z/Architecture provides a thin interrupt class called *Coupling Thin Interrupts*<sup>7</sup>. The capabilities that are provided by hardware, firmware, and software support the generation of coupling-related “Thin Interrupts” when the following situations occur:

- ▶ On the coupling facility (CF) side:
  - A CF command or a CF signal (arrival of a CF-to-CF duplexing signal) is received by a shared-engine CF image.
  - The completion of a CF signal that was previously sent by the CF occurs (completion of a CF-to-CF duplexing signal).
- ▶ On the z/OS side:
  - CF signal is received by a shared-engine z/OS image (arrival of a List Notification signal).
  - An asynchronous CF operation completes.

The interrupt causes the receiving partition to be dispatched by an LPAR if it is not dispatched. This process allows the request, signal, or request completion to be recognized and processed in a more timely manner.

After the image is dispatched, “poll for work” logic in CFCC and z/OS can be used largely as-is to locate and process the work. The new interrupt expedites the redispaching of the partition.

<sup>7</sup> It is the only option for shared processors in a CF image (whether they be ICFs or CPs) on IBM z16.



LPAR presents these Coupling Thin Interrupts to the guest partition, so CFCC and z/OS both require interrupt handler support that can deal with them. CFCC also changes to relinquish control of the processor when all available pending work is exhausted, or when the LPAR undispatches it off the shared processor, whichever comes first.

## CF processor combinations

A CF image can have one of the following combinations that are defined in the image profile:

- ▶ Dedicated ICFs
- ▶ Shared ICFs
- ▶ Shared CPs

Shared ICFs add flexibility. However, running only with shared coupling facility PUs (ICFs or CPs) is not a preferable production configuration. It is preferable for a production CF to operate by using dedicated ICFs.

In Figure 3-20, the CPC on the left participates in two parallel sysplexes (Production and Test), and each has one z/OS and one coupling facility image. The coupling facility images share an ICF.

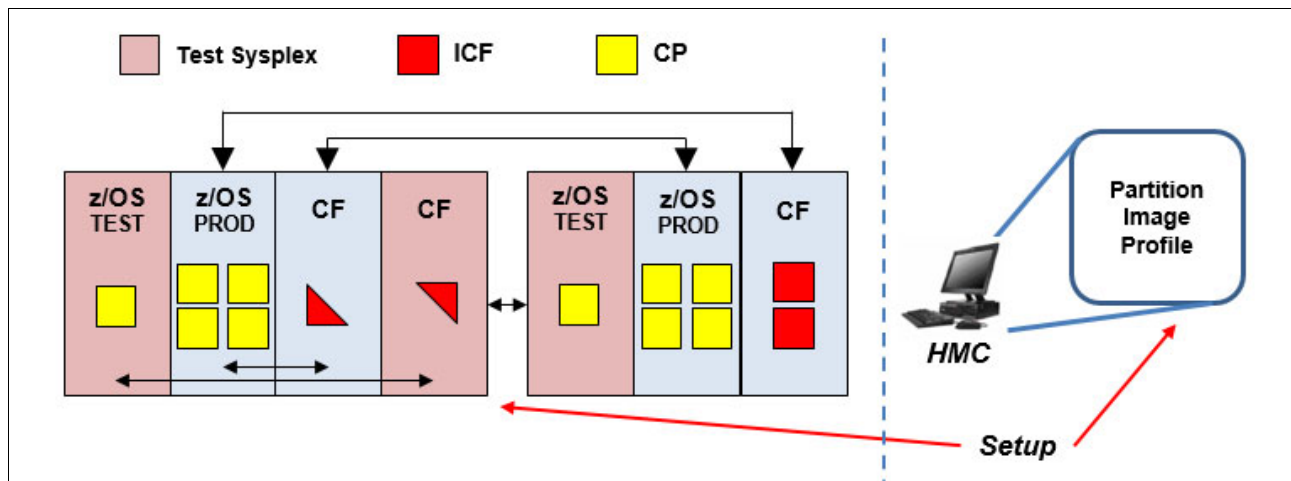


Figure 3-20 ICF options - shared ICFs

The LPAR processing weights are used to define how much processor capacity each CF image is entitled to. The capped option also can be set for a test CF image to protect the production environment.

Connections between these z/OS and CF images can use internal coupling links to avoid the use of real (external) coupling links, and get the best link bandwidth available.

## Dynamic CF dispatching

The *dynamic coupling facility dispatching* (DYNDISP) function features a dispatching algorithm that you can use to define a backup CF in an LPAR on the system. When this LPAR is in backup mode, it uses few processor resources.

DYNDISP allows more environments with multiple CF images to coexist in a server, and to share CF engines with reasonable performance. DYNDISP THIN is the only option for CF images that use shared processors on IBM z17. For more information, see 3.9.3, “Dynamic CF dispatching” on page 142.



### **Coupling Facility Processor scalability**

CF work management and dispatcher changed to improve efficiency as processors are added to scale up the capacity of a CF image.

With IBM z17, the maximum number of CF processors in an LPAR increases from 16 to 32. With this increase, customers might be able to consolidate the CF workload across fewer CF images. This could reduce complexity with fewer coupling links, and logical CHPIDs to define and manage for connectivity, and so on.

For more information about CFCC Level 26 enhancements, see 3.9.1, “CF Control Code (CFCC)” on page 135.

## **3.5.5 IBM Z Integrated Information Processor (FC 1653)**

A zIIP<sup>8</sup> reduces the standard processor (CP) capacity requirements for z/OS Java, XML system services applications, and a portion of work of z/OS Communications Server and Db2 UDB for z/OS Version 8 or later, which frees up capacity for other workload requirements.

A zIIP enables eligible z/OS workloads to have a portion of them directed for execution to a processor that is characterized as a zIIP. Because the zIIPs do not increase the MSU value of the processor, they do not affect the IBM software license charges.

IBM z17 is the fifth generation of IBM Z processors to support SMT. IBM z17 servers implement two threads per core on IFLs and zIIPs. SMT must be enabled at the LPAR level and supported by the z/OS operating system. SMT was enhanced for IBM z17 and it is enabled for SAPs by default (no customer intervention required).

Introduced in z/OS V2R4, the z/OS Container Extensions<sup>9</sup> allows deployment of Linux on IBM Z software components, such as Docker Containers in a z/OS system, in direct support of z/OS workloads without requiring a separately provisioned Linux server. It also maintains overall solution operational control within z/OS and with z/OS qualities of service. Workload deployed in z/OS Container Extensions is zIIP eligible.

### **How zIIPs work**

zIIPs are designed for supporting designated z/OS workloads. One of the workloads is Java code execution. When Java code must be run (for example, under control of IBM WebSphere), the z/OS JVM calls the function of the zIIP. The z/OS dispatcher then suspends the JVM task on the CP that it is running on and dispatches it on an available zIIP. After the Java application code execution is finished, z/OS redispaches the JVM task on an available CP. After this process occurs, normal processing is resumed.

This process reduces the CP time that is needed to run Java WebSphere applications, which frees that capacity for other workloads.

---

<sup>8</sup> IBM z Systems Application Assist Processors (zAAPs) are not available since IBM z14 servers. A zAAP workload is dispatched to available zIIPs (zAAP on zIIP capability).

<sup>9</sup> z/OS Container Extensions that are running on IBM z16 require IBM Container Hosting Foundation for z/OS software product (5655-HZ1).



The logical flow of Java code that is running on an IBM z17 that has a zIIP available is shown in Figure 3-21. When JVM starts the execution of a Java program, it passes control to the z/OS dispatcher that verifies the availability of a zIIP.

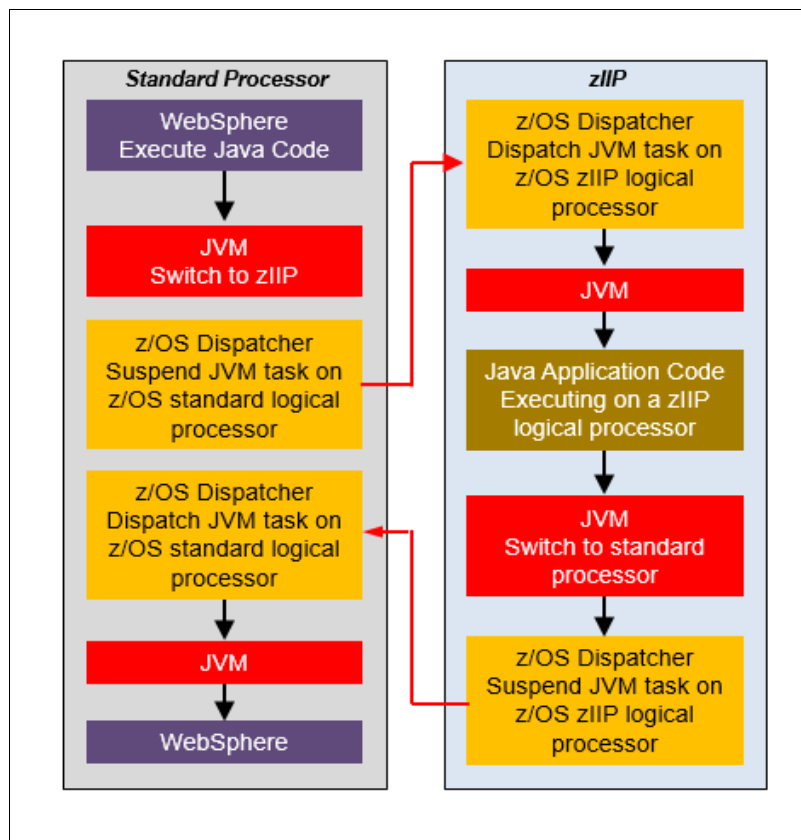


Figure 3-21 Logical flow of Java code execution on a zIIP

The availability is treated in the following manner:

- ▶ If a zIIP is available (not busy), the dispatcher suspends the JVM task on the CP and assigns the Java task to the zIIP. When the task returns control to the JVM, it passes control back to the dispatcher. The dispatcher then reassigns the JVM code execution to a CP.
- ▶ If no zIIP is available (all busy), the z/OS dispatcher allows the Java task to run on a standard CP. This process depends on the option that is used in the OPT statement in the IEAOPTxx member of SYS1.PARMLIB.

A zIIP runs IBM authorized code only. This IBM authorized code includes the z/OS JVM in association with parts of system code, such as the z/OS dispatcher and supervisor services. A zIIP cannot process I/O or clock comparator interruptions. It also does not support operator controls, such as IPL.

Java application code can run on a CP or a zIIP. The installation can manage the use of CPs so that Java application code runs only on CPs or zIIPs, or on both.



The following execution options for zIIP-eligible code execution are available and supported for z/OS<sup>10</sup>. These options are user-specified in IEAOPTxx and can be dynamically altered by using the **SET OPT** command:

► Option 1: Java dispatching by priority (IIPHONORPRIORITY=YES)

This option is the default option and specifies that CPs must not automatically consider zIIP-eligible work for dispatching on them. The zIIP-eligible work is dispatched on the zIIP engines until Workload Manager (WLM) determines that the zIIPs are overcommitted.

WLM then requests help from the CPs. When help is requested, the CPs consider dispatching zIIP-eligible work on the CPs based on the dispatching priority relative to other workloads. When the zIIP engines are no longer overcommitted, the CPs stop considering zIIP-eligible work for dispatch.

This option runs as much zIIP-eligible work on zIIPs as possible. It also allows it to spill over onto the CPs only when the zIIPs are overcommitted.

► Option 2: Java dispatching by priority (IIPHONORPRIORITY=NO)

zIIP-eligible work runs on zIIPs only while at least one zIIP engine is online. zIIP-eligible work is not normally dispatched on a CP, even if the zIIPs are overcommitted and CPs are unused. The exception is that zIIP-eligible work can sometimes run on a CP to resolve resource conflicts.

Therefore, zIIP-eligible work does not affect the CP utilization that is used for reporting through the subcapacity reporting tool (SCRT), no matter how busy the zIIPs are.

If zIIPs are defined to the LPAR but are not online, the zIIP-eligible work units are processed by CPs in order of priority. The system ignores the IIPHONORPRIORITY parameter in this case and handles the work as though it had no eligibility to zIIPs.

zIIPs provide the following benefits:

- Potential software cost savings.
- Simplification of infrastructure as a result of the colocation and integration of new applications with their associated database systems and transaction middleware, such as Db2, IMS, or CICS. Simplification can happen, for example, by introducing a uniform security environment, and by reducing the number of TCP/IP programming stacks and system interconnect links.
- Prevention of processing latencies that occur if Java application servers and their database servers are deployed on separate server platforms.

The following Db2 UDB for z/OS V8 or later workloads can run in Service Request Block (SRB) mode:

- Query processing of network-connected applications that access the Db2 database over a TCP/IP connection by using IBM Distributed Relational Database Architecture (DRDA).

DRDA enables relational data to be distributed among multiple systems. It is native to Db2 for z/OS, which reduces the need for more gateway products that can affect performance and availability. The application uses the DRDA requester or server to access a remote database. IBM Db2 Connect is an example of a DRDA application requester.

- Star schema query processing, which is mostly used in business intelligence work.

A *star schema* is a relational database schema for representing multidimensional data. It stores data in a central fact table and is surrounded by more dimension tables that hold information about each perspective of the data. For example, a star schema query joins various dimensions of a star schema data set.

<sup>10</sup> z/OS V2R4 and later (older z/OS versions are out of support)



- Db2 utilities that are used for index maintenance, such as LOAD, REORG, and REBUILD. Indexes allow quick access to table rows. However, the databases become less efficient over time and must be maintained as data in large databases is manipulated.

The zIIP runs portions of eligible database workloads, which helps to free computer capacity and lower software costs. Not all Db2 workloads are eligible for zIIP processing. Db2 UDB for z/OS V8 and later gives z/OS the information to direct portions of the work to the zIIP. The result is that in every user situation, different variables determine how much work is redirected to the zIIP.

On an IBM z17, the following workloads also can benefit from zIIPs:

- z/OS Communications Server uses the zIIP for eligible Internet Protocol Security (IPSec) network encryption workloads. Portions of IPSec processing take advantage of the zIIPs, specifically end-to-end encryption with IPSec. The IPSec function moves a portion of the processing from the general-purpose processors to the zIIPs. In addition, to run the encryption processing, the zIIP also handles the cryptographic validation of message integrity and IPSec header processing.
- z/OS Global Mirror, formerly known as *Extended Remote Copy* (XRC), also uses the zIIP. Most z/OS Data Facility Storage Management Subsystem (DFSMS) system data mover (SDM) processing that is associated with z/OS Global Mirror can run on the zIIP.
- The first IBM user of z/OS XML system services is Db2 V9. For Db2 V9 before the z/OS XML System Services enhancement, z/OS XML System Services non-validating parsing was partially directed to zIIPs when used as part of a distributed Db2 request through DRDA. This enhancement benefits Db2 by making all z/OS XML System Services non-validating parsing eligible to zIIPs. This configuration is possible when processing is used as part of any workload that is running in enclave SRB mode.
- z/OS Communications Server also allows the HiperSockets Multiple Write operation for outbound large messages (originating from z/OS) to be run by a zIIP. Application workloads that are based on XML, HTTP, SOAP, and Java, and traditional file transfer can benefit.
- During the SRB boost period, ANY work in a boosting image is eligible to run on a zIIP processor associated with the image (LPAR).

Many more workloads and software can use zIIP processors, such as the following examples:

- IBM z/OS Container Extensions (zCX)
- IBM z/OS CIM monitoring
- IBM z/OS Management Facility (z/OSMF)
- System Display and Search Facility (SDSF)
- IBM z/OS Connect EE components
- IBM Sterling® Connect:Direct®
- IBM Z System Automation:
- Java components of IBM Z SMS and SAS
- IBM Z NetView RESTful API server
- IBM Z Workload Scheduler & Dynamic Workload Console (under WebSphere Liberty)
- IMS workloads (DRDA, SOAP, MSC, ISC)
- Db2 for z/OS Data Gate
- Db2 Sort for z/OS
- Db2 Analytics Accelerator Loader for z/OS
- Db2 Utilities Suite for z/OS
- Db2 Log Analysis Tool for z/OS
- Data Virtualization Manager for z/OS (DVM)
- IzODA (Apache Spark workloads)
- Watson Machine Learning for z/OS (WMLz) for MLeap and Spark workloads



- ▶ IBM Z Common Data Provider (CDP)
- ▶ IBM Omegamon Portfolio components
- ▶ IBM RMF (Monitor III work)
- ▶ IBM Developer for z/OS Enterprise Edition components.

For more information about zIIP and eligible workloads, see [IBM zIIP web page](#).

## zIIP installation

One CP must be installed with or before any zIIP is installed. Since z16, the zIIP-to-CP ratio of 2:1<sup>11</sup> has been eliminated, which means for z17 up to 207 zIIPs on feature Max208 can be characterized.

## Unassigned zIIPs

Since z16 a zIIP that is purchased but not activated is registered as an unassigned zIIP (FC 1656). When the system is later upgraded with another zIIP, the system recognizes that an zIIP was purchased and is present.

The allowable number of zIIPs for each model is listed in Table 3-3.

Table 3-3 zIIPs and Unassigned zIIPs per feature

Features	Max43	Max90	Max136	Max183	Max208
Maximum of zIIPs FC 1653	42	89	135	182	207
Maximum of Unassigned zIIPs FC 1656	42	89	135	182	207

zIIPs are orderable by using FC 1653. At least one CP must be configured in order to add zIIPs to the system configuration. If the installed CPC drawer has no remaining unassigned PUs, the assignment of the next zIIP might require the installation of another CPC drawer.

PUs that are characterized as zIIPs within a configuration are grouped into the zIIP pool. This configuration allows zIIPs to have their own processing weights, independent of the weight of parent CPs. The zIIP pool can be seen on the hardware console.

The number of temporary zIIPs cannot exceed the number of permanent zIIPs.

## zIIPs and logical partition definitions

zIIPs are dedicated or shared, depending on whether they are part of an LPAR with dedicated or shared CPs. In an LPAR, at least one CP must be defined before zIIPs for that partition can be defined. The number of zIIPs that are available in the system is the number of zIIPs that can be defined to an LPAR.

**LPAR:** In an LPAR, as many zIIPs as are available can be defined together with at least one CP.

<sup>11</sup> The 2:1 ratio can be exceeded (during boost periods) if System Recovery Boost Upgrade (FC 9930 and FC 6802) is used for activating temporary zIIP capacity.



### 3.5.6 System assist processors

A system assist processor (SAP) is a PU that runs the channel subsystem LIC to control I/O operations. All SAPs run I/O operations for all LPARs. As with IBM z16 servers, in IBM z17 SMT is enabled<sup>12</sup> for all SAPs, except 1. All features include standard SAPs configured. The number of standard SAPs depends on the IBM z17 feature, as listed in Table 3-4.

Table 3-4 SAPs per feature

Features	Max43	Max90	Max136	Max183	Max208
Standard SAPs	5	10	16	21	24

Additional SAPs cannot be ordered with the IBM z17 ME1.

### 3.5.7 Reserved processors

*Reserved processors* are defined by PR/SM to allow for a nondisruptive capacity upgrade. Reserved processors are similar to spare logical processors and can be shared or dedicated. Reserved CPs can be defined to an LPAR dynamically to allow for nondisruptive image upgrades.

Reserved processors can be dynamically configured online by an operating system that supports this function if enough unassigned PUs are available to satisfy the request. The PR/SM rules that govern logical processor activation remain unchanged.

By using reserved processors, you can define more logical processors than the number of available CPs, IFLs, ICFs, and zIIPs in the configuration to an LPAR. This process makes it possible to nondisruptively configure online more logical processors after more CPs, IFLs, ICFs, and zIIPs are made available concurrently. They can be made available with one of the capacity on-demand options.

The maximum number of reserved processors that can be defined to an LPAR depends on the number of logical processors that are defined. A maximum of 208 logical processors plus reserved processors can be used. If the operating system in the LPAR supports the logical processor add function, reserved processors are no longer needed.

Do not define more active and reserved processors than the operating system for the LPAR can support. For more information about logical processors and reserved processors and their definitions, see 3.7, “Logical partitioning” on page 121.

### 3.5.8 Integrated firmware processors

Integrated Firmware Processors (IFP) are allocated from the pool of PUs and are available for the entire system. Unlike other characterized PUs, IFPs are standard on all IBM z17 models and not defined by the client.

The two PUs that are characterized as IFP are dedicated to supporting firmware functions that are implemented in Licensed Internal Code (LIC); for example, the resource groups (RGs) that are used for managing the following native Peripheral Component Interconnect Express (PCIe) feature:

- ▶ Coupling Express3 Long Reach

<sup>12</sup> Enabled by default, cannot be changed or altered by user.



IFPs also are initialized at POR. They support Resource Group (RG) LIC<sup>13</sup> to provide native PCIe I/O feature management and virtualization functions.

### 3.5.9 Processor unit assignment

The processor unit assignment of characterized PUs is done at POR time when the system is initialized. The initial assignment rules keep PUs of the same characterization type grouped as much as possible in relation to PU chips and CPC drawer boundaries to optimize shared cache usage.

The IBM z17 ME1 PU assignment is based on CPC drawer plug order (*not* “ordering”). Feature upgrade provides more processor (CPC) drawers. Max136 cannot be upgraded because the supposed targeted features (Max183 and Max208) are factory built only.

The CPC drawers are populated from the bottom up. This process defines the low-order and the high-order CPC drawers:

- ▶ CPC drawer 0 (CPC 0 at position A10): Plug order 1 (low-order CPC drawer)
- ▶ CPC drawer 1 (CPC 1 at position A15): Plug order 2
- ▶ CPC drawer 2 (CPC 2 at position A20): Plug order 3
- ▶ CPC drawer 3 (CPC 3 at position B10): Plug order 4 (high-order CPC drawer)

The assignment rules comply with the following order:

- ▶ SAPs: Spread across CPC drawers and high PU chips. Each CPC drawer includes at least five standard SAPs. Start with the highest PU chip high core, then the next highest PU chip high core.
- ▶ CPs and zIIPs: Assign CPs and zIIPs to cores on chips in lower CPC drawers working upward.
- ▶ IFLs and ICFs: Assign IFLs and ICFs to cores on chips in higher CPC drawers working downward.
- ▶ IFP: Two IFPs are assigned to CPC drawer 0<sup>14</sup> for a Max43; CPC drawer 1 for Max90 and Max136 and CPC drawer 2 for Max183 and Max208. If, for some reason, drawer 0 is not present the IFPs will be located in other drawers.
- ▶ Spares: Two spares are assigned per model. On z17 there are no specific rules to place them. Any “unassigned PU” in any drawer can be assigned as spare. For additional information about spare PUs use, refer to 3.5.10, “Sparing rules” on page 116.

The rules above are intended to isolate processors that are used by different operating systems as much as possible on different CPC drawers and even on different PU chips. This configuration ensures that different operating systems do not use the same shared caches. For example, CPs and zIIPs are all used by z/OS, and can benefit by using the same shared caches. However, IFLs are used by z/VM and Linux, and ICFs are used by CFCC.

This initial PU assignment, which is done at POR, can be dynamically rearranged by an LPAR by swapping an active core to a core in a different PU chip in a different CPC drawer to improve system performance. For more information, see “LPAR dynamic PU reassignment” on page 127.

When a CPC drawer is added concurrently after POR and new LPARs are activated, or processor capacity for active partitions is dynamically expanded, the extra PU capacity can be assigned from the new CPC drawer. The processor unit assignment rules consider the newly installed CPC drawer dynamically.

<sup>13</sup> IBM zHyperLink Express2.0 is not managed by Resource Groups LIC.

<sup>14</sup> For a layout of CPC drawers' locations, refer to 3.5.11, “CPC drawer numbering” on page 116



### 3.5.10 Sparing rules

On an IBM z17 ME1 system, two PUs are reserved as spares. The spare PUs are available to replace any two characterized PUs, whether they are CP, IFL, ICF, zIIP, SAP, or IFP.

Systems with a failed PU for which no spare is available *call home* for a replacement. A system with a failed PU that is spared and requires a DCM to be replaced (referred to as a *pending repair*) can still be upgraded when sufficient PUs are available.

#### Transparent CP, IFL, ICF, zIIP, SAP, and IFP sparing

Depending on the model, sparing of CP, IFL, ICF, zIIP, SAP, and IFP is transparent and does not require operating system or operator intervention.

With *transparent sparing*, the status of the application that was running on the failed processor is preserved. The application continues processing on a newly assigned CP, IFL, ICF, zIIP, SAP, or IFP (allocated to one of the spare PUs) without client intervention.

#### Application preservation

If no spare PU is available, *application preservation* (z/OS only) is started. The state of the failing processor is passed to another active processor that is used by the operating system. Through operating system recovery services, the task is resumed successfully (in most cases, without client intervention).

#### Dynamic SAP and IFP sparing and reassignment

*Dynamic recovery* is provided if a failure of the SAP or IFP occurs. If the SAP or IFP fails, and if a spare PU is available, the spare PU is dynamically assigned as a new SAP or IFP. If no spare PU is available, and more than one CP is characterized, a characterized CP is reassigned as an SAP or IFP. In either case, client intervention is not required. This capability eliminates an unplanned outage and allows a service action to be deferred to a more convenient time.

### 3.5.11 CPC drawer numbering

IBM z17 ME1 CPC drawer numbering starts with CPC 0, the first installed CPC drawer. It is in frame A at A10. The second one, in the same frame, at A15. The third one, in the same frame at A20. The fourth CPC drawer is in frame B (location B10).

Figure 3-22 on page 117 shows CPC drawer numbering.



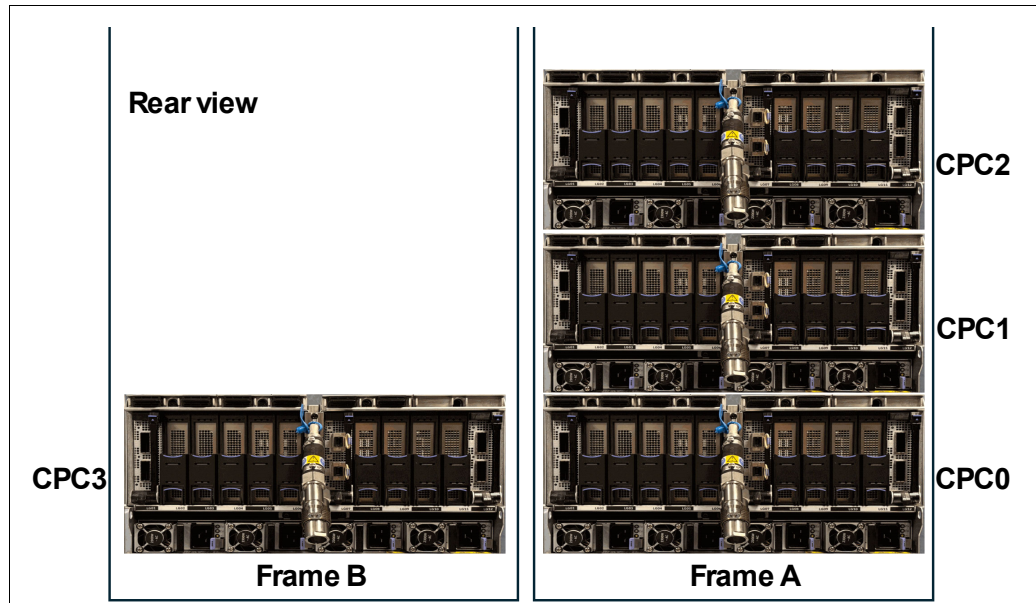


Figure 3-22 CPC drawer numbering

## 3.6 Memory design

Various considerations of the IBM z17 memory design are described in this section.

### 3.6.1 Overview

The IBM z17 ME1 memory design provides flexibility, high availability, and the following capabilities:

- ▶ Concurrent memory upgrades if the physically installed capacity is not yet reached  
IBM z17 servers can have more physically installed memory than the initial available capacity. Memory upgrades within the physically installed capacity can be done concurrently by LIC, and no hardware changes are required. However, memory upgrades *cannot* be done through CBU or On/Off CoD.
- ▶ Concurrent memory upgrades if the physically installed capacity is reached  
Physical memory upgrades require a processor drawer to be removed and reinstalled after replacing the memory cards in the processor drawer. Except for the feature Max43, the combination of enhanced drawer availability and the flexible memory option allows you to concurrently add memory to the system. For more information, see 2.5.5, “Drawer replacement and memory” on page 52, and 2.5.7, “Flexible Memory Option” on page 53.

When the total capacity that is installed has more usable memory than required for a configuration, the Licensed Internal Code Configuration Control (LICCC) determines how much memory is used from each processor drawer. The sum of the LICCC provided memory from each CPC drawer is the amount that is available for use in the system.

### Memory allocation

When the system is activated by a POR, PR/SM determines the total installed memory and the customer enabled memory. Later in the process, during LPAR activation, PR/SM assigns and allocates each partition memory according to their image profile.



PR/SM controls all physical memory, and can make physical memory available to the configuration when a CPC drawer is added.

In older IBM Z processors, memory allocation was striped across the available CPC drawers because relatively fast connectivity (that is, fast relative to the processor clock frequency) existed between the drawers. Splitting the work between all of the memory controllers allowed a smooth performance variability.

The memory allocation algorithm changed starting with IBM z13. For IBM z17, PR/SM tries to allocate memory into a single CPC drawer. If memory does not fit into a single drawer, PR/SM tries to allocate the memory into the CPC drawer with the most processor entitlement.<sup>15</sup>

The PR/SM memory and logical processor resources allocation goal is to place all partition resources on a single CPC drawer, if possible. The resources, such as memory and logical processors, are assigned to the logical partitions at the time of their activation. Later on, when all partitions are activated, PR/SM can move memory between CPC drawers to benefit the performance of each LPAR, without operating system knowledge. This process was done on the previous families of IBM Z servers only for PUs that use PR/SM dynamic PU reallocation.

With IBM z17 servers, this process occurs whenever the configuration changes, such as in the following circumstances:

- ▶ Activating or deactivating an LPAR
- ▶ Changing the LPARs' processing weights
- ▶ Upgrading the system through a temporary or permanent record
- ▶ Downgrading the system through deactivation of a temporary record

PR/SM schedules a global reoptimization of the resources in use. It does so by reviewing all the partitions that are active and prioritizing them based on their processing entitlement and weights, which creates a high- and low-priority rank. Then, the resources, such as logical processors and memory, can be moved from one CPC drawer to another to address the priority ranks that were created.

When partitions are activated, PR/SM tries to find a home assignment CPC drawer, home assignment node, and home assignment chip for the logical processors that are defined to them. The PR/SM goal is to allocate all the partition logical processors and memory to a single CPC drawer (the home drawer for that partition).

If all logical processors can be assigned to a home drawer and the partition-defined memory is greater than what is available in that drawer, the exceeding memory amount is allocated on another CPC drawer. If all the logical processors cannot fit in one CPC drawer, the remaining logical processors spill to another CPC drawer. When that overlap occurs, PR/SM stripes the memory (if possible) across the CPC drawers where the logical processors are assigned.

The process of reallocating memory is based on the *memory copy/reassign* function, which is used to allow enhanced drawer availability (EDA) and concurrent drawer replacement (CDR)<sup>16</sup>. This process was enhanced starting with z13 and IBM z13s® to provide more efficiency and speed to the process without affecting system performance.

IBM z17 ME1 implements a faster dynamic memory reallocation mechanism, which is especially useful during service operations (EDA and CDR). PR/SM controls the reassignment of the content of a specific physical memory array in one CPC drawer to a physical memory array in another CPC drawer. To accomplish this task, PR/SM uses all the

---

<sup>15</sup> Entitlement is based on PR/SM shares in all the pools for which the LPAR has logical processors. For example, an LPAR with 5 GCP's worth of share has a greater entitlement than one with 2 GCP's and 1 zIIP's worth of share.

<sup>16</sup> In previous IBM Z generations (before z13), these service operations were known as enhanced book availability (EBA) and concurrent book repair (CBR).



available physical memory in the system. This memory includes the memory that is not in use by the system that is available but not purchased by the client, and the planned memory options, if installed.

Because of the memory allocation algorithm, systems that undergo many miscellaneous equipment specification (MES) upgrades for memory can have different memory mixes and quantities in all processor drawers of the system. If the memory fails, it is technically feasible to run a POR of the system with the remaining working memory resources. After the POR completes, the memory distribution across the processor drawers is different, as is the total amount of available memory.

## **Large page support**

By default, page frames are allocated with a 4 KB size. IBM z17 servers also support large page sizes of 1 MB or 2 GB. The first z/OS release that supports 1 MB pages is z/OS V1R9. Linux on IBM Z 1 MB pages support is available in SUSE Linux Enterprise Server 10 SP2 and Red Hat Enterprise Linux (RHEL) 5.2 and later.

## **Large page support and Translation Lookaside Buffer (TLB)**

The TLB reduces the amount of time that is required to translate a virtual address to a real address. This translation is done by dynamic address translation (DAT) when it must find the correct page for the correct address space.

Each TLB entry represents one page. As with other buffers or caches, lines are discarded from the TLB on a least recently used (LRU) basis.

The worst-case translation time occurs when a TLB miss occurs and the segment table (which is needed to find the page table) and the page table (which is needed to find the entry for the particular page in question) are not in cache. This case involves two complete real memory access delays plus the address translation delay. The duration of a processor cycle is much shorter than the duration of a memory cycle, so a TLB miss is relatively costly.

It is preferable to have addresses in the TLB. With 4 K pages, holding all of the addresses for 1 MB of storage takes 256 TLB lines. When 1 MB pages are used, it takes only one TLB line. Therefore, large page size users have a much smaller TLB footprint.

Large pages allow the TLB to better represent a large working set and suffer fewer TLB misses by allowing a single TLB entry to cover more address translations.

Users of large pages are better represented in the TLB and are expected to see performance improvements in elapsed time and processor usage. These improvements are because DAT and memory operations are part of processor busy time, even though the processor waits for memory operations to complete without processing anything else in the meantime.

To overcome the processor usage that is associated with creating a 1 MB page, a process must run for some time. It also must maintain frequent memory access to keep the pertinent addresses in the TLB.

Short-running work does not overcome the processor usage. Short processes with small working sets are expected to receive little or no improvement. Long-running work with high memory-access frequency is the best candidate to benefit from large pages.

Long-running work with low memory-access frequency is less likely to maintain its entries in the TLB. However, when it does run, few address translations are required to resolve all of the memory it needs. Therefore, a long-running process can benefit even without frequent memory access.



Weigh the benefits of whether something in this category must use large pages as a result of the system-level costs of tying up real storage. A balance exists between the performance of a process that uses large pages and the performance of the remaining work on the system.

On IBM z17 server, 1 MB large pages become pageable if Virtual Flash Memory<sup>17</sup> is available and enabled. They are available only for 64-bit virtual private storage, such as virtual memory that is above 2 GB.

It is easy to assume that increasing the TLB size is a feasible option to deal with TLB-miss situations. However, this process is not as simple as it seems. As the size of the TLB increases, so does the processor usage that is involved in managing the TLB's contents. Correct sizing of the TLB is subject to complex statistical modeling to find the optimal tradeoff between size and performance.

### 3.6.2 Main storage

Main storage consists of memory space that is addressable by programs and storage that is not directly addressable by programs. Nonaddressable storage includes the hardware system area (HSA).

Main storage provides the following functions:

- ▶ Data storage and retrieval for PUs and I/O
- ▶ Communication with PUs and I/O
- ▶ Communication with and control of optional expanded storage
- ▶ Error checking and correction

Main storage can be accessed by all processors, but cannot be shared between LPARs. Any system image (LPAR) must include a defined main storage size. This defined main storage is allocated exclusively to the LPAR during partition activation.

### 3.6.3 Hardware system area

The HSA is a nonaddressable storage area that contains system LIC and configuration-dependent control blocks. On IBM z17 ME1 servers, the HSA has a fixed size of 884 GB and is not part of the purchased memory that you order and install.

The fixed size of the HSA eliminates planning for future expansion of the HSA because the hardware configuration definition (HCD)/input/output configuration program (IOCP) always reserves space for the following items:

- ▶ Six channel subsystems (CSSs)
- ▶ A total of 15 LPARs in CSSs 1 through 5, and 10 LPARs for the sixth CSS for a total of 85 LPARs
- ▶ Subchannel set 0 with 63.75-K devices in each CSS
- ▶ Subchannel set 1 with 64-K devices in each CSS
- ▶ Subchannel set 2 with 64-K devices in each CSS
- ▶ Subchannel set 3 with 64-K devices in each CSS

The HSA features sufficient reserved space to allow for dynamic I/O reconfiguration changes to the maximum capability of the processor.

---

<sup>17</sup> Virtual Flash Memory replaced IBM zFlash Express. No carry forward of zFlash Express exists.



### 3.6.4 Virtual Flash Memory (FC 0644)

IBM Virtual Flash Memory (VFM, FC 0566) is the replacement for the Flash Express features that were available on the IBM zEC12 and IBM z13. No application changes are required to change from IBM Flash Express to VFM.

For IBM z17 ME1, IBM VFM provides up to 6.0 TB of virtual flash memory in 512 GB increments. The minimum is 0, while the maximum is 12 features. The number of VFM features ordered reduces the maximum orderable memory for the IBM z17.

## 3.7 Logical partitioning

The logical partitioning features are described in this section.

### 3.7.1 Overview

Logical partitioning is a function that is implemented by PR/SM. IBM z17 servers can run in LPAR mode, or in DPM mode. DPM provides a GUI for PR/SM to manage I/O resources dynamically.

PR/SM is aware of the processor drawer structure on IBM z17 servers. LPARs have **logical** resources that are allocated to them from various **physical** resources. From a systems standpoint, LPARs have no control over these physical resources, but the PR/SM functions do have this control.<sup>18</sup> PR/SM manages and optimizes allocation and the dispatching of work on the physical topology.

PR/SM's job - in modern systems - is exceedingly complex. Its overall goal is to allocate resources according to policy and in a way that delivers the expected capacity, while optimizing the use of resources. It does this for up to 16 DCMs in four processor drawers, and up to 85 diverse LPARs, with often conflicting constraints.

As described in 3.5.9, "Processor unit assignment" on page 115, PU's are initially assigned during POR using algorithms to optimize cache usage. This step is the "physical" step, where cores are characterized as CPs, zIIPs, IFLs, ICFs, and SAPs in the appropriate processor drawers.

When an LPAR is activated, PR/SM builds logical processors and allocates memory for the LPAR.

PR/SM attempts to assign home addresses for an LPAR's logical processors to one CPC drawer.

With HiperDispatch PR/SM cooperates with the operating system to concentrate a unit of work's dispatching on the same logical processor, which is in turn concentrated on the same physical processor. "Concentrated" because it cannot be guaranteed a logical processor will always be dispatched on the same physical. Most particularly, logical processors without a full processor's weight can be dispatched on different physical processors. Nor can it be guaranteed that a unit of work is always dispatched on the same logical.

All processor types of an IBM z17 can be dynamically reassigned, except IFPs.

<sup>18</sup> Starting with z16, information on all logical processor home addresses in the server is available to LPARs. This is returned in the SYSIB control block (specifically SYSIB 15.1.2). Resource Measurement Facility (RMF) exposes this information in the Logical Processor Data Section of SMF Record Type 70 Subtype 1 and Type 74 Subtype 4.



In z16 memory allocation changed from the previous IBM Z servers, and this change persists with z17. Partition memory is now allocated based on processor drawer affinity. For more information, see “Memory allocation” on page 117.

Logical processors are dispatched by PR/SM on physical processors. The assignment used by PR/SM to dispatch logical processors on physical PUs is also based on cache usage optimization.

Processor drawers assignment is more important because they optimize virtual L4 cache usage. Therefore, logical processors from a specific LPAR are packed into a processor drawer as much as possible.

PR/SM optimizes chip assignments within the assigned processor drawer (or drawers) to maximize virtual L3 cache efficiency. Logical processors from an LPAR are dispatched on physical processors on the same PU chip as much as possible.

PR/SM also tries to redispach a logical processor on the same physical processor to optimize private cache (L1 and L2) usage.

## HiperDispatch

PR/SM and z/OS work in tandem to use processor resources more efficiently. HiperDispatch is a function that combines the dispatcher actions and the knowledge that PR/SM has about the topology of the system.

Performance can be optimized by redispaching units of work to the same subset of an LPAR’s logical processors (known as an **affinity node**), which keeps work units running near their cached instructions and data, and minimizes transfers of data ownership among processors and processor drawers.

The nested topology is returned to z/OS by the Store System Information (STSI) instruction. HiperDispatch uses the information to concentrate logical processors around shared caches (virtual L3 and virtual L4 caches at drawer level), and dynamically optimizes the assignment of logical processors and units of work.

The z/OS dispatcher manages multiple queues, called **affinity queues**, with a small number of processors per queue, which fits well onto a single PU chip. These queues are used to assign work to as few logical processors as are needed for an LPAR workload. Therefore, even if the LPAR is defined with many logical processors, HiperDispatch optimizes this number of processors to be near the required capacity. The optimal number of processors to be used is kept within a processor drawer boundary, when possible.

**Tip:** z/VM V7R1 and later also support HiperDispatch.

## Logical partitions

PR/SM enables IBM z17 ME1 systems to be initialized for a logically partitioned operation, supporting up to 85 LPARs. Each LPAR can run its own operating system image in any image mode, independently from the other LPARs.

An LPAR can be added, removed, activated, or deactivated at any time. Changing the number of LPARs is not disruptive and does not require a POR. Certain facilities might not be available to all operating systems because the facilities might have software corequisites.

Each LPAR has the following resources that are the same as a real CPC:

- Processors



Called *logical processors*, they can be defined as CPs, IFLs, ICFs, or zIIPs. They can be dedicated to an LPAR or shared among LPARs. When shared, a processor weight can be defined to provide the required level of processor resources to an LPAR. Also, the capping option can be turned on, which prevents an LPAR from acquiring more than its defined weight and limits its processor consumption.

LPARs for z/OS can have CP and zIIP logical processors. The logical processor types can be defined as all dedicated or all shared. The zIIP support is available in z/OS.

The weight and number of online logical processors of an LPAR can be dynamically managed by the LPAR CPU Management function of the Intelligent Resource Director (IRD). These functions can be used to achieve the defined goals of this specific partition and of the overall system. The provisioning architecture of IBM z17 systems adds a dimension to the dynamic management of LPARs, as described in Chapter 8, “System upgrades” on page 353.

PR/SM supports an option to limit the amount of physical processor capacity that is used by an individual LPAR when a PU is defined as a general-purpose processor (CP) or an IFL that is shared across a set of LPARs.

This capability is designed to provide a physical capacity limit that is enforced as an absolute (versus relative) limit. It is not affected by changes to the logical or physical configuration of the system. This physical capacity limit can be specified in units of CPs or IFLs. The Change LPAR Controls and Customize Activation Profiles tasks on the HMC were enhanced to support this new function.

For the z/OS Workload License Charges (WLC) pricing metric and metrics that are based on it, such as Advanced Workload License Charges (AWLC), an LPAR *defined capacity* can be set. This defined capacity enables the soft capping function. Workload charging introduces the capability to pay software license fees that are based on the processor utilization of the LPAR on which the product is running, rather than on the total capacity of the system.

Consider the following points:

- In support of WLC, the user can specify a defined capacity in millions of service units (MSUs) per hour. The defined capacity sets the capacity of an individual LPAR when soft capping is selected.

The defined capacity value is specified on the Options tab in the Customize Image Profiles window.

- WLM keeps a four-hour rolling average of the processor usage of the LPAR. When the four-hour average processor consumption exceeds the defined capacity limit, WLM dynamically activates LPAR capping (soft capping). When the rolling four-hour average returns below the defined capacity, the soft cap is removed.

For more information about WLM, see *System Programmer's Guide to: Workload Manager*, [SG24-6472](#).

For more information about software licensing, see 7.8, “Software licensing” on page 348.

**Weight settings:** When defined capacity is used to define an uncapped LPAR's capacity, carefully consider the weight settings of that LPAR. If the weight is much smaller than the defined capacity, PR/SM uses a discontinuous cap pattern to achieve the defined capacity setting. This configuration means PR/SM alternates between capping the LPAR at the MSU value that corresponds to the relative weight settings, and no capping at all. It is best to avoid this scenario and instead attempt to establish a defined capacity that is equal or close to the relative weight.



► **Memory**

Memory (main storage) must be dedicated to an LPAR. The defined storage must be available during the LPAR activation; otherwise, the LPAR activation fails.

*Reserved storage* can be defined to an LPAR, which enables nondisruptive memory addition to and removal from an LPAR by using the LPAR dynamic storage reconfiguration (z/OS and z/VM). For more information, see 3.7.5, “LPAR dynamic storage reconfiguration” on page 131.

► **Channels**

Channels can be shared between LPARs by including the partition name in the partition list of a channel-path identifier (CHPID). I/O configurations are defined by the IOCP or the HCD with the CHPID mapping tool (CMT). The CHPID Mapping Tool (CMT) is an optional tool that is used to map CHPIDs onto physical channel IDs (PCHIDs). PCHIDs represent the physical location of a port on a card in a PCIe I/O drawer.

IOCP is available on the z/OS, z/VM, and VSE<sup>n</sup> V6.3.1 (21<sup>st</sup> Century Software) operating systems, and as a stand-alone program on the hardware console. For more information, see *Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7177. HCD is available on the z/OS and z/VM operating systems. Review the suitable 9175DEVICE Preventive Service Planning (PSP) buckets before implementation.

Fibre Channel connection (FICON) channels can be managed by the Dynamic CHPID Management (DCM) function of the Intelligent Resource Director. DCM enables the system to respond to ever-changing channel requirements by moving channels from heavily used control units to lesser-used control units, as needed.

## Modes of operation

The modes of operation are listed in Table 3-5. All available mode combinations, including their operating modes and processor types, operating systems, and addressing modes, also are listed. Only the currently supported versions of operating systems are considered.

Table 3-5 IBM z17 modes of operation

Image mode	PU type	Operating system	Addressing mode
General <sup>a</sup>	CP and zIIP	<ul style="list-style-type: none"> <li>► z/OS</li> <li>► z/VM</li> </ul>	64-bit
	CP	<ul style="list-style-type: none"> <li>► VSE<sup>n</sup> V6.3.1 (21<sup>st</sup> CS)</li> <li>► Linux on IBM Z</li> <li>► z/TPF</li> </ul>	64-bit
Coupling facility	ICF or CP	CFCC	64-bit
Linux only	IFL or CP	<ul style="list-style-type: none"> <li>► Linux on IBM Z (64-bit)</li> <li>► z/VM</li> </ul>	64-bit
		Linux on IBM Z (31-bit)	31-bit
z/VM	CP, IFL, zIIP, or ICF	z/VM	64-bit
SSC <sup>b</sup>	IFL or CP	Linux-based appliance <sup>c</sup>	64 bit

a. General mode uses 64-bit z/Architecture

b. IBM Secure Service Container

c. IBM Db2 Analytics Accelerator (IDAA), Hyper Protect Virtual Servers (HPVS), and others

The 64-bit z/Architecture mode has no special operating mode because the architecture mode is not an attribute of the definable images operating mode. The 64-bit operating systems are in 31-bit mode at IPL and change to 64-bit mode during their initialization. The



operating system is responsible for taking advantage of the addressing capabilities that are provided by the architectural mode.

For information about operating system support, see Chapter 7, “Operating systems support” on page 261.

### Logically partitioned mode

If the IBM z17 ME1 system runs in LPAR mode, each of the 85 LPARs can be defined to operate in one of the following image modes:

- ▶ General mode to run the following systems:
  - A z/Architecture operating system, on dedicated or shared CPs
  - A Linux on IBM Z operating system, on dedicated or shared CPs
  - z/OS, on any of the following processor units:
    - Dedicated or shared CPs
    - Dedicated CPs *and* dedicated zIIPs
    - Shared CPs *and* shared zIIPs

**zIIP usage:** zIIPs can be defined to General mode or z/VM mode image, as listed in Table 3-5 on page 124. However, zIIPs are used by z/OS only. Other operating systems cannot use zIIPs, even if they are defined to the LPAR. z/VM V7R1 and later support real and virtual zIIPs to guest z/OS systems.

- ▶ General mode also is used to run the z/TPF operating system on dedicated or shared CPs
- ▶ CF mode, by loading the CFCC code into the LPAR that is defined as one of the following types:
  - Shared CPs
  - Dedicated or shared ICFs
- ▶ Linux only mode to run the following systems:
  - A Linux on IBM Z operating system, on either of the following types:
    - Dedicated or shared IFLs
    - Dedicated or shared CPs
  - A z/VM operating system, on either of the following types:
    - Dedicated or shared IFLs
    - Dedicated or shared CPs
- ▶ z/VM mode to run z/VM on dedicated or shared CPs or IFLs, plus zIIPs and ICFs
- ▶ IBM SSC mode LPAR can run on dedicated or shared:
  - CPs
  - IFLs



All LPAR modes, required characterized PUs, operating systems, and the PU characterizations that can be configured to an LPAR image are listed in Table 3-6. The available combinations of dedicated (DED) and shared (SHR) processors are also included. For all combinations, an LPAR also can include reserved processors that are defined, which allows for nondisruptive LPAR upgrades.

Table 3-6 LPAR mode and PU usage

LPAR mode	PU type	Operating systems	PUs usage
General	CPs	<ul style="list-style-type: none"> <li>▶ z/Architecture operating systems</li> <li>▶ Linux on IBM Z</li> </ul>	CPs DED or CPs SHR
	CPs and zIIPs	<ul style="list-style-type: none"> <li>▶ z/OS</li> <li>▶ z/VM (guest exploitation)</li> </ul>	CPs DED or zIIPs DED or CPs SHR or zIIPs SHR
General	CPs	z/TPF	CPs DED or CPs SHR
Coupling facility	ICFs or CPs	CFCC	ICFs DED or ICFs SHR or CPs SHR
Linux only	IFLs or CPs	<ul style="list-style-type: none"> <li>▶ Linux on IBM Z</li> <li>▶ z/VM</li> </ul>	IFLs DED or IFLs SHR or CPs DED or CPs SHR
z/VM	CPs, IFLs, zIIPs, or ICFs	z/VM (V7R1 and later)	All PUs must be SHR or DED
SSC <sup>a</sup>	IFLs, or CPs	Linux-based appliance	IFLs DED or IFLs SHR or CPs DED or CPs SHR

a. IBM Secure Service Container

### Dynamically adding or deleting a logical partition name

Dynamically adding or deleting an LPAR name is the ability to add or delete LPARs and their associated I/O resources to or from the configuration without a POR.

The extra channel subsystem and multiple image facility (MIF) image ID pairs (CSSID/MIFID) can be later assigned to an LPAR for use (or later removed). This process can be done through dynamic I/O commands by using the HCD. At the same time, required channels must be defined for the new LPAR.

**Partition profile:** Cryptographic coprocessors are not tied to partition numbers or MIF IDs. They are set up with Adjunct Processor (AP) numbers and domain indexes. These numbers are assigned to a partition profile of a specific name. The client assigns these AP numbers and domains to the partitions and continues to have the responsibility to clear them out when their profiles change.

### Adding logical processors to a logical partition

Logical processors can be concurrently added to an LPAR by defining them as reserved in the image profile and later configuring them online to the operating system by using the suitable console commands.



Logical processors also can be concurrently added to a logical partition dynamically by using the Support Element (SE) “Logical Processor Add” function under the CPC Operational Customization task. This SE function allows the initial and reserved processor values to be dynamically changed. The operating system must support the dynamic addition<sup>19</sup> of these resources.

### **Adding a crypto feature to a logical partition**

You can plan the addition of supported Crypto Express features to an LPAR on the crypto page in the image profile by defining the Cryptographic Candidate List, and the Usage and Control Domain indexes, in the partition profile. By using the Change LPAR Cryptographic Controls task, you can add crypto adapters dynamically to an LPAR without an outage of the LPAR. Also, dynamic deletion or moving of these features does not require planning. Support is provided in z/OS, z/VM, VSE<sup>®</sup> V6.3.1 (21<sup>st</sup> Century Software), IBM SSC (based on appliance requirements), and Linux on IBM Z.

### **LPAR dynamic PU reassignment**

The system configuration is enhanced to optimize the PU-to-CPC drawer assignment of physical processors dynamically. The initial assignment of client-usable physical processors to physical processor drawers can change dynamically to better suit the LPAR configurations that are in use.

For more information, see 3.5.9, “Processor unit assignment” on page 115.

Swapping of specialty engines and general processors with each other, with spare PUs, or with both, can occur as the system attempts to compact LPAR configurations into physical configurations that span the least number of processor drawers.

LPAR dynamic PU reassignment can swap client processors of different types between processor drawers. For example, reassignment can swap an IFL on processor drawer 1 with a CP on processor drawer 2. Swaps can also occur between PU chips within a processor drawer or a node and can include spare PUs. The goals are to pack the LPAR on fewer processor drawers and also on fewer PU chips, based on the IBM z17 processor drawers’ topology. The effect of this process is evident in dedicated and shared LPARs that use HiperDispatch.

LPAR dynamic PU reassignment is transparent to operating systems.

### **LPAR group capacity limit (LPAR group absolute capping)**

The group capacity limit feature allows the definition of a group of LPARs on an IBM z17 system, and limits the combined capacity usage by those LPARs. This process allows the system to manage the group so that the group capacity limits in MSUs per hour are not exceeded. To take advantage of this feature, you must be running z/OS V2R4 or later in all LPARs in the group.

PR/SM and WLM work together to enforce the capacity that is defined for the group and the capacity that is optionally defined for each individual LPAR.

### **LPAR absolute capping**

Absolute capping is a logical partition control that was first made available with zEC12 and is supported on IBM z17 systems. With this support, PR/SM and the HMC are enhanced to support a new option to limit the amount of physical processor capacity that is used by an

<sup>19</sup> In z/OS, this support is available since Version 1 Release 10 (z/OS V1R10), while z/VM supports this addition since z/VM V5R4, and z/VSE<sup>®</sup> since V4R3. However, IBM z17 supports z/OS V2R4 and later, VSE<sup>®</sup> V6.3.1 (21<sup>st</sup> Century Software) and z/VM V7R3 and later.



individual LPAR when a PU is defined as a general-purpose processor (CP), zIIP, or an IFL processor that is shared across a set of LPARs.

Unlike traditional LPAR capping, absolute capping is designed to provide a physical capacity limit that is enforced as an absolute (versus relative) value that is not affected by changes to the virtual or physical configuration of the system.

Absolute capping provides an optional maximum capacity setting for logical partitions that is specified in the absolute processors capacity (for example, 5.00 CPs or 2.75 IFLs). This setting is specified independently by processor type (namely CPs, zIIPs, and IFLs) and provides an enforceable upper limit on the amount of the specified processor type that can be used in a partition.

Absolute capping is ideal for processor types and operating systems that the z/OS WLM cannot control. Absolute capping is not intended as a replacement for defined capacity or group capacity for z/OS, which are managed by WLM.

Absolute capping can be used with any z/OS, z/VM, or Linux on IBM Z LPAR (that is running on an IBM Z server). If specified for a z/OS LPAR, absolute capping can be used concurrently with defined capacity or group capacity management for z/OS. When used concurrently, the absolute capacity limit becomes effective before other capping controls.

### Dynamic Partition Manager mode

DPM is an IBM Z server operation mode that provides a simplified approach to create and manage virtualized environments, which reduces the barriers of its adoption for new and existing customers.

The implementation provides built-in integrated capabilities that allow advanced virtualization management on IBM Z servers. With DPM, you can use your Linux and virtualization skills while taking advantage of the full value of IBM Z hardware, robustness, and security in a workload optimized environment.

DPM provides facilities to define and run virtualized computing systems by using a firmware-managed environment that coordinate the physical system resources that are shared by the partitions. The partitions' resources include processors, memory, network, storage, crypto, and accelerators.

DPM provides a new mode of operation for IBM Z servers that provide the following services:

- ▶ Facilitates defining, configuring, and operating PR/SM LPARs in a similar way to how these tasks are performed on another platform.
- ▶ Lays the foundation for a general IBM Z new user experience.

DPM is not another hypervisor for IBM Z servers. DPM uses the PR/SM hypervisor infrastructure and provides an intelligent interface on top of it that allows customers to define, use, and operate the platform virtualization without IBM Z experience or skills.

## 3.7.2 Storage operations

In IBM z17 ME1 systems, memory can be assigned as main storage, supporting up to 85 LPARs. Before you activate an LPAR, main storage must be defined to the LPAR. All installed storage can be configured as main storage.

For more information about operating system main storage support, see the *PR/SM Planning Guide*, SB10-7178.



Memory *cannot* be shared between system images (LPARs). It is possible to dynamically reallocate storage resources for z/Architecture LPARs that run operating systems that support dynamic storage reconfiguration (DSR). This process is supported by z/OS, and z/VM. z/VM, in turn, virtualizes this support to its guests.

For more information, see 3.7.5, “LPAR dynamic storage reconfiguration” on page 131.

Operating systems that run as guests of z/VM can use the z/VM capability of implementing virtual memory to guest virtual machines. The z/VM dedicated real storage can be shared between guest operating systems.

## LPAR main storage allocation and usage

The IBM z17 storage allocation and usage possibilities depend on the image mode and the operating system that is deployed in the LPAR.

**Important:** The memory allocation and usage depends on the operating system architecture and tested (documented for each operating system) limits.

While the maximum supported memory per LPAR for IBM z17 is 32 TB, each operating system has its own support specifications.

For more information about general guidelines, see the *PR/SM Planning Guide*, SB10-7178, which is available at the [IBM Resource Link website](#) (log in required).

The following modes are provided:

### ► z/Architecture

In z/Architecture (General) mode, storage addressing is 64-bit, which allows for virtual addresses up to 16 exabytes (16 EB). However, the current main storage limit for LPARs on an IBM z17 ME1 is 32 TB of main storage.

The operating system that runs in z/Architecture mode must support the real storage. z/OS V2R4 supports up to 4 TB of real storage, while z/OS V2R5 and later supports up to 16 TB.

### ► CF

In CF mode, storage addressing is 64-bit for a CF image that runs at CFCC. This configuration allows for an addressing range up to 16 EB. However, the current IBM z17 ME1 definition limit for CF LPARs is 32 TB of storage.

The following CFCC levels are supported in a Sysplex with IBM z17:

- CFCC Level 26, available on IBM z17 (Driver level 61)
- CFCC Level 25, available on IBM z16 (Driver level 51)
- CFCC Level 24, available on IBM z15 (Driver level 41)

For more information, see 3.9.1, “CF Control Code (CFCC)” on page 135.

Only IBM CFCC can run in CF mode.

### ► Linux only

In Linux only mode, storage addressing can be 31-bit or 64-bit, depending on the operating system architecture and the operating system configuration.

Only Linux and z/VM operating systems can run in Linux only mode. Linux on IBM Z 64-bit distributions:

- (SUSE SLES 16.1 (Post GA),
- SUSE SLES 15.6 (GA),



- ▶ SUSE SLES 12.5 (Post GA),
- ▶ Red Hat RHEL 10.0 (Post GA),
- ▶ Red Hat RHEL 9.4,
- ▶ Red Hat RHEL 8.10,
- ▶ Red Hat RHEL 7.9 (Post GA),
- ▶ Canonical Ubuntu 24.04 LTS (Post GA),
- ▶ Canonical Ubuntu 22.04 LTS (Post GA),
- ▶ Canonical Ubuntu 20.04 LTS (Post GA)

use 64-bit addressing and operate in z/Architecture mode. z/VM also uses 64-bit addressing and operates in z/Architecture mode.

**Note:** For information about the (kernel) supported amount of memory, check the Linux Distribution specific documentation.

#### ▶ z/VM

In z/VM mode, specific types of processor units can be defined within one LPAR. This feature increases flexibility and simplifies systems management by allowing z/VM to run the following tasks in the same z/VM LPAR:

- Manage guests to operate Linux on IBM Z on IFLs
- Operate z/VSE (or VSE<sup>n</sup> V6.3.1 - Century Link Software) and z/OS on CPs
- Offload z/OS system software processor usage, such as Db2 workloads on zIIPs
- Provide an economical Java execution environment under z/OS on zIIPs

#### ▶ IBM SSC

In IBM SSC mode, storage addressing is 64-bit for an embedded product. The amount of usable main storage by the appliance code that is deployed in the SSC LPAR is documented by the appliance code supplier.

### 3.7.3 Reserved storage

Reserved storage can be optionally defined to an LPAR, which allows a nondisruptive image memory upgrade for this partition. Reserved storage can be defined to central storage, and to any image mode except CF mode.

An LPAR must define an amount of main storage:

- ▶ The *initial value* is the storage size that is allocated to the partition when it is activated.
- ▶ The *reserved value* is another storage capacity that is beyond its initial storage size that an LPAR can acquire dynamically. The reserved storage sizes that are defined to an LPAR do not have to be available when the partition is activated. Instead, they are predefined storage sizes to allow a storage increase, from an LPAR perspective.

Without the reserved storage definition, an LPAR storage upgrade is a disruptive process that requires the following steps:

1. Partition deactivation.
2. An initial storage size definition change.
3. Partition activation.

The extra storage capacity for an LPAR upgrade can come from the following sources:

- ▶ Any unused available storage
- ▶ Another partition that features released storage
- ▶ A memory upgrade



A concurrent LPAR storage upgrade uses DSR. z/OS uses the reconfigurable storage unit (RSU) definition to add or remove storage units in a nondisruptive way.

z/VM V7R3 and later releases support the dynamic addition of memory to a running LPAR by using reserved storage. It also virtualizes this support to its guests.

Removing memory from a z/VM guest is not disruptive to the z/VM LPAR.

z/VM V7R2 and later also support Dynamic Memory Downgrade (DMD), which allows the removal of up to 50% of the real storage from a running z/VM system.

SLES 12 and later supports concurrent add and remove.

### 3.7.4 Logical partition storage granularity

Granularity of main storage for an LPAR depends on the largest main storage amount that is defined for initial or reserved main storage, as listed in Table 3-7<sup>20</sup>.

Table 3-7 Logical partition main storage granularity (IBM z17)

Logical partition: Largest main storage amount	Logical partition: Main storage granularity
Main storage amount <= 512 GB	1 GB
512 GB < main storage amount <= 1 TB	2 GB
1 TB < main storage amount <= 2 TB	4 GB
2 TB < main storage amount <= 4 TB	8 GB
4 TB < main storage amount <= 8 TB	16 GB
8 TB < main storage amount <= 16 TB	32 GB
16 TB < main storage amount <= 32 TB	64 GB

LPAR storage granularity information is required for LPAR image setup and for z/OS RSU definition. On IBM z17 ME1, LPARs are limited to a maximum of 16 TB of main storage. However, the maximum amount of memory that is supported by z/OS V2R4 is 4 TB. z/OS V2R5 and later supports up to 16 TB. For z/VM V7R3 and V7R4 and later the limit is 4 TB.

### 3.7.5 LPAR dynamic storage reconfiguration

Dynamic storage reconfiguration on IBM z17 systems allows an operating system that is running on an LPAR to add (nondisruptively) its reserved storage amount to its configuration. This process can occur only if unused storage exists. This unused storage can be obtained when another LPAR releases storage, or when a concurrent memory upgrade occurs.

With dynamic storage reconfiguration, the unused storage does not have to be continuous.

When an operating system that is running on an LPAR assigns a storage increment to its configuration, PR/SM determines whether any free storage increments are available. PR/SM then dynamically brings the storage online.

PR/SM dynamically takes offline a storage increment and makes it available to other partitions when an operating system running on an LPAR releases a storage increment.

<sup>20</sup> When defining an LPAR on the HMC, the 2 G boundary must be followed in PR/SM.



## 3.8 Intelligent Resource Director

Intelligent Resource Director (IRD) is an IBM Z capability that is used only by z/OS. IRD is a function that optimizes processor and channel resource utilization across LPARs within a single IBM Z server.

This feature extends the concept of goal-oriented resource management. It does so by grouping system images that are on the same IBM z17 server that is running in LPAR mode and in the same Parallel Sysplex into an *LPAR cluster*. This configuration allows WLM, in cooperation with PR/SM, to manage resources (processor and I/O) across the entire cluster of system images and not only in one single image.

An LPAR cluster is shown in Figure 3-23. It contains three z/OS images and one Linux image that is managed by the cluster. Included as part of the entire Parallel Sysplex is another z/OS image and a CF image. In this example, the scope over which IRD has control is the defined LPAR cluster.

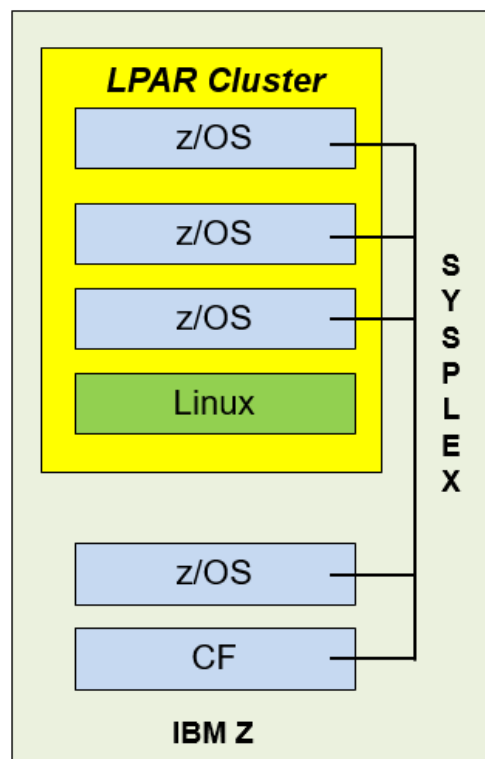


Figure 3-23 IRD LPAR cluster example



IRD features the following characteristics:

- IRD processor management

WLM dynamically adjusts the number of logical processors within an LPAR and the processor weight based on the WLM policy. The ability to move the processor weights across an LPAR cluster provides processing power where it is most needed, based on WLM goal mode policy.

The processor management function is automatically deactivated when HiperDispatch is active. However, the LPAR weight management function remains active with IRD with HiperDispatch. For more information about HiperDispatch, see 3.7, “Logical partitioning” on page 121.

HiperDispatch manages the number of logical CPs in use. It adjusts the number of logical processors within an LPAR to achieve the optimal balance between CP resources and the requirements of the workload.

HiperDispatch also adjusts the number of logical processors. The goal is to map the logical processor to as few physical processors as possible. This configuration uses the processor resources more efficiently by trying to stay within the local cache structure. Doing so makes efficient use of the advantages of the high-frequency microprocessors, and improves throughput and response times.

If logical weight management adjusts the weights within an LPAR cluster it can cause PR/SM to reassign resources - both logical processor home addresses and memory, potentially even to a different drawer. This can be quite disruptive to cache efficiency and in the movement of the LPAR's memory. PR/SM is, however, cautious in resource reassignment.

- Dynamic channel path management (DCM)

DCM moves FICON channel bandwidth between disk control units to address current processing needs. IBM z17 systems support DCM within a channel subsystem.

- Channel subsystem priority queuing

This function on IBM z17 and IBM Z allows the priority queuing of I/O requests in the channel subsystem and the specification of relative priority among LPARs. When running in goal mode, WLM sets the priority for an LPAR and coordinates this activity among clustered LPARs.

For more information about implementing LPAR processor management under IRD, see *z/OS Intelligent Resource Director*, [SG24-5952](#).

## 3.9 Clustering technology

Parallel Sysplex is the clustering technology that is used with IBM Z servers. The components of a Parallel Sysplex as implemented within the z/Architecture are shown in Figure 3-24 on page 134. The example in Figure 3-24 on page 134 shows one of many possible Parallel Sysplex configurations.



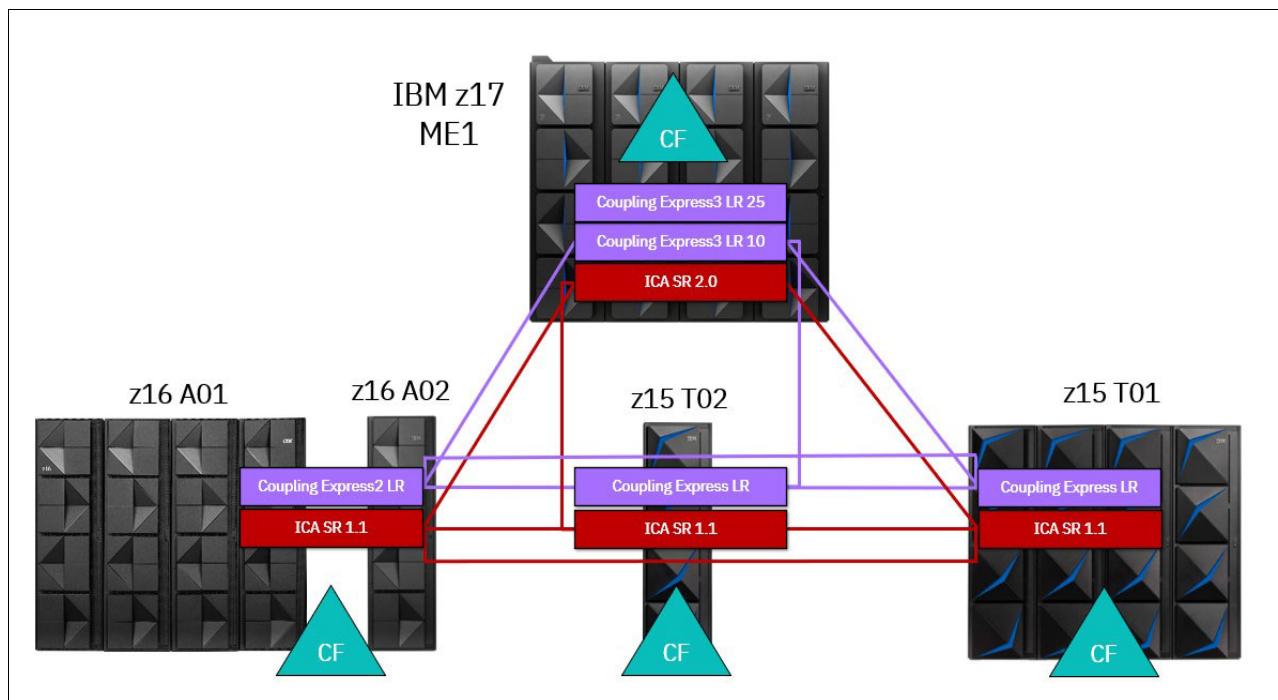


Figure 3-24 Sysplex Connectivity hardware overview

Figure 3-24 shows an IBM z17 model ME1 system that contains multiple z/OS sysplex partitions. It contains an internal CF, an IBM z15 model T02 system that contains a stand-alone CF, and an IBM z16 model A01 that contains multiple z/OS sysplex partitions.

STP over coupling links provides time synchronization to all systems. Selecting the suitable CF link technology Coupling Express3 Long Reach (CE3 LR) or Integrated Coupling Adapter Short Reach (ICA SR2.0) depends on the system configuration and how distant they are physically.

For more information about link technologies, see “Coupling links” on page 199.

Parallel Sysplex is an enabling technology that allows highly reliable, redundant, and robust IBM Z technology to achieve near-continuous availability. A Parallel Sysplex consists of one or more (z/OS) operating system images that are coupled through one or more Coupling Facility LPARs.

A correctly configured Parallel Sysplex cluster maximizes availability in the following ways:

- ▶ Continuous availability: Changes can be introduced, such as software upgrades, one image at a time, while the remaining images continue to process work. For more information, see *Parallel Sysplex Application Considerations*, [SG24-6523](#).
- ▶ High capacity: 1 - 32 z/OS images in a Parallel Sysplex that is operating as a single system.
- ▶ Dynamic workload balancing: because it is viewed as a single logical resource, work can be directed to any operating system image in a Parallel Sysplex cluster that has available capacity.
- ▶ Systems management: The architecture defines the infrastructure to satisfy client requirements for continuous availability. It also provides techniques for achieving simplified systems management consistent with this requirement.



- ▶ **Resource sharing:** Several base z/OS components use CF shared storage. This configuration enables sharing of physical resources with significant improvements in cost, performance, and simplified systems management.
- ▶ **Single logical system:** The collection of system images in the Parallel Sysplex is displayed as a single entity to the operator, user, and database administrator. A single system view means reduced complexity from operational and definition perspectives.
- ▶ **N-2 support:** Multiple hardware generations (normally three) are supported in the same Parallel Sysplex. This configuration provides for a gradual evolution of the systems in the Parallel Sysplex without changing all of them simultaneously. Software support for multiple releases or versions also is supported.

**Note:** Parallel sysplex coupling and timing links connectivity for IBM z17 (M/T 9175) is supported to N-2 generation CPCs (IBM z17, IBM z16, and IBM z15).

Through state-of-the-art cluster technology, the power of multiple images can be harnessed to work in concert on common workloads. The IBM Z Parallel Sysplex cluster takes the commercial strengths of the platform to improved levels of system management, competitive price performance, scalable growth, and continuous availability.

### 3.9.1 CF Control Code (CFCC)

The LPAR that is running the CFCC can be on IBM z17, IBM z16, or IBM z15 systems. For more information about CFCC requirements for supported systems, see “Coupling links” on page 199.

**Consideration:** IBM z17, IBM z16, and IBM z15 servers can coexist in the same sysplex

#### CFCC Level 26

CFCC level 26 is delivered on the IBM z17 (M/T9175) with Driver 61 adds the following features and capabilities<sup>21</sup>:

- ▶ Number of coupling CHPIDs per CEC:
  - CHPIDs per coupling link port, and long-reach coupling links
  - 384 coupling CHPIDs (of all types) per CEC

#### *Coupling Express3 Long Reach*

- Provides physical connectivity for cross-site GDPS configurations
- New higher-bandwidth 25Gb long-reach coupling links allow for more capacity/throughput for long-distance coupling link connectivity
- Coupling Express3 Long Reach (up to 10km unrepeated, and up to 100km with DWDM)
  - CL5 link type: Coupling Express3 LR with 10Gb optics : 2-port
  - CL6 link type: Coupling Express3 LR with 25Gb optics : 2-port
- 32 primary send buffers / CHPID
  - 32 or 8 subchannels/devices may be defined and used
- Four logical CHPIDs / port
- ▶ Physical connectivity limit for long-reach (CE3 LR) coupling links:

<sup>21</sup> Features and capabilities supported by CFCC Level 26 at GA1.



- Max 32 adapters (64 ports) per CEC
- No changes to cabling
- CL5 10Gb is compatible with CE LR and CE2 LR on z16 and z15
- ▶ Note that existing CE LR and CE2 LR adapters on current machines cannot be carried-forward to IBM z17.

#### ***Short-reach connectivity (ICA SR2.0):***

- 48 adapters, 96 ports
- ▶ On IBM z17, the ICA SR adapter hardware changed from ICA SR 1.1 to a new ICA SR 2.0 adapter
  - ICA SR1.1 (FC 0172) cannot be carried forward to IBM z17
  - IBM z17 uses ICA SR 2.0 for short-reach coupling
    - ICA SR 2.0 adapter short-reach coupling links remain link type CS5 and are fully compatible with ICA SR 1.1 coupling links on IBM z15 and IBM z16

#### ***Internal Coupling - ICP***

- ▶ Increased the number of ICP buffers per CHPID
  - IBM z17 will always use 8 buffers
  - Improve capacity/throughput for internal coupling channels
  - HCD/IOCP update to allow 7 or 8 subchannels/devices per ICP CHPID (default to 8)

#### ***Coupling facility configuration options***

- ▶ Simplification of coupling facility configuration options via deprecation of support for Dedicated GPs and Virtual Flash Memory (VFM) for CF images:
  - Dedicated GPs are not supported in CF images on IBM z17.
  - Virtual Flash Memory: VFM is not supported in CF images on IBM z17.
  - SE/HMC will disallow definition of unsupported resource types for CF partitions on IBM z17 CECs; will also handle migration from LPAR profiles on older hardware that still contain resource definitions for these resources.

**Note:** DYNDISP=THIN option is the only available behavior for shared-engine CF dispatching.

- ▶ For additional information please refer to [Considerations for Coupling Facility levels](#).

### **CFCC Level 25**

CFCC level 25 is delivered on the IBM z16 (M/T3931 and M/T3932) with driver level 51 adds the following features:

- ▶ New cache residency time metrics for directory/data entries are available.

These new metrics allow exploiters (such as Db2) to provide direct, useful feedback on the CF cache structure “cache effectiveness”. They also provide improved recommendations for making structure sizing changes or retargeting work from specific table spaces or data sets to other cache structures.

Consider the following points:

- The metrics show how long data entries or directory entries remain resident in the cache structure from the time they are created until the time they are eventually “reclaimed” out of existence.



- They provide moving weighted average directory entry and data area residency times, in microseconds.
- They allow monitoring of effects of cache-unfriendly batch processes, such as image copy, reorganization, and update-intensive workloads.
- Reclaims from all causes are included in the creation of directory entries or data areas, “structure alter” contractions or reapportionments, incidental reclaims of data areas that are caused by reclaim of a directory entry, and so on.
- Residency times are accounted for only at the time of reclaim (not while the cache objects are still in use).
- Specific deletions of these objects do not factor into the residency time metrics.
- These metrics were implemented as new fields within the CF Storage Class controls:  
They are retrieved by using the **IXLCACHE READ\_STGSTATS** command or IXLGM/IXLYAMDA services that are requesting CF Storage Class controls.  
They are also available in CF structure memory dumps that capture CF Storage Class controls.
- The metrics are included in Db2 Performance Manager statistics and used for improved cache structure management (cache sizing, castout processing, reclaim management, and so on). The inclusion of these metrics in sysplex SMF/RMF data is not planned, but can be added later. APAR OA60650 is required on z/OS V2R4, and V2R5 to support the new metrics.

► New cache retry buffer support was added for IFCC retry idempotency:

Initially, the CF cache structure architecture was defined to be idempotent (commands can be tried again after link glitches, such as IFCCs); therefore, no specific accommodations were available for retrying, such as retry buffer support.

However, the list structure architecture was always recognized as nonidempotent, and a rather sophisticated retry buffer mechanism was incorporated to allow z/OS to retrieve the results of commands (even after link glitches occurred) so that such glitches always were well recovered.

Over time, constructs were added to the cache and lock structure architecture that made them become not perfectly retrievable (nonidempotent), but retry buffers were not added to the architecture to mitigate the lack of retrievability:

- Cache structure serialization objects, such as castout locks and cache entry version numbers
- Performance-optimized lock structure commands with no retry buffer support

z/OS software provided simple retry logic to provide IFCC recovery for these nonidempotent commands, but inevitably cases existed in which z/OS cannot provide unambiguous command results to callers. Users might not handle this ambiguity well.

CF cache users who use of these nonidempotent constructs experienced occasional customer problems based on it. The only approach that cleanly and completely addresses the issue is to provide retry buffers for the small subset of cache and lock commands that manipulate objects in a nonidempotent way, along with the accompanying transparent z/OS retry buffer use. z/OS transparently provides all the required recovery support and no user participation is needed.

Down-level systems can continue to use the “old” software retry support until they are upgraded, while up-level systems that use the same CF structure can take full advantage of the new retry buffers for improved IFCC recovery. APAR OA60275 is required on z/OS V2R4, and V2R5.

► New lock record data reserved entries for structure full recovery.



Some lock structure users use “special” lock structure locks to serialize their own processing, such as management of open data sets and table space interest across the sysplex. Not all locks have anything to do with serialization of database updates or user database data or transactions.

When lock structures use up all of the modify lock “record data entries” that track held locks, users might need to perform special back-out or recovery processing to recover from this structure full condition. At times, that processing requires them to obtain more “special” lock structure locks, which are needed to perform the recovery that can lead to a paradoxical situation: They must use more “record data entries” to recover from being out of record data entries.

CFCC level 25 on IBM z16 provided improved use support for handling of lock structure “record data full” conditions by:

- Thresholding record data structure full conditions to occur when less than 100% full, reserving a special “for emergency use only” pool of record data entries for critical recovery purposes (user-specified threshold)
- Providing new APIs that allow exploiters to make use of this new reserved pool only when needed for recovery actions, but not for normal database locking purposes

z/OS APAR OA60650 and VSAM RLS APAR OA62059 are required in z/OSV2R4, and V2R5.

- ▶ DYNDISP=ONIOFF is deprecated on IBM z16, keeping only THIN option for shared-engine CF images, (also valid for IBM z17).

Coupling Facility images can run with shared or dedicated processors. Dedicated processors are recommended for best performance and production use (continuous polling model). Shared processors are recommended for test and development use in which a CF image requires significantly less than one processor’s worth of capacity, which encourages sharing CF processors across multiple CF images, or for less-performance critical production usage.

In shared-processor mode, the CF can currently use several different Dynamic Dispatching (DYNDISP) models:

- DYNDISP=OFF

LPAR time-slicing completely controls the CF processor; the processor polls the entire time that it is dispatched to a CF image by LPAR. The CF image never voluntarily gives up control of the shared processor. This option provides the least efficient sharing, and worst shared-engine CF performance.

- DYNDISP=ON

An optimization over pure LPAR time-slicing; the CF image sets timer interrupts to give the LPAR the initiative to redispach it, and the CF image voluntarily gives up control of the shared processor. This option provides the most efficient sharing, and better shared-engine CF performance.

- DYNDISP=THIN

An interrupt-driven model in which the CF processor is dispatched in response to a set of events that generate Thin Interrupts, runs until it runs out of work and then, gives up control voluntarily. This option provides the most efficient sharing, and best overall shared-engine CF performance.

DYNDISP=THIN support to use Thin Interrupts was available since zEC12, and proved to be efficient and well-performing in several shared-engine coupling facility configurations. IBM z15 made DYNDISP=THIN the default mode of operation for shared-engine coupling facility images, but supported the other options OFF and ON for continued “legacy” use.



**Note:** In IBM z16, DYNDISP=THIN option is the only available behavior for shared-engine CF dispatching.

Specifying OFF or ON in CF commands and the CF configuration file are preserved for compatibility; however, a warning message is issued to indicate that these options are no longer supported and that DYNDISP=THIN behavior is to be used.

## CFCC Level 24

CFCC level 24 is delivered on the IBM z15 (M/T 8561 and 8562) with driver level 41. CFCC level 24 adds the following features:

### ► CFCC Fair Latch Manager

This feature is an enhancement to the internals of the Coupling Facility (CFCC) dispatcher to provide CF work management efficiency and processor scalability improvements, and improve the “fairness” of arbitration for internal CF resource latches across tasks.

The tasks that are waiting for CF latches are not placed on the global suspend queue at all; instead, they are placed on latch-specific waiter queues for the exact instance of the latch they are requesting, and in the exact order in which they requested the latch. As a result, the global suspend queue is much less heavily used, and thus is much less a source of global contention or cache misses in the CF.

Also, when a latch is released, the specific latch’s latch waiter queue is used to transfer ownership of the latch directly to the next request in line (or multiple requests, in the case of a shared latch), and make that task (or tasks) ready to run, with the transferred latch already held. No possibility exists of any unfairness or “cutters” in line between the time that the latch is released versus when is obtained again.

For managing latches correctly for structures that are System-Managed (SM) synchronous duplexing, it is now important for the CF to understand which of the duplexed pair of requests operates as the “master” versus “slave” from a latching perspective, which requires more SM duplexing setup information from z/OS.

z/OS XCF/XES toleration APAR support is required to provide this enhancement.

### ► Message Path SYID Resiliency Enhancement

When a z/OS system IPLs, message paths are supposed to be deactivated by using system reset, and their SYIDs are supposed to be cleared in the process. During the IPL, z/OS then reactivates the message paths with a new SYID that represents the new instance of z/OS that uses the paths.

On rare occasions, a message path might not be deactivated during system reset or IPL processing, which leaves the message path active with the z/OS image’s OLD, now-obsolete SYID. Because the path erroneously remained active, z/OS does not see any need to reactivate it with a new, correct SYID.

From the perspective of the CF, the incorrect SYID persists and prevents delivery of signals to the z/OS image that uses that message path.

With IBM z15, CFCC provides a new resiliency mechanism that transparently recovers for this “missing” message path deactivate (if and when that situation ever occurs).

The CF provides more information to z/OS about every message path that appears active; namely, the current SYID with which the message path is registered in the CF. Whenever z/OS interrogates the state of the message paths to the CF, z/OS checks this SYID information for currency and correctness. If an obsolete or incorrect SYID exists in the message path for any reason, z/OS performs the following steps:



- i. Requests nondisruptive gathering of diagnostic information for the affected message paths and CF image.
- ii. Reactivates the message path with the correct SYID for the current z/OS image to seamlessly correct the problem.

This enhancement requires z/OS XCF/XES use APAR support for IBM z15.

► Shared-Engine CF Default is changed to “DYNDISP=THIN”

The CF operates with the following Dynamic Dispatching (DYNDISP) models:

- DYNDISP=OFF: LPAR time-slicing completely controls the CF processor. The processor polls the entire time that it is dispatched by LPAR, and it is idle (not dispatched) when undischpatched by LPAR. The result is least efficient sharing, worst shared-engine performance.
- DYNDISP=ON: An optimization over pure LPAR time slicing, in which the CFCC code judiciously sets timer interrupts to give LPAR the initiative to redispach it, and the CF sometimes voluntarily gives up control of the shared processor when it runs out of work to do. The result is more efficient sharing and better shared-engine performance. This setting is the default setting for CF LPAR running on IBM z15.
- DYNDISP=THIN: An interrupt-driven model in which the CF processor is dispatched in response to a set of events that generate Thin Interrupts and runs until it runs out of things to do and then, gives up control voluntarily (until the next interrupt causes it to get dispatched again). This model is the most efficient sharing, best shared-engine performance.

Thin Interrupt support is available since zEC12/zBC12, and proved to be efficient and performant in numerous different test and customer shared-engine coupling facility configurations.

For CFCC running on IBM z15, DYNDISP=THIN is now the default mode of operation for coupling facility images that use shared processors.

► CF monopolization avoidance

With IBM z15 T01/T02, the CF dispatcher monitors in real-time the number of CF tasks that have a command assigned to them for a specific structure on a structure by structure basis.

When the number of CF tasks that are used by any structure exceeds a model-dependent CF threshold, and a global threshold on the number of active tasks is also exceeded, the structure is considered to be “monopolizing” the CF, and z/OS is informed of this monopolization.

New support in z/OS observes the monopolization state for a structure, and starts to selectively queue and throttle incoming requests to the CF, on a structure-specific basis while other requests for other “nonmonopolizing” structures and workloads are unaffected.

z/OS dynamically manages the queue of requests for the “monopolizing” structures to limit the number of active CF requests (parallelism) to them. It monitors the monopolization state information of the CF to observe the structure becoming “nonmonopolized” again so that request processing can eventually revert to a nonthrottled mode of operation.

The overall goal of z/OS anti-monopolization support is to protect the ability of ALL well-behaved structures and workloads to access the CF, and get their requests processed in the CF in a timely fashion while implementing queuing and throttling mechanisms in z/OS to hold back the specific abusive workloads that are causing problems for other workloads.

z/OS XCF/XES use APAR support is required to provide this function.



To support an upgrade from one CFCC level to the next, different levels of CFCC can be run concurrently while the CF LPARs are running on different servers. CF LPARs that run on the same server share the CFCC level.

IBM z17 servers (CFCC level 26) can coexist in a sysplex with CFCC levels 25, and 24, nevertheless, the latest Coupling Facility Control Code and MCLs levels are always recommended for best performance and availability:

- ▶ On IBM z16 (MT 3931 and 3932): CFCC 25 - Service Level 02.51.2, Bundle S18 released April 2023.
- ▶ On IBM z15 (M/T 8561 and 8562): CFCC 24 - Service level 00.22, Bundle S48 released in August 2021.

For a CF LPAR with dedicated processors, the CFCC is implemented by using the active wait technique. This technique means that the CFCC is always running (processing or searching for service) and never enters a wait state. Therefore, the CF Control Code uses all the processor capacity that is available for the CF LPAR.

If the LPAR that is running the CFCC includes only dedicated processors (CPs or ICFs), the use of all processor capacity (cycles) is not an issue. However, this configuration can be an issue if the LPAR that is running the CFCC includes shared processors. On IBM z17, Thin Interrupts is the only valid option for shared engines in a CF LPAR (Thin Interrupts is also the only valid option on the IBM z16).

**Performance consideration:** Dedicated processor CF still provides the best CF image performance for production environments.

CF structure sizing changes are expected when moving to CFCC Level 2x. Always review the CF structure size by using the CFSizer tool when changing CFCC levels.

For more information about the recommended CFCC levels, see the current exception letter that is published on [IBM Resource Link®](#).

### 3.9.2 Coupling Thin Interrupts

CFCC Level 19 introduced Coupling Thin Interrupts to improve performance in environments that share CF engines. Although dedicated engines are preferable to obtain the best CF performance, Coupling Thin Interrupts helps facilitate the use of a shared pool of engines, which helps to lower hardware acquisition costs.

The interrupt causes a shared logical processor CF partition to be dispatched by PR/SM (if it is not already dispatched), which allows the request or signal to be processed in a more timely manner. The CF relinquishes control when work is exhausted or when PR/SM takes the physical processor away from the logical processor.



On IBM z17, the use of Coupling Thin Interrupts (DYNDISP=THIN) is now the only option that is available for shared engines in a CF LPAR. Specification of OFF or ON in CF commands and the CF configuration file will be preserved, for compatibility, but a warning message will be issued to indicate that these options are no longer supported, and that DYNDISP=THIN behavior will be used.

### 3.9.3 Dynamic CF dispatching

With the introduction of the Coupling Thin Interrupt support (only available option on IBM z17 and IBM z16), which is used only when the CF partition uses shared engines, the CFCC code is changed to handle these interrupts correctly. CFCC was also changed to relinquish voluntarily control of the processor whenever it runs out of work to do. It relies on Coupling Thin Interrupts to dispatch the image again in a timely fashion when new work (or new signals) arrives at the CF to be processed.

With IBM z17 and IBM z16, **DYNDISP=THIN** is the only mode of operation for CF images that use shared processors.

This capability allows ICF engines to be shared by several CF images. In this environment, it provides faster and far more consistent CF service times. It can also provide performance that is reasonably close to dedicated-engine CF performance.

The use of Thin Interrupts allows a CF to run by using a shared processor while maintaining good performance. The shared engine is allowed to be undispached when no more work exists, as in the past. The Thin Interrupt gets the shared processor that is dispatched when a command or duplexing signal is presented to the shared engine.

This function saves processor cycles and is an excellent option to be used by a production backup CF or a testing environment CF. This function is activated by default when a CF processor is shared.

The CPs can run z/OS operating system images and CF images. For software charging reasons, generally use only ICF processors to run CF images.

For more information about CF configurations, see the following resources:

- ▶ *Coupling Facility Configuration Options*, GF22-5042
- ▶ This [IBM Support web page](#)

## 3.10 Virtual Flash Memory

The Virtual Flash Memory feature code is 0566 on IBM z17 ME1.

### 3.10.1 IBM Z Virtual Flash Memory overview

VFM replaced the PCIe Flash Express feature with support that is based on main memory.

The “storage class memory” that is provided by Flash Express adapters is replaced with memory that is allocated from main memory (VFM).

VFM helps improve availability and handling of paging workload spikes when running z/OS. With this support, z/OS is designed to help improve system availability and responsiveness by using VFM across transitional workload events, such as market openings and diagnostic data collection.



z/OS also helps improve processor performance by supporting middleware use of pageable large (1 MB) pages, and eliminates delays that can occur when collecting diagnostic data during failures.

VFM also can be used in CF images to provide extended capacity and availability for workloads that use IBM WebSphere MQ Shared Queues structures.

### 3.10.2 VFM feature

A VFM feature (FC 0566) is 512 GB of memory on IBM z17 ME1. The maximum number of VFM features is 12 per IBM z17 ME1 system.

Ordered VFM memory reduces the maximum orderable memory for the model.

Simplification in its management is of great value because no hardware adapter is needed to manage. It also has no hardware repair and verify. It has a better performance because no I/O to attached adapter occurs. Finally, because this feature is part of memory, it is protected by RAIM and ECC.

VFM provides physical memory DIMMs that are needed to support activation of all customer purchased memory and HSA on a multiple drawer IBM z17 ME1 with one drawer that is done for the following features:

- ▶ Scheduled concurrent drawer upgrade, such as memory adds
- ▶ Scheduled concurrent drawer maintenance, such N+1 repair
- ▶ Concurrent repair of an out-of-service CPC drawer “fenced” during Activation (POR)

**Note:** All of these features can be done without VFM. However, all customer-purchased memory is not available for use in most cases. Some work might need to be shut down or not restarted.

### 3.10.3 VFM administration

The allocation and definition information of VFM for all partitions is viewed through the Storage Information panel that is under the Operational Customization panel.

The information is relocated during CDR in a manner that is identical to the process that was used for expanded storage. VFM is much simpler to manage (HMC task) and no hardware repair and verify (no cables and no adapters) are needed. Also, because this feature is part of internal memory, VFM is protected by RAIM and ECC and can provide better performance because no I/O to an attached adapter occurs.

**Note:** Use cases for Flash did not change (for example, z/OS paging and CF shared queue overflow). Instead, they transparently benefit from the changes in the hardware implementation.

No option is available for VFM plan ahead. The only option is to always include VFM plan ahead when Flexible Memory option is selected.

## 3.11 IBM Secure Service Container

Client applications are subject to several security risks in a production environment. These risks might include external risks (cyber hacker attacks) or internal risks (malicious software,



system administrators that use their privileged rights for unauthorized access and many others).

The IBM Secure Service Container (SSC) is a container technology through which you can more quickly and securely deploy software appliances on IBM z17.

An IBM SSC partition is a specialized container for installing and running specific appliances. An *appliance* is an integration of operating system, middleware, and software components that work autonomously and provide core services and infrastructures that focus on usability and security.

IBM SSC hosts most sensitive client workloads and applications. It acts as a highly protected and secured digital vault, enforcing security by encrypting the entire stack: memory, network, and data (both in-flight and at-rest). Applications that are running inside IBM SSC are isolated and protected from outsider and insider threats.

IBM SSC combines hardware, software, and middleware and is unique to IBM Z platform. Though it is called a container, it should not be confused with purely software open source containers (such as Kubernetes or Docker).

IBM SSC is a part of the Pervasive Encryption concept that was introduced with IBM z14, which is aimed at delivering best IBM Security® hardware and software enhancements, services, and practices for 360-degree infrastructure protection.

LPAR is defined as IBM SSC by using the HMC.

The IBM SSC solution includes the following key advantages:

- ▶ Applications require zero changes to use IBM SSC; software developers do not need to write any IBM SSC-specific programming code.
- ▶ End-to-end encryption (in-flight and at-rest data):
  - Automatic Network Encryption (TLS, IPsec): Data-in-flight.
  - Automatic File System Encryption (LUKS): Data-at-rest.
  - Linux Unified Key Setup (LUKS) is the standard way in Linux to provide disk encryption. SSC encrypts all data with a key that is stored within the appliance.
  - Protected memory: Up to 16 TB can be defined per IBM SSC LPAR.
- ▶ Encrypted Diagnostic Data

All diagnostic information (debug memory dump data, logs, and so on) are encrypted and do not contain any user or application data.
- ▶ No operating system access

After the IBM SSC appliance is built, Secure Shell (SSH) and the command line-interface (CLI) are disabled, which ensures that even system administrators cannot access the contents of the IBM SSC and do not know which application is running there.
- ▶ Applications that run inside IBM SSC are being accessed externally by REST APIs only, in a transparent way to user.
- ▶ Tamper-proof SSC Secure Boot:
  - IBM SSC-eligible applications are booted into IBM SSC by using verified booting sequence, where only trusted and digitally signed and verified by IBM software code is uploaded into the IBM SSC.
  - Vertical workload isolation, certified by EAL5+ Common Criteria Standard, which is the highest level that ensures workload separation and isolation.



- Horizontal workload isolation: Separation from the rest of the host environment.

IBM z17 technology provides built-in data encryption with excellent vertical scalability and performance that protects against data breach threats and data manipulation by privileged users. IBM SSC is a powerful IBM technology for providing the extra protection of the most sensitive workloads.

The following IBM solutions and offerings, and more to come, can be deployed in an IBM SSC environment:

- ▶ IBM Hyper Protect Virtual Servers (HPVS) solution is available for running Linux-based virtual servers with sensitive data and applications delivering a confidential computing environment to address your top security concerns.  
For more information, see this IBM Cloud® [web page](#).
- ▶ IBM Db2 Analytics Accelerator (IDAA) is a high-performance component that is tightly integrated with Db2 for z/OS. It delivers high-speed processing for complex Db2 queries to support business-critical reporting and analytic workloads. The accelerator transforms the mainframe into a hybrid transaction and analytic processing (HTAP) environment.  
For more information, see this IBM [web page](#).
- ▶ IBM Cloud Hyper Protect Data Base as a Service (DBaaS) for PostgreSQL or MongoDB offers enterprise cloud database environments with high availability for sensitive data workloads.  
For more information, see this IBM Cloud [web page](#).
- ▶ IBM Cloud Hyper Protect Crypto Services is a key management service and cloud hardware security module (HSM) that supports industry standards such as PKCS #11.  
For more information, see this IBM Cloud [web page](#).
- ▶ IBM Security Guardium® Data Encryption (GDE) consists of a unified suite of products that are built on a common infrastructure. These highly scalable solutions provide data encryption, tokenization, data masking, and key management capabilities to help protect and control access to data across the hybrid multicloud environment.  
For more information, see this [web page](#).
- ▶ IBM Blockchain® platform can be deployed on an IBM z16 by using IBM SSC to host the IBM Blockchain network.  
For more information, see this [web page](#).
- ▶ IBM Hyper Protect Data Controller (formerly IBM Data Privacy Passports) provided end-to-end, data-centric encryption and privacy to keep your data protected no matter where it travels in your enterprise. It maintains the suitable use of data, revokes future access at any time, and keeps an audit trail, which permits only authorized users to extract value from your data.

**Note:** IBM Hyper Protect Data Controller 1.2.x was withdrawn from support as of August 31, 2023.

For more information, see this [web page](#).







## 4



# Central processor complex I/O structure

This chapter describes the I/O system structure and connectivity options that are available on the IBM z17.

This chapter includes the following topics:

- ▶ 4.1, “Introduction to I/O infrastructure” on page 170
- ▶ 4.2, “I/O system overview” on page 172
- ▶ 4.3, “PCIe+ I/O drawer” on page 175
- ▶ 4.4, “CPC drawer fan-outs” on page 178
- ▶ 4.5, “I/O features” on page 181
- ▶ 4.6, “Connectivity” on page 184
- ▶ 4.7, “Cryptographic functions” on page 204
- ▶ 4.8, “Integrated Firmware Processor” on page 208



## 4.1 Introduction to I/O infrastructure

This section describes the I/O features that are available on the IBM z17.

### Notes:

- ▶ IBM z17 systems support PCIe+ I/O drawers only. I/O cage, I/O drawer, and PCIe I/O drawer are not supported.
- ▶ PCIe+ is an enhanced version of PCIe architecture which provides higher data rates, improved power efficiency, improved error correction, and improved virtualization capabilities. Throughout this chapter, the terms *adapter* and *card* refer to a PCIe I/O feature that is installed in a PCIe+ I/O drawer.

### 4.1.1 IBM z17 I/O infrastructure at a glance

IBM Z I/O is based on industry-standard Peripheral Component Interconnect Express Generation 4 (PCIe Gen4) I/O infrastructure. The PCIe I/O infrastructure that is provided by the central processor complex (CPC) enhances I/O capability and flexibility, while allowing for the future integration of PCIe adapters and features.

The PCIe I/O infrastructure in IBM z17 ME1 consists of the following components:

- ▶ PCIe+ Gen5 dual port fan-outs that support 32 GBps I/O bus for CPC drawer connectivity to the PCIe+ I/O drawers. The I/O PCIe hubs support Gen5 x16 into the fanout and will drive Gen4 x16 (Bifurcated to 2 at Gen4 x8) out of the fanout to the Gen4 in the PCIe+ I/O drawers. It connects to the PCIe Interconnect.
- ▶ Integrated Coupling Adapters Short Reach (ICA SR2.0), which are PCIe Gen4 features. Although the card hardware is Gen4 capable it will remain a PCIe Gen3 2 port PCIe Optical Coupling I/O hub card (150m short range transceivers). The ICA SR2.0 features has two ports, each port supporting 8 GBps.
- ▶ The 8U, 16-slot, and 2-domain PCIe+ I/O drawer for PCIe I/O features.

### Installed features in the PCIe+ I/O drawer

The I/O infrastructure of IBM z17 ME1 provides the following benefits:

- ▶ The bus connecting the CPC drawer to the I/O domain in the PCIe+ I/O drawer bandwidth is 16 GBps.
- ▶ Depending on the I/O card installed, up to 64 channels (16 PCIe I/O cards) can be supported in the PCIe+ I/O drawer.
- ▶ Storage connectivity:
  - Storage Area Network (SAN) connectivity:
    - FICON Express32-4P (FC 0387 LX and FC 0388 SX, 4 ports - new build)
    - FICON Express32S (FC 0461 LX and FC 0462 SX, 2 ports - carry forward)
    - FICON Express16SA (carry forward)

These cards provide two or four channels per feature for Fibre Channel connection (FICON), High-Performance FICON on Z (zHPF), and Fibre Channel Protocol (FCP) storage area networks.
  - IBM zHyperLink Express 2.0 (FC 0351)



Two ports per feature (new build) with ultra high-speed, direct connection to Select DS8000; works in tandem with FICON Express channels

- ▶ Local area network (LAN) connectivity- Open System Adapter (OSA):
  - The following features include two ports each:
    - OSA-Express7S 1.2 GbE (FC 0454, LX and FC 0455, SX)
    - OSA-Express7S 1000BASE-T (FC 0446 - Carry forward from z15 only)
  - The following features have one port each:
    - OSA-Express7S 1.2 GbE (FC 0456, LR and FC 0457, SR)
    - OSA-Express7S 1.2 25 GbE (FC 0459, SR and FC 0460, LR)
- ▶ Native PCIe features (plugged into the PCIe+ I/O drawer):
  - Network Express 10G (FC 0524, SR and FC 0525, LR): two ports per feature
  - Network Express 25G (FC 0526, SR and FC 0527, LR): two ports per feature
  - Coupling Express3 Long Reach (CE3 LR): two ports per feature
  - Crypto Express8S (single/dual HSM)
  - Crypto Express7S (single/dual HSM, carry forward)

#### 4.1.2 PCIe+ Generation 4 I/O Fanout - 2 Port

IBM z17 fanout card uses a PCIe Gen5 processor which provides two PCIe x16 Gen4 buses which interfaces to the PCIe+ I/O drawer. The interfaces to PCIe+ I/O Drawer is provided via a new Gen4 CDFP<sup>1</sup> passive copper I/O cable with a new Gen4 header. It can only be paired with the new Gen4 PCIe+ I/O Drawer switch card.

The PCIe Generation 4 uses 128b/130b encoding for data transmission. This configuration reduces the encoding overhead to approximately 1.54% versus the PCIe Generation 2 overhead of 20% that uses 8b/10b encoding.

The PCIe standard uses a low-voltage differential serial bus. Two wires are used for signal transmission, and a total of four wires (two for transmit and two for receive) form a lane of a PCIe link, which is full-duplex. Multiple lanes can be aggregated into a larger link width. PCIe supports link widths of 1, 2, 4, 8, 12, 16, and 32 lanes (x1, x2, x4, x8, x12, x16, and x32).

The data transmission rate of a PCIe link is determined by the link width (numbers of lanes), the signaling rate of each lane, and the signal encoding rule. The signaling rate of one PCIe Generation 4 lane is 16 gigatransfers per second (GTps). This results in a total bandwidth of 32 GB/s for a 16-lane (x16) configuration, compared to 16 GB/s in Gen 3.

**Note:** I/O infrastructure for IBM z17 is implemented as a combination of PCIe Gen3 and Gen4. The PU chip PCIe interface for IBM z17 is PCIe Generation 5 (x16 @32 GBps), but the CPC I/O fan-out infrastructure provides external connectivity as PCIe Generation 4 @16GBps

For example, a PCIe Gen3 x16 link features the following data transmission rates:

- ▶ The maximum theoretical data transmission rate per lane:
 
$$8 \text{ Gbps} * 128/130 \text{ bit (encoding)} = 7.87 \text{ Gbps} = 984.6 \text{ MBps}$$
- ▶ The maximum theoretical data transmission rate per link:
 
$$984.6 \text{ MBps} * 16 \text{ (lanes)} = 15.75 \text{ GBps}$$

<sup>1</sup> CDFP is short for 400 (CD in Roman numerals) Form factor Pluggable designed for high performance computing.



Considering that the PCIe link works in full-duplex mode, the data throughput rate of a PCIe Gen3 x16 link is 31.5 GBps (15.75 GBps in both directions).

**Link performance:** The link speeds do not represent the performance of the link. The performance depends on many factors, including latency through the adapters, cable lengths, and the type of workload.

PCIe Gen4 x16 links are used in IBM z17 servers for driving the PCIe+ I/O drawers, and for coupling links (ICA SR 2.0) for CPC to CPC communications. See Figure 4-1.

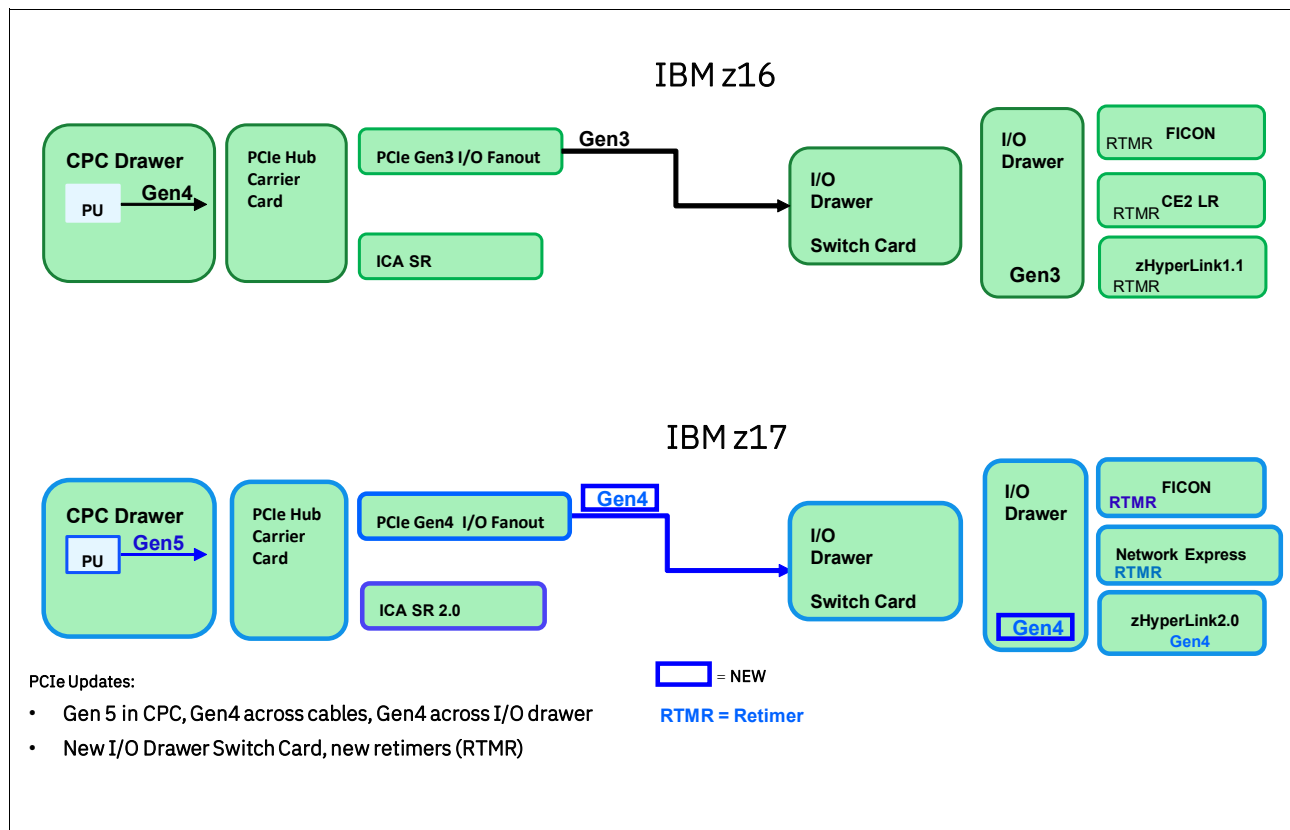


Figure 4-1 IBM z17 PCIe I/O infrastructure

## 4.2 I/O system overview

The IBM z17 I/O characteristics and supported features are described in this section.

### 4.2.1 Characteristics

The IBM z17 ME1 I/O subsystem provides great flexibility, high availability, and the following excellent performance characteristics:

- High bandwidth

IBM z17 servers use PCIe Gen4 protocol to drive PCIe+ I/O drawers and CPC to CPC (coupling) connections. PCI Gen4 doubles the data rate of PCIe Gen3.



For more information about coupling link connectivity, see 4.6.4, “Parallel Sysplex connectivity” on page 199.

- ▶ Connectivity options:
  - IBM z17 servers can be connected to an extensive range of interfaces, such as FICON/FCP for SAN connectivity, OSA and Network Express features for LAN connectivity, and zHyperLink Express for storage connectivity (low latency compared to FICON).
  - For CPC to CPC connections, IBM z17 servers use Integrated Coupling Adapter (ICA SR2.0 and the Coupling Express3 Long Reach (CE3 LR).
- ▶ Network Express Adapter
  - The IBM z17 Network Express adapter provides support for OSA, RoCE and Coupling LR. A CHPID type parameter is used to distinguish between the Network Express (OSA / RoCE) and Coupling LR features.
  - A single port in the Network Express can simultaneously have two “personalities”
    - The OSH CHPID type supports all legacy functions available with OSD, but implements Enhanced QDIO (EQDIO) architecture while OSD uses QDIO
    - NETH PFID type supports SMC-R RDMA for Linux native usage (TCP/IP, etc.)
    - Each port can be configured to provide support for a single host protocol (EQDIO or native PCIe) or a combination of host protocols
- ▶ Concurrent I/O upgrade
  - You can concurrently add I/O features to IBM z17 servers if unused I/O slot positions are available in the PCIe+ I/O or CPC drawer (for ICA SR2.0)
- ▶ Concurrent PCIe+ I/O drawer upgrade
  - More PCIe+ I/O drawers can be installed concurrently if free frame slots for the PCIe+ I/O drawers and PCIe fan-outs in the CPC drawer are available
- ▶ Dynamic I/O configuration
  - Dynamic I/O configuration supports the dynamic addition, removal, or modification of the channel path, control units, and I/O devices without a planned outage
- ▶ Remote Dynamic I/O Activation
  - Remote dynamic I/O activation is supported for CPCs running Stand-alone CFs, Linux on Z and z/TPF. IBM z17 provides a remote Dynamic I/O capability for driving hardware-only I/O configuration changes from a “driving” instance of z/OS Hardware Configuration Definition (HCD) on one CPC to a remote “target” standalone Coupling Facility CPC, to a CPC which hosts Linux on Z or to z/TPF images.
  - This new support is applicable only when both, the driving CPC and the target CPC are IBM z17 or IBM z16 with the required firmware support, and the driving systems z/OS is at level 2.4 or higher with APAR OA65559
- ▶ Pluggable optics:
  - The following features include Small Form-Factor Pluggable (SFP) optics<sup>2</sup>:
    - FICON Express32-4P
    - FICON Express32S
    - FICON Express16SA
    - OSA Express7S 1.2
    - OSA Express7S
    - Network Express

<sup>2</sup> SFP stands for Small Form-factor Pluggable, and it's a standardized format for optical transceivers used in network communication.



- Coupling Express3

These optics allow each channel to be individually serviced in a fiber optic module failure. The traffic on the other channels on the same feature can continue to flow if a channel requires servicing.

- The zHyperLink Express and ICA SR2.0 features uses fiber optics cable with MTP<sup>3</sup> connector and the cable uses a CXP connection to the adapter. The CXP<sup>4</sup> optics are provided with the adapter.

- ▶ Concurrent I/O card maintenance

Every I/O card that is plugged in a PCIe+ I/O drawer supports concurrent card replacement during a repair action.

## 4.2.2 Supported I/O features

The following I/O features are supported on an IBM z17 ME1 system (maximum for each individual adapter type):

- ▶ FICON Express32-4P
- ▶ FICON Express32S
- ▶ FICON Express16SA
- ▶ OSA-Express7S 1.2 25 GbE
- ▶ OSA-Express7S 1.2 10 GbE
- ▶ OSA-Express7S 1.2 GbE
- ▶ OSA-Express7S 1000BASE-T
- ▶ Network Express SR 10G
- ▶ Network Express LR 10G
- ▶ Network Express SR 25G
- ▶ Network Express LR 25G
- ▶ zHyperLink 2.0 Express
- ▶ ICA SR2.0
- ▶ Coupling Express3 LR10Gb
- ▶ Coupling Express3 LR 25Gb

For more details on the maximum number for each specific feature, the number of ports and definitions, please refer to Table 4-6 on page 184

**Notes:** Consider the following points:

- ▶ The number of I/O features depends on the number of installed PCIe+ I/O drawers. IBM z17 ME1 supports a maximum of 12 PCIe+ I/O drawers;
- ▶ A maximum of 384 coupling CHPIDs are available for an IBM z17, as a combination of the following examples (not all combinations are possible; subject to I/O configuration options):
  - Up to 48 ICA SR2.0 features (96 ports)
  - Up to 64 CE3 LR features (128 ports)
  - Up to 64 IC connections

<sup>3</sup> Multifiber Termination Push-On.

<sup>4</sup> For more information, see <https://cw.infinibandta.org/document/dl/7157>.



## 4.3 PCIe+ I/O drawer

The PCIe+ I/O drawers (see Figure 4-2) are attached to the CPC drawer through a PCIe cable and use PCIe Gen4 as the infrastructure bus within the drawer. The PCIe Gen4 I/O bus infrastructure data rate is up to 32 GBps.

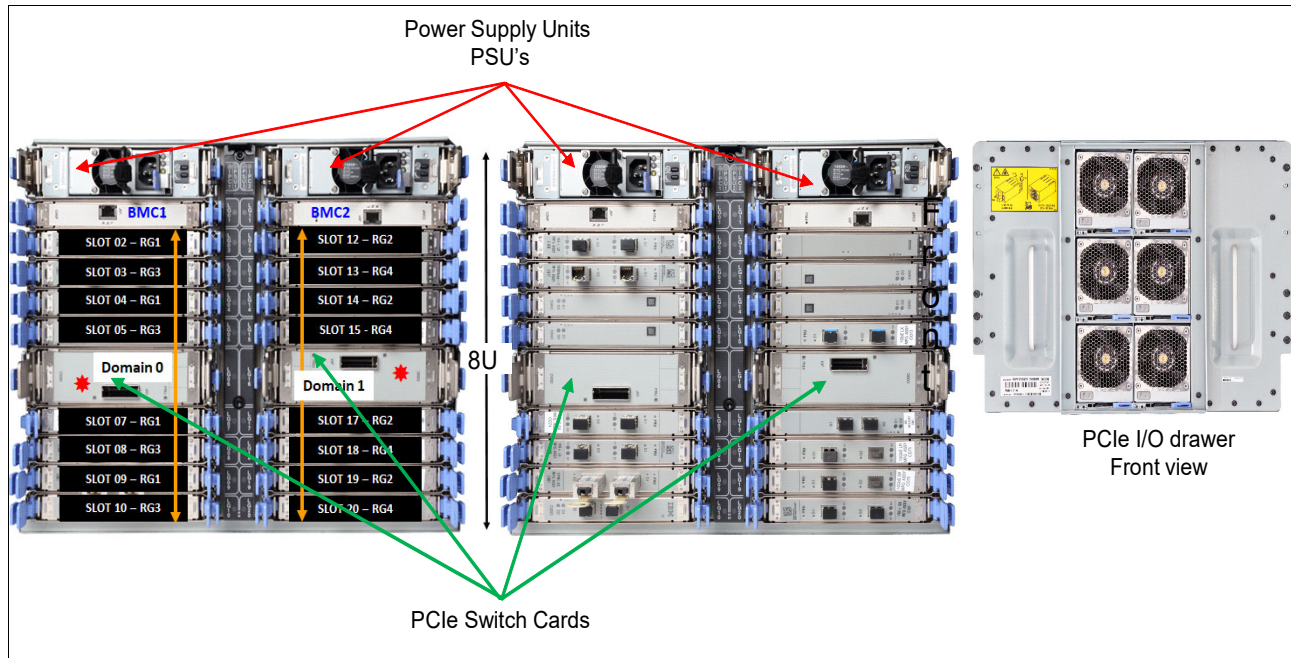


Figure 4-2 Rear and Front view of PCIe+ I/O drawer components

PCIe switch application-specific integrated circuits (ASICs) are used to fan out the host bus from the CPC drawer through the PCIe+ I/O drawer to the individual I/O features. Maximum 16 PCIe I/O features (up to 32 channels) per PCIe+ I/O drawer are supported.

The PCIe+ I/O drawer is a one-sided drawer (all I/O cards on one side, in the rear of the drawer) that is 8U high. The PCIe+ I/O drawer contains the 16 I/O slots for PCIe features, two switch cards, and two power supply units (PSUs) to provide redundant power, as shown in Figure 4-3 on page 176.



The PCIe+ I/O drawer slots numbers and Region Groups are shown in Figure 4-3.

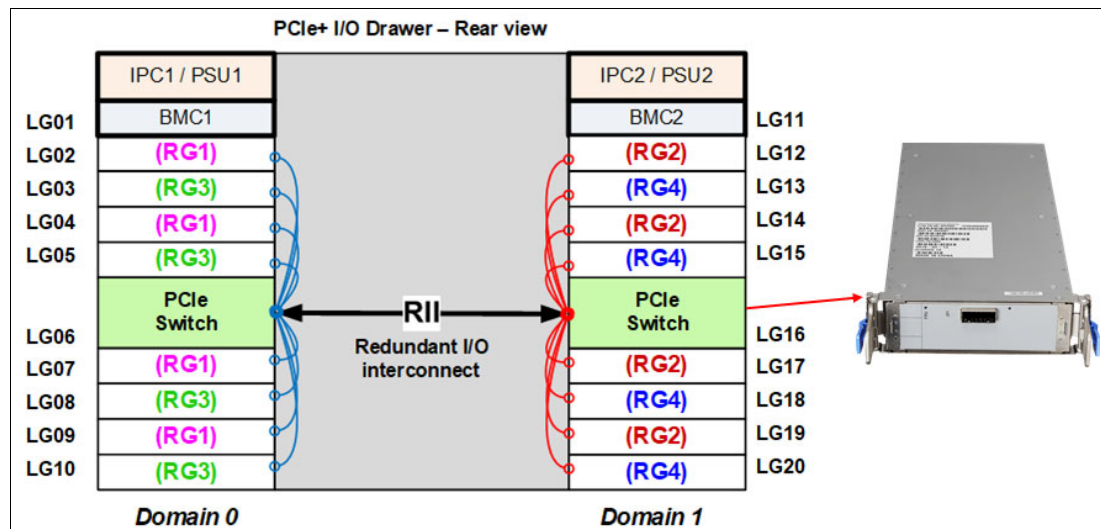


Figure 4-3 PCIe+ I/O drawer slots numbers and Region Groups

The I/O structure in an IBM z17 CPC is shown in Figure 4-4.

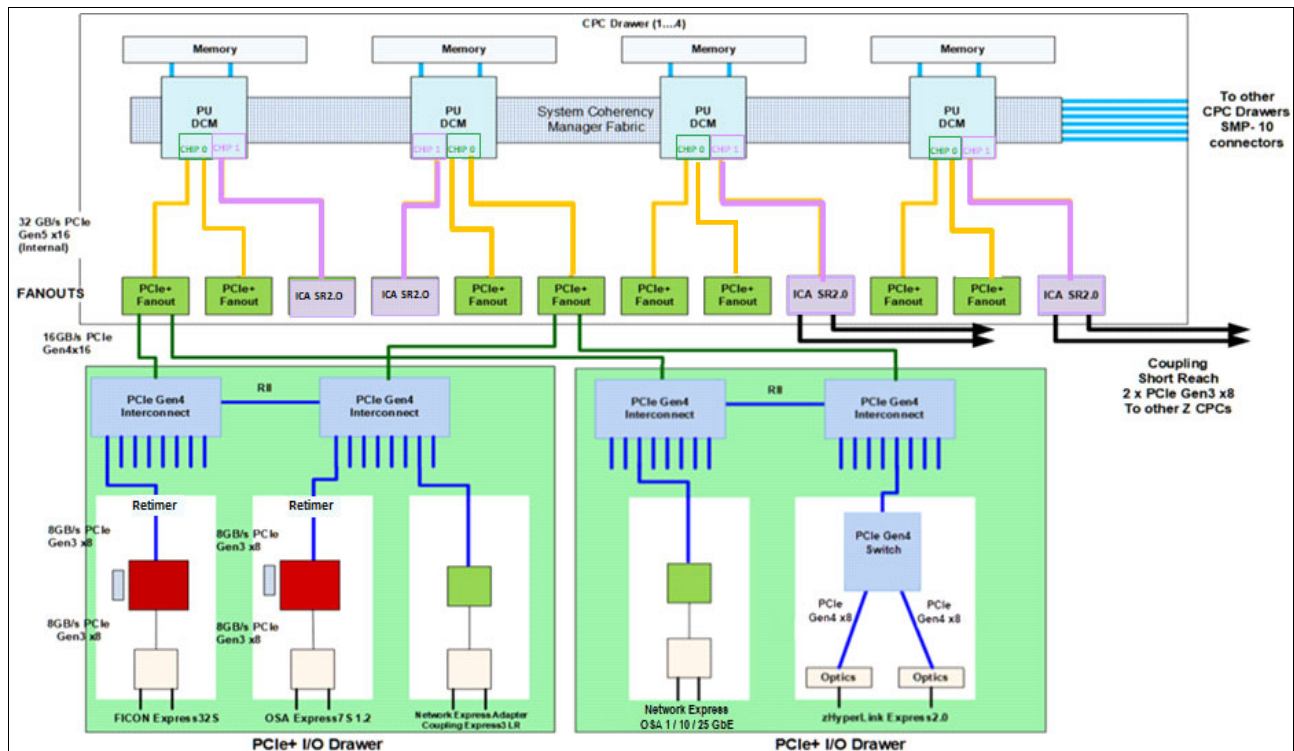


Figure 4-4 IBM z17 I/O connectivity(Max43 feature with two PCIe+ I/O drawers represented)

The PCIe switch card provides the fan-out from the high-speed x16 PCIe host bus to eight individual card slots. The PCIe switch card is connected to the CPC drawer through a single x16 PCIe Gen4 bus from a PCIe fan-out card (PCIe fan-out cards on IBM z17 have two PCIe Gen4 x16 ports/busses/links).



In the PCIe+ I/O drawer, The PCIe+ I/O drawer supports concurrent add and replace I/O features with which you can increase I/O capability as needed, depending on the CPC drawer.

The PCIe slots in a PCIe+ I/O drawer are organized into two I/O domains.(Figure 4-2 on page 175). The I/O feature cards that directly attach to the switch card constitute an I/O domain. Each I/O domain supports up to eight features and is driven through a PCIe switch card. Two PCIe switch cards always provide a backup path for each other through the passive connection in the PCIe+ I/O drawer backplane.

During a PCIe fan-out card or cable failure, 16 I/O cards in two domains can be driven through a single PCIe switch card. It is not possible to drive 16 I/O cards after one of the PCIe switch cards is removed.

The two switch cards are interconnected through the PCIe+ I/O drawer board (Redundant I/O Interconnect, or RII). In addition, switch cards in same PCIe+ I/O drawer are connected to PCIe fan-outs across clusters in CPC drawer for higher availability.

The RII design provides a failover capability during a PCIe fan-out card failure. Both domains in one of these PCIe+ I/O drawers are activated with two fan-outs. The Base Management Cards (BMCs) are used for system control.

The domains and their related I/O slots are shown in Figure 4-3 on page 176.

Each I/O domain supports up to eight features (FICON, OSA, Crypto, and so on.) All I/O cards connect to the PCIe switch card through the backplane board. The I/O domains and slots are listed in Table 4-1.

*Table 4-1 I/O domains of PCIe+ I/O drawer*

Domain	I/O slot in the domain
0	LG02, LG03, LG04, LG05, LG07, LG08, LG09, and LG10
1	LG12, LG13, LG14, LG15, LG17, LG18, LG19, and LG20

### 4.3.1 PCIe+ I/O drawer offering

Up to 12 PCIe+ I/O drawers can be installed on IBM z17 for supporting up to 192 PCIe I/O features.

Only the following PCIe I/O features can be carried forward for an upgrade to IBM z17 servers:

- ▶ FICON Express32S (SX and LX)
- ▶ FICON Express16SA (SX and LX)
- ▶ OSA-Express7S GbE SX - OSD; from IBM z15 only
- ▶ OSA-Express7S GbE LX - OSD; from IBM z15 only
- ▶ OSA-Express7S 10 GbE SR - OSD; from IBM z15 only
- ▶ OSA-Express7S 10 GbE LX - OSD; from IBM z15 only
- ▶ OSA-Express7S 1000 Base-T - OSD; from IBM z15 only
- ▶ OSA-Express7S 1.2 GbE SX - OSC,OSD
- ▶ OSA-Express7S 1.2 GbE LX - OSC,OSD
- ▶ OSA-Express7S 1.2 10 GbE SR - OSD
- ▶ OSA-Express7S 1.2 10 GbE LR - OSD
- ▶ OSA-Express7S 1.2 25 GbE SR - OSD
- ▶ OSA-Express7S 1.2 25 GbE LR - OSD
- ▶ Crypto Express7S(one or two ports/HSMs)



**Note:** On an IBM z17 system, only PCIe+ I/O drawers are supported. No older generation drawers can be carried forward.

IBM z17 server supports the following PCIe I/O new features that are hosted in the PCIe+ I/O drawers:

- ▶ FICON Express32-4P (SX and LX)
- ▶ OSA-Express7S 1.2 25 GbE (SR and LR)
- ▶ OSA-Express7S 1.2 10 GbE (SR and LR)
- ▶ OSA-Express7S 1.2 GbE (SX and LX)
- ▶ Crypto Express8S (one or two HSMs)
- ▶ Coupling Express3 Long Reach (CE3 LR) - 10Gb and 25Gb
- ▶ zHyperLink Express2.0
- ▶ Network Express SR 10G (SR and LR)
- ▶ Network Express SR 25G (SR and LR)

## 4.4 CPC drawer fan-outs

The IBM z17 server uses PCIe+ Gen4 fan-out cards to connect the CPC drawer to the I/O subsystem in the PCIe+ I/O drawers. The fan-out cards also include the ICA SR 2.0 coupling links for Parallel Sysplex. All fan-out cards support concurrent add, remove, and move.

The IBM z17 CPC drawer I/O infrastructure consists of the following features:

- ▶ The PCIe Generation 4 fan-out cards: Two ports per card (feature) that connect to PCIe+ I/O drawers.
- ▶ ICA SR2.0 fan-out cards: Two ports per card (feature) that connect to other (external) CPCs.

The PCIe fan-outs cards are installed in the rear of the CPC drawers. Each CPC drawer features 12 PCIe+ Gen4 fan-out slots.

The PCIe fan-outs and ICA SR2.0 fan-outs are installed in locations LG01 - LG12 at the rear in the CPC drawers (see Figure 2-7 on page 29).

On the CPC drawer, two BMC/OSC cards are available, each being a combination of a Base Management (BMC) card and an Oscillator Card (OSC) card. Each BMC/OSC card features one PPS port and one ETS port (RJ45 Ethernet) for both PTP and NTP.

An I/O connection diagram is shown in Figure 4-4 on page 176.

### 4.4.1 PCIe+ Gen4 fan-out (FC 0315)

The PCIe+ Gen4 fan-out card provides connectivity to a PCIe+ I/O drawer by using a copper cable. This PCIe fan-out card supports a link rate of 16 GBps (with two links per card).

A 16x PCIe copper cable of 1.5 meters (4.92 feet) - 4.0 meters (13.1 feet) is used for connection to the PCIe switch card in the PCIe+ I/O drawer. PCIe fan-out cards are always plugged in pairs and provide redundancy for I/O domains within the PCIe+ I/O drawer.

**Note:** The PCIe+ fan-out is used exclusively for I/O and cannot be shared for any other purpose.



## 4.4.2 PCIe Gen3 and PCIe Gen4 differences

PCIe Gen 4 doubles the data rate of PCIe Gen 3, allowing PCIe Gen 4 devices to transfer data at much faster speeds. PCIe Gen 3 operates at 8 GT/s (gigatransfers per second) which roughly translates to 1 GB/s per PCIe lane. By comparison, PCIe Gen 4 operates at 16 GT/s, or around 2 GB/s (gigabytes per second) per PCIe lane.

To understand the maximum bandwidth of a PCIe Gen 4 device, you must know the number of PCIe lanes that it supports. PCIe devices use “lanes” for transmitting and receiving data, so the more lanes a PCIe device can use, the greater the bandwidth can be. The number of lanes that a PCIe device supports is typically expressed like “x4” for 4 lanes, “x8” for 8 lanes, and so on. Refer to Table 4-2.

Table 4-2 PCIe Gen 3 and Gen4 data transfer speeds

	x1 lanes	x2lanes	x4 lanes	x8 lanes	x16 lanes
<b>PCIe Gen3</b> bandwidth	1 GB/s	2 GB/s	4 GB/s	8 GB/s	16 GB/s
<b>PCIe Gen4</b> bandwidth	2 GB/s	4 GB/s	8 GB/s	16 GB/s	32 GB/s
<b>PCIe Gen5</b> bandwidth	4 GB/s	8 GB/s	16 GB/s	32 GB/s	64 GB/s

**Note:** PCIe generations are backward compatible, so for example a PCIe Gen 3 device connected to a PCIe Gen 4 system will function normally at PCIe Gen 3 speeds.

## 4.4.3 Integrated Coupling Adapter - ICA SR2.0 (FC 0216)

The IBM ICA SR2.0 (FC 0216) is a two-port fan-out feature that is used for short distance coupling connectivity and uses channel type CS6. For IBM z17, the new build feature is ICA SR2.0. No carry forward of FCs 0172 and FC 0176 - ICA SR and ICA SR1.1 respectively.

The ICA SR2.0 has been updated with new Gen4 re-timers. It will also use the new CXP16 Gen4 optical module. Although the card hardware is Gen4 capable it will remain a PCIe Gen3 2 port PCIe Optical Coupling I/O hub card (150m short range transceivers).

The card is designed to drive distances up to 150 meters (492 feet) with a link data rate of eight GBps. ICA SR2.0 supports up to four channel-path identifiers (CHPIDs) per port and eight subchannels (devices) per CHPID.

The coupling links can be defined as shared between images (z/OS) within a CSS. They also can be spanned across multiple CSSs in a CPC. For ICA SR features, a maximum four CHPIDs per port can be defined.

When STP<sup>5</sup> (FC 1021) is available, ICA SR coupling links can be defined as timing-only links to other IBM z17, IBM z16, or IBM z15 systems.

These two fan-out features are housed in the PCIe+ Gen4 I/O fan-out slot on the IBM z17 CPC drawers. Up 48 ICA SR2.0 features (up to 96 ports) are supported on an IBM z17 ME1 system.

<sup>5</sup> Server Time Protocol



OM3 fiber optic can be used for distances up to 100 meters (328 feet). OM4 fiber optic cables can be used for distances up to 150 meters (492 feet). For more information, see the following publications:

- ▶ *Planning for Fiber Optic Links*, GA23-1409
- ▶ *9175 Installation Manual for Physical Planning*, GC28-7049.

#### 4.4.4 Fan-out considerations

Fan-out slots in the CPC drawer can be used to plug different fan-outs. On IBM z17 ME1, the CPC drawers can hold up to 48 PCIe fan-out cards (four-CPC drawers configuration).

##### Adapter ID number assignment

PCIe fan-outs and ports are identified by an Adapter ID (AID) that is initially dependent on their physical locations, which is unlike channels that are installed in a PCIe+ I/O drawer. Those channels are identified by a physical channel ID (PCHID) number that is related to their physical location. This AID must be used to assign a CHPID to the fan-out in the IOCDS definition. The CHPID assignment is done by associating the CHPID to an AID port (see Table 4-3).

Table 4-3 Fan-out locations and their AIDs for the CPC drawer (IBM z17 ME1)

Fan-out locations	CPC0 Location A10 AID (Hex)	CPC1 Location A15 AID (Hex)	CPC2 Location A20 AID (Hex)	CPC3 Location B10 AID (Hex)
LG01	00	0C	18	24
LG02	01	0D	19	25
LG03	02	0E	1A	26
LG04	03	0F	1B	27
LG05	04	10	1C	28
LG06	05	11	1D	29
LG07	06	12	1E	2A
LG08	07	13	1F	2B
LG09	08	14	20	2C
LG10	09	15	21	2D
LG11	0A	16	22	2E
LG12	0B	17	23	2F

##### Fan-out slots

The fan-out slots are numbered LG01 - LG12 (from left to right), as listed in Table 4-3. All fan-out locations and their AIDs for the CPC drawer are shown for reference only.

**Important:** The AID numbers that are listed in Table 4-3 are valid for a new build system only. If a fan-out is moved, the AID follows the fan-out to its new physical location.

The AID assignment is listed in the PCHID REPORT that is provided for each new server or for an MES upgrade on servers. Part of a PCHID REPORT for an IBM z17 is shown in



Example 4-1. In this example, four fan-out cards are installed at locations A10/LG05, A10/LG06, A15/LG05, and A15/LG06 with AIDs 04, 05, 10 and 11, respectively.

Example 4-1 **T**AID assignments PCHID REPORT sample

CHPIDSTART							
31463036				PCHID REPORT		Nov 10,2024	
Machine: 9175-ME1 SN1							
-----							
Source	Drwr	Slot	F/C	PCHID/Ports or AID	CHPIDs	Comment	
A10/LG05	A10B	LG05	0216	AID=04	N/A	ICA SR2.0	
A10/LG06	A10B	LG06	0216	AID=05	N/A	ICA SR2.0	
A15/LG05	A15B	LG05	0216	AID=10	N/A	ICA SR2.0	
A15/LG06	A15B	LG06	0216	AID=11	N/A	ICA SR2.0	

Fan-out features that are supported by the IBM z17 are listed in Table 4-4, which includes the feature type, feature code, and information about the link that is supported by the fan-out feature.

Table 4-4 Fan-out summary

Fan-out feature	Feature Code	Use	Cable type	Connector type	Maximum distance	Link data rate <sup>a</sup>
PCIe+ Gen4 fan-out	0315	PCIe I/O drawer conn.	Copper	N/A	4 m (13.1 ft.)	16 GBps
ICA SR2.0	0216	Coupling link	OM4	MTP	150 m (492 ft.)	8 GBps
			OM3	MTP	100 m (328 ft.)	8 GBps

a. The link data rates do not represent the performance of the link. The performance depends on many factors, including latency through the adapters, cable lengths, and the type of workload.

## 4.5 I/O features

I/O features (adapters) include ports<sup>6</sup> to connect the IBM z17 to external devices, networks, or other servers. I/O features are plugged into the PCIe+ I/O drawer, based on the configuration rules for the server. Different types of I/O cards are available: one for each channel or link type. I/O cards can be installed or replaced concurrently.

### 4.5.1 I/O feature card ordering information

The I/O features that are supported by IBM z17 servers and the ordering information for them are listed in Table 4-5.

Table 4-5 I/O features and ordering information

Channel feature	Feature code	New build	Carry-forward
FICON Express32-4P LX (4 port)	0387	Y	N/A
FICON Express32-4P SX (4 port)	0388	Y	N/A
FICON Express32S LX (2 port)	0461	N	Y

<sup>6</sup> Certain I/O features do not have external ports, such as Crypto Express.



Channel feature	Feature code	New build	Carry-forward
FICON Express32S SX (2 port)	0462	N	Y
FICON Express16SA LX	0436	N	Y
FICON Express16SA SX	0437	N	Y
OSA-Express7S 1.2 25 GbE LR	0460	Y	Y
OSA-Express7S 1.2 25 GbE SR	0459	Y	Y
OSA-Express7S 1.2 10 GbE LR	0456	Y	Y
OSA-Express7S 1.2 10 GbE SR	0457	Y	Y
OSA-Express7S 1.2 GbE LX	0454	Y	Y
OSA-Express7S 1.2 GbE SX	0455	Y	Y
OSA-Express7S 1000BASE-T	0446	N	Y
PCIe+ Gen4 fan-out <sup>a</sup>	0315	Y	N/A
Integrated Coupling Adapter (ICA SR2.0) <sup>b</sup>	0216	Y	N/A
Coupling Express3 LR 10Gb	0498	Y	N/A
Coupling Express3 LR 25Gb	0499	Y	N/A
Crypto Express8S (dual HSM)	0908	Y	Y
Crypto Express8S (single HSM)	0909	Y	Y
Crypto Express7S (2 ports)	0898	N	Y
Crypto Express7S (1 port)	0899	N	Y
Network Express SR 10G	0524	Y	N/A
Network Express LR10G	0525	Y	N/A
Network Express SR 25G	0526	Y	N/A
Network Express LR 25G	0527	Y	N/A
IBM Adapter for NVMe1.1	0448	Y	N/A
IDAA Internal Storage NVMe 15TB SSD	0528	Y	N/A
IBM Spyre AI Accelerator Adapter <sup>c</sup>	0463	Y	N/A
IBM Spyre AI Accelerator Reserve Slots <sup>d</sup>	0061	Y	N/A
zHyperLink Express2.0	0351	Y	N/A

a. Installed in the CPC Drawer; provides connectivity for the PCIe+ I/O Drawer.

b. Installed in the CPC Drawer; provides coupling connectivity (short distance: up to 150m).

c. The IBM Spyre AI Adapter is configured in sets of 8 adapters

d. This feature reserves slots in the PCIe+ drawer for future IBM Spyre cards (in sets of 8 slots)

**Coupling links connectivity support:** Consider the following points:

- z14 and z14 ZR1 and older systems are not supported in same Parallel Sysplex or STP CTN with IBM z17.



## 4.5.2 Physical channel ID (PCHID) report

A physical channel ID (PCHID) reflects the physical location of a channel-type interface. A PCHID number is based on the following factors:

- ▶ PCIe+ I/O drawer location
- ▶ Channel feature slot number
- ▶ Port number of the channel feature

A CHPID does not directly correspond to a hardware channel port. Instead, it is assigned to a PCHID in the hardware configuration definition (HCD) or IOCP.

A PCHID REPORT is created for each new build server and for upgrades on servers. The report lists all I/O features that are installed, the physical slot location, and the assigned PCHID. A portion of a sample PCHID REPORT is shown in Example 4-2.

*Example 4-2 PCHID REPORT*

---

```

CHPIDSTART
  31463036                                PCHID REPORT                                Apr 10,2025
Machine: 9175-ME1  SN1
-----
Source      Drwr  Slot  F/C    PCHID/Ports or AID      Comment
A10/LG06    A10B  LG06  0216   AID=05
A15/LG06    A15B  LG06  0216   AID=11
A15/LG12/J02 Z01B  02    0461   100/D1 101/D2
A15/LG12/J02 Z01B  05    0457   10C/D1
A15/LG12/J02 Z01B  07    0457   110/D1
A15/LG12/J02 Z01B  08    0459   114/D1
A15/LG12/J02 Z01B  09    0908   118/P00 119/P01
A15/LG12/J02 Z01B  10    0524   11C/D1D2      RG3
A20/LG12/J02 Z01B  12    0461   120/D1 121/D2
A20/LG12/J02 Z01B  13    0462   124/D1 125/D2
A20/LG12/J02 Z01B  18    0457   134/D1
A20/LG12/J02 Z01B  19    0457   138/D1
.....<< snippet >>.....

```

---

The PCHID REPORT that is shown in Example 4-2 includes the following components:

- ▶ Feature Code 0216 (Integrated Coupling Adapters (ICA SR2.0) is installed in the CPC drawer (location A10, slot LG06, and A15 LG06), and has AIDs 05 and 11 assigned.
- ▶ Feature 0461 (FICON Express32S LX) is installed in PCIe+ I/O drawer 1: Location Z01B, slot 02 with PCHIDs 100/D1 and 101/D2 assigned.
- ▶ Two feature codes 0457 (OSA-Express7S 1.2 10 GbE SR) installed in PCIe+ I/O drawer 1 in slots 05 and 07, with PCHIDs 10C/D1 and 110/D1, respectively.

A resource group (RG) parameter also is shown in the PCHID REPORT for native PCIe features. A balanced plugging of native PCIe features exists between four resource groups (RG1, RG2, RG3, and RG4).

The preassigned PCHID number of each I/O port relates directly to its physical location (jack location in a specific slot).



## 4.6 Connectivity

I/O channels are part of the CSS. They provide connectivity for data exchange between servers, between servers and external control units (CUs) and devices, or between networks.

For more information about connectivity to external I/O subsystems (for example, disks), see 4.6.2, “Storage connectivity” on page 186.

For more information about communication to LANs, see 4.6.3, “Network connectivity” on page 191.

Communication between servers is implemented by using CE LR, ICA SR, or channel-to-channel (FICON CTC) connections. For more information, see 4.6.4, “Parallel Sysplex connectivity” on page 199.

### 4.6.1 I/O feature support and configuration rules

The supported I/O features are listed in Table 4-6. Also listed in Table 4-6 are the number of ports per card, port increments, the maximum number of feature cards, and the maximum number of channels for each feature type. The CHPID definitions that are used in the IOCDs also are listed

Table 4-6 IBM z17 ME1 supported I/O features

I/O feature	Ports per card	Port increments	Maximum ports	Maximum I/O slots	PCHID	CHPID definition
<b>Storage access</b>						
FICON Express32-4P LX/SX (FCs 0387, 0388)	4	4	384	96	Yes	FC, FCP
FICON Express32S LX/SX (FCs 0461, 0462)	2	2	384	192	Yes	FC, FCP
FICON Express16SA LX/SX (FCs 0436, 0437)	2	2	384	192	Yes	FC, FCP
zHyperLink Express 2.0	2	2	32	16	Yes	N/A <sup>a</sup>
<b>OSA-Express features<sup>b</sup></b>						
OSA-Express7S 1.2 25 GbE LR/SR (FCs 0459, 0460)	1	1	48	48	Yes	OSD
OSA-Express7S 1.2 10 GbE LR/SR (FCs 0456, 0457)	1	1	48	48	Yes	OSD
OSA-Express7S 1.2 GbE LX/SX (FCs 0454, 0455)	2	2	96	48	Yes	OSC, OSD
OSA-Express 7S GbE LX/SX (FCs 0443, 0442)	2	2	96	48	Yes	OSD
OSA-Express 7S 10GbE LR/SR (FCs 0444, 0445)	1	1	48	48	Yes	OSD
OSA-Express7S 1000BASE-T (FC 0446)	2	2	96	48	Yes	OSD <sup>c</sup>



I/O feature	Ports per card	Port increments	Maximum ports	Maximum I/O slots	PCHID	CHPID definition
<b>Network Express features<sup>d</sup></b>						
Network Express SR 10G	2	2	96	48	Yes	OSH, NETH
Network Express LR 10G	2	2	96	48	Yes	OSH, NETH
Network Express SR 25G	2	2	96	48	Yes	OSH, NETH
Network Express LR 25G	2	2	96	48	Yes	OSH, NETH
<b>Coupling Express3 features<sup>e</sup></b>						
Coupling Express3 LR 10Gb	2	2	64	32	Yes	CL5
Coupling Express3 LR 25Gb	2	2	64	32	Yes	CL6
<b>Integrated Coupling Adapter</b>						
ICA SR2.0 <sup>f</sup>	2	2	96	48	N/A <sup>g</sup>	CS5

a. These features are defined by using Virtual Functions IDs (FIDs).

b. The OSA Express7S cards (non v1.2) are only Carry Forward from z15

c. OSA-Express7S 1000BASE-T (carry forward from z15) no longer supports CHPID type OSC

d. IBM z17 multi-function Network Adapter (supports OSH CHPID using Enhanced QDIO(EQDIO) and NETH PFID)

e. Coupling Express3 features are supported by the Network Express adapter

f. Installed in the CPC drawer.

g. ICA SR2.0 features are characterized by Adapter ID (AID).

At least one I/O feature (FICON) or one coupling link feature (ICA SR or CE LR) must be present in the minimum configuration.

The following features can be shared and spanned:

- ▶ FICON channels that are defined as FC or FCP
- ▶ Network Express features that are defined as NETH or OSH
- ▶ Coupling links that are defined as CL5 or CL6
- ▶ HiperSockets that are defined as IQD

The following features are plugged into a PCIe+ I/O drawer and do not require the definition of a CHPID and CHPID type:

- ▶ Each Crypto Express (8S/7S) feature occupies one I/O slot, but does not include a PCHID type. However, LPARs in all CSSs can access the features. Each Crypto Express adapter can support up to 85 domains.
- ▶ Each zHyperlink Express2.0 feature occupies one I/O slot but does not include a CHPID type. However, LPARs in all CSSs can access the feature. The zHyperLink Express adapter works as native PCIe adapter and can be shared by multiple LPARs. Each port supports up to 127 Virtual Functions (VFs), with one or more VFs/PFIDs being assigned to each LPAR. This support gives a maximum of 254 VFs per adapter.

## I/O feature cables and connectors

The IBM Facilities Cabling Services fiber transport system offers a total cable solution service to help with cable ordering requirements. These services can include the requirements for all of the protocols and media types that are supported (for example, FICON, Coupling Links, and OSA). The services can help whether the focus is the data center, SAN, LAN, or the end-to-end enterprise.



**Cables:** All fiber optic cables, cable planning, labeling, and installation are client responsibilities for new IBM z17 installations and upgrades. Fiber optic conversion kits and mode conditioning patch cables are not orderable as features on IBM z17 servers. All other cables must be sourced separately.

For more details on the cable specifications for these features, please refer to Appendix D, “Channel options” on page 547.

## 4.6.2 Storage connectivity

Connectivity to external I/O subsystems (for example, disks) is provided by FICON channels and zHyperLink<sup>7</sup>.

### FICON channels

IBM z17 supports the following FICON features:

- ▶ FICON Express32-4P LX and SX - 4 Ports/Feature (FC 0387/0388 - NB<sup>8</sup>)
- ▶ FICON Express32S LX and SX - 2 Ports/Feature (FC 0461/0462 - CF<sup>9</sup>)
- ▶ FICON Express16SA LX and SX - 2 Ports/Feature (FC 0436/0437 - CF)

These FICON features conform to the following architectures:

- ▶ Fibre Connection (FICON)
- ▶ High Performance FICON on Z (zHPF)
- ▶ Fibre Channel Protocol (FCP)

The FICON features provide connectivity between any combination of servers, directors, switches, and devices (control units, disks, tapes, and printers) in a SAN.

Each FICON Express feature occupies one I/O slot in the PCIe+ I/O drawer. Each feature includes two or four ports, each supporting an LC Duplex connector, with one PCHID and one CHPID that is associated with each port.

Each FICON Express feature uses SFP+ optics that allow for concurrent repairing or replacement for each SFP. The data flow on the unaffected channels on the same feature can continue. A problem with one FICON Express port does not require replacement of a complete feature.

Each FICON Express feature also supports cascading, which is the connection of two FICON Directors in succession. This configuration minimizes the number of cross-site connections and helps reduce implementation costs for disaster recovery applications, IBM Geographically Dispersed Parallel Sysplex (GDPS), and remote copy.

IBM z17 servers support 32 K devices per FICON channel for all FICON features.

Each FICON Express channel can be defined independently for connectivity to servers, switches, directors, disks, tapes, and printers, by using the following CHPID types:

- ▶ CHPID type FC: The FICON, zHPF, and FCTC protocols are supported simultaneously.
- ▶ CHPID type FCP: Fibre Channel Protocol that supports attachment to SCSI devices directly or through Fibre Channel switches or directors.

<sup>7</sup> zHyperLink feature operates with a FICON channel.

<sup>8</sup> NB - New Build

<sup>9</sup> CF- Carry forward



FICON channels (CHPID type FC or FCP) can be shared among LPARs and defined as spanned. All ports on a FICON feature must be of the same type (LX or SX). The features are connected to a FICON capable control unit (point-to-point or switched point-to-point) through a Fibre Channel switch.

### ***FICON Express32S and FICON Express32-4P***

The FICON Express32S feature is installed in the PCIe+ I/O drawer. Each of the two (or four - depending on the FC) independent ports is capable of 8 Gbps, 16 Gbps, or 32 Gbps. The link speed depends on the capability of the attached switch or device. The link speed is auto-negotiated, point-to-point, and is transparent to users and applications.

These FICON Express32S and FICON Express32-4P features support LC Duplex optical transceivers.

**Consideration:** FICON Express32S features do not support auto-negotiation to a data link rate of 2 or 4 Gbps (only 8, 16, or 32 Gbps) for point-to-point connections. A compatible switch must be used to connect to lower speed devices.

### ***FICON Express16SA***

The FICON Express16SA feature is installed in the PCIe+ I/O drawer. Each of the two independent ports is capable of 8 or 16 Gbps. The link speed depends on the capability of the attached switch or device. The link speed is auto-negotiated, point-to-point, and is transparent to users and applications.

These FICON Express16SA features support LC Duplex optical transceivers.

**Consideration:** FICON Express16SA features do not support auto-negotiation to a data link rate of 2 or 4 Gbps (only 8 or 16 Gbps) for point-to-point connections. A compatible switch must be used to connect to lower speed devices.

For more information, see the FICON Express chapter in the *IBM Z Connectivity Handbook*, [SG24-5444](#).

## **FICON features and built-in functions**

In combination with the FICON Express32-4P, FICON Express32S, and FICON Express16SA, IBM z17 servers provide enhancements for FICON in functional and performance aspects with IBM Endpoint Security solution.

### ***IBM Fibre Channel Endpoint Security***

IBM z17 Model ME1 supports IBM Fibre Channel Endpoint Security Enablement feature (FC 1146). FC 1146 provides FC/FCP link encryption and endpoint authentication. It is an end-to-end solution that helps ensure all data flowing on the Fibre Channel links within and across data centers flows between trusted entities.

More information about Fibre Channel Endpoint Security can be found in 4.7.3, “IBM Fibre Channel Endpoint Security” on page 206

### ***Forward Error Correction***

Forward Error Correction (FEC) is a technique that is used for reducing data errors when transmitting over unreliable or noisy communication channels (improving signal to noise ratio). By adding redundancy error-correction code (ECC) to the transmitted information, the receiver can detect and correct several errors without requiring retransmission. This process features improve signal reliability and bandwidth use by reducing retransmissions because of



bit errors, especially for connections across long distance, such as an inter-switch link (ISL) in a GDPS Metro Mirror environment.

The FICON Express32-4P, FICON Express32S, and FICON Express16SA support FEC coding on top of its 64b/66b data encoding for 16 and 32 Gbps connections. This design can correct up to 11 bit errors per 2112 bits transmitted.

Therefore, while connected to devices that support FEC at 16 Gbps connections, the FEC design allows FICON Express32-4P, FICON Express32S, and FICON Express16SA channels to operate at higher speeds, over longer distances, with reduced power and higher throughput. At the same time, the same reliability and robustness for which FICON channels traditionally known are maintained.

With the IBM DS8870 or newer, IBM z17 servers can extend the use of FEC to the fabric N\_Ports for a completed end-to-end coverage of 32 Gbps FC links.

### ***FICON dynamic routing***

Starting with IBM z14 and with newer servers, FICON channels are no longer restricted to the use of static SAN routing policies for ISLs for cascaded FICON directors. The IBM Z servers now support dynamic routing in the SAN with the FICON Dynamic Routing (FIDR) feature. It is designed to support the dynamic routing policies that are provided by the FICON director manufacturers; for example, Brocade's exchange-based routing (EBR) and Cisco's originator exchange ID (OxID)<sup>10</sup> routing.

A static SAN routing policy normally assigns the ISL routes according to the incoming port and its destination domain (port-based routing), or the source and destination ports pairing (device-based routing).

The port-based routing (PBR) assigns the ISL routes statically that is based on "first come, first served" when a port starts a fabric login (FLOGI) to a destination domain. The ISL is round-robin that is selected for assignment. Therefore, I/O flow from same incoming port to same destination domain always is assigned the same ISL route, regardless of the destination port of each I/O.

This setup can result in some ISLs becoming overloaded while some are under-used. The ISL routing table is changed whenever IBM Z server undergoes a power-on-reset (POR), so the ISL assignment is unpredictable.

Device-based routing (DBR) assigns the ISL routes statically that is based on a hash of the source and destination port. That I/O flow from same incoming port to same destination is assigned to same ISL route. Compared to PBR, the DBR is more capable of spreading the load across ISLs for I/O flow from the same incoming port to different destination ports within a destination domain.

When a static SAN routing policy is used, the FICON director features limited capability to assign ISL routes that are based on workload. This limitation can result in unbalanced use of ISLs (some might be overloaded, while others are under-used).

The dynamic routing ISL routes are dynamically changed based on the Fibre Channel exchange ID, which is unique for each I/O operation. ISL is assigned at I/O request time; therefore, different I/Os from same incoming port to same destination port are assigned different ISLs.

With FIDR, IBM z17 servers feature the following advantages for performance and management in configurations with ISL and cascaded FICON directors:

---

<sup>10</sup> Check with the switch provider for their support statement.



- ▶ Support sharing of ISLs between FICON and FCP (PPRC or distributed)
- ▶ I/O traffic is better balanced between all available ISLs
- ▶ Improved use of FICON director and ISL
- ▶ Easier to manage with a predictable and repeatable I/O performance

FICON dynamic routing can be enabled by defining dynamic routing-capable switches and control units in HCD. Also, z/OS implemented a health check function for FICON dynamic routing.

### ***Improved zHPF I/O execution at distance***

By introducing the concept of pre-deposit writes, zHPF reduces the number of round trips of standard FCP I/Os to a single round trip. Originally, this benefit is limited to writes that are less than 64 KB.

zHPF on IBM z14 and newer servers were enhanced to allow all large write operations (> 64 KB) at distances up to 100 kilometers (62 miles) to be run in a single round trip to the control unit. This improvement avoids elongating the I/O service time for these write operations at extended distances.

### ***Read Diagnostic Parameter Extended Link Service support***

To improve the accuracy of identifying a failed component without unnecessarily replacing components in a SAN fabric, a new Extended Link Service (ELS) command called Read Diagnostic Parameters (RDP) was added to the Fibre Channel T11 standard. This command allows IBM Z servers to obtain extra diagnostic data from the SFP optics that are throughout the SAN fabric.

Starting with IBM z14, IBM Z servers can read this extra diagnostic data for all the ports that are accessed in the I/O configuration and make the data available to an LPAR. For z/OS LPARs that use FICON channels, z/OS displays the data with a [new message and display](#) command. For Linux on IBM Z, z/VM, and VSE<sup>n</sup> V6.3.1 (21<sup>st</sup> Century Software), and LPARs that use FCP channels, this diagnostic data is available in a new window in the SAN Explorer tool.

### ***N\_Port ID Virtualization***

N\_Port ID Virtualization (NPIV) allows multiple system images (in LPARs or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. First introduced with IBM z9<sup>®</sup> EC, this feature can be used with earlier FICON features that were carried forward from earlier servers.

By using the FICON Express as an FCP channel with NPIV enabled, the maximum numbers of the following aspects for one FCP physical channel are doubled:

- ▶ Maximum number of NPIV hosts defined: Increased from 32 to 64
- ▶ Maximum number of remote N\_Ports communicated: Increased from 512 to 1024
- ▶ Maximum number of addressable LUNs: Increased from 4096 to 8192
- ▶ Concurrent I/O operations: Increased from 764 to 1528

For more information about operating systems that support NPIV, see [“N\\_Port ID Virtualization”](#).

### ***Export and import physical port WWPNs for FCP Channels***

IBM Z automatically assign worldwide port names (WWPNs) to the physical ports of an FCP channel that is based on the PCHID. This WWPN assignment changes when an FCP channel is moved to a different physical slot position.



IBM z15 and newer servers allow for the modification of these default assignments, which also allows FCP channels to keep previously assigned WWPNNs, even after being moved to a different slot position. This capability can eliminate the need for reconfiguration of the SAN in many situations, and is especially helpful during a system upgrade (FC 0099 - WWPNN Persistence).

### ***FICON support for multiple-hop cascaded SAN configurations***

Before the introduction of z13 and z13s servers, IBM Z FICON SAN configurations supported a single ISL (a single hop) in a cascaded FICON SAN environment only.

IBM z14 and newer servers support up to three hops in a cascaded FICON SAN environment. This support allows clients to more easily configure a three- or four-site disaster recovery solution.

For more information about the FICON multi-hop, see the [FICON Multihop: Requirements and Configurations white paper](#) at the IBM Techdocs Library website.

### **FICON cable specifications**

The FICON feature codes, cable type, maximum unrepeatd distance, and the link data rate on an IBM z17 ME1 server are listed in Appendix D, "Channel options" on page 547. All FICON features use LC Duplex connectors.

### **zHyperLink Express2.0 (FC 0351)**

zHyperLink is a technology that provides up to 5x reduction in I/O latency times for Db2 read requests with the qualities of service IBM Z clients expect from I/O infrastructure for Db2 v11 plus fixes (for read support) and v12 plus fixes (for write support) with z/OS.

The following z/OS versions are supported for zHyperLink:

- ▶ z/OS V2.4 with PTFs
- ▶ z/OS V2.5 with PTFs
- ▶ z/OS V3.1

The zHyperLink Express2.0 feature (FC 0351) provides a low latency direct connection between IBM z17 and DS8000 storage system.

The zHyperLink Express2.0 is the result of new business requirements that demand fast and consistent application response times. It dramatically reduces latency by interconnecting the IBM z17 directly to I/O Bay of the DS8k by using PCIe Gen3 x 8 physical link (up to 150 meters [492 feet]). A new transport protocol is defined for reading and writing IBM CKD data records<sup>11</sup>, as documented in the zHyperLink interface specification.

On IBM z17, zHyperLink Express2.0 card is a PCIe Gen4 adapter with updated Gen4 retimer, which is installed in the PCIe+ I/O drawer. HCD definition support was added for new PCIe function type with PORT attributes.

### ***Requirements of zHyperLink Express2.0***

The zHyperLink Express2.0 feature is available on IBM z17 servers, and includes the following requirements:

- ▶ z/OS 2.4 or later
- ▶ 150 m (492 feet) maximum distance in a point-to-point configuration
- ▶ Supported DS8000 (see *Getting Started with IBM zHyperLink for z/OS*, [REDP-5493](#))
- ▶ zHyperLink Express2.0 adapter (FC 0351) installed
- ▶ FICON channel as a driver

---

<sup>11</sup> CKD data records are handled by using IBM Enhanced Count Key Data (ECKD) command set.



- ▶ Only ECKD supported
- ▶ z/VM is not supported

Up to 16 zHyperLink Express2.0 adapters can be installed in an IBM z17 (up to 32 links).

The zHyperLink Express2.0 is virtualized as a native PCIe adapter and can be shared by multiple LPARs. Each port can support up to 127 Virtual Functions (VFs), with one or more VFs/PFIDs being assigned to each LPAR. This configuration gives a maximum of 254 VFs per adapter.

The zHyperLink Express requires the following components:

- ▶ zHyperLink connector on DS8k I/O Bay

For DS8880 firmware R8.3 or newer, the I/O Bay planar is updated to support the zHyperLink interface. This update includes the update of the PEX 8732 switch to PEX8733 that includes a DMA engine for the zHyperLink transfers, and the upgrade from a copper to optical interface by a CXP connector (provided).

- ▶ Cable

The zHyperLink Express2.0 uses optical cable with MTP connector. Maximum supported cable length is 150 meters (492 feet).

### 4.6.3 Network connectivity

Communication for LANs is provided by the OSA-Express7S (1.2) and Network Express Adapters (only available on z17). All these features are installed in a PCIe+ Drawer.

#### OSA-Express7S features

##### ***OSA-Express7S 1.2 25 GbE SR (FC 0459)***

OSA-Express7S 1.2 25 Gigabit Ethernet Short Reach (SR) feature includes one PCIe Gen3 adapter and one port per feature. The port supports CHPID type OSD.

The OSA-Express7S 1.2 25 GbE SR feature supports attachment to a multimode fiber 25 Gbps Ethernet LAN or Ethernet switch that is capable of 25 Gbps. The port can be defined as a spanned channel and shared among LPARs within and across logical channel subsystems.

The OSA-Express7S 1.2 25 GbE SR feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device includes an SR transceiver. The transceivers at both ends must be the same (SR to SR).

The OSA-Express7S 1.2 25 GbE SR feature does *not* support auto-negotiation to any other speed and runs in full duplex mode only.

A 50 µm multimode fiber optic cable that ends with an LC Duplex connector is required for connecting this feature to the selected device.

##### ***OSA-Express7S 1.2 25 GbE LR (FC 0460)***

The OSA-Express7S 1.2 25 Gigabit Ethernet (GbE) Long Reach (LR) feature includes one PCIe Gen3 adapter and one port per feature. The port supports CHPID types OSD.

The OSA-Express7S 1.2 25 GbE LR feature supports attachment to a single-mode fiber 25 Gbps Ethernet LAN or Ethernet switch that is capable of 25 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.



The OSA-Express7S 1.2 25 GbE LR feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device includes an LR transceiver. The transceivers at both ends must be the same (LR to LR).

The OSA-Express7S 1.2 25 GbE LR feature does *not* support auto-negotiation to any other speed and runs in full duplex mode only.

A 9  $\mu$ m single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting this feature to the selected device.

### ***OSA-Express7S 1.2 10 GbE LR (FC 0456)***

The OSA-Express7S 1.2 10 Gigabit Ethernet (GbE) Long Reach (LR) feature includes one PCIe Gen3 adapter and one port per feature. The port supports CHPID type OSD. The 10 GbE feature is designed to support attachment to a single-mode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.

The OSA-Express7S 1.2 10 GbE LR feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device includes an LR transceiver. The transceivers at both ends must be the same (LR to LR).

The OSA-Express7S 1.2 10 GbE LR feature does *not* support auto-negotiation to any other speed and runs in full duplex mode only.

A 9  $\mu$ m single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting this feature to the selected device.

### ***OSA-Express7S 1.2 10 GbE SR (FC 0457)***

The OSA-Express7S 1.2 10 GbE Short Reach (SR) feature includes one PCIe Gen3 adapter and one port per feature. The port supports CHPID type OSD. The 10 GbE feature is designed to support attachment to a multimode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and shared among LPARs within and across logical channel subsystems.

The OSA-Express7S 1.2 10 GbE SR feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device includes an SR transceiver. The sending and receiving transceivers must be the same (SR to SR).

The OSA-Express7S 1.2 10 GbE SR feature does *not* support auto-negotiation to any other speed and runs in full duplex mode only.

A 50 or a 62.5  $\mu$ m multimode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

### ***OSA-Express7S 1.2 GbE LX (FC 0454)***

The OSA-Express7S 1.2 GbE LX feature includes one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID types OSD or OSC). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and can be shared among LPARs and across logical channel subsystems.

The OSA-Express7S 1.2 GbE LX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an LX transceiver. The sending and receiving transceivers must be the same (LX to LX).

A 9  $\mu$ m single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables



are being reused, a pair of Mode Conditioning Patch cables is required, with one cable for each end of the link.

### ***OSA-Express7S 1.2 GbE SX (FC 0455)***

The OSA-Express7S 1.2 GbE SX feature includes one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID types OSD and OSC). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and shared among LPARs and across logical channel subsystems.

The OSA-Express7S 1.2 GbE SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an SX transceiver. The sending and receiving transceivers must be the same (SX to SX).

A multi-mode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

### **OSA carry forward adapters from z15**

The following other OSA-Express7S features can be installed on IBM z17 servers (when carried forward from an IBM z15 upgrade):

- ▶ OSA-Express7S 10 Gigabit Ethernet LR, FC 0444
- ▶ OSA-Express7S 10 Gigabit Ethernet SR, FC 0445
- ▶ OSA-Express7S Gigabit Ethernet LX, FC 0442
- ▶ OSA-Express7S Gigabit Ethernet SX, FC 0443
- ▶ OSA-Express7S 1000BASE-T Ethernet, FC 0446 - (OSD only)

The supported OSA-Express7S features are listed in Table 4-6 on page 184.

### ***OSA-Express7S 10 Gigabit Ethernet LR (FC 0444) - (Carry forward from z15)***

The OSA-Express7S 10 Gigabit Ethernet (GbE) Long Reach (LR) feature includes one PCIe Gen3 adapter and one port per feature. The port supports CHPID types OSD. The 10 GbE feature is designed to support attachment to a single-mode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.

The OSA-Express7S 10 GbE LR feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device includes an LR transceiver. The transceivers at both ends must be the same (LR to LR).

The OSA-Express7S 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 9 µm single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting this feature to the selected device.

### ***OSA-Express7S 10 Gigabit Ethernet SR (FC 0445) - (Carry forward from z15)***

The OSA-Express7S 10 GbE Short Reach (SR) feature includes one PCIe Gen3 adapter and one port per feature. The port supports CHPID types OSD. The 10 GbE feature is designed to support attachment to a multimode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and shared among LPARs within and across logical channel subsystems.

The OSA-Express7S 10 GbE SR feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device has an SR transceiver. The sending and receiving transceivers must be the same (SR to SR).



The OSA-Express7S 10 GbE SR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 50 or a 62.5  $\mu\text{m}$  multimode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

#### ***OSA-Express7S Gigabit Ethernet LX (FC 0442) - (Carry forward from z15)***

The OSA-Express7S GbE LX feature includes one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID types OSD and OSC). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and can be shared among LPARs and across logical channel subsystems.

The OSA-Express7S GbE LX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device includes an LX transceiver. The sending and receiving transceivers must be the same (LX to LX).

A 9  $\mu\text{m}$  single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables are being reused, a pair of Mode Conditioning Patch cables is required, with one cable for each end of the link.

#### ***OSA-Express7S Gigabit Ethernet SX (FC 0443) - (Carry forward from z15)***

The OSA-Express7S GbE SX feature includes one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID types OSD and OSC). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and shared among LPARs and across logical channel subsystems.

The OSA-Express7S GbE SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an SX transceiver. The sending and receiving transceivers must be the same (SX to SX).

A multi-mode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

#### ***OSA-Express7S 1000BASE-T Ethernet (FC 0446) - (Carry forward from z15)***

This adapter occupies one slot in the PCIe+ I/O drawer and features two ports that connect to a 1000 Mbps (1 Gbps) Ethernet LAN. Each port has an SFP+ with an RJ-45 receptacle for cabling to an Ethernet switch. The RJ-45 receptacle is required to be attached by using an EIA/TIA Category 5 or Category 6 UTP cable with a maximum length of 100 meters (328 feet). The SFP allows a concurrent repair or replace action.

The OSA-Express7S 1000BASE-T Ethernet feature does not support auto-negotiation. It supports links at 1000 Mbps in full duplex mode only.

**Note:** On IBM z17, the OSA-Express7S 1000BASE-T Ethernet feature can no longer be configured as CHPID type OSC and is *OSD only*.

On IBM z17 CHPID type OSC is only supported on the OSA-Express7S 1.2 GbE SX/LX features.

### **Network Express features**

The Network Express was introduced with IBM z17 and enables converging the legacy OSA and ROCE into one hardware feature.

A single port on the Network Express feature can simultaneously have two functionalities:



- ▶ OSH channel, which is a new OSH CHPID type for OSA-style I/O, which will support all legacy functions available with OSD, but uses the EQDIO architecture while OSD uses QDIO. This includes providing support for the network interface of a single OS instance to operate in promiscuous mode.
- ▶ NETH PFID for SMC-R RDMA or Linux native usage.

Each port on the adapter can be configured to provide support for a single host protocol (EQDIO or native PCIe) or a combination of both and can be independently (de)configured. If there is a requirement for legacy QDIO architecture (CHPID type OSD) OSA-Express7S 1.2 adapters should be configured.

### ***Network Express SR 25G (FC 0526)***

The Network Express SR (Short Reach) 25G feature includes one PCIe Gen4 adapter and two ports per feature. The port supports CHPID type OSH and PFID NETH as described above.

The Network Express SR 25G feature supports attachment to a single-mode fiber 25 Gbps Ethernet LAN or Ethernet switch that is capable of 25 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.

The Network Express SR 25G feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device includes an SR transceiver. The transceivers at both ends must be the same (SR to SR).

The Network Express SR 25G feature does *not* support auto-negotiation to any other speed and runs in full duplex mode only.

A 9  $\mu$ m single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting this feature to the selected device.

### ***Network Express LR 25G (FC 0527)***

The Network Express LR (Long Reach) 25G feature includes one PCIe Gen4 adapter and two ports per feature. The port supports CHPID type OSH and PFID NETH as described above.

The Network Express LR 25G feature supports attachment to a single-mode fiber 25 Gbps Ethernet LAN or Ethernet switch that is capable of 25 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.

The Network Express LR 25G feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device includes an LR transceiver. The transceivers at both ends must be the same (LR to LR).

The Network Express LR 25G feature does *not* support auto-negotiation to any other speed and runs in full duplex mode only.

A 9  $\mu$ m single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting this feature to the selected device.

### ***Network Express SR 10G (FC 0524)***

The Network Express SR (Short Reach) 10G feature includes one PCIe Gen4 adapter and two ports per feature. The port supports CHPID type OSH and PFID NETH as described above.

The Network Express SR 10G feature supports attachment to a single-mode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.



The Network Express SR 10G feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device includes an SR transceiver. The transceivers at both ends must be the same (SR to SR).

The Network Express SR 10G feature does *not* support auto-negotiation to any other speed and runs in full duplex mode only.

A 9  $\mu$ m single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting this feature to the selected device.

### ***Network Express LR 10G (FC 0525)***

The Network Express LR (Long Reach) 10G feature includes one PCIe Gen4 adapter and two ports per feature. The port supports CHPID type OSH and PFID NETH as described above.

The Network Express LR 10G feature supports attachment to a single-mode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.

The Network Express LR 10G feature supports the use of an industry standard small form factor (SFP+) LC Duplex connector. Ensure that the attaching or downstream device includes an LR transceiver. The transceivers at both ends must be the same (LR to LR).

The Network Express LR 10G feature does *not* support auto-negotiation to any other speed and runs in full duplex mode only.

A 9  $\mu$ m single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting this feature to the selected device.

**Note:** Consider the following Network Express adapter requirements:

- ▶ The Network Express feature does not support auto-negotiation to any other than their designate speed (either 10Gb or 25Gb) and runs in full duplex mode only.
- ▶ 10 GbE/25 GbE Network Express features should not be mixed in a z/OS SMC-R Link Group.

## **OSA-Express and Network Express cable specifications**

The OSA-Express and Network Express feature for cabling (cable type, connector type, maximum unrepeat distance, and the link rate) on an IBM z17 are listed in Appendix D, “Channel options” on page 547.

## **Shared Memory Communications functions**

The Shared Memory Communication (SMC) capabilities of the IBM z17 help optimize the communications between applications for server-to-server (SMC-R) or LPAR-to-LPAR (SMC-D) connectivity.

### ***Shared Memory Communications Version 1***

The following versions of SMC are available:

- ▶ SMC-R

SMC-R provides application transparent use of the Network Express feature. This feature reduces the network overhead and latency of data transfers, which effectively offers the benefits of optimized network performance across processors.



► SMC-D

SMC-D was used with the introduction of the Internal Shared Memory (ISM) virtual PCI function. ISM is a virtual PCI network adapter that enables direct access to shared virtual memory, which provides a highly optimized network interconnect for IBM Z intra-CPC communications.

SMC-D maintains the socket-API transparency aspect of SMC-R so that applications that use TCP/IP communications can benefit immediately without requiring any application software or IP topology changes. SMC-D completes the overall SMC solution, which provides synergy with SMC-R.

SMC-R and SMC-D use shared memory architectural concepts, which eliminate the TCP/IP processing in the data path, yet preserves TCP/IP Qualities of Service for connection management purposes.

### ***Internal Shared Memory***

ISM is a function that is supported by IBM z17, IBM z16, and IBM z15 systems. It provides connectivity by using shared memory access between multiple operating system images within the same CPC. ISM creates virtual adapters with shared memory that is allocated for each operating system image.

ISM is defined by the FUNCTION statement with a virtual CHPID (VCHID) in HCD/IOCDs. Identified by the PNETID parameter, each ISM VCHID defines an isolated, internal virtual network for SMC-D communication, without any hardware component required. Virtual adapters are defined by virtual function (VF) statements. Multiple LPARs can access the same virtual network for SMC-D data exchange by associating their VF with same VCHID.

Applications that use HiperSockets can realize network latency and CPU reduction benefits and performance improvement by using the SMC-D over ISM.

IBM z17 servers support up to 32 ISM VCHIDs per CPC. Each VCHID supports up to 255 VFs, with a total maximum of 8,000 VFs.

### ***Shared Memory Communications Version 2***

Shared Memory Communications v2 is available in z/OS V2R4 (with PTFs) and later.

The initial version of SMC was limited to TCP/IP connections over the same layer 2 network; therefore, it was not routable across multiple IP subnets. The associated TCP/IP connection was limited to hosts within a single IP subnet that requires the hosts to have direct access to the same physical layer 2 network (that is, the same Ethernet LAN over a single VLAN ID). The scope of eligible TCP/IP connections for SMC was limited to and defined by the single IP subnet.

SMC Version 2 (SMCv2) provides support for SMC over multiple IP subnets for SMC-D and SMC-R and is referred to as SMC-Dv2 and SMC-Rv2. SMCv2 requires updates to the underlying network technology. SMC-Dv2 requires ISMv2 and SMC-Rv2 requires RoCEv2.

The SMCv2 protocol is downward compatible, which allows SMCv2 hosts to continue to communicate with SMCv1 down-level hosts.

Although SMCv2 changes the SMC connection protocol enabling multiple IP subnet support, SMCv2 does not change how user TCP socket data is transferred, which preserves the benefits of SMC to TCP workloads.

TCP/IP connections that require IPsec are not eligible for any form of SMC.



## HiperSockets

The HiperSockets function of IBM z17 servers provides up to 32 high-speed virtual LAN attachments.

HiperSockets can be customized to accommodate varying traffic sizes. Because HiperSockets does not use an external network, it can free up system and network resources. This advantage can help eliminate attachment costs and improve availability and performance.

HiperSockets eliminates the need to use I/O subsystem operations and traverse an external network connection to communicate between LPARs in the same IBM z17 CPC. HiperSockets offers significant value in server consolidation when connecting many virtual servers. It can be used instead of certain coupling link configurations in a Parallel Sysplex.

HiperSockets internal networks support the following transport modes:

- ▶ Layer 2 (link layer)
- ▶ Layer 3 (network or IP layer)

Traffic can be IPv4 or IPv6, or non-IP, such as AppleTalk, DECnet, IPX, NetBIOS, or SNA.

HiperSockets devices are protocol-independent and Layer 3-independent. Each HiperSockets device (Layer 2 and Layer 3 mode) features its own Media Access Control (MAC) address. This address allows the use of applications that depend on the existence of Layer 2 addresses, such as Dynamic Host Configuration Protocol (DHCP) servers and firewalls.

Layer 2 support helps facilitate server consolidation, and can reduce complexity and simplify network configuration. It also allows LAN administrators to maintain the mainframe network environment similarly to non mainframe environments.

Packet forwarding decisions are based on Layer 2 information instead of Layer 3. The HiperSockets device can run automatic MAC address generation to create uniqueness within and across LPARs and servers. The use of Group MAC addresses for multicast is supported, and broadcasts to all other Layer 2 devices on the same HiperSockets networks.

Datagrams are delivered only between HiperSockets devices that use the same transport mode. A Layer 2 device cannot communicate directly to a Layer 3 device in another LPAR network. A HiperSockets device can filter inbound datagrams by VLAN identification, the destination MAC address, or both.

Analogous to the Layer 3 functions, HiperSockets Layer 2 devices can be configured as primary or secondary connectors, or multicast routers. This configuration enables the creation of high-performance and high-availability link layer switches between the internal HiperSockets network and an external Ethernet network. It also can be used to connect to the HiperSockets Layer 2 networks of different servers.

HiperSockets Layer 2 is supported by Linux on IBM Z, and by z/VM for Linux guest use.

IBM z17 supports the HiperSockets Completion Queue function that is designed to allow HiperSockets to transfer data synchronously (if possible) and asynchronously, if necessary. This feature combines ultra-low latency with more tolerance for traffic peaks.



With the asynchronous support, data can be temporarily held until the receiver has buffers that are available in its inbound queue during high volume situations. The HiperSockets Completion Queue function requires the following *minimum* OS support releases<sup>12</sup>:

- ▶ z/OS V2.4 with PTFs
- ▶ Linux Minimum distributions:
  - SUSE SLES 16.1 (Post GA)
  - SUSE SLES 15.6 (GA)
  - SUSE SLES 12.5 (Post GA)
  - Red Hat RHEL 10.0 (Post GA)
  - Red Hat RHEL 9.4
  - Red Hat RHEL 8.10
  - Red Hat RHEL 7.9 (Post GA)
  - Canonical Ubuntu 24.04 LTS (Post GA)
  - Canonical Ubuntu 22.04 LTS (Post GA)
  - Canonical Ubuntu 20.04 LTS (Post GA)
- ▶ VSE<sup>n</sup> V6.3.1 (21<sup>st</sup> Century Software)
- ▶ z/VM V7.3 with maintenance

The z/VM virtual switch function transparently bridges a guest virtual machine network connection on a HiperSockets LAN segment. This bridge allows a single HiperSockets guest virtual machine network connection to communicate directly with the following systems:

- ▶ Other guest virtual machines on the virtual switch
- ▶ External network hosts through the virtual switch OSA UPLINK port

#### 4.6.4 Parallel Sysplex connectivity

Coupling links are required in a Parallel Sysplex configuration to provide connectivity from the z/OS images to the coupling facility (CF). A properly configured Parallel Sysplex provides a highly reliable, redundant, and robust IBM Z technology solution to achieve near-continuous availability. A Parallel Sysplex is composed of one or more z/OS operating system images that are coupled through one or more Coupling Facilities (CFs).

This section describes coupling link features supported in a Parallel Sysplex in which an IBM z17 can participate.

##### Coupling links

The type of coupling link that is used to connect a CF to an operating system LPAR is important. The link performance significantly affects response times and coupling processor usage. For configurations that extend over large distances, the time that is spent on the link can be the largest part of the response time.

IBM z16, and IBM z15 support the following coupling link types:

- ▶ Integrated Coupling Adapter Short Reach (ICA SR1.1 and ICA SR) links connect directly to the CPC drawer and are intended for short distances between CPCs of up to 150 meters (492 feet).
- ▶ Coupling Express2 Long Reach (CE2 LR) adapters for IBM z16 and Coupling Express Long Reach (CE LR) are in the PCIe+ drawer and support unrepeated distances of up to 10 km (6.2 miles) or up to 100 km (62.1 miles) over qualified WDM services.
- ▶ Internal Coupling (IC) links are for internal links within a CPC.

<sup>12</sup> Minimum OS support for IBM z17 can differ. For more information, see Chapter 7, “Operating systems support” on page 261.



IBM z17 supports the following coupling link types:

- ▶ Integrated Coupling Adapter Short Reach 2.0 (ICA SR2.0) links connect directly to the CPC drawer and are intended for short distances between CPCs of up to 150 meters (492 feet).
- ▶ Coupling Express3 Long Reach (CE3 LR) adapters for IBM z17 are in the PCIe+ drawer and support unrepeated distances of up to 10 km (6.2 miles) or up to 100 km (62.1 miles) over qualified DWDM services.

**Note:** Parallel Sysplex supports connectivity between systems that differ by up to two generations (n-2). For example, an IBM z17 can participate in an IBM Parallel Sysplex cluster with IBM z16, and IBM z15 systems.

Only Integrated Coupling Adapter Short Reach 2.0 (ICA SR2.0) Feature Code 0216 and Coupling Express3 Long Reach (CE3 LR) Feature Codes 0498 and 0499 are supported on IBM z17.

Figure 4-5 shows the supported Coupling Link connections for the IBM z17. Only ICA SR and CE LR links are supported on IBM z17, IBM z16, and IBM z15 systems.

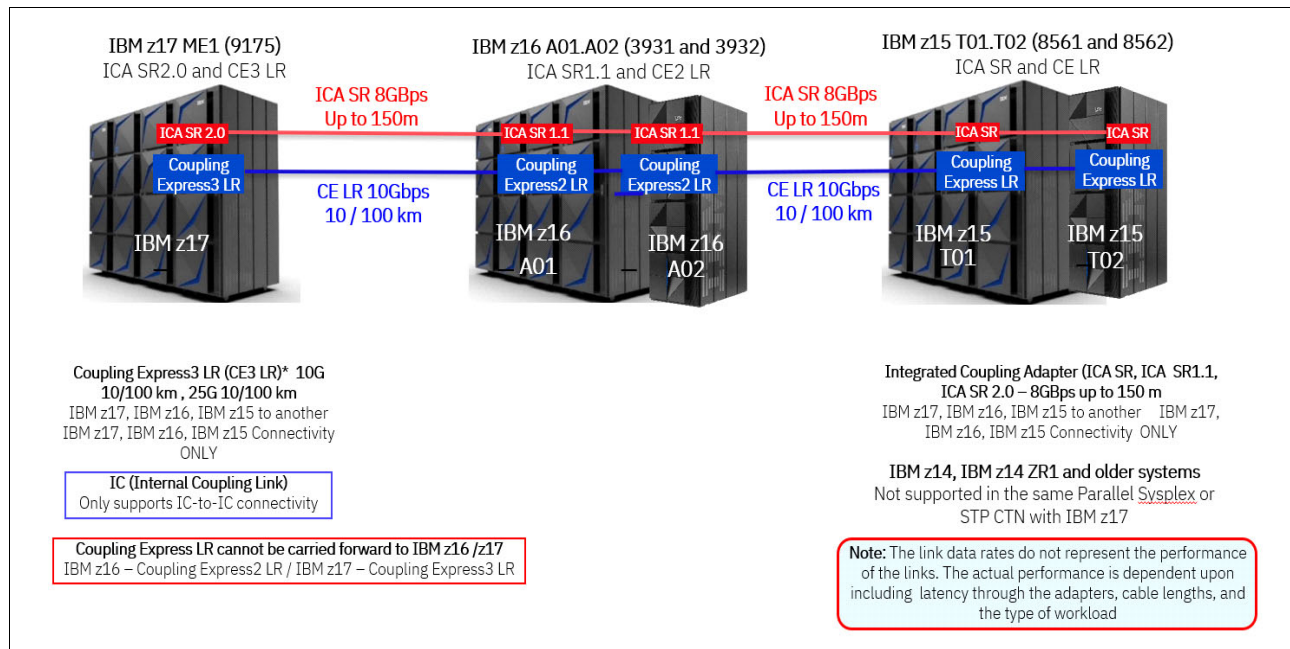


Figure 4-5 Parallel Sysplex connectivity options



The coupling link options are listed in Table 4-10. Also listed is the coupling link support for each IBM Z platform. Restrictions on the maximum numbers can apply, depending on the configuration. Always check with your IBM support team for more information.

Table 4-7 Coupling link (Coupling Express Long-Reach- CE LR) options

Channel feature	Feature Code	Link rate	Max unrepeat- ed distance	Maximum number of supported links		
				IBM z17	IBM z16	IBM z15
Coupling Express3 LR 10 GB (CL5)	0498	10 Gbps	10 km (6.2 miles)	128	N/A	N/A
Coupling Express3 LR 25 GB (CL6)	0499	25 Gbps	10 km (6.2 miles)	128	N/A	N/A
Coupling Express2 LR	0434	10 Gbps	10 km (6.2 miles)	N/A	64	N/A
Coupling Express LR	0433	10 Gbps	10 km (6.2 miles)	N/A	N/A	64
ICA SR2.0	0216	8 GBps	150 meters (492 feet)	96	N/A	N/A
ICA SR1.1	0176	8 GBps	150 meters (492 feet)	N/A	96	96
ICA SR	0172	8 GBps	150 meters (492 feet)	N/A	80	80
IC	N/A	Internal speeds	N/A	64	64	64

The maximum number of combined external coupling links (active CE LR, ICA SR links) is 160 per IBM z17 ME1 system. IBM z17 systems support up to 384 coupling CHPIDs per CPC. An IBM z17 coupling link support summary is shown in Table 4-7.

Consider the following points:

- ▶ The maximum supported links depends on the IBM Z model or capacity feature code.
- ▶ IBM z17 ICA SR2.0 maximum depends on the number of CPC drawers. A total of 12 PCIe+ fan-outs are used per CPU drawer, which gives a maximum of 24 ICA SR ports. IBM z17 supports up to 96 ICA SR2.0 ports.

For more information about distance support for coupling links, see *IBM Z End-to-End Extended Distance Guide*, [SG24-8047](#).

## Internal Coupling link

IC links are Licensed Internal Code-defined links that are used to connect a CF to a z/OS logical partition in the same CPC. These links are available on all IBM Z platforms. The IC link is an IBM Z coupling connectivity option that enables high-speed, efficient communication between a CF partition and one or more z/OS logical partitions that are running on the same CPC. The IC is a linkless connection (implemented in LIC) and does not require any hardware or cabling.

An IC link is a fast coupling link that uses memory-to-memory data transfers. IC links do not have PCHID numbers, but do require CHPIDs.



IC links have the following attributes:

- ▶ They provide the fastest connectivity that is significantly faster than external link alternatives.
- ▶ They result in better coupling efficiency than with external links, which effectively reduces the CPU cost that is associated with Parallel Sysplex.
- ▶ They can be used in test or production configurations, reduce the cost of moving into Parallel Sysplex technology, and enhance performance and reliability.
- ▶ They can be defined as spanned channels across multiple channel subsystems.
- ▶ They are available at no extra hardware cost (no feature code). Employing ICFs with IC links results in considerable cost savings when configuring a cluster.

IC links are enabled by defining CHPID type ICP. A maximum of 64 IC links can be defined on an IBM z16.

## Integrated Coupling Adapter Short Reach - ICA SR2.0

IBM z17 introduces ICA SR2.0.

ICA SR2.0 are two-port, short-distance coupling features that allow the supported IBM Z to connect to each other. ICA SR2.0 use coupling channel type: CS6.

The ICA SR2.0 uses PCIe Gen4 technology, with x16 lanes that are bifurcated into x8 lanes for coupling. The ICA SR2.0 is designed to drive distances up to 150 m (492 feet) and supports a link data rate of 8 GBps. It is designed to support up to four CHPIDs per port and eight subchannels (devices) per CHPID. It allows the supported IBM Z to connect to each other over extended distance. The ICA SR2.0 (FCs 0216) is a two-port card that uses coupling channel type CS5.

**Note:** Previous versions of the ICA SR (FC 0172), introduced with the IBM z13, ICA SR1.1 (FC 0176) introduced with IBM z15 are not available on IBM z17.

For more information, see *IBM Z Planning for Fiber Optic Links (FICON/FCP, Coupling Links, and Open System Adapters)*, GA23-1409. This publication is available in [the Library section of Resource Link](#) (log-in required).

## Coupling Express3 Long Reach

The Coupling Express3LR (FC 0498, 10GB and FC 0499, 25GB) occupies one slot in an IBM z17 PCIe+ I/O drawer<sup>13</sup>. It allows the supported IBM Z to connect to each other over extended distance. The Coupling Express3LR (FCs 0498) is a two-port card that uses coupling channel type CL5. The Coupling Express3LR (FCs 0499) is a two-port card that uses coupling channel type CL6.

The Coupling Express3 LR designed to drive distances up to 10 km (6.21 miles) unrepeated and support a link data rate of 10 Gbps (FC 0489) or 25 Gbps (FC 0499). For distance requirements greater than 10 km (6.21 miles), clients must use a Dense Wavelength Division Multiplexer (DWDM). The DWDM vendor must be qualified by IBM Z.

Coupling Express3 LR is designed to support up to four CHPIDs per port, 32 buffers (that is, 32 subchannels) per CHPID. The Coupling Express3 LR feature installs in the PCIe+ I/O drawer on IBM z17.

<sup>13</sup> PCIe+ I/O drawer FC 4011 on IBM z17 (and FC 4023 IBM z16, and FC 4021 on IBM z15) is installed in a 19-inch rack. PCIe+ I/O Drawers contains and can host up to 16 PCIe I/O features (adapters). FCs 4023 and 4011 are not carried forward to IBM z17.



For more information, see *IBM Z Planning for Fiber Optic Links (FICON/FCP, Coupling Links, Open Systems Adapters, and zHyperLink Express)*, GA23-1409. This publication is available in [the Library section of Resource Link](#) (log-in required).

### Coupling link cable specifications

For more details on the cable specifications for these features, please refer to Appendix D, “Channel options” on page 547.

### Extended distance support

For more information about extended distance support, see *System z End-to-End Extended Distance Guide*, [SG24-8047](#).

### Migration considerations

Upgrading from previous generations of IBM Z systems in a Parallel Sysplex to IBM z17 servers in that same Parallel Sysplex requires proper planning for coupling connectivity. Planning is important because of the change in the supported type of coupling link adapters and the number of available fan-out slots of the IBM z17 CPC drawers.

The ICA SR fan-out features provide short-distance connectivity to another IBM z17, IBM z16, or IBM z15 server.

The CE LR adapter provides long-distance connectivity to another IBM z17, IBM z16, or IBM z15 server.

#### Notes:

- ▶ The new ICA SR2.0 in IBM z17 adapter is fully compatible with ICA SR and ICA SR1.1 in IBM z15 and IBM z16
- ▶ Only Coupling Express3 LR 10Gb (CHPID type CL5) is fully compatible with Coupling Express LR and Coupling Express2 LR in IBM z15 and IBM z16.
- ▶ The Coupling Express3 LR 25Gb (CHPID type CL6) *is not* compatible with previous generation of CE LR features and can not be used to connect to IBM z15 or IBM z16

The IBM z17 fan-out slots in the CPC drawer provide coupling link connectivity through the ICA SR fan-out cards. In addition to coupling links for Parallel Sysplex, the fan-out cards provide connectivity for the PCIe+ I/O drawer (PCIe+ Gen4 fan-out).

Up to 12 PCIe fan-out cards can be installed in an IBM z17 CPC drawer.

To migrate from an older generation machine to an IBM z17 without disruption in a Parallel Sysplex environment requires that the older machines are no more than n-2 generation (namely, at least IBM z15) and that they carry enough coupling links to connect to the existing systems while also connecting to the new machine. This requirement is necessary to allow individual components (z/OS LPARs and CFs) to be shut down and moved to the target machine and continue to be connected to the remaining systems.

It is beyond the scope of this book to describe all possible migration scenarios. Always consult with subject matter experts to help you to develop your migration strategy.

### Coupling links and Server Time Protocol

All external coupling links can be used to pass time synchronization signals by using Server Time Protocol (STP). STP is a message-based protocol in which timing messages are passed over data links between servers. The same coupling links can be used to exchange time and CF messages in a Parallel Sysplex.



The use of the coupling links to exchange STP messages has the following advantages:

- ▶ By using the same links to exchange STP messages and CF messages in a Parallel Sysplex, STP can scale with distance. Servers that are exchanging messages over short distances (ICA SR links), can meet more stringent synchronization requirements than servers that exchange messages over long distance (CE3 LR links), with distances up to 100 kilometers (62 miles)<sup>14</sup>. This advantage is an enhancement over the IBM Sysplex Timer implementation, which does not scale with distance.
- ▶ Coupling links provide the connectivity that is necessary in a Parallel Sysplex. Therefore, a potential benefit can be realized of minimizing the number of cross-site links that is required in a multi-site Parallel Sysplex.

Between any two servers that are intended to exchange STP messages, configure each server so that at least two coupling links exist for communication between the servers. This configuration prevents the loss of one link from causing the loss of STP communication between the servers. If a server does not have a CF LPAR, timing-only links can be used to provide STP connectivity.

### IBM z17 Precision Time Protocol support

Precision Time Protocol (PTP) is introduced as an alternative to NTP. Consider the following points:

- ▶ PTP provides more accurate timestamps to connected devices
- ▶ Initially used for Power Distribution Systems, Telecommunications, and Laboratories
- ▶ Requires Customer Network Infrastructure to be PTP capable
- ▶ IBM z17 provides PTP connectivity direct to the CPC

## 4.7 Cryptographic functions

Cryptographic functions are provided by the CP Assist for Cryptographic Function (CPACF) and the PCI Express cryptographic adapters. IBM z17 servers support the Crypto Express8S, and as carry forward, Crypto Express7S.

### 4.7.1 CPACF functions (FC 3863)

FC 3863 is required to enable Cryptographic functions (subject to export regulations).

### 4.7.2 Crypto Express features

The following two generations of the Peripheral Component Interconnect Express (PCIe) cryptographic coprocessors, which are optional features, are available on the IBM z17:

- ▶ Crypto Express8S feature (FC 0908 and FC 0909) - New Build & Carry forward
- ▶ Crypto Express7S feature (FC 0898 and FC 0899) - Carry forward only

The Crypto Express8S represents the newest generation that was introduced with the IBM z16, *the rest of the paragraph is applicable for both generations (7S and 8S)*.

These coprocessors are Hardware Security Modules (HSMs) that provide high-security cryptographic processing as required by banking and other industries. These features provide a secure programming and hardware environment wherein crypto processes are performed.

---

<sup>14</sup> 10 km (6.2 miles) without DWDM extenders; 100 km (62 miles) with certified DWDM equipment.



Each cryptographic coprocessor includes general-purpose processors, nonvolatile storage, and specialized cryptographic electronics, which are all contained within a tamper-sensing and tamper-responsive enclosure that eliminates all keys and sensitive data on any attempt to tamper with the device. The security features of the HSM are designed to meet the requirements of FIPS 140-2, Level 4, which is the highest-defined security level.

A Crypto Express (2 HSM) feature includes two IBM PCIe Cryptographic Co-processors (PCIeCC); a Crypto Express (1 HSM) feature includes one PCIeCC per feature. For availability reasons, a minimum of two features is required. Up to 30 Crypto Express (2 HSM) features are supported. The maximum number of the 1 HSM features is 16. A Crypto Express feature occupies one I/O slot in a PCIe+ I/O drawer.

Each adapter can be configured as a Secure IBM CCA coprocessor, a Secure IBM Enterprise PKCS #11 (EP11) coprocessor, or as an accelerator.

These Crypto Express features provide domain support for up to 85 logical partitions.

The accelerator function is designed for maximum-speed Secure Sockets Layer and Transport Layer Security (SSL/TLS) acceleration, rather than for specialized financial applications for secure, long-term storage of keys or secrets. The Crypto Express feature can also be configured as one of the following configurations:

- ▶ The Secure IBM CCA coprocessor includes secure key functions with emphasis on the specialized functions that are required for banking and payment card systems. It is optionally programmable to add custom functions and algorithms by using User Defined Extensions (UDX).

Payment Card Industry (PCI) PIN<sup>®</sup> Transaction Security (PTS) Hardware Security Module (HSM) (PCI-HSM), is available for Crypto Express7S and newer in CCA mode. PCI-HSM mode simplifies compliance with PCI requirements for hardware security modules.

- ▶ The Secure IBM Enterprise PKCS #11 (EP11) coprocessor implements an industry-standardized set of services that adheres to the PKCS #11 specification v2.20 and more recent amendments. It was designed for extended FIPS and Common Criteria evaluations to meet industry requirements

This cryptographic coprocessor mode introduced the PKCS #11 secure key function.

**TKE feature:** The Trusted Key Entry (TKE) Workstation feature is required for supporting the administration of the Crypto Express7S and Crypto Express8S when configured as an Enterprise PKCS #11 coprocessor or managing the CCA mode PCI-HSM.

When the Crypto Express PCI Express adapter is configured as a secure IBM CCA co-processor, it still provides accelerator functions. However, up to 3x better performance for those functions can be achieved if the Crypto Express PCI Express adapter is configured as an accelerator.

CCA enhancements include the ability to use triple-length (192-bit) Triple-DES (TDES) keys for operations, such as data encryption, IBM PIN processing, and key wrapping to strengthen security. CCA also extended the support for the cryptographic requirements of the German Banking Industry Committee Deutsche Kreditwirtschaft (DK).

Several features that support the use of the AES algorithm in banking applications also were added to CCA. These features include the addition of AES-related key management features and the AES ISO Format 4 (ISO-4) PIN blocks as defined in the ISO 9564-1 standard. PIN block translation is supported and use of AES PIN blocks in other CCA callable services.



IBM continues to add enhancements as AES finance industry standards are released.

More about the cryptographic capabilities of the IBM z17 can be found in Chapter 6, “Cryptographic features” on page 221.

### 4.7.3 IBM Fibre Channel Endpoint Security

IBM z17 Model ME1 supports IBM Fibre Channel Endpoint Security Enablement feature (FC 1146). FC 1146 provides Fibre Channel Endpoint Authentication and Encryption of Data in Flight for FICON and FC connections to IBM DS8000 storage systems. It is an end-to-end solution that helps ensure all data flowing on the Fibre Channel links within and across data centers flows between trusted entities.

IBM Fibre Channel Endpoint Security is designed to provide a means to help ensure the integrity and confidentiality of all data that flows on FC links between trusted entities within and across data centers. The trusted entities are IBM z17 and the IBM Storage subsystem (select IBM DS8000 storage systems). No application or middleware changes are required. Fibre Channel Endpoint Security supports all data in-flight from any operating system.

IBM Z Feature Code 1146, Endpoint Security Enablement turns on the Fibre Channel Endpoint Security panels on the HMC so setup can be done.

Based tightly on the Fibre Channel–Security Protocol-2 (FC-SP-2) standard, which provides various means of authentication and essentially maps IKEv2 constructs for security association management and derivation of encryption keys to Fibre Channel Extended Link Services, the IBM Fibre Channel Endpoint Security implementation uses the IBM solution for key server infrastructure in the storage system (for data at-rest encryption).

IBM Security Guardium Key Lifecycle Manager acts as a trusted authority for key generation operations and as an authentication server. It provides shared secret key generation in a relationship between an FC initiator (IBM Z server) and the IBM Storage target. The solution implements an authentication and key management solution that is called IBM Secure Key Exchange (SKE), as illustrated in Figure 4-6 on page 207.



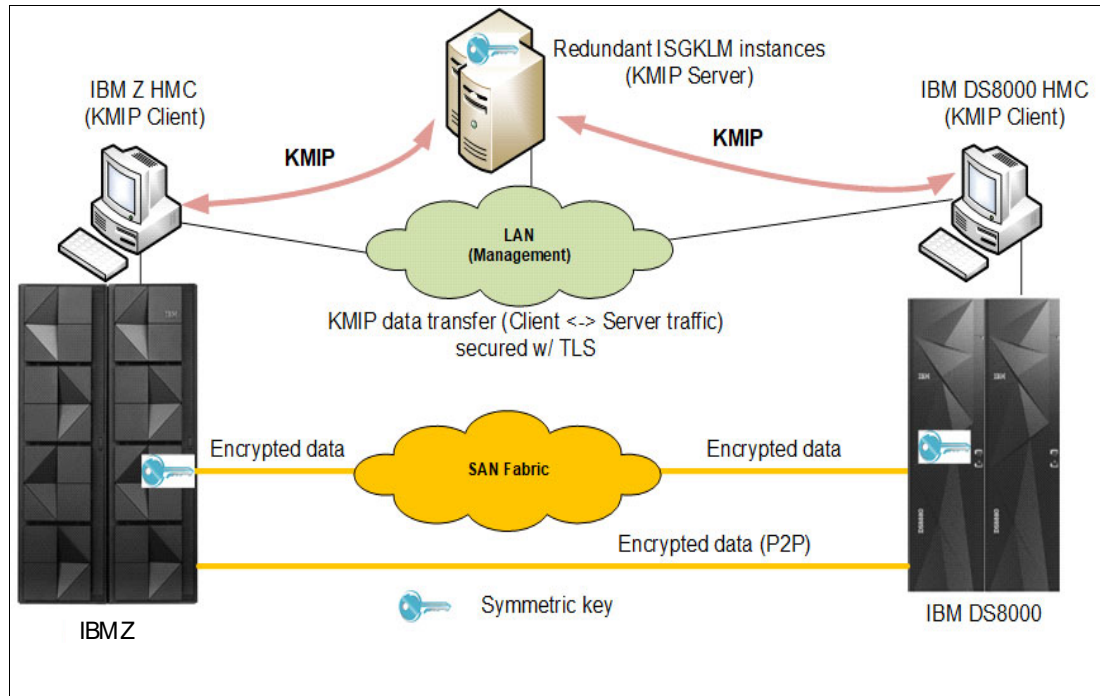


Figure 4-6 Fibre Channel Endpoint Security

Before establishing the connection, each link must be authenticated, and if successful, then it becomes a trusted connection. A policy sets the rules, for example, enforcing trusted connections only. If the link goes down, the authentication process starts again. The secure connection can be enabled automatically if both the IBM Z and IBM Storage endpoints are encryption-capable.

Data in-flight (from and to IBM Z and IBM Storage servers) is encrypted when it leaves either endpoint (source), and then it is decrypted at the destination. Encryption and decryption are done at the FC adapter level. The operating system that is running on the host (IBM Z server) is not involved in Fibre Channel Endpoint Security related operations. Tools are provided at the operating system level for displaying information about the encryption status.

The following prerequisites must be met for this optional priced feature:

- ▶ FICON Express32-4P LX/SX (FC0387/FC0388), FICON Express32S LX/SX (FC0461/FC0462) or FICON Express16SA LX/SX (FC0436/0437) for both link encryption and endpoint authentication
- ▶ DS8910, DS8890 (Power 9 based) or newer with 32GFC (encryption) Host Bus Adapter or 16GFC (authentication-only) Host Bus Adapter
- ▶ Key server: IBM Security Guardium Key Lifecycle Manager (ISGKLM)  
Latest version: 4.2, minimum: IBM Security Key Lifecycle Manager V3.0.1
- ▶ IBM z17 Endpoint Security Enablement: FC 1146
- ▶ CPACF enablement (FC 3863)

For more information and implementation details, see *IBM Fibre Channel Endpoint Security for IBM DS8900F and IBM Z*, SG24-8455 and *IBM DS8000 Encryption for Data at Rest, Transparent Cloud Tiering, and Endpoint Security (DS8000 Release 10.0)*, REDP-4500.

On April 23, 2024 IBM published a Statement of Direction (AD24-0457):



“

Given the increasing importance of providing the highest level of data protection to IBM Z clients, IBM intends to require the use of IBM Fibre Channel Endpoint Security for all FICON connected devices starting with the release of IBM zNext+1. This direction will require investment by IBM Infrastructure teams, FICON storage vendors and IBM Z clients as an important step towards continuing to secure the most mission critical workloads. In support of this direction, all new FICON-connected storage systems introduced after December 31, 2024, will be required to support IFCES to connect to zNext+1.

## 4.8 Integrated Firmware Processor

The Integrated Firmware Processor (IFP) was introduced with the zEC12 and zBC12 servers. The IFP is used for managing PCIe native features.

The Coupling Express3 Long Reach (CE LR) is installed in the PCIe+ I/O drawer and is the native PCIe feature in the IBM z17 and belongs to a Resource Group managed by the IFP.

The native PCIe features should be ordered in pairs for redundancy. The features are assigned to one of the four resource groups (RGs) that are running on the IFP according to their physical location in the PCIe+ I/O drawer, which provides management functions and virtualization functions.

If two features of the same type are installed, one always is managed by resource group 1 (RG 1) or resource group 3 (RG3); the other feature is managed by resource group 2 (RG 2) or resource group 4 (RG 4). This configuration provides redundancy if one of the features or resource groups needs maintenance or fails.

The IFP and RGs support the following infrastructure management functions:

- ▶ Firmware update of adapters and resource groups
- ▶ Error recovery and failure data collection
- ▶ Diagnostic and maintenance tasks





## Central processor complex channel subsystem

This chapter describes the concepts of the IBM z17 ME1 channel subsystem, including multiple channel subsystems and multiple subchannel sets. It also describes the technology, terminology, and implementation aspects of the channel subsystem.

This chapter includes the following topics:

- ▶ 5.1, “Channel subsystem” on page 210
- ▶ 5.2, “I/O configuration management” on page 218
- ▶ 5.3, “Channel subsystem summary” on page 219
- ▶ 5.4, “IBM z17 Data Processing Unit (DPU)” on page 219



## 5.1 Channel subsystem

*Channel subsystem* (CSS) is a collective name of facilities that IBM Z servers use to control I/O operations.

The channel subsystem directs the flow of information between I/O devices and main storage. It allows data processing to proceed concurrently with I/O processing, which relieves data processors (central processor (CP) and Integrated Facility for Linux [IFL]) of the task of communicating directly with I/O devices.

The channel subsystem includes subchannels, I/O devices that are attached through control units, and channel paths between the subsystem and control units. For more information about the channel subsystem, see 5.1.1, “Multiple logical channel subsystems”.

The design of IBM Z servers offers considerable processing power, memory size, and I/O connectivity. In support of the larger I/O capability, the CSS structure is scaled up by introducing the multiple logical channel subsystem (LCSS) since IBM z990, and multiple subchannel sets (MSS) since IBM z9.

An overview of the channel subsystem for IBM z17 servers is shown in Figure 5-1. IBM z17 ME1 systems are designed to support up to six logical channel subsystems, each with four subchannel sets and up to 256 channels.

IBM z17 ME1					
HSA fixed at 884 GB					
LCSS 0 Up to 15 Logical Partitions	LCSS 1 Up to 15 Logical Partitions	LCSS 2 Up to 15 Logical Partitions	LCSS 3 Up to 15 Logical Partitions	LCSS 4 Up to 15 Logical Partitions	LCSS 5 Up to 10 Logical Partitions
Subchannel Sets: SS 0 – 63.75 k SS 1 – 64 k SS 2 – 64 k SS 3 – 64 k	Subchannel Sets: SS 0 – 63.75 k SS 1 – 64 k SS 2 – 64 k SS 3 – 64 k	Subchannel Sets: SS 0 – 63.75 k SS 1 – 64 k SS 2 – 64 k SS 3 – 64 k	Subchannel Sets: SS 0 – 63.75 k SS 1 – 64 k SS 2 – 64 k SS 3 – 64 k	Subchannel Sets: SS 0 – 63.75 k SS 1 – 64 k SS 2 – 64 k SS 3 – 64 k	Subchannel Sets: SS 0 – 63.75 k SS 1 – 64 k SS 2 – 64 k SS 3 – 64 k
Up to 256 Channels	Up to 256 Channels	Up to 256 Channels	Up to 256 Channels	Up to 256 Channels	Up to 256 Channels

Figure 5-1 Multiple channel subsystem and multiple subchannel sets

All channel subsystems are defined within a single configuration, which is called *I/O configuration data set* (IOCDS). The IOCDS is loaded into the hardware system area (HSA) during a central processor complex (CPC) power-on reset (POR) to start all of the channel subsystems.

On IBM z17 A01 systems, the HSA is pre-allocated in memory with a fixed size of 884 GB, which is in addition to the customer purchased memory. This fixed size memory for HSA eliminates the requirement for more planning of the initial I/O configuration and planning for future I/O expansions.

**CPC drawer repair:** The HSA can be moved from one CPC drawer to a different drawer in an enhanced availability configuration as part of a concurrent CPC drawer repair (CDR) action.

The following objects are always reserved in the IBM z17 ME1 HSA during POR, regardless of whether they are defined in the IOCDS for use:

- Six LCSSs



- ▶ A total of 15 LPARs in each LCSS0 to LCSS4
- ▶ A total of 10 LPARs in LCSS5
- ▶ Subchannel set 0 with 63.75 K devices in each LCSS
- ▶ Subchannel set 1 with 64 K minus one device in each LCSS
- ▶ Subchannel set 2 with 64 K minus one device in each LCSS
- ▶ Subchannel set 3 with 64 K minus one device in each LCSS

### 5.1.1 Multiple logical channel subsystems

In the z/Architecture, a single channel subsystem can have up to 256 channel paths that are defined, which limited the total numbers of I/O connectivity on older IBM Z servers to 256.

The introduction of multiple LCSSs enabled an IBM Z server to have more than one channel subsystems logically, while each logical channel subsystem maintains the same manner of I/O processing. Also, a logical partition (LPAR) is now attached to a specific logical channel subsystem, which makes the extension of multiple logical channel subsystems not apparent to the operating systems and applications. The multiple image facility (MIF) in the structure enables resource sharing across LPARs within a single LCSS or across the LCSSs.

The multiple LCSS structure extended the IBM Z servers' total number of I/O connectivity to support a balanced configuration for the growth of processor and I/O capabilities.

A one-digit number ID starting from 0 (CSSID) is assigned to an LCSS, and a one-digit hexadecimal ID (MIF ID) starting from 0 is assigned to an LPAR within the LCSS.

**Note:** The phrase *channel subsystem* has same meaning as *logical channel subsystem* in this section, unless otherwise stated.

#### Subchannels

A *subchannel* provides the logical appearance of a device to the program and contains the information that is required for sustaining a single I/O operation. Each device is accessible by using one subchannel in a channel subsystem to which it is assigned according to the active IOCDs of the IBM Z server.

A *subchannel set* (SS) is a collection of subchannels within a channel subsystem. The maximum number of subchannels of a subchannel set determines how many devices are accessible to a channel subsystem.

In z/Architecture, the first subchannel set of an LCSS can have 63.75 K subchannels (with 0.25 K reserved), with a subchannel set ID (SSID) of 0. By enabling the multiple subchannel sets, extra subchannel sets are available to increase the device addressability of a channel subsystem.

For more information about multiple subchannel sets, see 5.1.2, “Multiple subchannel sets” on page 212.

#### Channel paths

A *channel path* provides a connection between the channel subsystem and control units that allows the channel subsystem to communicate with I/O devices. Depending on the type of connections, a channel path might be a physical connection to a control unit with I/O devices, such as FICON, or an internal logical control unit, such as HiperSockets.

Each channel path in a channel subsystem features a unique 2-digit hexadecimal identifier that is known as a *channel-path identifier* (CHPID), which ranges 00 - FF. Therefore, a total of



256 CHPIDs are supported by a CSS, and a maximum of 1536 CHPIDs are available on an IBM z17 with six logical channel subsystems.

By assigning a CHPID to a physical port of an I/O feature adapter, such as FICON Express32S, or a fan-out adapter (ICA SR) port, the channel subsystem connects to the I/O devices through these physical ports.

A port on an I/O feature card features a unique physical channel identifier (PCHID) according to the physical location of this I/O feature adapter, and the sequence of this port on the adapter.

In addition, a port on a fan-out adapter has a unique adapter identifier (AID), according to the physical location of this fan-out adapter, and the sequence of this port on the adapter.

A CHPID is assigned to a physical port by defining the corresponding PCHID or AID in the I/O configuration definitions.

### Control units

A *control unit* provides the logical capabilities that are necessary to operate and control an I/O device. It adapts the characteristics of each device so that it can respond to the standard form of control that is provided by the CSS.

A control unit can be housed separately or can be physically and logically integrated with the I/O device, channel subsystem, or within the IBM Z server.

### I/O devices

An *I/O device* provides external storage, a means of communication between data-processing systems, or a means of communication between a system and its environment. In the simplest case, an I/O device is attached to one control unit and is accessible through one or more channel paths that are connected to the control unit.

## 5.1.2 Multiple subchannel sets

A *subchannel set* is a collection of subchannels within a channel subsystem. The maximum number of subchannels of a subchannel set determines how many I/O devices that a channel subsystem can access. This number also determines the number of addressable devices to the program (for example, an operating system) that is running in the LPAR.

Each subchannel has a unique four-digit hexadecimal number 0x0000 - 0xFFFF. Therefore, a single subchannel set can address and access up to 64 K I/O devices.

The IBM z17 ME1 systems support four subchannel sets for each logical channel subsystem. The IBM z17 ME1 system can access a maximum of 255.74 K devices for a logical channel subsystem and a logical partition and the programs that are running on it.

**Note:** Do not confuse the multiple subchannel sets function with multiple channel subsystems.

### Subchannel number

The *subchannel number* is a four-digit hexadecimal number 0x0000 - 0xFFFF, which is assigned to a subchannel within a subchannel set of a channel subsystem. Subchannels in each subchannel set are always assigned subchannel numbers within a single range of contiguous numbers.



The lowest-numbered subchannel is subchannel 0; the highest-numbered subchannel includes a subchannel number equal to one less than the maximum numbers of subchannels that are supported by the subchannel set. Therefore, a subchannel number is always unique within a subchannel set of a channel subsystem and depends on the sequence of assigning.

With the subchannel numbers, a program that is running on an LPAR (for example, an operating system) can specify all I/O functions relative to a specific I/O device by designating a subchannel that is assigned to the I/O devices.

Normally, subchannel numbers are used only in communication between the programs and the channel subsystem.

### **Subchannel set identifier**

While introducing the MSS, the channel subsystem is extended to assign a value 0 - 3 for each subchannel set, which is the SSID. A subchannel can be identified by its SSID and subchannel number.

### **Device number**

A device number is an arbitrary number 0x0000 - 0xFFFF, which is defined by a system programmer in an I/O configuration for naming an I/O device. The device number must be unique within a subchannel set of a channel subsystem. It is assigned to the corresponding subchannel by channel subsystem when an I/O configuration is activated. Therefore, a subchannel in a subchannel set of a channel subsystem includes a device number together with a subchannel number for designating an I/O operation.

The device number provides a means to identify a device that is independent of any limitations that are imposed by the system model, configuration, or channel-path protocols.

A device number also can be used to designate an I/O function to a specific I/O device. Because it is an arbitrary number, it can easily be fit into any configuration management and operating management scenarios. For example, a system administrator can set all of the z/OS systems in an environment to device number 1000 for their system RES volumes.

With multiple subchannel sets, a subchannel is assigned to a specific I/O device by the channel subsystem with an automatically assigned subchannel number and a device number that is defined by user. An I/O device can always be identified by an SSID with a subchannel number or a device number. For example, a device with device number AB00 of subchannel set 1 can be designated as 1AB00.

Normally, the subchannel number is used by the programs to communicate with the channel subsystem and I/O device, whereas the device number is used by a system programmer, operator, and administrator.

### **Device in subchannel set 0 and extra subchannel sets**

An LCSS always includes the first subchannel set (SSID 0), which can have up to 63.75 K subchannels with 256 subchannels that are reserved by the channel subsystem. Users can always define their I/O devices in this subchannel set for general use.

For the extra subchannel sets enabled by the MSS facility, each has 65535 subchannels (64 K minus one) for specific types of devices. These extra subchannel sets are referred as *alternative subchannel sets* in z/OS.

Also, a device that is defined in an alternative subchannel set is considered a special device, which normally features a special device type in the I/O configuration.



Currently, an IBM z17 ME1 system that is running z/OS defines the following types of devices in another subchannel set, with proper APAR or PTF installed:

- ▶ Alias devices of the parallel access volumes (PAV).
- ▶ Secondary devices of GDPS Metro Mirror Copy Service (formerly Peer-to-Peer Remote Copy [PPRC]).
- ▶ IBM FlashCopy® SOURCE and TARGET devices with program temporary fix (PTF) OA46900.
- ▶ Db2 data backup volumes with APAR OA24142.

The use of another subchannel set for these special devices helps reduce the number of devices in the subchannel set 0, which increases the growth capability for accessing more devices.

### **Initial program load from an alternative subchannel set**

IBM z17 ME1 systems support initial program load (IPL) from alternative subchannel sets in addition to subchannel set 0. Devices that are used early during IPL processing now can be accessed by using subchannel set 1, subchannel set 2, or subchannel set3 on an IBM z17.

This configuration allows the users of Metro Mirror (formerly PPRC) secondary devices that are defined by using the same device number and a new device type in an alternative subchannel set to be used for IPL, an I/O definition file (IODF), and stand-alone memory dump volumes, when needed.



## The display ios,config command

The z/OS **display ios,config(all)** command that is shown in Figure 5-2 includes information about the MSSs.

```
D IOS,CONFIG(ALL)
IOS506I 14.26.53 I/O CONFIG DATA 606
ACTIVE IODF DATA SET = SYS9.IODF63
CONFIGURATION ID = ITS0          EDT ID = 01
TOKEN:  PROCESSOR DATE      TIME      DESCRIPTION
SOURCE: PAVO      25-02-17 11:35:59 SYS9      IODF63
ACTIVE CSS: 3      SUBCHANNEL SETS CONFIGURED: 0, 1, 2, 3
CHANNEL MEASUREMENT BLOCK FACILITY IS ACTIVE
SUBCHANNEL SET FOR PPRC PRIMARY: INITIAL = 0    ACTIVE = 0
HYPERSWAP FAILOVER HAS OCCURRED: NO
LOCAL SYSTEM NAME (LSYSTEM): PAVO
HARDWARE SYSTEM AREA AVAILABLE FOR CONFIGURATION CHANGES
PHYSICAL CONTROL UNITS          8061
CSS 0 - LOGICAL CONTROL UNITS    4027
  SS 0  SUBCHANNELS              59510
  SS 1  SUBCHANNELS              65535
  SS 2  SUBCHANNELS              65535
  SS 3  SUBCHANNELS              65535
CSS 1 - LOGICAL CONTROL UNITS    4005
  SS 0  SUBCHANNELS              59218
  SS 1  SUBCHANNELS              65535
  SS 2  SUBCHANNELS              65535
  SS 3  SUBCHANNELS              65535
CSS 2 - LOGICAL CONTROL UNITS    4025
  SS 0  SUBCHANNELS              59410
  SS 1  SUBCHANNELS              65535
  SS 2  SUBCHANNELS              65535
  SS 3  SUBCHANNELS              65535
CSS 3 - LOGICAL CONTROL UNITS    4026
  SS 0  SUBCHANNELS              60906
  SS 1  SUBCHANNELS              65535
  SS 2  SUBCHANNELS              65535
  SS 3  SUBCHANNELS              65535
CSS 4 - LOGICAL CONTROL UNITS    4043
  SS 0  SUBCHANNELS              61266
  SS 1  SUBCHANNELS              65535
  SS 2  SUBCHANNELS              65535
  SS 3  SUBCHANNELS              65535
CSS 5 - LOGICAL CONTROL UNITS    4088
  SS 0  SUBCHANNELS              65280
  SS 1  SUBCHANNELS              65535
  SS 2  SUBCHANNELS              65535
  SS 3  SUBCHANNELS              65535
ELIGIBLE DEVICE TABLE LATCH COUNTS
      0 OUTSTANDING BINDS ON PRIMARY EDT
```

Figure 5-2 Output for display ios,config(all) command with MSS



### 5.1.3 Channel path spanning

With the implementation of multiple LCSSs, a channel path can be available to LPARs as dedicated, shared, and spanned.

Although a shared channel path can be shared by LPARs within a same LCSS, a spanned channel path can be shared by LPARs within and across LCSSs.

By assigning the same CHPID from different LCSSs to the same channel path (for example, a PCHID), the channel path can be accessed by any LPARs from these LCSSs at the same time. The CHPID is spanned across those LCSSs. The use of spanned channels paths decreases the number of channels that are needed in an installation of IBM Z servers.

A sample of channel paths that are defined as dedicated, shared, and spanned is shown in Figure 5-3.

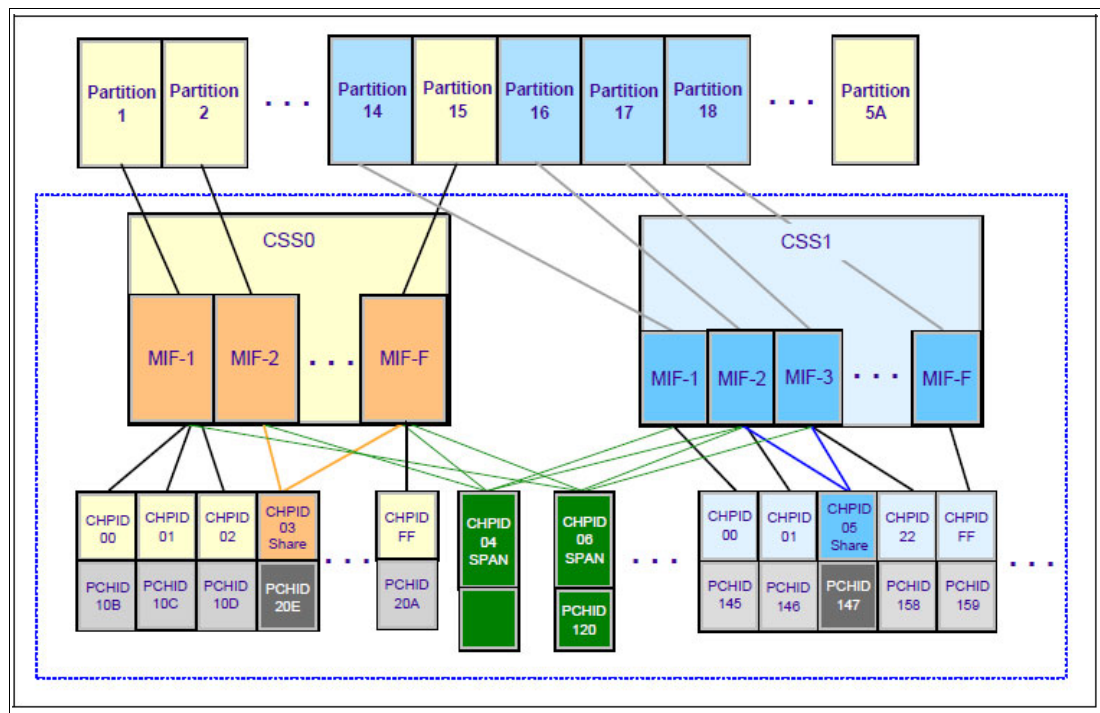


Figure 5-3 IBM Z CSS: Channel subsystems with channel spanning

The following definitions of a channel path are shown in Figure 5-3:

- ▶ CHPID FF, assigned to PCHID 20A, is dedicated access for partition 15 of LCSS0. The same applies to CHPID 00,01,02 of LCSS0, and CHPID 00,01,FF of LCSS1.
- ▶ CHPID 03, assigned to PCHID 20E, is shared access for partition 2, and 15 of LCSS0. The same applies to CHPID 05 of LCSS1.
- ▶ CHPID 06, assigned to PCHID 120 is spanned access for partition 1, 15 of LCSS0, and partition 16, 17 of LCSS1. The same applies to CHPID 04.

Channel spanning is supported for internal links (HiperSockets and IC links) and for specific types of external links. External links that are supported on IBM z17 ME1 systems include FICON Express 32G, FICON Express32S, FICON Express16SA, OSA-Express7S 1.2, OSA-Express7S, Network Express, and Coupling Links.



The definition of LPAR name, MIF image ID, and LPAR ID are used to identify an LPAR by the channel subsystem to identify I/O functions from different LPARs of multiple LCSSs, which support the implementation of these dedicated, shared, and spanned paths.

An example of definition of these LPAR-related identifications is shown in Figure 5-4.

CSS0			CSS1			CSS2		CSS3		CSS4		CSS5		Specified in HCD / IOCP
Logical Partition Name			Logical Partition Name			LPAR Name		LPAR Name		LPAR Name		LPAR Name		
TST1	PROD1	PROD2	TST2	PROD3	PROD4	TST3	TST4	PROD5	PROD6	TST5	PROD7	PROD8	TST6	Specified in Image Profile
Logical Partition ID			Logical Partition ID			LPAR ID		LPAR ID		LPAR ID		LPAR ID		
02	04	0A	14	16	1D	22	26	35	3A	44	47	56	5A	Specified in HCD / IOCP
MIF ID			MIF ID			MIF ID		MIF ID		MIF ID		MIF ID		
2	4	A	4	6	D	2	6	5	A	4	7	6	A	Specified in HCD / IOCP

Figure 5-4 CSS, LPAR, and identifier example

## LPAR name

The LPAR name is defined as partition name parameter in the RESOURCE statement of an I/O configuration. The LPAR name must be unique across the server.

## MIF image ID

The MIF image ID is defined as a parameter for each LPAR in the RESOURCE statement of an I/O configuration. It ranges 1 - F, and must be unique within an LCSS. However, duplicates are allowed in different LCSSs.

If a MIF image ID is not defined, an arbitrary ID is assigned when the I/O configuration activated. The IBM z17 server supports a total of 85 LPARs that can be defined.

Each LCSS of an IBM z17 ME1 system can support the following numbers of LPARs:

- ▶ LCSS0 to LCSS4 support 15 LPARs each, and the MIF image ID is 1 - F.
- ▶ LCSS5 supports 10 LPARs, and the MIF image IDs are 1 - A.

## LPAR ID

The LPAR ID is defined by a user in an image activation profile for each LPAR. It is a 2-digit hexadecimal number 00 - 7F. The LPAR ID must be unique across the server.

Although it is arbitrarily defined by the user, an LPAR ID often is the CSS ID concatenated to its MIF image ID, which makes the value more meaningful for the system administrator. For example, an LPAR with LPAR ID 1A defined in that manner means that the LPAR is defined in LCSS1, with the MIF image ID A.



## 5.2 I/O configuration management

The following tools are available to help maintain and optimize the I/O configuration:

- ▶ IBM Configurator for e-business (eConfig)

The eConfig tool is used by your IBM representative. It is used to create configurations or upgrades of a configuration, and maintains tracking to the installed features of those configurations. eConfig produces reports that help you understand the changes that are being made for a new system, or a system upgrade, and what the target configuration looks like.

- ▶ Hardware configuration definition (HCD)

HCD supplies an interactive dialog to generate the IODF, and later the IOCDs. Generally, use HCD or Hardware Configuration Manager (HCM) to generate the I/O configuration rather than writing I/O configuration program (IOCP) statements. The validation checking that HCD runs against a IODF source file helps minimize the risk of errors before an I/O configuration is activated.

HCD support for multiple channel subsystems is available with z/VM and z/OS. HCD provides the capability to make dynamic hardware and software I/O configuration changes.

**Note:** Specific functions might require specific levels of an operating system, PTFs, or both.

Consult the suitable fix categories:

- IBM z17 ME1:IBM.Device.Server.IBM z17-9175
- IBM z16 A01:IBM.Device.Server.IBMz16-3931
- IBM z16 A02:IBM.Device.Server.IBMz16-3932
- IBM z15 T01: IBM.Device.Server.IBM z15-8561
- IBM z15 T02: IBM.Device.Server.IBM z15-8562

- ▶ HCM

HCM is a priced optional feature that supplies a graphical interface of HCD. It is installed on a PC and allows you to manage the physical and logical aspects of a mainframe's hardware configuration.

- ▶ CHPID Mapping Tool (CMT)

The CMT helps to map CHPIDs onto PCHIDs that are based on an IODF source file and the eConfig configuration file of a mainframe. It provides a CHPID to PCHID mapping with high availability for the targeted I/O configuration. It also features built-in mechanisms to generate a mapping according to customized I/O performance groups. More enhancements are implemented in CMT to support IBM z17 servers.

The CMT is available for download from the [IBM Resource Link](#) (log-in required).

The configuration file for a new machine or upgrade is also available from the [IBM Resource Link](#) (log-in required). Contact your IBM technical sales representative for the name of the file to download.



## 5.3 Channel subsystem summary

IBM z17 ME1 systems support the channel subsystem features of multiple LCSS, MSS, and the channel spanning that is described in this chapter. The channel subsystem capabilities of IBM z17 ME1 systems are listed in Table 5-1.

*Table 5-1 IBM z17 ME1 CSS overview*

<b>Maximum number of CSSs</b>	6
<b>Maximum number of LPARs per CSS</b>	CSS0 - CSS4: 15 CSS5: 10
<b>Maximum number of LPARs per system</b>	85
<b>Maximum number of subchannel sets per CSS</b>	4
<b>Maximum number of subchannels per CSS</b>	255.74 K SS0: 65280 SS1 - SS3: 65535
<b>Maximum number of CHPIDs per CSS</b>	256

## 5.4 IBM z17 Data Processing Unit (DPU)

The I/O firmware stack has evolved on Application Specific Integrated Circuits (ASICs) platforms over the past 30 years. In IBM z17 system, we are moving that specialised I/O ASIC functionality, across the PCI bus, and into the mainframe processor chip – much like the on-chip AIU engine in z16, and the on-chip compression acceleration engine in z15. Figure 5-5 on page 220 shows the implementation of the new IBM z17 I/O infrastructure.

This will allow IBM z17 and future servers to:

- ▶ Provide better I/O Performance/Latency/RAS
- ▶ Use higher I/O density (4-port FICON cards and converged Network Express adapter)
- ▶ Be agile in delivery of new IO feature function every generation (not just when new I/O adapters are announced).



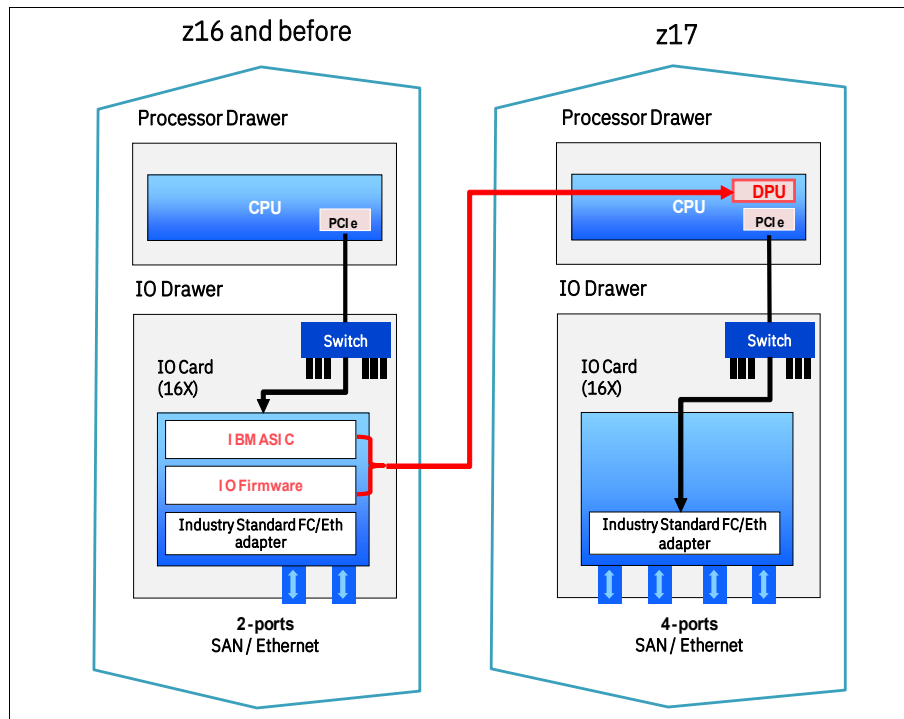


Figure 5-5 IBM z17 I/O Engine Infrastructure

IBM z17 DPU encompasses a comprehensive refactoring of the I/O subsystem. The goals of this effort are to deliver improved IBM Z platform efficiencies, to improve peak I/O start rates and reduce latencies, to provide focused per port recovery for the most common types of failures, to improve recurring networking costs for customers by providing integrated RoCE SMC-R and OSA support, to provide single port serviceability for all DPU managed I/O adapters, and to reduce dependence on the PCI support partition by providing physical function support for PCIe Native use cases.

The following protocols are supported and will run on the DPU:

- ▶ Legacy Mode FICON
- ▶ HPF (High Performance FICON)
- ▶ FCP (SCSI over fiber channel)
- ▶ OSA (Open Systems Adapter)
- ▶ OSA-ICC (Open Systems Adapter - Integrated Console Controller)
- ▶ Physical function support for Native Ethernet exploitation.

This support also allows a port to be shared between a PCIe Native protocol and OSA.dndn



## 6



# Cryptographic features

This chapter describes the hardware cryptographic functions that are available on IBM z17. The CP Assist for Cryptographic Function (CPACF), together with the IBM Peripheral Component Interconnect Express (PCIe) cryptographic coprocessors (PCIeCC), offer a balanced use of processing resources and unmatched scalability for fulfilling pervasive encryption demands.

This chapter also introduces the principles of cryptography and describes the implementation of cryptography in the hardware and software architecture of IBM Z. It also describes the features that IBM z17 offers.

Finally, the cryptographic features and required software are summarized.

This chapter includes the following topics:

- ▶ 6.1, “Cryptography enhancements on IBM z17” on page 222
- ▶ 6.2, “Cryptography overview” on page 223
- ▶ 6.3, “Cryptography on IBM z17” on page 227
- ▶ 6.4, “CP Assist for Cryptographic Functions” on page 231
- ▶ 6.5, “Crypto Express8S” on page 236
- ▶ 6.6, “Trusted Key Entry workstation” on page 253
- ▶ 6.7, “Cryptographic functions comparison” on page 256
- ▶ 6.8, “Cryptographic operating system support for IBM z17” on page 258



## 6.1 Cryptography enhancements on IBM z17

IBM recognizes that with any new technology, new threats exist, and as such, suitable counter measures must be taken. Quantum technology can be used for incredible good, but in the hands of an adversary, it has the potential to weaken or break core cryptographic primitives that were used to secure systems and communications. Quantum-safe cryptography aims to provide protection against attacks that can be started by quantum computers.

The IBM z17 uses quantum-safe technologies to help protect your business-critical infrastructure and data from potential attacks.

The IBM z17 delivers a transparent and consumable approach that enables extensive (pervasive) encryption of data in flight and at rest, with the goal of substantially simplifying data security and reducing the costs that are associated with protecting data while achieving compliance mandates.

**Naming:** The IBM z17, Machine Type 9175 (M/T 9175), Model ME1 is further identified as *IBM z17*, unless otherwise specified.

IBM z16 introduced the new PCI Crypto Express8S feature (that can be managed by a new Trusted Key Entry (TKE) workstation) together with a further improved CPACF Coprocessor. In addition, the IBM Common Cryptographic Architecture (CCA) and the IBM Enterprise PKCS #11 (EP11) Licensed Internal Code (LIC) were enhanced.

The new features support new standards and meet the following compliance requirements:

- ▶ Payment Card Industry (PCI) Hardware Security Module (HSM) certification to strength the cryptographic standards for attack resistance in the payment card systems area.  
PCI HSM certification is available for Crypto Express8S, Crypto Express7S, and Crypto Express6S.
- ▶ National Institute of Standards and Technology (NIST) through the Federal Information Processing Standard (FIPS) standard to implement guidance requirements.
- ▶ Common Criteria EP11 EAL4.
- ▶ German Banking Industry Commission (GBIC).
- ▶ Visa Format Preserving Encryption (VFPE) for credit card numbers.
- ▶ Enhanced public key Elliptic Curve Cryptography (ECC) for users such as Chrome, Firefox, and Apple's iMessage.
- ▶ Accredited Standards Committee X9 Inc Technical Report-34 (ASC X9 TR-34)

For the HSM, IBM z17 uses a new adapter released as a break in replacement to the current design. The new adapter fix end of life issues on multiple components and provides a fix for the battery power circuit. No firmware update will be included and the adapter is backward compatible to the current design. The Crypto Express8S adapter is used by IBM z17 with the following enhancements:

- COP: HMAC support / acceleration via hardware
- SHA3, SHAKE improved ICV, OCV handling
- COP: KM-XTS performance improvement
- True Random Number Generator (TRNG) - Entropy Speed-up and Enhancements

All IBM z16 and IBM z17 enhancements are described in this chapter.



IBM z16 and IBM z17 include standard cryptographic hardware and optional cryptographic features for flexibility and growth capability. IBM has a long history of providing hardware cryptographic solutions. This history stretches from the development of the Data Encryption Standard (DES) in the 1970s to the Crypto Express tamper-sensing and tamper-responding programmable features.

Crypto Express meets the US Government's highest security rating of FIPS 140-3 Level 4 certification<sup>1</sup>. It also meets several other security ratings, such as the Common Criteria for Information Technology Security Evaluation, the PCI HSM criteria, and the criteria for German Banking Industry Commission (formerly known as Deutsche Kreditwirtschaft evaluation).

The cryptographic functions include the full range of cryptographic operations that are necessary for local and global business and financial institution applications. User Defined Extensions (UDX) allow you to add custom cryptographic functions to the functions that IBM z17 systems offer.

## 6.2 Cryptography overview

Throughout history, a need existed for secret communication between people that cannot be understood by outside parties.

Also, it is necessary to ensure that a message cannot be corrupted (message integrity), while ensuring that the sender and the receiver really are the persons who they claim to be. Over time, several methods were used to achieve these objectives, with more or less success. Many procedures and algorithms for encrypting and decrypting data were developed that are increasingly complicated and time-consuming.

### 6.2.1 Modern cryptography

With the development of computing technology, the encryption and decryption algorithms can be performed by computers, which enables the use of complicated mathematical algorithms. Most of these algorithms are based on the prime factorization of large numbers.

Cryptography is used to meet the following requirements:

- ▶ Data protection

The protection of data usually is the main concept that is associated with cryptography. Only authorized persons should be able to read the message or get information about it. Data is encrypted by using a known algorithm and secret keys, such that the intended party can de-scramble the data, but an interloper cannot. This concept is also referred to as *confidentiality*.

- ▶ Authentication (identity validation)

This process decides whether the communication partners are who they claim to be, which can be done by using certificates and signatures. It must be possible to clearly identify the owner of the data or the sender and the receiver of the message.

- ▶ Message (data) Integrity

The verification of data ensures that what was received is identical to what was sent. It must be proven that the data is complete and was not altered during the moment it was transmitted (by the sender) and the moment it was received (by the receiver).

- ▶ Nonrepudiation

---

<sup>1</sup> FIPS 140-3 certification - Security Requirements for Cryptographic Modules.



It must be impossible for the owner of the data or the sender of the message to deny authorship. Nonrepudiation ensures that both sides of a communication know that the other side agreed to what was exchanged, and not someone else. This specification implies a legal liability and contractual obligation, which is the same as a signature on a contract.

These goals should all be possible without unacceptable overhead to the communication. The goal is to keep the system secure, manageable, and productive.

The basic data protection method is achieved by encrypting and decrypting the data, while hash algorithms, message authentication codes (MACs), digital signatures, and certificates are used for authentication, data integrity, and nonrepudiation.

When encrypting a message, the sender transforms the clear text into a secret text. Doing so requires the following main elements:

- ▶ The *algorithm* is the mathematical or logical formula that is applied to the key and the clear text to deliver a ciphered result, or to take a ciphered text and deliver the original clear text.
- ▶ The *key* ensures that the result of the encrypting data transformation by the algorithm is only the same when the same key is used. That decryption of a ciphered message results only in the original clear message when the correct key is used. Therefore, the receiver of a ciphered message must know which algorithm and key must be used to decrypt the message.

## 6.2.2 Kerckhoffs' principle

In modern cryptography, the algorithm is published and known to everyone, whereas the keys are kept secret. This configuration corresponds to Kerckhoffs' principle, which is named after Auguste Kerckhoffs, a Dutch cryptographer, who formulated it in 1883:

"A system should not depend on secrecy, and it should be able to fall into the enemy's hands without disadvantage."

In other words, the security of a cryptographic system should depend on the security of the key, so the key must be kept secret. Therefore, the secure management of keys is the primal task of modern cryptographic systems.

Adhering to Kerckhoffs' Principle is done for the following reasons:

- ▶ It is much more difficult to keep an algorithm secret than a key.
- ▶ It is harder to exchange a compromised algorithm than to exchange a compromised key.
- ▶ Secret algorithms can be reconstructed by reverse engineering software or hardware implementations.
- ▶ Errors in public algorithms can generally be found more easily, when many experts examine it.
- ▶ In history, most secret encryption methods proved to be weak and inadequate.
- ▶ When a secret encryption method is used, it is possible that a back door was built in.
- ▶ If an algorithm is public, many experts can form an opinion about it. Also, the method can be more thoroughly investigated for potential weaknesses and vulnerabilities.



### 6.2.3 Keys

The keys that are used for the cryptographic algorithms often are sequences of numbers and characters, but can also be any other sequence of bits. The length of a key influences the security (strength) of the cryptographic method. The longer the used key, the more difficult it is to compromise a cryptographic algorithm.

For example, the DES (symmetric key) algorithm uses keys with a length of 56 bits, Triple-DES (TDES) uses keys with a length of 112 bits, and Advanced Encryption Standard (AES) uses keys of 128, 192, or 256 bits. The asymmetric key RSA algorithm (named after its inventors Rivest, Shamir, and Adleman) uses keys with a length of 1024, 2048, or 4096 bits.

In modern cryptography, keys must be kept secret. Depending on the effort that is made to protect the key, keys are classified into the following levels:

- ▶ A *clear key* is a key that is transferred from the application in clear text to the cryptographic function. The key value is stored in the clear (at least briefly) somewhere in unprotected memory areas. Therefore, the key can be made available to someone under specific circumstances who is accessing this memory area.

This risk must be considered when clear keys are used. However, many applications exist where this risk can be accepted. For example, the transaction security for the (widely used) encryption methods Secure Sockets Layer (SSL) and Transport Layer Security (TLS) is based on clear keys.

- ▶ The value of a *protected key* is stored only in clear in memory areas that cannot be read by applications or users. The key value does not exist outside of the physical hardware, although the hardware might not be tamper-resistant. The principle of protected keys is unique to IBM Z. For more information, see 6.4.2, “CPACF protected key” on page 234.
- ▶ For a *secure key*, the key value does not exist in clear format outside of a special hardware device (HSM), which must be secured and tamper-resistant. A secure key is protected from disclosure and misuse, and can be used for the trusted execution of cryptographic algorithms on highly sensitive data. If used and stored outside of the HSM, a secure key must be encrypted with a *master key*, which is created within the HSM and never leaves the HSM.

Because a secure key must be handled in a special hardware device, the use of secure keys usually is far slower than the use of clear keys, as shown in Figure 6-1 on page 226.



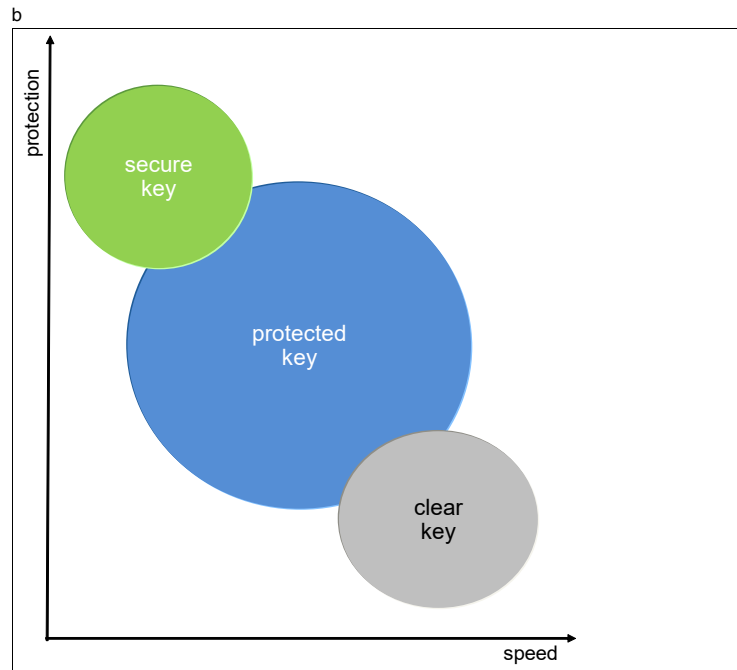


Figure 6-1 Three levels of protection with three levels of speed

## 6.2.4 Algorithms

The following algorithms of modern cryptography are differentiated based on whether they use the same key for the encryption of the message as for the decryption:

- *Symmetric algorithms* use the same key to encrypt and decrypt data. The function that is used to decrypt the data is the opposite of the function that is used to encrypt the data. Because the same key is used on both sides of an operation, it must be negotiated between both parties and kept secret. Therefore, symmetric algorithms are also known as *secret key algorithms*.

The main advantage of symmetric algorithms is that they are fast and therefore can be used for large amounts of data, even if they are not run on specialized hardware. The disadvantage is that the key must be known by both sender and receiver of the messages, which implies that the key must be exchanged between them. This key exchange is a weak point that can be attacked.

Prominent examples for symmetric algorithms are DES, TDES, and AES.

- *Asymmetric algorithms* use two distinct but related keys: the *public key* and the *private key*. As the names imply, the private key must be kept secret, whereas the public key is shown to everyone. However, with asymmetric cryptography, it is not important who sees or knows the public key. Whatever is done with one key can be undone by the other key only.

For example, data that is encrypted by the public key can be decrypted by the associated private key only, and vice versa. Unlike symmetric algorithms, which use distinct functions for encryption and decryption, only one function is used in asymmetric algorithms. Depending on the values that are passed to this function, it encrypts or decrypts the data. Asymmetric algorithms are also known as *public key algorithms*.

Asymmetric algorithms use complex calculations and are relatively slow (about 100 - 1000 times slower than symmetric algorithms). Therefore, such algorithms are not used for the encryption of bulk data.



Because the private key is never exchanged between the parties in communication, they are less vulnerable than symmetric algorithms. Asymmetric algorithms mainly are used for authentication, digital signatures, and for the encryption and exchange of secret keys, which in turn are used to encrypt bulk data with a symmetric algorithm.

Examples for asymmetric algorithms are RSA and the elliptic curve algorithms.

- *One-way algorithms* are not cryptographic functions. They do not use keys, and they can scramble data only, not de-scramble it. These algorithms are used extensively within cryptographic procedures for digital signing and tend to be developed and governed by using the same principles as cryptographic algorithms. One-way algorithms are also known as *hash algorithms*.

The most prominent one-way algorithms are the Secure Hash Algorithms (SHA).

## 6.3 Cryptography on IBM z17

In principle, cryptographic algorithms can run on processor hardware. However, these workloads are compute-intensive, and the handling of secure keys also requires special hardware protection. Therefore, IBM Z offer several cryptographic hardware features, which are specialized to meet the requirements for cryptographic workload.

The cryptographic hardware that is supported on IBM z17 is shown in Figure 6-2. These features are described in this chapter.

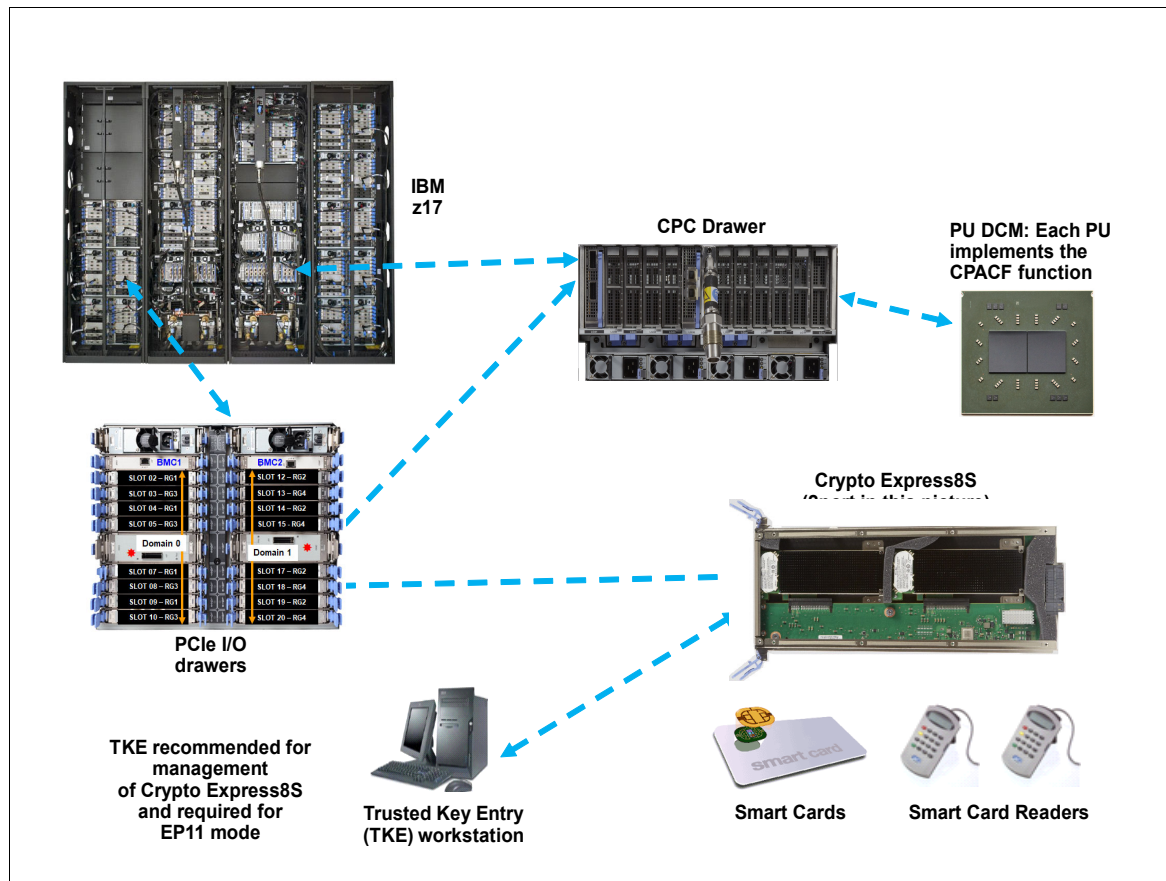


Figure 6-2 Cryptographic hardware that is supported in IBM z17



Implemented in every processor unit (PU) or core in a central processor complex (CPC) is a cryptographic coprocessor that can be used<sup>2</sup> for cryptographic algorithms that uses clear keys or protected keys. For more information, see 6.4, “CP Assist for Cryptographic Functions” on page 231.

The Crypto Express8S adapter is an HSM that is placed in the PCIe+ I/O drawer of IBM z17. It also supports cryptographic algorithms by using secret keys. For more information, see 6.5, “Crypto Express8S” on page 236.

Finally, a TKE workstation is required for entering keys in a secure way into the Crypto Express8S HSM, which often also is equipped with smart card readers. For more information, see 6.6, “Trusted Key Entry workstation” on page 253.

The feature codes and purpose of the cryptographic hardware features that are available for IBM z17 are listed in Table 6-1.

*Table 6-1 Cryptographic features for IBM z17 ME1*

Feature Code	Description
3863	CP Assist for Cryptographic Function (CPACF) enablement  This feature is a prerequisite to use CPACF (except for SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512) and the PCIe Crypto Express features.
0908	Crypto Express8S feature (dual HSM) <sup>a</sup>  These features are optional. The 2-port feature contains two IBM 4770 PCIe Cryptographic Coprocessors (HSMs), which can be independently defined as Coprocessor or Accelerator. New feature. Not supported on previous generations IBM Z systems.  A TKE, smart card Reader and latest available level smart cards are required to operate the Crypto adapter card in EP11 mode.
0909	Crypto Express8S feature (single HSM) <sup>a</sup>  These features are optional. The 2-port feature contains two IBM 4770 PCIe Cryptographic Coprocessors (HSMs), which can be independently defined as Coprocessor or Accelerator. New feature. Not supported on previous generations IBM Z systems.  A TKE, Smart Card Reader and latest available level smart cards are required to operate the Crypto adapter card in EP11 mode.
0898	Crypto Express7S feature (2-port) <sup>a</sup>  Carry forward from IBM z16. This feature contains two IBM 4769 PCIe Cryptographic Coprocessors (HSMs), which can be independently defined as Coprocessor or Accelerator.
0899	Crypto Express7S feature (1-port) <sup>a</sup>  Carry forward from IBM z16. This feature contains one IBM 4769 PCIe Cryptographic Coprocessor (HSM), which can be defined as Coprocessor or Accelerator.

<sup>2</sup> CPACF enablement feature must be ordered (FC 3863).



Feature Code	Description
0058	<p>TKE tower workstation</p> <p>A TKE provides basic key management (key identification, exchange, separation, update, and backup) and security administration. It is optional for running a Crypto Express feature in CCA mode in non PCI-compliant environment. It is required for running in EP11 mode and CCA mode with full PCI compliance. The TKE workstation has one 1000BASE-T Ethernet port, and supports connectivity to an Ethernet local area network (LAN). Up to 10 features combined (0057/0058) can be ordered per IBM z17.</p>
0057	<p>TKE rack-mounted workstation</p> <p>The rack-mounted version of the TKE, which needs a customer-provided standard 19-inch rack. It features a 1u TKE unit and an (optional) 1u console tray (screen, keyboard, and pointing device). When smart card readers are used, another customer-provided tray is needed. Up to 10 features combined (0057/0058) can be ordered per IBM z17.</p>
0883	<p>TKE 10.1 Licensed Internal Code (LIC)</p> <p>Included with the TKE tower workstation FC 0058 and the TKE rack-mounted workstation FC 0057 for IBM z17. Earlier versions of TKE features (feature codes: 0087, 0088, 0085, and 0086) also can be upgraded to TKE 10.1 LIC, adding FC 0851 (IBM 4770 PCIeCC) if the TKE is assigned to an IBM z17 and manages Crypto Express8S.</p>
0851	<p>4770 TKE Crypto Adapter (IBM PCIeCC)</p> <p>The stand-alone crypto adapter is required for TKE upgrade from FC 0086 and FC 0088 TKE tower, or FC 0085 and FC 0087 TKE Rack Mount when carry forward these features to IBM z17.</p>
0144	<p>TKE Tower carry forward to IBM z17</p> <p>TKE Tower FC 0088 can be carried forward to IBM z17. It requires IBM 4770 PCIeCC (FC 0851) for compatibility with TKE LIC 10.1 (FC 0883) and for managing Crypto Express8S (FC 0144 = FC 0088 + FC 0851 + FC 0883).</p>
0145	<p>TKE Rack Mount carry forward to IBM z17</p> <p>TKE Rack Mount FC 0087 can be carried forward to IBM z17. It requires IBM 4770 PCIeCC (FC 0851) for compatibility with TKE LIC 10.1 (FC 0883) and for managing Crypto Express8S (FC 0145 = FC 0087 + FC 0851 + FC 0883).</p>
0233	<p>TKE Rack Mount carry forward to IBM z17</p> <p>TKE Rack Mount FC 0085 can be carried forward to IBM z17. It requires IBM 4770 PCIeCC (FC 0851) for compatibility with TKE LIC 10.1 (FC 0883) and for managing Crypto Express8S (FC 0233 = FC 0085 + FC 0851 + FC 0883).</p>
0234	<p>TKE Tower Carry forward to IBM z17</p> <p>TKE Tower FC 0086 can be carried forward to IBM z17. It requires IBM 4770 PCIeCC (FC 0851) for compatibility with TKE LIC 10.1 (FC 0883) and for managing Crypto Express8S (FC 0234 = FC 0086 + FC 0851 + FC 0883).</p>



Feature Code	Description
0891	TKE Smart Card Reader  Access to information in the smart card is protected by a PIN. One feature code includes two smart card readers, two cables to connect to the TKE workstation, and 20 smart cards.
0885	TKE Smart Card Reader carry forward  Access to information in the smart card is protected by a PIN. Carry forward with existing cards (non-FIPS).
0900	New TKE smart cards (10 pack)  This card allows the TKE to support zones with EC 521 key strength (EC 521 strength for Logon Keys, Authority Signature Keys, and EP11 signature keys).
0892	More TKE smart cards  Carry forward only to IBM z17. Ten smart cards are included.

- a. The maximum number of combined features of all types cannot exceed 60 HSMs on an IBM z17 ME1 (any combination of single and dual HSM Crypto Express features). Therefore, the maximum number for Feature Code 0908 is 30; for all other (single HSM) types, it is 16 for an IBM z17 system.

A TKE includes support for AES encryption algorithm with 256-bit master keys and key management functions to load or generate master keys to the cryptographic coprocessor.

If the TKE workstation is chosen to operate the Crypto Express8S adapter in an IBM z17, TKE workstation with the TKE 10.1 LIC is required. For more information, see 6.6, “Trusted Key Entry workstation” on page 253.

**Important:** Products that include any of the cryptographic feature codes contain cryptographic functions that are subject to special export licensing requirements by the US Department of Commerce. It is your responsibility to understand and adhere to these regulations when you are moving, selling, or transferring these products.

To access and use the cryptographic hardware devices that are provided by IBM z17, the application must use an application programming interface (API) that is provided by the operating system. In z/OS, the Integrated Cryptographic Service Facility (ICSF) provides the APIs and is managing the access to the cryptographic devices, as shown in Figure 6-3 on page 231.



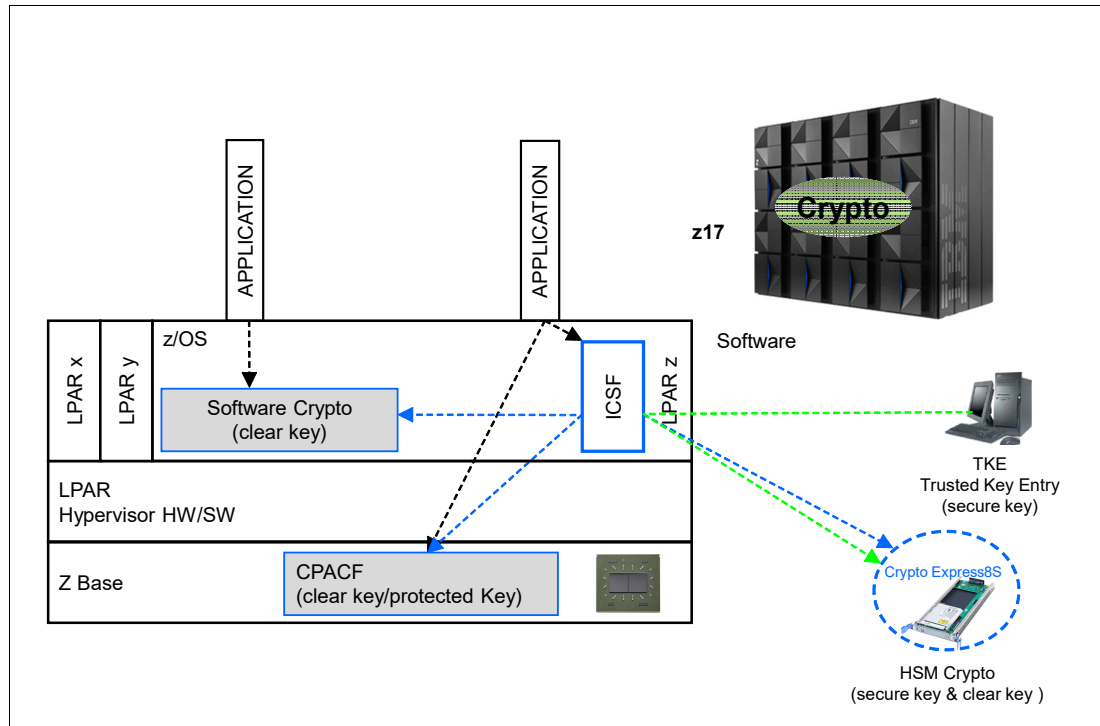


Figure 6-3 IBM z17 Cryptographic Support in z/OS

ICSF is a software component of z/OS. ICSF works with the hardware cryptographic features and the Security Server (IBM Resource Access Control Facility [IBM RACF®] element) to provide secure, high-speed cryptographic services in the z/OS environment. ICSF provides the APIs by which applications request the cryptographic services, and from the CPACF and the Crypto Express features.

ICSF transparently routes application requests for cryptographic services to one of the integrated cryptographic engines (CPACF or a Crypto Express feature), depending on performance or requested cryptographic function. ICSF is also the means by which the secure Crypto Express features are loaded with master key values, which allows the hardware features to be used by applications.

The cryptographic hardware that is installed in IBM z7 determines the cryptographic features and services that are available to the applications.

The users of the cryptographic services call the ICSF API. Some functions are performed by the ICSF software without starting the cryptographic hardware features. Other functions result in ICSF going into routines that contain proprietary IBM Z crypto instructions. These instructions are run by a CPU engine and result in a work request that is generated for a cryptographic hardware feature.

## 6.4 CP Assist for Cryptographic Functions

Attached to every PU (core) of an IBM z17 system is an independent engine for performing the computational intensive operations for sort, compression and cryptographic functions, as shown in Figure 6-4 on page 232.



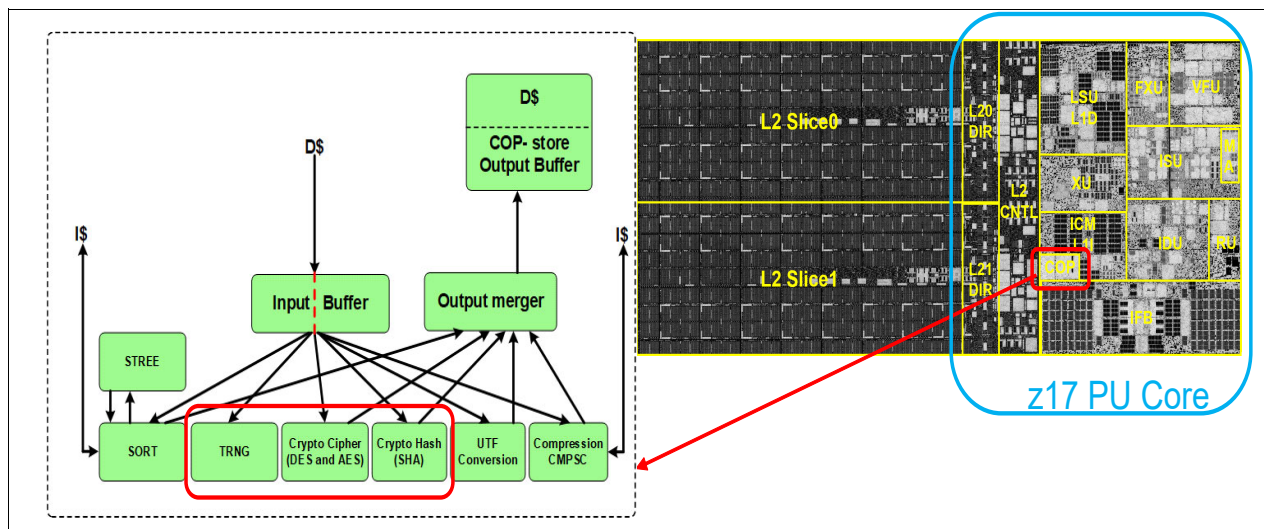


Figure 6-4 The cryptographic coprocessor CPACF

This cryptographic coprocessor, known as the CP assist cryptographic function, *CPACF*, is not qualified as an HSM; therefore, it is not suitable for handling algorithms that use secret keys. However, the coprocessor can be used for cryptographic algorithms using clear keys or protected keys. The CPACF works synchronously with the PU, which means that the owning processor is busy when its coprocessor is busy. This setup provides a fast device for cryptographic services.

CPACF supports pervasive encryption. Simple policy controls allow businesses to enable encryption to protect data in mission-critical databases without the need to stop the database or re-create database objects. Pervasive encryption includes z/OS Dataset Encryption, z/OS Coupling Facility Encryption, z/VM encrypted hypervisor paging, and z/TPF transparent database encryption, which use performance enhancements in the hardware.

The CPACF offers a set of symmetric cryptographic functions that enhances the encryption and decryption performance of clear key operations. These functions are for SSL, virtual private network (VPN), and data-storing applications that do not require FIPS 140-2 Level 4 security.

CPACF is designed to facilitate the privacy of cryptographic key material when used for data encryption through key wrapping implementation. It ensures that key material is not visible to applications or operating systems during encryption operations. For more information, see 6.4.2, “CPACF protected key” on page 234.

The CPACF feature provides hardware acceleration for the following cryptographic services:

- ▶ Symmetric ciphers:
  - DES
  - Triple-DES
  - AES-128
  - AES-192
  - AES-256 (all for clear and protected keys)
- ▶ Elliptic curves cryptography (ECC):
  - ECDSA, ECDH, support for the NIST P256, NIST P386, NIST P521
  - EdDSA for Ed25519, Ed448 Curves
  - ECDH for X25519, X448 Curves
  - Key generation for NIST, Ed, and X curves



- ▶ Hashes/MACs:
  - SHA-1
  - SHA-224 (SHA-2 or SHA-3 standard)
  - SHA-256 (SHA-2 or SHA-3 standard)
  - SHA-384 (SHA-2 or SHA-3 standard)
  - SHA-512 (SHA-2 or SHA-3 standard)
  - SHAKE-128
  - SHAKE-256
  - GHASH
- ▶ Random number generator:
  - PRNG (3DES based)
  - DRNG (NIST SP-800-90A SHA-512 based)
  - TRNG (true random number generator)

It provides high-performance hardware encryption, decryption, hashing, and random number generation support. The following instructions support the cryptographic assist function:

- ▶ KMAC: Compute Message Authentic Code
- ▶ KM: Cipher Message
- ▶ KMC: Cipher Message with Chaining
- ▶ KMF: Cipher Message with CFB
- ▶ KMCTR: Cipher Message with Counter
- ▶ KMO: Cipher Message with OFB
- ▶ KIMD: Compute Intermediate Message Digest
- ▶ KLMD: Compute Last Message Digest
- ▶ PCKMO: Provide Cryptographic Key Management Operation

These functions are provided as problem-state *z/Architecture* instructions that are directly available to application programs. These instructions are known as *Message-Security Assist* (MSA). When enabled, the CPACF runs at processor speed for every CP, IFL, and zIIP.

For more information about MSA instructions, see *z/Architecture Principles of Operation*, SA22-7832.

For activating these functions, the CPACF must be enabled by using Feature Code (FC) 3863, which is available at no charge. Support for hashing algorithms SHA-1, SHA-256, SHA-384, and SHA-512 always is enabled.

### 6.4.1 Cryptographic synchronous functions

Because the CPACF works synchronously with the PU, it provides cryptographic synchronous functions. For IBM and client-written programs, CPACF functions can be started by using the MSA instructions. z/OS ICSF callable services on z/OS, in-kernel crypto APIs, and a libica cryptographic functions library that is running on Linux on IBM Z also can start CPACF synchronous functions.

The following tools might benefit from the throughput improvements for IBM z17 CPACF:

- ▶ Db2/IMS encryption tool
- ▶ Db2 built-in encryption
- ▶ z/OS Communication Server: IPsec/IKE/AT-TLS
- ▶ z/OS System SSL
- ▶ z/OS Network Authentication Service (Kerberos)
- ▶ DFDSS Volume encryption
- ▶ z/OS Java SDK



- ▶ z/OS Encryption Facility
- ▶ Linux on IBM Z: Kernel, openSSL, openCryptoki, and GSKIT

The IBM z17 hardware includes the implementation of algorithms as hardware synchronous operations. This configuration holds the PU processing of the instruction flow until the operation completes.

IBM z17 offers the following synchronous functions:

- ▶ Data encryption and decryption algorithms for data privacy and confidentiality:
  - Data Encryption Standard (DES):
    - Single-length key DES
    - Double-length key DES
    - Triple-length key DES (also known as Triple-DES)
  - Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys
- ▶ Hashing algorithms for data integrity, such as SHA-1 and SHA-2.

New since IBM z14 ZR1 is SHA-3 support for SHA-224, SHA-256, SHA-384, and SHA-512 and the two extendable output functions as described by the standard SHAKE-128 and SHAKE-256.
- ▶ Message authentication code (MAC):
  - Single-length key MAC
  - Double-length key MAC
- ▶ Pseudo-Random Number Generator (PRNG), Deterministic Random Number Generation (DRNG), and True Random Number Generation (TRNG) for cryptographic key generation.
- ▶ Galois Counter Mode (GCM) encryption, which is enabled by a single hardware instruction.

For the SHA hashing algorithms and the random number generation algorithms, only clear keys are used. For the symmetric encryption and decryption DES and AES algorithms and clear keys, protected keys also can be used. On IBM z17, protected keys require a Crypto Express adapter that is running in CCA mode.

For more information, see 6.5.2, “Crypto Express8S as a CCA coprocessor” on page 240.

The hashing algorithms SHA-1, SHA-2, and SHA-3 support for SHA-224, SHA-256, SHA-384, and SHA-512, are enabled on all systems and do not require the CPACF enablement feature. For all other algorithms, the no-charge CPACF enablement feature (FC 3863) is required.

The CPACF functions are implemented as processor instructions and require operating system support for use. Operating systems that use the CPACF instructions include z/OS, z/VM, VSE<sup>3</sup> V6.3.1 – 21st Century Software, z/TPF, and Linux on IBM Z.

## 6.4.2 CPACF protected key

IBM z17 supports the protected key implementation. Secure keys are processed on the PCIeCC adapters (HSMs)<sup>3</sup>. This process requires an asynchronous operation to move the data and keys from the general-purpose central processor (CP) to the crypto adapters.

Clear keys process faster than secure keys because the process is done synchronously on the CPACF. Protected keys blend the security of Crypto Express8S, or Crypto Express7S

<sup>3</sup> PCIeCC: IBM PCIe Cryptographic Coprocessor is the Hardware Security Module (HSM).



coprocessors and the performance characteristics of the CPACF. This process allows it to run closer to the speed of clear keys.

CPACF facilitates the continued privacy of cryptographic key material when used for data encryption. In Crypto Express8S, or Crypto Express7S coprocessors, a secure key is encrypted under a master key. However, a protected key is encrypted under a wrapping key that is unique to each LPAR.

Because the wrapping key is unique to each LPAR, a protected key cannot be shared with another LPAR. By using key wrapping, CPACF ensures that key material is not visible to applications or operating systems during encryption operations.

CPACF code generates the wrapping key and stores it in the protected area of the hardware system area (HSA). The wrapping key is accessible only by firmware. It cannot be accessed by operating systems or applications.

DES/T-DES and AES algorithms are implemented in CPACF code with the support of hardware assist functions. Two variations of wrapping keys are generated: one for DES/T-DES keys and another for AES keys.

Wrapping keys are generated during the clear reset each time an LPAR is activated or reset. No customizable option is available at Support Element (SE) or Hardware Management Console (HMC) that permits or avoids the wrapping key generation. This function flow for the Crypto Express8S, and Crypto Express7S adapters is shown in Figure 6-5.

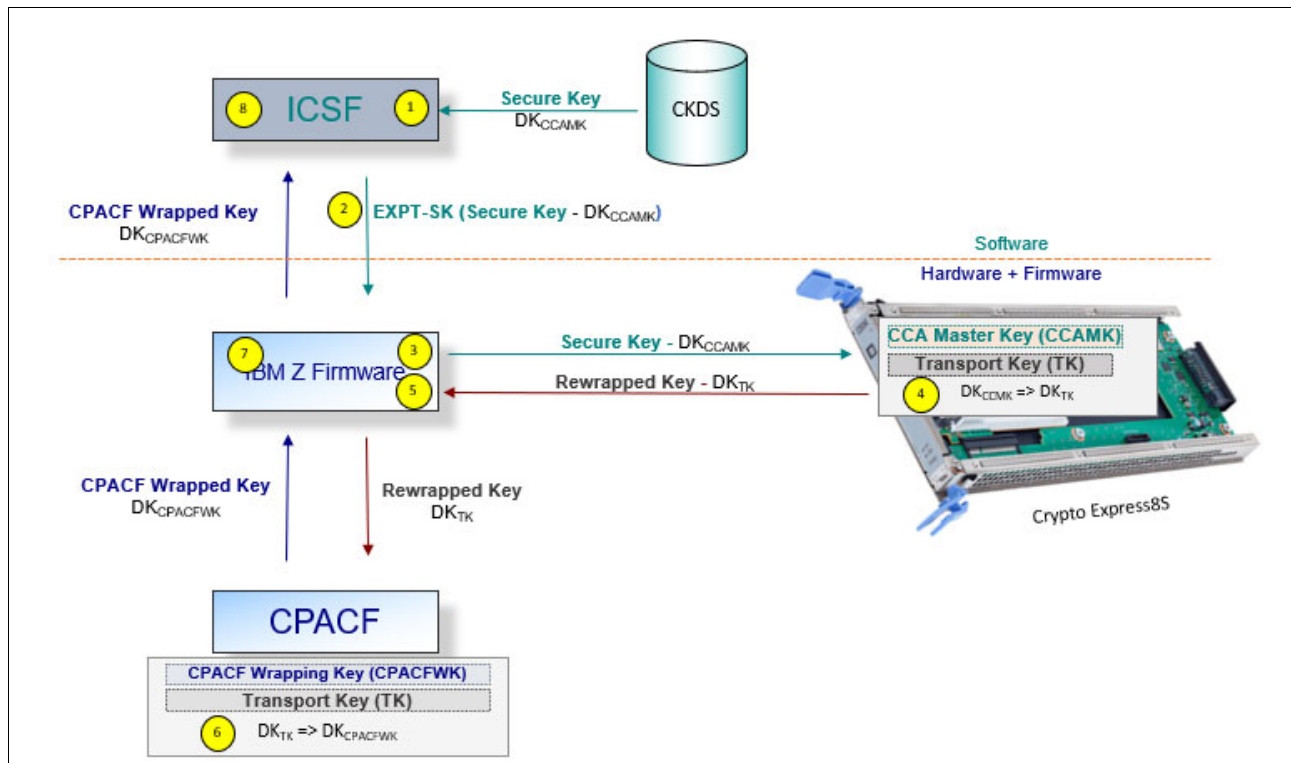


Figure 6-5 CPACF key wrapping for Crypto Express8S, and Crypto Express7S

The CPACF Wrapping Key and the Transport Key for use with Crypto Express8S, and Crypto Express7S are in a protected area of the HSA that is not visible to operating systems or applications.



If a Crypto Express coprocessor (CEX8C, or CEX7C) is available, a protected key can begin its life as a secure key. Otherwise, an application is responsible for creating or loading a clear key value and then, uses the PCKMO instruction to wrap the key. ICSF is not called by the application if the CEXxC is not available.

A new segment in the profiles of the CSFKEYS class in IBM RACF restricts which secure keys can be used as protected keys. By default, all secure keys are considered not eligible to be used as protected keys. The process that is shown in Figure 6-5 on page 235 considers a secure key as the source of a protected key.

The source key in this case is stored in the ICSF Cryptographic Key Data Set (CKDS) as a secure key, which was encrypted under the master key. This secure key is sent to CEX8C, or CEX7C, to be deciphered and then, sent to the CPACF in clear text.

At the CPACF, the key is wrapped under the LPAR wrapping key, and is then returned to ICSF. After the key is wrapped, ICSF can keep the protected value in memory. It then passes it to the CPACF, where the key is unwrapped for each encryption or decryption operation.

The protected key is designed to provide substantial throughput improvements for a large volume of data encryption and low latency for encryption of small blocks of data. A high-performance secure key solution, also known as a *protected key solution*, requires the ICSF HCR7770 as a minimum release.

### CPACF Enhancements

- ▶ Adding several HMAC algorithms to KMAC
  - KMAC supports taking a clear key and converting it to a protected key using the PCKMO instruction
- ▶ Added new function codes for optimized AES-XTS functionality to KM

### CPACF usage tracking

Deliver a new hardware managed counter set to track crypto algorithms, bit lengths and key security (e.g., AES 256 encrypted).

- ▶ Provides evidence for:
  - compliance (i.e., which crypto is used)
  - performance (i.e., frequency of crypto use)
  - configuration (i.e., proof of change)
  - Correlated with the workload that invoked the CPACF function (e.g., PCI workload)

## 6.5 Crypto Express8S

The Crypto Express8S feature (FC 0908 or FC 0909) is an optional feature that is exclusive to IBM z16 and IBM z17 systems. Each feature FC 0909 has one IBM 4770 PCIe cryptographic adapter (hardware security module [HSM]), whereas FC 0908 has two IBM 4770 PCIe cryptographic adapters (two HSMs). The Crypto Express8S (CEX8S) feature occupies one I/O slot in an IBM z17 PCIe+ I/O drawer. This feature provides one or two HSMs and for a secure programming and hardware environment on which crypto processes are run.

Each cryptographic coprocessor includes a general-purpose processor, nonvolatile storage, and specialized cryptographic electronics. The Crypto Express8S feature provides tamper-sensing and tamper-responding, high-performance cryptographic operations.



Each Crypto Express8S PCI Express adapter (HSM) is available in one of the following configurations:

- ▶ Secure IBM CCA coprocessor (CEX8C) for FIPS 140-2 Level 4 certification. This configuration includes secure key functions. It is optionally programmable to deploy more functions and algorithms by using UDX.

For more information, see 6.5.2, “Crypto Express8S as a CCA coprocessor” on page 240.

A TKE workstation is required to support the administration of the Crypto Express8S when it is configured in CCA mode when in full Payment Card Industry (PCI)-compliant mode for the necessary certificate management in this mode. The TKE is optional in all other use cases for CCA.

- ▶ Secure IBM Enterprise PKCS #11 (EP11) coprocessor (CEX8P) implements an industry-standardized set of services that adheres to the PKCS #11 specification V2.20 and more recent amendments. It was designed for extended FIPS and Common Criteria evaluations to meet public sector requirements. This new cryptographic coprocessor mode introduced the PKCS #11 secure key function.

For more information, see 6.5.3, “Crypto Express8S as an EP11 coprocessor” on page 247.

A TKE workstation is always required to support the administration of the Crypto Express7S when it is configured in EP11 mode.

- ▶ Accelerator (CEX8A) for acceleration of public key and private key cryptographic operations that are used with SSL/TLS processing.

For more information, see 6.5.4, “Crypto Express8S as an accelerator” on page 248.

These modes can be configured by using the SE. The PCIe adapter must be configured offline to change the mode.

**Attention:** Switching between configuration modes erases all adapter secrets. The exception is when you are switching from Secure CCA to accelerator, and vice versa.

The Crypto Express8S feature is released for enhanced cryptographic performance. Customers who migrated to variable-length AES key tokens cannot take advantage of faster encryption speeds by using CPACF. Support is being added to translate a secure variable-length AES CIPHER token to a protected key token (protected by the system wrapping key). This support allows for faster AES encryption speeds when variable-length tokens are used while maintaining strong levels of security.

The Crypto Express8S feature does not include external ports and does not use optical fiber or other cables. It does not use channel path identifiers (CHPIDs), but requires one slot in the PCIe I/O drawer and one physical channel ID (PCHID) for each PCIe cryptographic adapter. Removal of the feature or adapter *zeroizes* its content. Access to the PCIe cryptographic adapter is controlled through the setup in the image profiles on the SE.

**Adapter:** Although PCIe cryptographic adapters include no CHPID type and are not identified as external channels, all logical partitions (LPARs) in all channel subsystems can access the adapter. In IBM z17, up to 85 LPARs are supported per adapter (HSM). Accessing the adapter requires a setup in the image profile for each partition. The adapter must be in the candidate list.

Each IBM z17 ME1 supports up to 60 HSMs in total (combination of Crypto Express8S (1 or 2 HSMs), and Crypto Express7S (1 or 2 HSMs). Crypto Express7S (1 or 2 ports) are *not* orderable for a new build IBM z17 A01 system, but can be carried forward from an IBM z16 or



IBM z15 by using an MES. Configuration information for Crypto Express8S is listed in Table 6-2.

Table 6-2 *Crypto Express8S features*

Feature	Quantity
Minimum number of orderable features 0908 for IBM z17 ME1	2
Minimum number of orderable features 0909 for IBM z17 <sup>a</sup> ME1	2
Order increment (above two features for features 0908 and 0909)	1
Maximum number of HSMs for IBM z17 (combining all CEX8S, and CEX7S)	60 <sup>b</sup>
Number of PCIe cryptographic adapters for each feature 0908 (coprocessor or accelerator)	2
Number of PCIe cryptographic adapters for each feature 0909 (coprocessor or accelerator)	1
Number of cryptographic domains at IBM z17 for each PCIe adapter <sup>c</sup>	85

- a. The minimum initial order of Crypto Express8S feature 0909 is two. After the initial order, more Crypto Express8S features 0909 can be ordered individually.
- b. Crypto Express8S (dual HSM) has two hardware security modules (HSMs) per feature. The HSM is one IBM 4770 PCIe Cryptographic Coprocessor (PCIeCC). The maximum number of HSMs per A01 system (combining all cryptographic features) is 60. The maximum number of single HSM (port) cryptographic features is 16 (CEX8S [single HSM], CEX7S [1 port]).
- c. More than one partition, which is defined to the same channel subsystem (CSS) or to different CSSs, can use the same domain number when assigned to different PCIe cryptographic adapters.

The concept of *dedicated processor* does not apply to the PCIe cryptographic adapter. Whether configured as a coprocessor or an accelerator, the PCIe cryptographic adapter is made available to an LPAR. It is made available as directed by the domain assignment and the candidate list in the LPAR image profile. This availability is not changed by the shared or dedicated status that is given to the PUs in the partition.

When installed non-concurrently, Crypto Express8S features are assigned PCIe cryptographic adapter numbers sequentially during the power-on reset (POR) that follows the installation. When a Crypto Express8S feature is installed concurrently, the installation can select an out-of-sequence number from the unused range. When a Crypto Express8S (or Crypto Express7S) feature is removed concurrently, the PCIe adapter numbers are automatically freed.

The definition of domain indexes and PCIe cryptographic adapter numbers in the candidate list for each LPAR must be planned to allow for nondisruptive changes. Consider the following points:

- ▶ Operational changes can be made by using the Change LPAR Cryptographic Controls task from the SE, which reflects the cryptographic definitions in the image profile for the partition. With this function, adding and removing the cryptographic feature without stopping a running operating system can be done dynamically.
- ▶ The same usage domain index can be defined more than once across multiple LPARs. However, the PCIe cryptographic adapter number that is coupled with the usage domain index that is specified must be unique across all active LPARs.

The same PCIe cryptographic adapter number and usage domain index combination can be defined for more than one LPAR (up to 85 for IBM z17). For example, you might define a configuration for backup situations. However, only one of the LPARs can be active at a time.

For more information, see 6.5.5, “Managing Crypto Express8S” on page 249.



## 6.5.1 Cryptographic asynchronous functions

The Crypto Express8S feature provides asynchronous cryptographic functions to IBM z17. More than 300 Cryptographic algorithms and modes are supported, including the following examples:

- ▶ DES/TDES with DES/TDES MAC/CMAC

The Data Encryption Standard is a widespread symmetrical encryption algorithm. DES, along with its double-length and triple length variations, TDES today are considered to be insufficiently secure for many applications. They were replaced by AES as the official US standard, but it is still used in the industry with the MAC and the Cipher-based Message Authentication Code (CMAC) for verifying the integrity of messages.

- ▶ AES, AESKW, AES GMAC, AES GCM, AES XTS, AES CIPHER mode, and CMAC

AES replaced DES as the official US standard in October 2000. The enhanced standards for AES Key Wrap (AEKW), AES Galois Message Authentication Code (AES GMAC), and Galois/Counter Mode (AES GCM), the XEX-based tweaked-codebook mode with ciphertext stealing (AES XTS), and CMAC are supported.

- ▶ MD5, SHA-1, SHA-2, or SHA-3<sup>4</sup> (224, 256, 384, and 512), and HMAC

The Secure Hash Algorithm (SHA-1 and the enhanced SHA-2 or SHA-3 for different block sizes), the older message-digest (MD5) algorithm, and the advanced keyed-hash method authentication code (HMAC) are used for verifying the data integrity and the authentication of a message.

- ▶ VFPE

A method of encryption in which the resulting cipher text features the same form as the input clear text, which is developed for use with credit cards.

- ▶ RSA (512, 1024, 2048, and 4096)

RSA was published in 1977. It is a widely used asymmetric public-key algorithm, which means that the encryption key is public whereas the decryption key is kept secret. It is based on the difficulty of factoring the product of two large prime numbers. The number describes the length of the keys.

- ▶ ECDSA (192, 224, 256, 384, and 521 Prime/NIST)

ECC is a family of asymmetric cryptographic algorithms that are based on the algebraic structure of elliptic curves. ECC can be used for encryption, pseudo-random number generation, and digital certificates. The Elliptic Curve Digital Signature Algorithm (ECDSA) Prime/NIST method is used for ECC digital signatures, which are recommended for government use by NIST.

- ▶ ECDSA (160, 192, 224, 256, 320, 384, and 512 BrainPool)

ECC Brainpool is a work group of companies and institutions that collaborate on developing ECC algorithms. The ECDSA algorithms that are recommended by this group are supported.

- ▶ ECDH (192, 224, 256, 384, and 521 Prime/NIST)

Elliptic Curve Diffie-Hellman (ECDH) is an asymmetric protocol that is used for key agreement between two parties by using ECC-based private keys. The recommendations by NIST are supported.

- ▶ ECDH (160, 192, 224, 256, 320, 384, and 512 BrainPool)

ECDH according to the Brainpool recommendations.

<sup>4</sup> SHA-3 was standardized by NIST in 2015. SHA-2 is still acceptable and no indication exists that SHA-2 is vulnerable or that SHA-3 is more or less vulnerable than SHA-2.



- ▶ Montgomery Modular Math Engine  
The Montgomery Modular Math Engine is a method for fast modular multiplication. Many crypto systems, such as RSA and Diffie-Hellman key Exchange, can use this method.
- ▶ Clear Key Fast Path (Symmetric and Asymmetric)  
This mode of operation gives a direct hardware path to the cryptographic engine and provides high performance for public-key cryptographic functions.
- ▶ Random Number Generator (RNG)
- ▶ Prime Number Generator (PNG)

Several of these algorithms require a secure key and must run on an HSM. Some of these algorithms can also run with a clear key on the CPACF. Many standards are supported only when Crypto Express8S is running in CCA mode. Others are supported only when the adapter is running in EP11 mode.

The three modes for Crypto Express8S are described next. For more information, see 6.7, “Cryptographic functions comparison” on page 256.

## 6.5.2 Crypto Express8S as a CCA coprocessor

A Crypto Express8S adapter that is running in CCA mode supports IBM CCA. CCA is an architecture and a set of APIs. It provides cryptographic algorithms, secure key management, and many special functions that are required for banking. Over 157 APIs with more than 1000 options are provided, with new functions and algorithms always being added.

The IBM CCA provides functions for the following tasks<sup>5</sup>:

- ▶ Encryption of data (DES/TDES/AES)
- ▶ Key management:
  - Using TDES or AES keys
  - Using RSA or Elliptic Curve keys
  - QSA keys
  - TR-31 Key Blocks (AES/TDES/HMAC, Comp-tag, operational/native and key exchange)
  - TR-34 / PKI-based ATM remote key load
- ▶ Message authentication for MAC/HMAC/AES-CMAC
- ▶ Key generation
- ▶ Digital signatures
  - RSA/ECC/EdDSA
- ▶ Quantum-Safe algorithms
  - Dilithium, Kyber
- ▶ Random number generation
- ▶ Hashing
  - SHA, MD5, and others
- ▶ ATM PIN generation and processing
  - TDES & AES (ISO-4), DUKPT & AES-DUKPT
- ▶ Credit card transaction processing
- ▶ Visa Data Secure Platform (DSP) Point to Point Encryption (P2PE)
- ▶ Format Preserving Encryption
  - (FPE): FF1, FF2, FF2.1
- ▶ Europay, MasterCard, and Visa (EMV) card transaction processing
- ▶ Card personalization
- ▶ Other financial transaction processing

<sup>5</sup> List is not exhaustive; new functions are added continuously



- ▶ Integrated role-based access control system
- ▶ Compliance support for:
  - All DES services
  - AES services
  - RSA services, including full use of X.509 certificates
- ▶ TR-34 Remote Key Load

### Quantum Safe Algorithm Updates

- ▶ Standardized support conforming to official NIST specifications.
- ▶ Support for ML-KEM (Module-Lattice-Based Key-Encapsulation Mechanism, NIST FIPS 203) and ML-DSA (Module-Lattice-Based Digital Signature Algorithm, NIST FIPS 204) key generation and use with digital signatures (ML-DSA) and Key Encapsulation (ML-KEM) in the CCA API.
- ▶ ML-KEM Key Sizes 768 + 1024.
- ▶ ML-DSA (6,5) & (8,7).

### User-defined extensions support

User-defined extension (UDX) allows a developer to add customized operations to IBM's CCA Support Program. UDXs to the CCA support customized operations that run within the Crypto Express features when defined as a coprocessor.

UDX is supported under a special contract through an IBM or approved third-party service offering. The Crypto Cards website directs your request to an IBM representative for your geographic location. A special contract is negotiated between IBM and you for the development of the UDX code by IBM according to your specifications and an agreed-upon level of the UDX.

A UDX toolkit for IBM Z is tied to specific versions of the CCA code and the related host code. UDX is available for the Crypto Express8S, and Crypto Express7S (Secure IBM CCA coprocessor mode only) features. An UDX migration is no more disruptive than a normal Microcode Change Level (MCL) or ICSF release migration.

In IBM z17, up to four UDX files can be imported. These files can be imported from a USB media stick or an FTP server. The UDX configuration window is updated to include a Reset to IBM Default button.

**Consideration:** CCA features a new code level starting with z13 systems, and the UDX clients require a new UDX.

On IBM z17 ME1, Crypto Express8S is delivered with CCA 8.4 firmware and Crypto Express7S with CCA7.6 firmware. A new set of cryptographic functions and callable services is provided by the IBM CCA LIC to enhance the functions that secure financial transactions and keys. The Crypto Express8S includes the following features:

- ▶ Greater than 16 domains support up to 85 LPARs on IBM z17 ME1.
- ▶ PCI PIN Transaction Security (PTS) HSM Certification that is available to IBM z17 in combination with CEX8S, or CEX7S, features.
- ▶ VFPE support, which was introduced with z13/z13s systems.
- ▶ AES PIN support for the German banking industry.
- ▶ PKA Translate UDX function into CCA.
- ▶ Verb Algorithm Currency.



## CCA improvements

With CCA 8.0, the following CCA improvements for Quantum Safe algorithms were made:

- ▶ CCA Quantum Safe Algorithm enhancements:
  - Updated support for Dilithium signatures:
    - Round 2: Level 2 (6 5) and 3 (8 7)
    - Round 3: Level 3 (6 5) and 5 (8 7)
  - Added support for Kyber key encapsulation in Round 2: Level 5 (1024)
- ▶ Quantum Safe protected key support for CCA

Host Firmware and CCA now employ a hybrid scheme combining ECDH and Kyber to accomplish a quantum safe transport key exchange for protected key import.

## CCA 8.4

On z17 and Crypto Express8S, CCA 8.4 content adds updates for Quantum Safe Algorithms:

- ▶ Standardized support conforming to official NIST specifications.
- ▶ Support for ML-KEM (Module-Lattice-Based Key-Encapsulation Mechanism, NIST FIPS 203) and ML-DSA (Module-Lattice-Based Digital Signature Algorithm, NIST FIPS 204) key generation and use with digital signatures (ML-DSA) and Key Encapsulation (ML-KEM) in the CCA API.
- ▶ ML-KEM Key Sizes 768 + 1024.
- ▶ ML-DSA (6,5) & (8,7).

Also on z17, CCA releases 8.4 and 7.6 add support for RSA keys with modulus bit lengths of up to 8192 bits.

- ▶ RSA key length 8192 is above the safety margin for AES-192 bit key transport (ignoring QC considerations)
- ▶ RSA key length extension is NOT expected to add safety margin for Post Quantum Computing considerations
- ▶ Support available in various CCA services including:
  - PKA Key token Build and Generate – CSNDPKB, CSNDPKG
  - Digital Signature Generate & Verify – CSNDDSG, CSNDDSV
  - RSA encryption/decryption with CSNDPKE, CSNDPKD

## CCA 8.2

CCA 8.2 content, released 01/2024 for z16, is described in this [blog post](#) and adds support for the following features:

- ▶ TR-31 Import / Export of AES K0-B and K1-B Key Blocks, enhancing key exchange with popular payment networks;
- ▶ Import RSA AES Key Wrapped Objects, enhancing key exchange with common cloud environments;
- ▶ New CCA service: Multi-MAC Scheme (CSNBMMS), adding a rich financial service interface for verified PIN change;
- ▶ RSA-OAEP v2.1 updates, adding flexibility with hash options for asymmetric encryption.

## CCA 8.0

The following CCA 8.0 improvements were made:



- ▶ Performance enhancement for mixed workloads: better performance when one partition focuses on RSA/ECC and another partition focuses on AES/DES/TDES or financial operations
- ▶ Hardware accelerated key unwrap for AES wrapped keys
  - Trusted Key Entry workstation (TKE) controlled selection of WRAPENH3 as the default TDES key token wrapping method for easier management.

### **CCA 7.2 and CCA 6.5**

The following CCA 7.2 and 6.5 improvements were made:

- ▶ AES DUKPT unique key per transaction for AES-based PIN and transaction protection
- ▶ ISO 4 enhanced support for adding AES protected PIN block support to all remaining services that process PIN blocks
- ▶ Format Preserving Encryption (FPE) supports three algorithms that are standardized by NIST and in X9.124: FF1, FF2, and FF2.1
- ▶ Elliptic Curve support for the Koblitz curve secp256k1, to all ECC services including native support for X.509 certificates

### **CCA 7.3, CCA 6.6**

WRAPENH3 Enhanced Wrapping Method for TDES key tokens was added.

### **CCA 7.4 and CCA 6.7**

The following CCA 7.4 and 6.7 improvements were made:

- ▶ German banking API updates for program currency
- ▶ X9.23 random padding for AES encryption for important cryptographic operation protection
- ▶ Enhanced triple length TDES PIN encryption key support for PIN change workloads
- ▶ New service to compare encrypted PINs, which is required for ISO 4 PIN block verification inside the HSM
- ▶ EC SDSA signature support, which is useful for new EMV certificate formats
- ▶ PKCS#11 update to CCA API export of AES and RSA keys by using RSA public key and AES ephemeral keys for key exchange with Cloud service key management APIs
- ▶ Australian Payment Network Acquirer function for key derivation and MAC chaining added for interoperability with Australian audited payment networks

### **CCA 7.1**

The following CCA 7.1 improvements were made:

- ▶ Supported curves:
  - NIST Prime Curves: P192, P224, P256, P384, and P521
  - Brainpool Curves: 160, 192, 224, 256, 320, 384, and 512
- ▶ Support in the CCA coprocessor for Edwards curves ED25519 (128-bit security strength) and ED448 (224-bit security strength)

Although ED25519 is faster, ED448 is more secure. Practically though, 128-bit security strength is very secure.

Edwards curves are used for digitally signing documents and verifying those signatures.

Edwards curves are less susceptible to side channel attacks when compared to Prime and Brainpool curves.

- ▶ ECC Protected Keys



Crypto Express8S and Crypto Express7S provide support in CCA coprocessors to take advantage of fast DES and AES data encryption speeds in CPACF while maintaining high levels of security for the secure key material. The key remains encrypted and the key encrypting key never appears in host storage.

When CCA ECC services are used, ICSF can now take advantage of ECC support in CPACF (protected key support) for the following curves:

- Prime: P256, P384, P521
- Edwards: ED25519, ED448

CPACF can achieve much faster crypto speeds compared to the coprocessor.

The translation to protected key happens automatically after the attribute is set in the key token. No application change is required.

► New signatures

Support for the Cryptographic Suite for Algebraic Lattices signatures algorithm with the largest key sizes (MODE=3):

- Public Key size: 1760 bytes
- Private Key Size: 3856 bytes
- Signature Size: 3366 bytes

Lattice-based cryptographic keys are protected by the 256-bit AES MK. The lattice-based key has a security strength of 128 bits.

► TR-31 for Hash-based Message Authentication Code (HMAC)

HMAC keys are used to verify the integrity and authenticity of a message. This support provides a standard method of exchanging HMAC keys with a partner that uses symmetric key techniques. The key is exchanged in the standard TR-31 key block format, which can be used by any crypto system that supports the standard.

## Greater than 16 domains support

IBM z17 ME1 supports up to 85 LPARs. The IBM Z crypto architecture supports 16 domains, which matched the LPAR maximum at the time. Before z13 systems, crypto workload separation can be complex in customer environments where the number of LPARs was larger than 16. These customers mapped a large set of LPARs to a small set of crypto domains.

Starting with IBM z14, the IBM Z crypto architecture can support up to 256 domains in an adjunct processor (AP) with the AP extended addressing (APXA) facility that is installed. As such, the Crypto Express adapters are enhanced to handle 256 domains.

The IBM Z firmware provides up to 85 domains for IBM z17 to customers (to match the current LPAR maximum). Customers can map individual LPARs to unique crypto domains or continue to share crypto domains across LPARs.

The following requirements must be met to support 85 domains:

- Hardware: IBM z17 ME1 and Crypto Express8S, or Crypto Express7S
- Operating systems:
  - z/OS:
    - New ICSF support is required to administer a CEX8 coprocessor by using a TKE workstation because of the exploitation of quantum algorithms. Otherwise, workloads run on IBM z17 without requiring ICSF support.
    - Exploitation of new function is supplied in ICSF PTFs on z/OS V2.2 V2.4 (Web deliverable HCR77D1) or V2.5 (base, which is HCR77D2).



- When using new Quantum Safe Algorithms and sharing a KDS in a sysplex, ensure all ICSF PTFs are installed on all systems.

**Tip:** All supported levels of ICSF automatically detect what hardware cryptographic capabilities are available where it is running; then, it enables functions accordingly. No toleration of new hardware is necessary. If you want to use new capabilities, ICSF support is necessary.

- z/VM Version 7.3 or newer for guest use.

## **Payment Card Industry-HSM certification**

PCI standards are developed to help ensure security in the PCI. PCI defines their standards as a set of security standards that is designed to ensure that all companies that accept, process, store, or transmit credit card information is maintained a secure environment.

Compliance with the PCI-HSM standard is valuable for customers, particularly those customers who are in the banking and finance industry. This certification is important to clients for the following fundamental reasons:

- ▶ Compliance is increasingly becoming mandatory
- ▶ The requirements in PCI-HSM make the system more secure.

### ***Industry requirements for PCI-HSM compliance***

The PCI organization cannot require compliance with its standards. Compliance with PCI standards is enforced by the payment card brands, such as Visa, MasterCard, American Express, JCB International, and Discover.

If you are a bank, acquirer, processor, or other participant in the payment card systems, the card brands can impose requirements on you if you want to process their cards. One set of requirements they are increasingly enforcing is the PCI standards.

The card brands work with PCI in developing these standards, and they focused first on the standards they considered most important, particularly the PCI Data Security Standard (PCI-DSS). Some of the other standards were written or required later, and PCI-HSM is one of the last standards to be developed. In addition, the standards themselves were increasing the strength of their requirements over time. Some requirements that were optional in earlier versions of the standards are now mandatory.

In general, the trend is for the card brands to enforce more of the PCI standards and to enforce them more rigorously. The trend in the standards is to impose more and stricter requirements in each successive version. The net result is that companies subject to these requirements can expect that they eventually must comply with all of the requirements.

### ***Improved security through use of PCI-HSM***

PCI-HSM was developed primarily to improve security in payment card systems. It imposes requirements in key management, HSM API functions, and device physical security. It also controls during manufacturing and delivery, device administration, and several other areas. It prohibits many things that were in common use for many years, but are no longer considered secure.

The result of these requirements is that applications and procedures often must be updated because they used some of the things that are now prohibited. Although this issue is inconvenient and imposes some costs, it does increase the resistance of the systems to attacks of various kinds. Updating a system to use PCI-HSM compliant HSMs is expected to reduce the risk of loss for the institution and its clients.



The following requirements must be met to use PCI-HSM:

- ▶ Hardware: IBM z17<sup>6</sup> systems and Crypto Express8S, or Crypto Express7S
- ▶ Operating systems:
  - z/OS - ICSF Web deliverable 19 (HCR77D1), unless otherwise noted. WD19 supports z/OS V2R4, V2R5, V3R1.
  - WD 20 supports z/OS V2R5 (base, which is HCR77D2)
  - z/VM Version 7.3 or newer for guest use

## Visa Format Preserving Encryption

VFPE refers to a method of encryption in which the resulting cipher text features the same form as the input clear text. The form of the text can vary according to use and application.

One of the classic examples is a 16-digit credit card number. After VFPE is used to encrypt a credit card number, the resulting cipher text is another 16-digit number. This process helps older databases contain encrypted data of sensitive fields without having to restructure the database or applications.

VFPE allows customers to add encryption to their applications in such a way that the encrypted data can flow through their systems without requiring a massive redesign of their application. In our example, if the credit card number is VFPE-encrypted at the point of entry, the cipher text still behaves as a credit card number. It can flow through business logic until it meets a back-end transaction server that can VFPE-decrypt it to get the original credit card number to process the transaction.

**Note:** VFPE technology forms part of Visa, Inc.'s, Data Secure Platform (DSP). The use of this function requires a service agreement with Visa. You must maintain a valid service agreement with Visa when you use DSP/FPE.

## AES PIN support for the German banking industry

The German banking industry organization, DK, defined a new set of PIN processing functions to be used on the internal systems of banks and their servers. CCA is designed to support the functions that are essential to those parts of the German banking industry that are governed by DK requirements. The functions include key management support for new AES key types, AES key derivation support, and several DK-specific PIN and administrative functions.

This support includes PIN method APIs, PIN administration APIs, new key management verbs, and new access control points support that is needed for DK-defined functions.

### ***Support for the updated German Banking standard (DK)***

Update support requires ICSF WD19 (HCR77D1) for z/OS V2R2, V2R3, and V2R4 or higher.

## PKA Translate UDX function into CCA

UDX is custom code that allows the client to add unique operations or extensions to the CCA firmware. Specific UDX functions are integrated into the base CCA code over time to accomplish the following tasks:

- ▶ Remove headaches and challenges that are associated with UDX management and currency.
- ▶ Make available popular UDX functions to a wider audience to encourage adoption.

<sup>6</sup> Always check the latest information about security certification status for your specific model.



UDX is integrated into the base CCA code to support translating an external RSA CRT key into new formats. These formats use tags to identify key components. Depending on which new rule array keyword is used with the PKA Key Translate callable service, the service TDES encrypts those components in CBC or ECB mode. In addition, AES CMAC support is delivered.

### Verb Algorithm Currency

Verb Algorithm Currency is a collection of CCA verb enhancements that are related to customer requirements, with the intent of maintaining currency with cryptographic algorithms and standards. It is also intended for customers who want to maintain the following latest cryptographic capabilities:

- ▶ Secure key support AES GCM encryption
- ▶ Key Check Value (KCV) algorithm for service CSNBKYT2 Key Test 2
- ▶ Key derivation options for CSNDEDH EC Diffie-Hellman service

## 6.5.3 Crypto Express8S as an EP11 coprocessor

A Crypto Express8S adapter that is configured in Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode provides PKCS #11 secure key support for public sector requirements. Before EP11, the ICSF PKCS #11 implementation supported only clear keys.

In EP11, keys can now be generated and securely wrapped under the EP11 Master Key. The secure keys never leave the secure coprocessor boundary decrypted.

The secure IBM Enterprise PKCS #11 (EP11) coprocessor runs the following tasks:

- ▶ Encrypt and decrypt (AES, DES, TDES, and RSA)
- ▶ Sign and verify (DSA, RSA, and ECDSA)
- ▶ Generate keys and key pairs (DES, AES, DSA, ECC, and RSA)
- ▶ HMAC (SHA1, SHA2 or SHA3 [SHA224, SHA256, SHA384, and SHA512])
- ▶ Digest (SHA1, SHA2 or SHA3 [SHA224, SHA256, SHA384, and SHA512])
- ▶ Wrap and unwrap keys
- ▶ Random number generation
- ▶ Get mechanism list and information
- ▶ Attribute values
- ▶ Key Agreement (Diffie-Hellman)

**Note:** The function extension capability through UDX is not available to the EP11.

When defined in EP11 mode, the TKE workstation is required to manage the Crypto Express features.

Enterprise PKCS #11 (EP11) with IBM z16 provided the following updates:

- ▶ Quantum Safe Algorithm enhancements provide:
  - Updated support for Dilithium signatures:
    - Round 2: Level 2 (6 5) and 3 (8 7)
    - Round 3: Level 2 (4 4), 3 (6 5) and 5 (8 7)
  - Add support for Kyber key encapsulation: Round 2: Level 3 (768) and 5 (1024)

- ▶ Quantum Safe protected key support for EP11

Host Firmware and EP11 now use a hybrid scheme that combines ECDH and Kyber to accomplish a quantum safe transport key exchange for protected key import.

- ▶ Quantum Safe host firmware management support for EP11



Host Firmware and EP11 now use a hybrid scheme for authenticating management functions that are started from the SE/HMC.

- ▶ EP11 for all of CEX8S (5.9.x), and CEX7S (4.9.x)):
  - Support for HSM backed Hierarchical Deterministic Wallets for Bitcoin (BIP 0032 and SLIP 0010)
  - Hash collision resistant Schnorr signature scheme BSI TR 03111, two variants:
    - Plain BSI TR 03111
    - With compressed keys and signing party's public key as extra input
  - Support for Edwards and Montgomery elliptic curves: EdDSA (Ed25519, Ed448) and ECDH (X25519, X448) (8s and 7s)
  - RSA OAEP with SHA 2 and SHA 3 (8s and 7s only)
  - Extensive IBM Cloud Crypto support:
    - Domains fully manageable by clients without cloud admins assistance
    - Do not Disturb: Actively prohibit cloud administrators from domain management
  - HSM internal re encrypt support for block-based cipher modes
- ▶ EP11 for CEX8S (5.9.x) only
  - Three new compliance modes: FIPS2021, FIPS2024, and Administrative FIPS2021 (first of its kind)
  - Enhanced concurrent update support now includes kernel modules
  - Enhanced maximum performance for digest and random number generation
  - Allow for regular extractable keys to be tagged as protected key exportable

Enterprise PKCS #11 (EP11) with IBM z17 provides the following updates for Quantum Safe Algorithms:

- ▶ Standardized support conforming to official NIST specifications
- ▶ Support for Module-Lattice-Base (ML, aka CRYSTALS) cryptography:
  - ML-KEM (Key-Encapsulation Mechanism, FIPS 203), Key Sizes: 512, 768, 1024
  - ML-DSA (Digital Signature Algorithm, FIPS 204), Strength Levels: (4,4), (6,5), (8,7)

If required for compliance reasons, the FIPS 140-2 certified version of EP11 can be used for selected Crypto Express8S and Express7S adapters.

For Ethereum support, three enabling enhancements for EP11 are available for Crypto Express8S and Express7S adapters on z17:

- ▶ Pairing-based BLS signature support
- ▶ BLS12-381 pairing-friendly elliptic curve
- ▶ EIP2333 for deterministic hierarchical wallets

#### 6.5.4 Crypto Express8S as an accelerator

A Crypto Express8S adapter that is running in accelerator mode supports only RSA clear key and SSL Acceleration. A request is processed fully in hardware.

The Crypto Express accelerator is a coprocessor that is reconfigured by the installation process so that it uses only a subset of the coprocessor functions at a higher speed.



Reconfiguration is disruptive to coprocessor and accelerator operations. The coprocessor or accelerator must be deactivated before you begin the reconfiguration.

FIPS 140-2 certification is not relevant to the accelerator because it operates with clear keys only. The function extension capability through UDX is not available to the accelerator.

The functions that remain available when the Crypto Express8S feature is configured as an accelerator are used for the acceleration of modular arithmetic operations. That is, the RSA cryptographic operations are used with the SSL/TLS protocol. The following operations are accelerated:

- ▶ PKA Decrypt (CSNDPKD) with PKCS-1.2 formatting
- ▶ PKA Encrypt (CSNDPKE) with zero-pad formatting
- ▶ Digital Signature Verify

The RSA encryption and decryption functions support key lengths of 512 - 4,096 bits in the Modulus-Exponent (ME) and Chinese Remainder Theorem (CRT) formats.

### 6.5.5 Managing Crypto Express8S

Each cryptographic coprocessor has 85 physical sets of registers or queue registers, which corresponds to the maximum number of LPARs that are running on an IBM z17 ME1, which is also 85. Each of these sets belongs to the following domains:

- ▶ A cryptographic domain index, in the range of 0 - 84 for IBM z17 ME1, is allocated to a logical partition in its image profile. The same domain also must be allocated to the ICSF instance that is running in the logical partition that uses the Options data set.
- ▶ Each ICSF instance accesses only the Master Keys or queue registers that correspond to the domain number that is specified in the logical partition image profile at the SE and in its Options data set. Each ICSF instance sees a logical cryptographic coprocessor that consists of the physical cryptographic engine and the unique set of registers (the domain) that is allocated to this logical partition.

The installation of CP Assist for Cryptographic Functions (CPACF) DES/TDES enablement (FC 3863) is required to use the Crypto Express8S feature.

Each Crypto Express8S FC 0908 includes two IBM 4770 PCIe Cryptographic Coprocessors (PCIeCC), which is a hardware security module (HSM); FC 0909 includes one IBM 4770 PCIeCC. The adapters are available in the following configurations:

- ▶ IBM Enterprise Common Cryptographic Architecture (CCA) Coprocessor (CEX8C)
- ▶ IBM Enterprise Public Key Cryptography Standards #11 (PKCS) Coprocessor (CEX8P)
- ▶ IBM Crypto Express8S Accelerator (CEX8A)

During the feature installation, the PCI-X adapter is configured by default as the CCA coprocessor.

The configuration of the Crypto Express8S adapter as EP11 coprocessor requires a TKE workstation (FC 0057/0058) with TKE 10.1 (FC 0883) LIC. The same requirement applies to CCA mode for a full PCI-compliant environment.

The Crypto Express8S feature does not use CHPIDs from the channel subsystem pool. However, the Crypto Express8S feature requires one slot in a PCIe I/O drawer, and one PCHID for each PCIe cryptographic adapter.



For enabling an LPAR to use a Crypto Express8S adapter, the following cryptographic resources in the image profile must be defined for each partition:

- ▶ Usage domain index
- ▶ Control domain index
- ▶ PCI Cryptographic Coprocessor Candidate List
- ▶ PCI Cryptographic Coprocessor Online List

This task is accomplished by using the Customize/Delete Activation Profile task, which is in the Operational Customization Group, from the HMC or from the SE. Modify the cryptographic initial definition from the Crypto option in the image profile, as shown in Figure 6-6.

**Important:** After this definition is modified, any change to the image profile requires a DEACTIVATE and ACTIVATE of the logical partition for the change to take effect. Therefore, this cryptographic definition is disruptive to a running system.

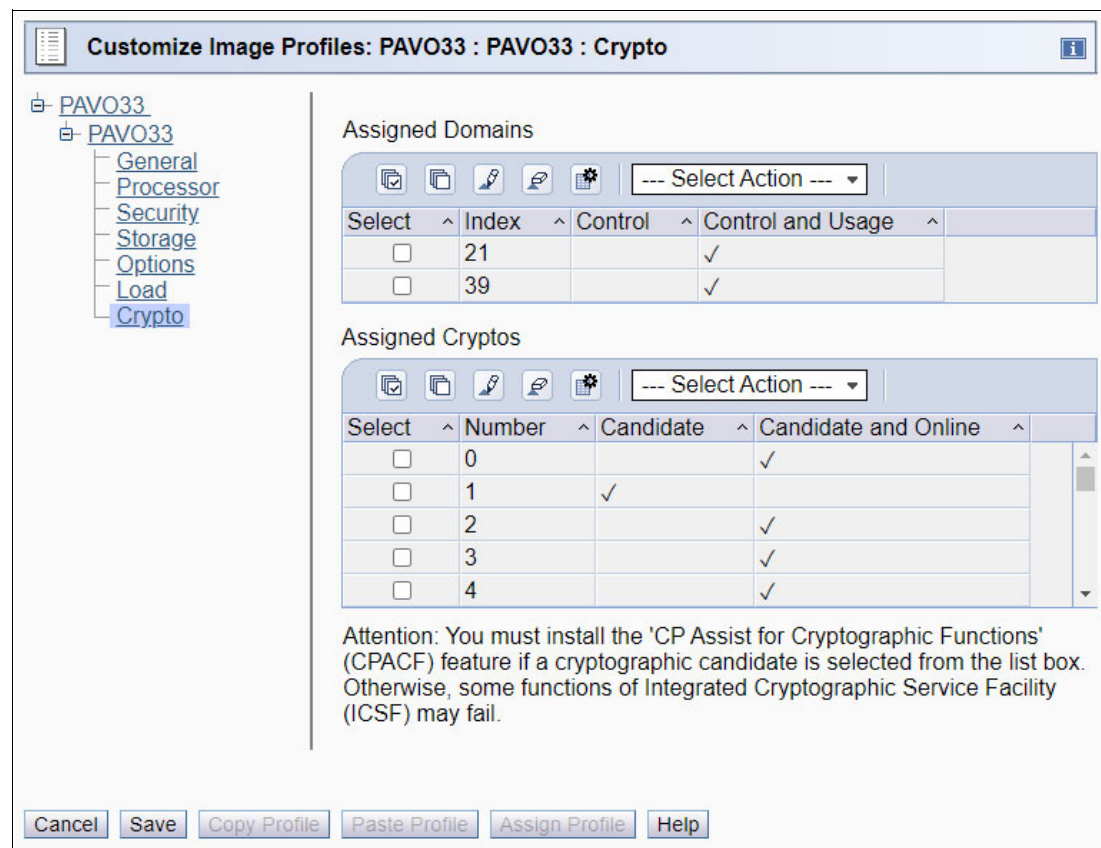


Figure 6-6 Customize Image Profiles: Crypto

The following cryptographic resource definitions are used:

- ▶ Control Domain

Identifies the cryptographic coprocessor domains that can be administered from this logical partition if it is set up as the TCP/IP host for the TKE.

If you are setting up the host TCP/IP in this logical partition to communicate with the TKE, the partition is used as a path to other domains' Master Keys. Indicate all the control



domains that you want to access (including this partition's own control domain) from this partition.

- ▶ **Control and Usage Domain**

Identifies the cryptographic coprocessor domains that are assigned to the partition for all cryptographic coprocessors that are configured on the partition. The usage domains cannot be removed if they are online. The numbers that are selected must match the domain numbers that are entered in the Options data set when you start this partition instance of ICSF.

The same usage domain index can be used by multiple partitions, regardless to which CSS they are defined. However, the combination of PCIe adapter number and usage domain index number must be unique across all active partitions.

- ▶ **Cryptographic Candidate list**

Identifies the cryptographic coprocessor numbers that can be accessed by this logical partition. From the list, select the coprocessor numbers (in the range 0 - 15) that identify the PCIe adapters to be accessed by this partition.

- ▶ **Cryptographic Online list**

Identifies the cryptographic coprocessor numbers that are automatically brought online during logical partition activation. The numbers that are selected in the online list must also be part of the candidate list.

After they are activated, the active partition cryptographic definitions can be viewed from the HMC. Select the CPC, and click **View LPAR Cryptographic Controls** in the CPC Operational Customization window. The resulting window displays the definition of Usage and Control domain indexes, and PCI Cryptographic candidate and online lists, as shown in Figure 6-7 on page 252.



**View LPAR Cryptographic Controls - PAVO**

Installed Crypto Express8S: 00 01 02 03 04 05 06 07

Cryptographic Candidates

Partition	Active	Crypto Numbers	Conflicts
PAVO31	Yes		
PAVO32	Yes		
PAVO33	Yes	0-4	
PAVO34	Yes		
PAVO35	Yes		
PAVO36	Yes		
PAVO37	No		
PAVO38	Yes		
PAVO39	Yes		
PAVO41	No		
PAVO42	No		
PAVO43	No		

Usage Domain Indexes

Partition	Active	Indexes	Conflicts
PAVO31	Yes		
PAVO32	Yes		
PAVO33	Yes	21, 39	
PAVO34	Yes		
PAVO35	Yes		
PAVO36	Yes		
PAVO37	No		
PAVO38	Yes		
PAVO39	Yes		
PAVO41	No		
PAVO42	No		
PAVO43	No		

Close Refresh Help

**Summary**

- PAVO01
- PAVO02
- PAVO3A
- PAVO31
- PAVO32
- PAVO33
- PAVO34
- PAVO35
- PAVO36
- PAVO38
- PAVO39

Figure 6-7 View LPAR Cryptographic Controls

Operational changes can be made by using the Change LPAR Cryptographic Controls task, which reflects the cryptographic definitions in the image profile for the partition. With this function, the cryptographic feature can be added and removed dynamically, without stopping a running operating system.

For more information about the management of Crypto Express8S, see *IBM z17 Configuration Setup*, SG24-8960.



## 6.6 Trusted Key Entry workstation

The TKE workstation is an optional feature that offers key management functions. It can be a TKE tower workstation (FC 0058) or TKE rack-mounted workstation (FC 0057) for IBM z17 systems to manage Crypto Express8S, or Crypto Express7S.

The TKE contains a combination of hardware and software. A mouse, keyboard, flat panel display, PCIe adapter, and a writable USB media to install the TKE Licensed Internal Code (LIC) are included with the system unit. The TKE workstation requires an IBM 4770 crypto adapter.

A TKE workstation is part of a customized solution for the use of the Integrated Cryptographic Service Facility for z/OS (ICSF for z/OS) or Linux for IBM Z. This program provides a basic key management system for the cryptographic keys of an IBM z17 system that has Crypto Express features installed.

The TKE provides a secure, remote, and flexible method of providing Master Key Part Entry, and to remotely manage PCIe cryptographic coprocessors. The cryptographic functions on the TKE run by one PCIe cryptographic coprocessor. The TKE workstation communicates with the IBM Z system through a TCP/IP connection. The TKE workstation is available with Ethernet LAN connectivity only. Up to 10 TKE workstations can be ordered.

TKE FCs 0057 and 0058 can be used to control any supported Crypto Express feature on IBM z17, IBM z16, IBM z15 systems, and the Crypto adapters on older, still supported systems.

The TKE 10.1 LIC (FC 0883) feature requires a 4770 HSM. The following features are supported:

- ▶ Managing the Crypto Express8S HSMs (CCA normal mode, CCA PCI mode, and EP11)
- ▶ Quantum Safe Cryptography (QSC) used when:
  - TKE authenticates Crypto Express8S HSMs
  - Deriving a Transport Key between TKE's HSM and target Crypto Express8S HSM
  - On-demand HSM dual validation check.
- ▶ Domain groups limitations All HSM in group must:
  - Support QSC (can include Crypto Express8S HSMs only)
  - Not support QSC (cannot include Crypto Express8S HSMs)
- ▶ Configuration migration tasks support:
  - Can collect and apply data to a Crypto Express8S HSM
  - Can apply data from a pre-Crypto Express8S HSM.
- ▶ New default wrapping method for the Crypto Express8S HSM
- ▶ New AES DUKPT key attribute on AES DKYGENKY parts

**Tip:** For more information about handling a TKE, see the [TKE Introduction video](#).



## 6.6.1 Logical partition, TKE host, and TKE target

If one or more LPARs are configured to use Crypto Express coprocessors, the TKE workstation can be used to manage DES, AES, ECC, and PKA master keys. This management can be done for all cryptographic domains of each Crypto Express coprocessor feature that is assigned to the LPARs that are defined to the TKE workstation.

Each LPAR in the same system that uses a domain that is managed through a TKE workstation connection is a TKE host or TKE target. An LPAR with a TCP/IP connection to the TKE is referred to as the *TKE host*; all other partitions are *TKE targets*.

The cryptographic controls that are set for an LPAR through the SE determine whether the workstation is a TKE host or a TKE target.

## 6.6.2 Optional smart card reader

An optional smart card reader (FC 0886) can be added to the TKE workstation. One FC 0886 includes two smart card readers, two cables to connect them to the TKE workstation, and 20 smart cards. The reader supports the use of smart cards that contain an embedded microprocessor and associated memory for data storage. The memory can contain the keys to be loaded into the Crypto Express features. These readers can be used only with smart cards that have applets that were loaded from a TKE 8.1 or later. These cards are FIPS certified.

Smart card readers from FC 0885 or FC 0891 can be carried forward. Smart cards can be used on TKE 10.1 with these readers. Access to and use of confidential data on the smart card are protected by a user-defined PIN. Up to 990 other smart cards can be ordered for backup (the extra smart card feature code is FC 0900). When one feature code is ordered, 10 smart cards are included. The order increment is 1 - 99 (10 - 990 blank smart cards).

If smart cards with applets that are not supported by the new smart card reader are reused, new smart cards on TKE 8.1 or later must be created and the content from the old smart cards to the new smart cards must be copied. The new smart cards can be created and copied on a TKE 8.1 system. If the copies are done on TKE 9.0, the source smart card must be placed in an older smart card reader from feature code 0885 or 0891.

A new smart card for the Trusted Key Entry (TKE) allows stronger Elliptic Curve Cryptography (ECC) levels. More TKE Smart Cards (FC 0900, packs of 10, FIPS certified blanks) require TKE 9.1 LIC or up.

## 6.6.3 TKE hardware support and migration information

The new TKE 10.1 LIC (FC 0883) is originally shipped with a new IBM z17 server. The following TKE workstations can be ordered with a new IBM z17:

- ▶ TKE 10.1 tower workstation (FC 0058)
- ▶ TKE 10.1 rack-mounted workstation (FC 0057)

**Note:** Several options for ordering the TKE with or without ordering Keyboard, Mouse, and Display are available. Ask your IBM Representative for more information about which option is the best option for you.



The TKE 10.1 LIC requires the 4770 crypto adapter. The TKE 9.x and TKE 8.x workstations can be upgraded to the TKE 10.1 tower workstation by purchasing a 4770 crypto adapter.

The Omnikey Cardman 3821 smart card readers can be carried forward to any TKE 10.0 workstation. Smart cards 45D3398, 74Y0551, and 00JA710 can be used on TKE 10.1.

When performing a MES upgrade from TKE 8.x, or TKE 9.x to a TKE 10.1 installation, the following steps must be completed:

1. Save Upgrade Data on an old TKE to USB memory to save client data.
2. Replace the TKE crypto adapter with the 4770 crypto adapter.
3. Upgrade the firmware to TKE 10.1.
4. Install the Frame Roll to apply Save Upgrade Data (client data) to the TKE 10.1 system.
5. Run the TKE Workstation Setup wizard.

### TKE upgrade considerations

The TKE LIC does *not* need to be upgraded if you are migrating your configuration with Crypto Express7S and TKE Release 9.x to an IBM z17.

**Note:** If your IBM z17 includes only Crypto Express7S, you can use TKE V9.2, which requires the 4768 cryptographic adapter.

For more information about TKE hardware support, see Table 6-3. For some functions, requirements must be considered; for example, the characterization of a Crypto Express adapter in EP 11 mode always requires the use of a TKE.

Table 6-3 TKE Compatibility Matrix

TKE workstation	TKE Release LIC	8.1	9.0	9.1	9.2	10.0	10.1
	HW Feature Code	0847 or 0097	0085 or 0086	0085 or 0086	0087 or 0088	0057 or 0058	0057 or 0058
	LICC	0878	0879	0880	0881	0882	0883
	Smart Card Reader	0891	0891	0891	0891	0891	0886
	Smart Card	0892	0892	0892	0892	0900	0889
Manage Host Crypto Module	CEX8C (CCA)	No	No	No	No	Yes	Yes
	CEX8P (EP11)	No	No	No	No	Yes	Yes
	CEX7C (CCA)	No	No	No	Yes	Yes	Yes
	CEX7P (EP11)	No	No	No	Yes	Yes	Yes
	CEX6C (CCA)	No	Yes	Yes	Yes	Yes	Yes
	CEX6P (EP11)	No	Yes	Yes	Yes	Yes	Yes

**Attention:** The TKE is unaware of the CPC type where the host crypto module is installed. That is, the TKE does not consider whether a Crypto Express is running on IBM z17, IBM 16, or IBM z15 system. Therefore, the LIC can support any CPC where the coprocessor is supported, but the TKE LIC must support the specific crypto module.



## 6.7 Cryptographic functions comparison

The functions or attributes on IBM z17 for the two cryptographic hardware features are listed in Table 6-4, where “X” indicates that the function or attribute is supported.

Table 6-4 Cryptographic functions on IBM z17

Functions or attributes	CPACF	CEX8C	CEX8P	CEX8A
Supports z/OS applications that use CSF	X	X	X	X
Supports Linux on IBM Z CCA applications	X	X	-	X
Encryption and decryption by using secret-key algorithm	-	X	X	-
Provides the highest SSL/TLS handshake performance	-	-	-	X
Supports SSL/TLS functions	X	X	X	X
Provides the highest symmetric (clear key) encryption performance	X	-	-	-
Provides the highest asymmetric (clear key) encryption performance	-	-	-	X
Provides the highest asymmetric (encrypted key) encryption performance	-	X	X	-
Nondisruptive process to enable <sup>a</sup>	-	X	X	X
Requires IOCDS definition	-	-	-	-
Uses CHPID numbers	-	-	-	-
Uses PCHIDs (one PCHID)	-	X	X	X
Requires CPACF enablement (FC 3863) <sup>b</sup>	X	X	X	X
Requires ICSF to be active	-	X	X	X
Offers UDX	-	X	-	-
Usable for data privacy: Encryption and decryption processing	X	X	X	-
Usable for data integrity: Hashing and message authentication	X	X	X	-
Usable for financial processes and key management operations	-	X	X	-
Crypto performance IBM RMF monitoring	-	X	X	X
Requires system master keys to be loaded	-	X	X	-
System (master) key storage	-	X	X	-
Retained key storage	-	X	-	-
Tamper-resistant hardware packaging	-	X	X	X <sup>c</sup>
Designed for FIPS 140-2 Level 4 certification	-	X	X	X



Functions or attributes	CPACF	CEX8C	CEX8P	CEX8A
Supports Linux applications that perform SSL handshakes	-	-	-	X
RSA functions	-	X	X	X
High-performance SHA-1, SHA-2, and SHA-3	X	X	X	-
Clear key DES or triple DES	X	-	-	-
Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys	X	X	X	-
True random number generator (TRNG)	X	X	X	-
Deterministic random number generator (DRNG)	X	X	X	-
Pseudo random number generator (PRNG)	X	X	X	-
Clear key RSA	-	-	-	X
Payment Card Industry (PCI) PIN Transaction (PTS) Hardware Security Module (HSM) PCI-HSM		X	X	
Europay, MasterCard, and Visa (EMV) support	-	X	-	-
Public Key Decrypt (PKD) support for Zero-Pad option for clear RSA private keys	-	X	-	-
Public Key Encrypt (PKE) support for Mod_Raised_to Power (MRP) function	-	X	X	-
Remote loading of initial keys in ATM	-	X	-	-
Improved key exchange with non-CCA systems	-	X	-	-
ISO 16609 CBC mode triple DES message authentication code (MAC) support	-	X	-	-
AES GMAC, AES GCM, AES XTS mode, CMAC	-	X	-	-
SHA-2, SHA-3 (384,512), HMAC	X	X	-	-
Visa Format Preserving Encryption	-	X	-	-
AES PIN support for the German banking industry	-	X	-	-
ECDSA (192, 224, 256, 384, 521 Prime/NIST)	X	X	-	-
ECDSA (160, 192, 224, 256, 320, 384, 512 BrainPool)	X	X	-	-
ECDH (192, 224, 256, 384, 521 Prime/NIST)	X	X	-	-
ECDH (160, 192, 224, 256, 320, 384, 512 BrainPool)	X	X	-	-

- a. To make adding the Crypto Express features nondisruptive, the logical partition must be predefined with the suitable PCI Express cryptographic adapter number. This number must be selected from its candidate list in the partition image profile.
- b. This feature is not required for Linux if only RSA clear key operations are used. DES or triple DES encryption requires CPACF to be enabled.
- c. This feature is physically present, but is not used when configured as an accelerator (clear key only).



## 6.8 Cryptographic operating system support for IBM z17

In this section, we provide an overview of the operating systems requirements in relation to cryptographic elements.

### 6.8.1 Crypto Express8S Toleration

Crypto Express8S (0908/0909) Toleration treats Crypto Express8S cryptographic coprocessors and accelerators as Crypto Express5 coprocessors and accelerators. The following minimum prerequisites must be met:

- ▶ z/OS 3.1; z/OS V2.5; specific functions require WD 20 (HCR77D2)
- ▶ z/VM V7.3 and V7.4 for guest use
- ▶ VSE<sup>n</sup> V6.3.1 – 21st Century Software
- ▶ z/TPF V1.1 with PTFs
- ▶ Linux on IBM Z: IBM is working with its Linux distribution partners to provide support by way of maintenance or future releases for the following distributions:
  - ▶ SUSE SLES 16.1 (Post GA)
  - ▶ SUSE SLES 15.6 (GA)
  - ▶ SUSE SLES 12.5 (Post GA)
  - ▶ Red Hat RHEL 10.0 (Post GA)
  - ▶ Red Hat RHEL 9.4
  - ▶ Red Hat RHEL 8.10
  - ▶ Red Hat RHEL 7.9 (Post GA)
  - ▶ Canonical Ubuntu 24.04 LTS (Post GA)
  - ▶ Canonical Ubuntu 22.04 LTS (Post GA)
  - ▶ Canonical Ubuntu 20.04 LTS (Post GA)
  - ▶ Red Hat Enterprise Linux (RHEL) 8 and Red Hat Enterprise Linux 7
  - ▶ Ubuntu 16.04 LTS (or higher)

The KVM hypervisor, which is offered supported Linux distributions. For more information about minimal and recommended distribution levels, see the [Tested platforms for Linux web page](#) of the IBM IT infrastructure website.

### 6.8.2 Crypto Express8S support of VFPE

The following minimum prerequisites must be met to use this element:

- ▶ z/OS V2.5, z/OS V2.4 with PTFs
- ▶ z/VM V7.4 for guest use
- ▶ z/VM V7.3 for guest use

Linux on IBM Z:

- ▶ See 6.8.1, “Crypto Express8S Toleration” on page 258
  - The support statements for IBM z17 also cover the KVM hypervisor on distribution levels that have KVM support.

For more information about the minimum required and recommended distribution levels, see the [IBM Z website](#).



### 6.8.3 Crypto Express8S support of greater than 16 domains

The following prerequisites must be met to support more than 16 domains:

- ▶ z/OS 3.1
- ▶ z/OS V2.5
- ▶ z/OS V2.4 with PTFs
- ▶ z/VM V7.3 and V7.4 for guest use
- ▶ VSE<sup>n</sup> V6.3.1 (21<sup>st</sup> Century Software)

Linux on IBM Z:

- ▶ See 6.8.1, “Crypto Express8S Toleration” on page 258

For more information about the minimum that is required and recommended distribution levels, see the [IBM Z website](#).

For more information about the software support levels for cryptographic functions, see Chapter 7, “Operating systems support” on page 261.









# Operating systems support

This chapter describes the minimum operating system requirements and support considerations for the IBM z17™ servers and their features. It addresses IBM z/OS, z/VM, z/TPF, Linux on IBM Z, VSE<sup>n</sup> V6.3.1 (21<sup>st</sup> Century Software), and the KVM hypervisor.

Because this information is subject to continuous updating, for the most current information check the hardware fix categories for IBM z17 Model ME1: `IBM.Device.Server.IBM.z17-9175.*`.

Support of IBM z17 functions depends on the operating system and version and release.

This chapter includes the following topics:

- ▶ 7.1, “Operating systems summary” on page 262
- ▶ 7.2, “Support by operating system” on page 263
- ▶ 7.3, “IBM z17 features and function support overview” on page 268
- ▶ 7.4, “Support by features and functions” on page 281
- ▶ 7.5, “z/OS migration considerations” on page 336
- ▶ 7.6, “z/VM migration considerations” on page 342
- ▶ 7.7, “VSE<sup>n</sup> migration considerations” on page 347
- ▶ 7.8, “Software licensing” on page 348
- ▶ 7.9, “References” on page 350



## 7.1 Operating systems summary

The minimum operating system levels that are required on IBM z17 servers are listed in Table 7-1.

**End of service operating systems:** Operating system levels that are no longer in service are not covered in this publication.

Table 7-1 IBM z17 minimum operating systems requirements

Operating systems	Supported Version and release on IBM z17 <sup>a</sup>
z/OS	V2R4 <sup>b</sup> , V2R5, V3R1
z/VM	V7R3, V7R4
z/TPF	V1R1
21 <sup>st</sup> Century Software VSE <sup>n c</sup>	V6.3.1
Linux on IBM Z <sup>d</sup>	See Table 7-2 on page 267

a. Service is required.

b. z/OS V2R4 - Toleration mode only. The IBM Software Support Services for z/OS V2R4 offer provides the ability for customers to purchase extended defect support service for z/OS V2.R4.

c. VSE<sup>n</sup> V6.3.1 is supported by 21<sup>st</sup> Century Software.

d. KVM hypervisor is supported by Linux distribution partners.

The use of specific features depends on the operating system. In all cases, program temporary fixes (PTFs) might be required with the operating system level that is indicated. Check the z/OS fix categories, or the subsets of the 9175DEVICE (IBM z17 ME1), and PSP buckets for z/VM.

**Important:** Starting with z17 Model PSP buckets are no longer available. See the following IBM Support Web page:

<https://www.ibm.com/support/pages/psp-bucket-information-ibm-z-products>

The fix categories are continuously updated. They contain the latest information about maintenance, and contain installation information, hardware and software service levels, service guidelines, and cross-product dependencies.

For more information about Linux on IBM Z distributions and KVM hypervisor, see the Linux distributor's support information.



## 7.2 Support by operating system

IBM z17 systems has introduced several new functions. This section describes the support of those functions by the current operating systems. Also included are some of the functions that were introduced with previous IBM Z server generations and carried forward or enhanced in support of IBM z17 servers.

Features and functions that are available on previous servers, but no longer supported by IBM z17 servers, are *not* documented here.

For more information about supported functions that are based on operating systems, see 7.3, “IBM z17 features and function support overview” on page 268. Tables are built by function and feature classification to help you determine, by a quick scan, what is supported and the minimum operating system level that is required.

### 7.2.1 z/OS

z/OS Version 2 Release 4 is the earliest release that supports IBM z17 servers. Consider the following points:

- ▶ Service support for z/OS Version 2 Release 4 is a fee-based extension for defect support (for up to three years), and can be obtained by ordering IBM Software Support Services - Service Extension for z/OS V2.R4.
- ▶ IBM z17 capabilities differ depending on the z/OS release. Toleration support is provided on z/OS V2R4. Basic exploitation is provided by z/OS V2R5 and exploitation support is provided on z/OS V3R1 and later only<sup>1</sup>.

#### How to get the latest fix information for z/OS systems

For the latest information about z/OS PTFs that apply to the IBM z17, consult the Fix Categories (FIXCATs).

Fixes are grouped into the following categories (for more information about IBM Fix Categories, see this IBM Support [web page](#)):

- ▶ Base support is provided by PTFs identified by:  
`IBM.Device.Server.z17-9175.RequiredService`  
Fixes that are required to run z/OS on the IBM z17 servers and must be installed before migration.
- ▶ Use of many functions is provided by PTFs identified by:  
`IBM.Device.Server.z17-9175.Exploitation`  
Fixes that are required to use the capabilities of the IBM z17. They must be installed only if you use the function.
- ▶ Recommended service is identified by:  
`IBM.Device.Server.z17-9175.RecommendedService`  
Fixes that are recommended to run z/OS on the IBM z17. These fixes also are listed in the Recommended Service section of the hardware PSP bucket. They represent fixes that were recommended by IBM Service. It is recommended that you review and install these PTFs.

---

<sup>1</sup> Use support for select features by way of PTFs. Toleration support for new hardware might also require PTFs.



All information needed to upgrade z/OS to support the IBM z17 is provided in the z/OS IBM z17 Upgrade Workflow.

For more information about supported functions and their minimum required support levels, see 7.3, “IBM z17 features and function support overview” on page 268.

## 7.2.2 z/VM

IBM z17 support is provided with PTFs for z/VM 7.3 and 7.4 including PTFs for IOCP, HCD, and HLASM<sup>2</sup>. IBM z17 support is requires PTFs for z/VM 7.3 and z/VM 7.4<sup>3</sup>.

### **Compatibility:**

- z/VM support for host / guests on IBM z17 at the z16 functional level with limited exploitation of new functions (some transparent, where applicable)
- Support available as PTFs applicable concurrently at IBM z17 general availability
- Includes PTFs for IOCP, HCD, and HLASM
- IBM z17 support is provided with PTFs for z/VM 7.3 and 7.4

### **Enable Guest Exploitation for the following new facilities:**

- Vector-Enhancements Facility 3:
  - new instructions intended to provide performance improvements
- Vector-Packed-Decimal-Enhancement 3:
  - Intended to provide performance improvements of COBOL programs when compiled using the NUMCHECK option to detect and convert data.
- Workload-Instrumentation:
  - provides a means of classifying and sampling workloads to enhance the z/OS pricing model.
- Message-Security-Assist Extensions:
  - enhancements which allow use of XTS and HMAC algorithms and allow for generation of XTS and HMAC encryption keys while using AES algorithms
- Reduced support for TX:
  - Non-constrained transactions will result in unconditionally aborting with a CC1 set and no TDB stored.
- Perform Lock Operation (PLO):
  - provides operations for managing locks in storage to replace capabilities previously provided by the constrained-transactional-execution facility (CTX)
- Concurrent-Functions:
  - provides new instructions intended to replace use of TX for software serialization

### **Guest Exploitation support for the following new features:**

- Network Express Adapter Hybrid (NETH<sup>4</sup>) and Direct (NETD) Virtual Function support:
  - allows guests to directly exploit RoCE functionality of the Network Express adapter;
- Networking Express Adapter EQDIO<sup>5</sup> OSA Hybrid (OSH) CHPID support;
  - allows guests to directly exploit OSH functionality of the Network Express adapter, and allows guests to exploit OSD simulated devices via the z/VM VSwitch to a real OSH device
- AI Accelerator Adapter:
  - allow guests to take advantage of and exploit the capabilities of the enhanced IBM Spyre Adapter

<sup>2</sup> IBM Z HLASM - High Level Assembler Language

<sup>3</sup> For z/VM 7.4 GA date - September 2024. See the [IBM z/VM continuous delivery page](#).

<sup>4</sup> Provides RoCE and TCP/IP support over the same Network Express adapter, no separate RoCE Express card required

<sup>5</sup> EQDIO - Enhanced Queued Direct Input/Output



### ***z/VM IBM z17 ME1 Host Support***

Compatibility Support; Host support for the following new facilities:

- Network Express Adapter EQDIO Support within the z/VM VSwitch:
  - Allows customers to configure the VSwitch to take advantage of lower latency and higher bandwidths provided by networking EQDIO devices within their datacenter. The VSwitch EQDIO exploitation includes QDIO to EQDIO translation allowing guests which do not support EQDIO to directly take advantage of this networking support.<sup>6</sup>
- Power Consumption metrics provided within z/VM monitor:
  - enhancements to the z/VM monitor to include entire CPC and LPAR specific level power consumption information. This information includes power metrics for CPU, I/O, and Memory usage. Consumers of z/VM monitor, such as the z/VM Performance Datapump<sup>7</sup>, can be enhanced to calculate/approximate guest level apportionment.
- CPU-Measurement Facility (CPU-MF enhancements):
  - provides CPU-MF specific support for IBM z17
- Data Processing Unit Next Generation I/O accelerator Instrumentation provided within z/VM monitor:
  - collect instrumentation data within z/VM monitor for the new IBM z17 I/O complex
- Dynamic I/O support and guest exploitation for the following:
  - 25G LR for Long Distance Coupling with CHPID type CL6 (Dynamic I/O support only; No guest exploitation)
  - Network Express Adapter CHPID type OSH and PCI Function Types NETH and NETD
  - IBM Spyre AI Accelerator Adapter PCI Function Type PAIA

z/VM Compatibility Support enables guest use for several additional facilities:

- ▶ Embedded Artificial Intelligence Acceleration
  - Designed to reduce the overall time required to execute CPU operations for neural networking processing functions, and help support real-time applications, such as fraud detection.
- ▶ Compliance-ready CPACF Counters support
  - Provides a means for guests to track crypto compliance and instruction usage.
- ▶ Breaking Event Address Register (BEAR) Enhancement Facility;
  - Facilitates debugging wild branches.
- ▶ Vector Packed Decimal Enhancements 2
  - New instructions intended to provide performance improvements.
- ▶ Reset DAT protection Facility
  - Provides a more efficient way to disable DAT protection, such as during copy-on-write or page change tracking operations.
- ▶ RoCE Express3 adapter
  - Allows guests to use Routable RoCE, Zero Touch RoCE, and SMC-R V2 support.
- ▶ Guest Enablement for the CEX8S crypto adapter and assorted crypto enhancements
  - Includes Quantum Safe API Guest Exploitation Support that is available to dedicated guests.
- ▶ CPU/Core topology location information within z/VM monitor data
  - Provides a better picture of the system for diagnostic and tuning purposes.
- ▶ Consolidated Boot Loader for guest IPL from SCSI

The following IBM z17 support is not apparent to z/VM:

<sup>6</sup> A z/VM VSwitch supporting Network Express OSH does not currently support z/OS guests exploiting an EQDIO uplink port. In the interim, clients will be required to use either a guest-attached OSH device or existing functionality available with OSA-Express7S adapters.

<sup>7</sup> <https://www.ibm.com/docs/en/zvm/7.4?topic=performance-zvm-data-pump>



- ▶ Dynamic Partition Mode (DPM) enhancements SMC-R, SMC-D
- ▶ Dynamic Partition Mode (DPM) FICON CTC support for LPARs running on the same CPC
  - FICON CTC support is a prerequisite for establishing z/VM Single System Image (SSI) clustering. DPM FICON CTC supports FCTC within a single CEC but not cross CECs.e CPC. SSI Clustering technology is required in order for customers to exploit Live Guest Relocation (LGR) to move running Linux guests across z/VM images.
- ▶ OSA-Express7S 1.2 GbE, 10 GbE, 1000BASE-T, OSA-Express7S 1.1 (25 GbE) Adapters
- ▶ Coupling Express3 LR Adapter
- ▶ 32 Gbps 2-port FICON Adapter
- ▶ 32 Gbps 4-port FICON Adapter
- ▶ Coupling facility scalability and performance improvements

For more information about supported functions and their minimum required support levels, see 7.3, “IBM z17 features and function support overview” on page 268.

### 7.2.3 z/TPF

IBM z17 support is provided by z/TPF V1R1 with PTFs. For more information about supported functions and their minimum required support levels, see 7.3, “IBM z17 features and function support overview” on page 268.

### 7.2.4 VSE<sup>n</sup>

IBM z17 VSE support is provided by VSE<sup>n</sup> V6.3.1 – 21st Century Software and later, with the following considerations:

- ▶ VSE runs in z/Architecture mode only
- ▶ VSE supports 64-bit real and virtual addressing

For more information about supported functions and their minimum required support levels, see 7.3, “IBM z17 features and function support overview” on page 268.

### 7.2.5 21<sup>st</sup> Century Software VSE<sup>n</sup> V6.3.1

21<sup>st</sup> Century Software VSE<sup>n</sup> V6.3 was announced in March 2022 and is based on an IBM licensed copy of IBM z/VSE. For more information about this product, visit see the 21<sup>st</sup> Century Software [website](#). VSE<sup>n</sup> V6.3.1 is the required version for the IBM z17.

### 7.2.6 Linux on IBM Z

Generally, a new machine is not apparent to Linux on IBM Z. For IBM z17, toleration support is required for the following functions and features:

- ▶ IPL in “z/Architecture” mode
- ▶ Crypto Express8S cards
- ▶ New Network Express Adapter (RoCE for Linux replacement)
- ▶ 8-byte LPAR offset



The service levels of SUSE, Red Hat, and Ubuntu releases that are supported at the time of this writing are listed in Table 7-2.

Table 7-2 Linux on IBM Z distributions

Linux on IBM Z distribution <sup>a</sup>	Supported Version and Release on IBM z17 <sup>b</sup>
SUSE Linux Enterprise Server	15 SP6 and later
SUSE Linux Enterprise Server	12 SP5 <sup>c</sup>
Red Hat RHEL	10.0
Red Hat RHEL	9.4
Red Hat RHEL	8.1 <sup>c</sup> and later with service
Red Hat RHEL	7.9 <sup>c</sup> with service
Ubuntu	24.04 LTS
Ubuntu	22.04 <sup>c</sup> LTS
Ubuntu	20.04.1 LTS <sup>c</sup>
KVM Hypervisor <sup>d</sup>	Offered with the supported Linux distributions.

a. Only z/Architecture (64-bit mode) is supported. IBM testing identifies the minimum required level and the recommended levels of the tested distributions.

b. Fixes required for toleration.

c. Maintenance is required.

d. For more information about minimal and recommended distribution levels, see this [Linux on IBM Z website](#).

For more information about supported Linux distributions on IBM Z servers, see the [Tested platforms for Linux page](#) of the IBM IT infrastructure website.

IBM is working with Linux distribution Business Partners to provide further use of selected IBM z17 functions in future Linux on IBM Z distribution releases.

Consider the following guidelines:

- ▶ Use SUSE Linux Enterprise Server 15, Red Hat RHEL 9, or Ubuntu 22.10 LTS or newer in any new projects for IBM z17 servers.
- ▶ Update any Linux distribution to the latest service level before migrating to IBM z17 servers.
- ▶ Adjust the capacity of any Linux on IBM Z and z/VM guests, in terms of the number of IFLs and CPs, real or virtual, according to the PU capacity of the IBM z17 servers.

## 7.2.7 KVM hypervisor

KVM is offered through our Linux distribution partners to help simplify delivery and installation. Linux and KVM are provided from a single source. With KVM being included in the Linux distribution, ordering and installing KVM is easier.

For more information about KVM support, see [IBM Z website](#).



## 7.3 IBM z17 features and function support overview

The following tables list the IBM z17 features and functions and their minimum required operating system support levels:

- ▶ Table 7-3 on page 269
- ▶ Table 7-4 on page 270
- ▶ Table 7-5 on page 272
- ▶ Table 7-6 on page 272
- ▶ Table 7-7 on page 273
- ▶ Table 7-8 on page 275
- ▶ Table 7-9 on page 276
- ▶ Table 7-10 on page 279
- ▶ Table 7-11 on page 280

Information about Linux on IBM Z refers exclusively to the suitable distributions of SUSE, Red Hat, and Ubuntu.

The tables in this section list but do not specifically mark all the features that require fixes that are required by the corresponding operating system for toleration or use.

All tables use the following conventions:

- ▶ **Y**: The function is supported.
- ▶ **N**: The function is not supported.
- ▶ **na** : The function is not applicable to that specific operating system.



### 7.3.1 Supported CPC functions

The supported Base CPC Functions or z/OS and z/VM are listed in Table 7-3.

**Note:** Consider the following points:

- ▶ In a future IBM Z hardware system family, the transactional execution and constrained transactional execution facility will NO longer be supported. Users of the facility on current servers must always check the facility indications before use.
- ▶ z/OS V2R4 - No new function is provided for the use of the new hardware features (toleration support only). Although extended (fee-based) support for z/OS V2.R4 can be obtained, support for z/OS V2.R4 is not covered extensively in this document.

Table 7-3 Supported Base CPC Functions or z/OS and z/VM

Function <sup>a</sup>	z/OS V3R1	z/OS V2R5	z/OS V2R4	z/VM V7R4	z/VM V7R3
IBM z17 servers	Y	Y	Y	Y	Y
Maximum processor unit (PUs) per system image	208	200	200	80 <sup>b</sup>	80 <sup>b</sup>
Maximum main storage size	16 TB	16 TB	4 TB	4 TB	4 TB <sup>c</sup>
Dynamic PU add	Y	Y	Y	Y	Y
Dynamic LPAR memory add	Y	Y	Y	Y	Y
Dynamic LPAR memory removal	Y	Y	Y	Y	Y
LPAR group absolute capping	Y	Y	Y	Y	Y
Capacity Provisioning Manager	Y	Y	Y	N	N
Program-directed re-IPL	Y	na	na	Y	Y
Transactional Execution <sup>d</sup>	Y	Y	Y	Y	Y <sup>e</sup>
Java Exploitation of Transactional Execution	Y	Y	Y	Y <sup>d</sup>	Y <sup>d</sup>
Simultaneous multithreading (SMT)	Y	Y	Y	Y	Y
Single Instruction Multiple Data (SIMD)	Y	Y	Y	Y	Y
2 GB large page support	Y	Y	Y	N	N
Large page (1 MB) support	Y	Y	Y	Y <sup>e</sup>	Y <sup>e</sup>
Db2 exploitation of IBM Z zAIU	Y	Y <sup>f</sup>	Y <sup>f</sup>	na	na
CPUMF (CPU measurement facility) for IBM z17	Y	Y	Y	Y	Y
Flexible Capacity	Y	Y	Y	Y	Y
IBM Virtual Flash Memory (VFM)	Y	Y	Y	N	N
1 MB pageable large pages	Y	Y	Y	N	N
Guarded Storage Facility (GSF)	Y	Y	Y	Y <sup>e</sup>	Y <sup>e</sup>
Instruction Execution Protection (IEP)	Y	Y	Y	Y <sup>e</sup>	Y <sup>e</sup>
Co-processor Compression Enhancements (CMPSC)	Y	Y	Y <sup>f</sup>	N	N



Function <sup>a</sup>	z/OS V3R1	z/OS V2R5	z/OS V2R4	z/VM V7R4	z/VM V7R3
IBM Integrated Accelerator for zEDC (on-chip compression)	Y	Y	Y	Y	Y
CPU MF Extended Counters	Y	Y	Y	Y	Y
CF level 26 Enhancements	Y	Y <sup>h</sup>	Y <sup>h</sup>	na	na
HiperDispatch Optimization	Y	Y	Y	Y <sup>e</sup>	Y <sup>e</sup>
System Recovery Boost Enhancements <sup>g</sup>	Y	Y <sup>h</sup>	Y <sup>h</sup>	N/A	N/A
IBM Integrated Accelerator for Z SORT	Y	Y	Y	Y <sup>e</sup>	Y <sup>e</sup>
ICSF Enhancements	Y	Y	Y	na	na
Support for Breaking Event Address Register (BEAR) Enhancement Facility	Y	Y	Y	Y	Y
Support for Reset DAT protection Facility	Y	Y	Y	Y	Y
zDNN library enablement for IBM Z Integrated Accelerator for AI	Y	Y	Y <sup>h</sup>	na	na

- a. PTFs might be required for toleration support or use of IBM z17 features and functions.
- b. 80-way without multithreading; 40-way with multithreading enabled.
- c. With Service.
- d. Limited support, as per Statement of Direction: In a future IBM Z hardware system family, the transactional execution and constrained transactional execution facility will no longer be supported. Users of the facility on current servers should always check the facility indications before use.
- e. Guest Exploitation support.
- f. With PTFs for use.
- g. Same enhancements as for z16; System Recovery Boost (SRB) Upgrade record offering is not offered on z17.
- h. With Required PTFs

The supported base CPC functions for VSE<sup>n</sup> V6.3.1, z/TPF, and Linux on IBM Z are listed in Table 7-4.

Table 7-4 Supported base CPC functions for z/VSE, z/TPF, and Linux on IBM Z

Function <sup>a</sup>	z/TPF V1R1	VSE <sup>n</sup> V6.3.1 <sup>b</sup>	Linux on IBM Z <sup>c</sup>
IBM z17 servers	Y	Y	Y
Maximum processor unit (PUs) per system image	86	10	208 <sup>d</sup>
Maximum main storage size	32 TB	32 GB	16 TB <sup>e</sup>
Dynamic PU add	N	Y	Y
Dynamic LPAR memory upgrade	N	N	Y
LPAR group absolute capping	N	Y	N
Program-directed re-IPL	N	Y	Y
HiperDispatch	Y	N	Y
IBM Z Integrated Information Processors (zIIPs)	N	N	N
Java Exploitation of Transactional Execution	N	N	Y
Simultaneous multithreading (SMT)	N	N	Y <sup>f</sup>



Function <sup>a</sup>	z/TPF V1R1	VSE <sup>n</sup> V6.3.1 <sup>b</sup>	Linux on IBM Z <sup>c</sup>
Single Instruction Multiple Data (SIMD)	N	Y	Y
Hardware decimal floating point <sup>g</sup>	N	N	Y
2 GB large page support	Y	N	Y
Large page (1 MB) support	Y	Y	Y
CPUMF (CPU measurement facility) for IBM z17	Y	N	Y <sup>h,i</sup>
AI accelerator exploitation	N	N	Y <sup>j</sup>
IBM Virtual Flash Memory (VFM)	N	N	Y
Guarded Storage Facility (GSF)	N	N	Y
Instruction Execution Protection (IEP)	N	N	Y
System Recovery Boost	Y <sup>k</sup>	Y <sup>k</sup>	N
Secure Boot (code integrity check)	na	na	Y <sup>l</sup>
Secure Execution Support for Linux	na	na	Y <sup>m</sup>
IBM Integrated Accelerator for zEDC (on-chip compression)	Y	N	Y <sup>n</sup>
IBM Integrated Accelerator for Z SORT	N	N	N

a. PTFs might be required for toleration support or exploitation of IBM z17 features and functions.

b. VSE<sup>n</sup> V6.3.1 - 21st Century Software: <https://21cs.com/vsen/>

c. Support statement varies based on Linux on IBM Z distribution and release.

d. For SLES12/RHEL7/Ubuntu 16.04 and later, Linux kernel supports 256 cores without SMT and 128 cores with SMT (= 256 threads).

e. IBM z17 ME1 supports defining up to 32 TB per LPAR (OS support is required).

f. On IFL only (not for CPs allocated to Linux LPARs).

g. Packed decimal conversion support.

h. Limited support - IBM is working with its Linux distribution Business Partners to provide this feature.

i. <https://www.ibm.com/docs/en/linux-on-systems?topic=p-cpu-measurement-facilities-1>

j. Delivered with Linux distributions as a new package: libzdnn.

k. Subcapacity CP speed boost (no zIIP boost).

l. For SCSI IPL.

m. For second-level guests that are running under KVM.

n. Requires Linux kernel exploitation support for gzip/zlib compression.



### 7.3.2 Coupling and clustering

The supported coupling and clustering functions for z/OS and z/VM are listed in Table 7-5.

**Note:** z/OS V2R4 support ended as of September 2024. No new function is provided for the use of the new hardware features (toleration support only). Extended (fee-based) support for z/OS V2.R4 can be ordered from IBM.

Table 7-5 Supported coupling and clustering functions for z/OS and z/VM

Function <sup>a</sup>	z/OS V3R1	z/OS V2R5	z/OS V2R4	z/VM V7R4	z/VM V7R3
CFCC Level 26 <sup>b</sup>	Y	Y	Y	Y	Y
CFCC Level 25 <sup>c</sup>	na	Y	Y	Y	Y
CFCC Level 24 <sup>d</sup>	na	Y	Y	Y	Y
CFCC Level 25 Coupling Facility Processor Scalability Enhancements	na <sup>e</sup>	Y	Y	Y	Y
RMF coupling channel reporting	Y	Y	Y	Y	N
Asynchronous CF Duplexing for lock structures	Y	Y	Y	Y	Y
Cache residency time metrics <sup>f</sup>	Y	Y	Y	Y	N
Dynamic I/O activation for stand-alone CF <sup>g</sup> and for stand-alone Linux on Z and z/TPF <sup>h</sup> CPCs	Y	Y	Y	Y	Y

a. PTFs are required for toleration support or use of IBM z17 features and functions.

b. CFCC Level 26 with Driver 61 (IBM z17)

c. CFCC Level 25 with Driver 51 (IBM z16).

d. CFCC Level 24 with Driver 41 (IBM z15).

e. Processor scalability enhancements apply to CFCC Level 26

f. With APAR OA60650.

g. Requires HMC 2.14.1(Driver 36) or newer and various OS fixes (HCD, HCM, IOS, IOCP)

h. Requires 2.4 or higher z/OS partition with APAR OA65559 applied, be running on an IBM z17, and Linux on Z and z/TPF also running on IBM z16. Both z17 CPCs require proper firmware level. This configuration continues to support stand-alone CF Dynamic I/O activations.

In addition to operating system support that is listed in Table 7-5, Server Time Protocol is supported on z/TPF V1R1 and Linux on IBM Z. Also, CFCC Level 23, Level 24, Level 25, and Level 26 are supported for z/TPF V1R1.

### 7.3.3 Storage connectivity

The supported storage connectivity functions for z/OS and z/VM are listed Table 7-6.

Table 7-6 Supported storage connectivity functions for z/OS and z/VM

Function <sup>a</sup>	z/OS R3V1	z/OS V2R5	z/OS V2R4	z/VM V7R4	z/VM V7R3
zHyperLink Read Support for Db2 and VSAM		Y	Y	N	N
zHyperLink Write Support for Db2 logs		Y	Y	N	N
zHyperLink Writes support for asynchronous mirroring <sup>b</sup>		Y	Y <sup>b</sup>	N	N



Function <sup>a</sup>	z/OS R3V1	z/OS V2R5	z/OS V2R4	z/VM V7R4	z/VM V7R3
zHyperLink Consistent read from metro mirror secondary	Y	Y <sup>c</sup>	Y <sup>b</sup>	N	N
z/VM Dynamic I/O support for FICON FC and FCP CHPIDs	na	na	na	Y	Y
<b>CHPID (Channel-Path Identifier) type FC</b>					
FICON support for zHPF (IBM Z High-Performance FICON)	Y	Y	Y		Y <sup>e</sup>
IBM Fibre Channel Endpoint Security <sup>d</sup>	Y	Y	Y	Y	Y
FICON when using CTC (channel-to-channel)	Y	Y	Y	Y <sup>e</sup>	Y <sup>e</sup>
IPL from an alternative subchannel set	Y	Y	Y	N	N
32 K subchannels for FICON Express	Y	Y	Y	Y	Y
Request node identification data (RNID)	Y	Y	Y	N	N
FICON link incident reporting	Y	Y	Y	N	N
<b>CHPID (Channel-Path Identifier) type FCP</b>					
FICON Express support of SCSI devices	na	na	na	Y	Y
FICON Express support of hardware data router	na	na	na	Y <sup>f</sup>	Y <sup>f</sup>
FICON Express support of T10 Data Integrity Field (DIF)	na	na	na	Y <sup>f</sup>	Y <sup>f</sup>
N_Port ID Virtualization (NPIV)	na	na	na	Y	Y
Worldwide port name tool	na	na	na	Y	Y

a. PTFs might be required for toleration support or exploitation of IBM z17 features and functions.

b. DS8900F only, 9.1 release with z/OS PTFs, does not include XRC.

c. DS8900F only, 9.2 release with z/OS PTFs.

d. FC 1146 (optional) is available for IBM z17 ME1, IBM z16 A01 and IBM z15 T01 only. Minimum, DS8910 or DS8890 storage, CPACF enablement and platform supported FICON Express adapters. This feature requires HMC 2.17.1 integration with IBM Guardium Security Key Lifecycle Manager (GSKLM).

e. CTC channel type within the same CEC is supported when CPC is managed in DPM mode.

f. For guest use

The supported storage connectivity functions for VSE<sup>n</sup> V6.3.1<sup>n</sup>, z/TPF, and Linux on IBM Z are listed in Table 7-7.

Table 7-7 Supported storage connectivity functions for z/TPF, and Linux on IBM Z

Function <sup>a</sup>	z/TPF V1R1	VSE <sup>n</sup> V6.3.1 <sup>b</sup>	Linux on IBM Z <sup>c</sup>
The 63.75-K subchannels	Y	N	Y
Six logical channel subsystems (LCSSs)	N	Y	Y
Four subchannel set per LCSS	N	Y	Y
<b>CHPID (Channel-Path Identifier) type FC</b>			
FICON Express support of zHPF (IBM Z High-Performance FICON) <sup>d</sup>	Y	Y	Y
IBM Fibre Channel Endpoint Security <sup>e</sup>	Y <sup>f</sup>	Y <sup>f</sup>	Y <sup>f</sup>
MIDAW (Modified Indirect Data Address Word)	N	N	N



Function <sup>a</sup>	z/TPF V1R1	VSE <sup>n</sup> V6.3.1 <sup>b</sup>	Linux on IBM Z <sup>c</sup>
FICON Express support for CTC (channel-to-channel)	Y	Y	Y <sup>g</sup>
IPL from an alternative subchannel set	N	N	N
32 K subchannels for FICON Express	N	N	Y
Request node identification data (RNID)	N	N	N
<b>CHPID (Channel-Path Identifier) type FCP</b>			
FICON Express support of SCSI devices	na	Y	Y
FICON Express support of hardware data router	N	N	Y
FICON Express support of T10 Data Integrity Field (DIF)	N	N	Y
N_Port ID Virtualization (NPIV)	N	Y	Y
Worldwide port name tool	na	na	Y

a. PTFs might be required for toleration support or exploitation of IBM z17 features and functions.

b. VSEn V6.3.1 - 21st Century Software: <https://21cs.com/vsen/>

c. Support statement varies based on Linux on IBM Z distribution and release.

d. Transparent to operating systems.

e. FC 1146 (optional) is available for IBM z17 ME1, IBM z16 A01 and IBM z15 T01 only.  
Minimum, DS8910 or DS8890 storage, CPACF enablement and FICON Express16SA or FICON Express32S LX/SX.

f. Feature is operating system-independent (that is transparent to the operating system); operating system support is needed only for displaying, configuration, and monitoring (fixes might be required).

g. CTC channel type is now supported when CPC is managed in DPM mode.



### 7.3.4 Network connectivity

The supported network connectivity functions for z/OS and z/VM are listed in Table 7-8.

**As per z16 Statements of Direction<sup>a</sup>, consider the following points:**

- ▶ IBM z17 does NOT support the OSE CHPID type.
- ▶ IBM z17 does NOT support OSA Express 1000BASE-T hardware adapters.
- ▶ z/OS V2R4 support ended as of September 2024. No new function is provided for the use of the hardware features (toleration support only). Extended (fee-based) support for z/OS V2.R4 can be ordered from IBM.

a. Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM.

Table 7-8 Supported network connectivity functions for z/OS and z/VM

Function <sup>a</sup>	z/OS V3R1	z/OS V2R5	z/OS V2R4	z/VM V7R4	z/VM V7R3
Checksum offload for IPV6 packets	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
Checksum offload for LPAR-to-LPAR traffic with IPv4 and IPv6	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
QDIO data connection isolation for z/VM	na	na	na	Y	Y
QDIO interface isolation for z/OS	Y	Y	Y	na	na
QDIO OLM (Optimized Latency Mode)	Y	Y	Y	na	na
QDIO Diagnostic Synchronization	Y	Y	Y	N	N
IWQ (Inbound Workload Queuing) for OSA	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
GARP VLAN Registration Protocol	Y	Y	Y	Y	Y
Link aggregation support for z/VM	na	na	na	Y	Y
Multi-vSwitch Link Aggregation	na	na	na	Y	Y
Large send for IPV6 packets	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
z/VM Dynamic I/O Support for OSA-Express OSD CHPIDs	na	na	na	Y	Y
OSA Dynamic LAN idle	Y	Y	Y	N	N
OSA Layer 3 virtual MAC for z/OS environments	Y	Y	Y	na	na
Network Traffic Analyzer	Y	Y	Y	N	N
<b>HiperSockets</b>					
HiperSockets <sup>c</sup>	Y	Y	Y	Y	Y
HiperSockets Completion Queue	Y	Y	Y	Y	Y
HiperSockets Virtual Switch Bridge	na	na	na	Y	Y
HiperSockets Multiple Write Facility	Y	Y	Y	N	N
HiperSockets support of IPV6	Y	Y	Y	Y	Y
HiperSockets Layer 2 support	Y	Y	Y	Y	Y



Function <sup>a</sup>	z/OS V3R1	z/OS V2R5	z/OS V2R4	z/VM V7R4	z/VM V7R3
<b>SMC-D and SMC-R</b>					
RoCE Network Express Adapter Hybrid (NETH) <sup>d</sup> Virtual Function support.	Y	Y <sup>h</sup>	N	Y <sup>e</sup>	Y <sup>e</sup>
SMC-D <sup>f</sup> over ISM (Internal Shared Memory)	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
SMC-R over 25 GbE and 10 GbE Network Express (NETH)	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
z/VM Dynamic I/O support for Network Express	na	na	na	Y	Y
Shared Network Express environment	Y	Y	Y	Y	Y
<b>Network Express Adapter (OSH CHPID<sup>g</sup>)</b>					
Network Express 10GbE and 25GbE (using Enhanced QDIO architecture - EQDIO) CHPID type OSH	Y	Y <sup>h</sup>	N	Y <sup>i</sup>	Y <sup>i</sup>
<b>Open Systems Adapter (OSA)<sup>j,k</sup></b>					
OSA-Express7S 1.2 25 GbE CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express7S 25-Gigabit Ethernet Short Reach (SR and SR1.1) CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express7S 1.2 10 GbE CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express7S 10-Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express7S 1.2 GbE CHPID type OSD and OSC	Y	Y	Y	Y	Y
OSA-Express7S GbE CHPID type OSD and OSC	Y	Y	Y	Y	Y

- a. PTFs might be required for toleration support or use of IBM z17 features and functions.
- b. For guest use or exploitation.
- c. On IBM z17, the CHPID statement of HiperSockets devices requires the keyword VCHID. If you are migrating from a zEC12 or earlier, the IOCP definitions must be migrated to support the HiperSockets definitions (from CHPID type IQD). VCHID specifies the virtual channel identification number that is associated with the channel path (valid range is 7C0 - 7FF).
- d. NETH FID requires OSH CHPID defined over the same PCHID.
- e. Allows guests to directly exploit RoCE functionality of the Network Express adapter.
- f. Shared Memory Communications - Direct Memory Access.
- g. A z/VM VSwitch supporting Network Express OSH does not currently support z/OS guests exploiting an EQDIO uplink port. In the interim, clients will be required to use either a guest-attached OSH device or existing functionality available with OSA-Express7S adapters.
- h. With PTFs.
- i. Guests can exploit OSD simulated devices via the z/VM VSwitch to a real OSH device
- j. CHPID types OSM, OSN, OSX, and OSE are no longer supported.
- k. OSA Express 1000BASE-T hardware adapters are not supported on IBM z17.

The supported network connectivity functions for VSE<sup>n</sup> V6.3.1, z/TPF, and Linux on IBM Z are listed in Table 7-9.

Table 7-9 Supported network connectivity functions for VSE<sup>n</sup> V6.3.1, z/TPF, and Linux on IBM Z

Function <sup>a</sup>	z/TPF V1R1	VSE <sup>n</sup> V6.3.1 <sup>b</sup>	Linux on IBM Z <sup>c</sup>
Checksum offload for IPV6 packets	N	N	Y



Function <sup>a</sup>	z/TPF V1R1	VSE <sup>n</sup> V6.3.1 <sup>b</sup>	Linux on IBM Z <sup>c</sup>
Checksum offload for LPAR-to-LPAR traffic with IPv4 and IPv6	N	N	Y
QDIO Diagnostic Synchronization	N	N	N
Inbound Workload Queuing (IWQ) for OSA	N	N	N
GARP VLAN Registration Protocol	N	N	Y <sup>d</sup>
Multi-vSwitch Link Aggregation	N	N	N
Large send for IPV6 packets	N	N	Y
OSA Dynamic LAN idle	N	N	N
<b>HiperSockets</b>			
HiperSockets <sup>e</sup>	N	Y	Y
HiperSockets Completion Queue	N	Y	Y
HiperSockets Virtual Switch Bridge	na	na	Y <sup>f</sup>
HiperSockets support of IPV6	N	Y	Y
HiperSockets Layer 2 support	N	Y	Y
HiperSockets Network Traffic Analyzer for Linux on IBM Z	N	N	Y
<b>SMC-D and SMC-R</b>			
RoCE Network Express Adapter Direct (NETD) Virtual Function support	N	N	N <sup>g</sup>
SMC-D <sup>h</sup> over ISM (Internal Shared Memory)	N	N	Y <sup>i</sup>
Shared Network Express environment (NETD FID)	N	N	Y
<b>Network Express Adapter (OSH)</b>			
Network Express 10GbE and 25GbE (EQDIO) CHPID type OSH	Y	N	N <sup>j</sup>
<b>Open Systems Adapter (OSA)</b>			
OSA-Express7S 1.2 25-GbE CHPID type OSD	Y	Y	Y
OSA-Express7S 25-Gigabit Ethernet Short Reach (SR and SR1.1) CHPID type OSD	Y	Y	Y
OSA-Express7S 1.2 10-GbE CHPID type OSD	Y	Y	Y
OSA-Express7S 10-Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y
OSA-Express7S 1.2 GbE CHPID type OSD and OSC	Y	Y	Y
OSA-Express7S GbE CHPID type OSD and OSC	Y	Y	Y

a. PTFs might be required for toleration support or use of IBM z17 features and functions.

b. VSE<sup>n</sup> V6.3 - 21st Century Software: <https://21cs.com/vsen/>

c. Support statement varies based on Linux on IBM Z distribution and release.

d. By using VLANs.



- e. On IBM z17 as for z16, the CHPID statement of HiperSockets devices requires the keyword VCHID. Therefore, the IOCP definitions must be migrated to support the HiperSockets definitions (CHPID type IQD). VCHID specifies the virtual channel identification number that is associated with the channel path (valid range is 7C0 - 7FF). VCHID is not valid on IBM Z servers before IBM z13.
- f. Applicable to guest operating systems.
- g. Support pending Linux Distribution adoption
- h. Shared Memory Communications - Direct Memory Access.
- i. SMC-R and SMC-D are supported on Linux kernel; see:  
<https://linux-on-z.blogspot.com/p/smc-for-linux-on-ibm-z.html>
- j. Linux support for NETD Virtual Function pending distribution partners adoption and testing. NETD FID supports both NETH and ETH (TCP/IP) Linux drivers. No need for OSA/QDIO driver.



### 7.3.5 Cryptographic functions

The IBM z17 supported cryptography functions for z/OS and z/VM are listed in Table 7-10.

**Note:** No new function is provided for leveraging the new HW features (toleration support only). Although extended (fee-based) support for z/OS V2.R4 can be obtained, support for z/OS V2.R4 is not covered extensively in this document.

Table 7-10 Supported cryptography functions for z/OS and z/VM

Function <sup>a</sup>	z/OS V3R1	z/OS V2R5	z/OS V2R4	z/VM V7R4	z/VM V7R3
CP Assist for Cryptographic Function (CPACF)	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
Support for 85 Domains	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
CPACF support for: ► Encryption (DES, TDES, AES) ► Hashing (SHA-1, SHA-2, SHA-3, SHAKE) ► Random Number Generation (PRNG, DRNG, TRNG)	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
Crypto Express8S	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
Crypto Express7S	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
Crypto Express7S Support for Visa Format Preserving Encryption	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
Crypto Express7S Support for Coprocessor in PCI-HSM Compliance Mode <sup>c</sup>	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
Elliptic Curve Cryptography (ECC)	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode	Y	Y	Y	Y <sup>b</sup>	Y <sup>b</sup>
z/OS Data Set Encryption	Y	Y	Y	na	na
z/VM Encrypted paging support	na	na	na	Y	Y
RMF Support for Crypto Express8, Express7, and Express6	Y	Y	Y	na	na
z/OS SMF Enhancements for CPACF	Y <sup>d</sup>	Y <sup>d</sup>	Y <sup>d</sup>	na	na
z/OS encryption readiness technology (zERT)	Y	Y	Y	na	na
ICFSF New Function Support	Y	Y	Y <sup>d</sup>	na	na
Quantum-Safe Cryptography (QSC) for signing and key negotiation	Y	Y	Y <sup>e</sup>	Y	Y

a. PTFs might be required for toleration support or use of IBM z17 features and functions.

b. For guest use or exploitation.

c. Requires TKE 9.1 or newer.

d. Requires z/OS Exploitation support by way of APAR

e. Requires Web deliverable HCR77D1



The IBM z17 supported cryptography functions for VSE<sup>n</sup> V6.3.1 - 21st Century Link Software, z/TPF, and Linux on IBM Z are listed in Table 7-11.

*Table 7-11 Supported cryptography functions for z/TPF, VSE<sup>n</sup> V6.3.1, and Linux on IBM Z*

Function <sup>a</sup>	z/TPF V1R1	VSE <sup>n</sup> V6.3.1 <sup>b</sup>	Linux on IBM Z <sup>c</sup>
CP Assist for Cryptographic Function (CPACF)	Y	Y	Y
Support for 85 Domains	N	Y	Y
CPACF support for: ▶ Encryption (DES, TDES, AES) ▶ Hashing (SHA-1, SHA-2, SHA-3, SHAKE) ▶ Random Number Generation (PRNG, DRNG, TRNG)	Y <sup>de</sup>	Y	Y
CPACF protected key	N	N	N
Crypto Express8S	Y	Y	Y
Crypto Express7S	Y	Y	Y
Crypto Support for Visa Format Preserving Encryption	N	N	N
Crypto Support for Coprocessor in PCI-HSM Compliance Mode	N	N	N
Elliptic Curve Cryptography (ECC)	Y	Y	Y
Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode	N	N	Y
z/TPF transparent database encryption	Y	na	na

a. PTFs might be required for toleration support or use of IBM z17 features and functions.

b. VSE<sup>n</sup> V6.3.1 - 21st Century Software - <https://www.21cs.com/vsen/>

c. Support statement varies based on Linux on IBM Z distribution and release.

d. z/TPF supports only AES-128 and AES-256.

e. z/TPF only supports SHA-1 and the SHA-2 standard for SHA-256, SHA-384 and SHA-512.



## 7.4 Support by features and functions

This section addresses operating system support by function. Only the currently in-support releases are covered.

Tables in this section use the following convention:

- ▶ N/A: Not applicable
- ▶ NA: Not available

### 7.4.1 LPAR Configuration and Management

A single system image can control multiple processor units (PUs), such as CPs, zIIPs, or IFLs.

**Note:** Extended (fee-based) support for z/OS V2.R4 can be obtained.

#### Maximum number of PUs per system image

The maximum number of PUs that is supported by each operating system image and by special-purpose LPARs are listed in Table 7-12.

Table 7-12 Maximum number of PUs per system image

Operating system	Maximum number of PUs per system image
z/OS V3R1	256 <sup>ab</sup>
z/OS V2R5	256 <sup>cd</sup>
z/OS V2R4	256 <sup>cd</sup>
z/VM V7R4	80 <sup>e</sup>
z/VM V7R3	80 <sup>f</sup>
VSE <sup>n</sup> V6.3.1 <sup>g</sup>	VSE Turbo Dispatcher can use up to 4 CPs, and tolerates up to 10-way LPARs
z/TPF V1R1	86 CPs
CFCC Level 26	16 CPs or ICFs (CPs and ICFs cannot be mixed)
Linux on IBM Z	<ul style="list-style-type: none"> <li>▶ SUSE Linux Enterprise Server 12 and later: 256 CPs or IFLs.</li> <li>▶ Red Hat RHEL 7 and later: 256 CPs or IFLs.</li> <li>▶ Ubuntu 20.04.1 LTS and later: 256 CPs or IFLs.</li> </ul>
KVM Hypervisor	The KVM hypervisor is offered with the following Linux distributions -- 256CPs or IFLs--: <ul style="list-style-type: none"> <li>▶ SLES 12 SP5 and later</li> <li>▶ RHEL 7.9 and later</li> <li>▶ Ubuntu 20.04.1 LTS and later</li> </ul>
Secure Service Container	80
GDPS Virtual Appliance	80

a. IBM z17 ME1 LPARs support 208-way without multithreading; 128-way with multithreading (SMT).

b. Total characterizable PUs, including zIIPs and CPs.

c. IBM z17 ME1 LPARs support 208-way without multithreading; 128-way with multithreading (SMT).



- d. Total characterizable PUs, including zIIPs and CPs.
- e. 80-way without multithreading and 40-way with multithreading enabled.
- f. 80-way without multithreading and 40-way with multithreading enabled.
- g. VSE<sup>n</sup> V6.3.1 - 21st Century Software: <https://21cs.com/vsen/>

## Maximum main storage size

The maximum amount of main storage that is supported by current operating systems is listed in Table 7-13. A maximum of 32 TB of main storage can be defined for an LPAR on an IBM z17 server.

Table 7-13 Maximum memory that is supported by the operating system

Operating system	Maximum supported main storage <sup>a</sup>
z/OS	z/OS V3R1 and V2R5 support 16 TB. (Prior z/OS releases support 4 TB)
z/VM	z/VM V7R3 and z/VM V7R4 support 4 TB
VSE <sup>n</sup> V6.3.1 <sup>b</sup>	supports 32 GB
z/TPF	z/TPF supports 32TB
CFCC	Levels 24, 25 and 26 support up to 3 TB
Secure Service Containers	Supports up to 16 TB <sup>a</sup>
Linux on IBM Z (64-bit)	16 TB <sup>a,c</sup>

a. On IBM z17 ME1, LPAR storage definition supports 32 TB (IBM z17 ME1 supports up to 64 TB of customer memory).

b. VSE<sup>n</sup> V6.3.1 - 21st Century Software: <https://21cs.com/vsen/>

c. Support might vary by distribution. Check with your distribution provider.

## IBM z17 Model ME1 - Up to 85 LPARs

The IBM z17 ME1 can be configured with up to 85 LPARs (same as previous models). Because channel subsystems can be shared by up to 15 LPARs, it is necessary to configure six channel subsystems to reach the 85 LPARs limit.

**Remember:** A virtual appliance that is deployed in a Secure Service Container runs in a dedicated LPAR. When activated, it reduces the maximum number of available LPARs by one.

## Dynamic PU add

Planning an LPAR configuration includes defining reserved PUs that can be brought online when extra capacity is needed. Operating system support is required to use this capability without an IPL; that is, nondisruptively. This support is available in z/OS for some time.

The dynamic PU add function enhances this support by allowing you to dynamically define and change the number and type of reserved PUs in an LPAR profile, which removes any planning requirements. The new resources are immediately made available to the operating system and in the case of z/VM, to its guests.

The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270.



## Dynamic LPAR memory upgrade

An LPAR can be defined with an initial and a reserved amount of memory. At activation time, the initial amount is made available to the partition and the reserved amount can be added later, partially or totally. Although these two memory zones do not have to be contiguous in real memory, they appear as logically contiguous to the operating system that runs in the LPAR.

z/OS can take advantage of this support and nondisruptively acquire and release memory from the reserved area. z/VM V7R1 and later can acquire memory nondisruptively and immediately make it available to guests.

z/VM virtualizes this support to its guests, which now also can increase their memory nondisruptively if supported by the guest operating system. Currently, releasing memory from z/VM is supported on z/VM V7.2 with PTFs<sup>8</sup>. Releasing memory from the z/VM guest depends on the guest's operating system support.

Linux on IBM Z also supports acquiring and releasing memory nondisruptively. This feature is enabled for SUSE Linux Enterprise Server 12 and RHEL 7.9 and later releases.

## LPAR group absolute capping

Group absolute capping allows you to limit the amount of physical processor capacity that is used by an individual LPAR when a PU that is defined as a CP or an IFL is shared across a set of LPARs. This facility is designed to provide a physical capacity limit that is enforced as an absolute (versus a relative) limit. It is not affected by changes to the logical or physical configuration of the system. This physical capacity limit can be specified in units of CPs or IFLs.

## Capacity Provisioning Manager

The provisioning architecture enables clients to better control the configuration and activation of the On/Off CoD. For more information, see Chapter 8, "System upgrades" on page 353. This process is inherently more flexible and can be automated. This capability can result in easier, faster, and more reliable management of the processing capacity.

The Capacity Provisioning Manager, which is a feature that was first available with z/OS V1R9, interfaces with z/OS Workload Manager (WLM) and implements capacity provisioning policies. Several implementation options are available, from an analysis mode that issues only guidelines, to an autonomic mode that provides fully automated operations.

Replacing manual monitoring with autonomic management or supporting manual operation with guidelines can help ensure that sufficient processing power is available with the least possible delay. The supported operating systems are listed in Table 7-3 on page 269.

## Program-directed re-IPL

Program directed re-IPL allows an operating system to re-IPL without operator intervention. This function is supported for SCSI and IBM extended count key data (IBM ECKD) devices.

The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270.

## IOCP

All IBM Z servers require a description of their I/O configuration. This description is stored in I/O configuration data set (IOCDs) files. The I/O configuration program (IOCP) allows for the

<sup>8</sup> z/VM Dynamic Memory Downgrade (releasing memory from z/VM LPAR) made available with PTFs for APAR VM66271. For more information, see: <http://www.vm.ibm.com/newfunction/#dmd>



creation of the IOCDs file from a source file that is known as the I/O configuration source (IOCS).

The IOCS file contains definitions of LPARs and channel subsystems. It also includes detailed information for each channel and path assignment, control unit, and device in the configuration.

IOCP for IBM z17 provides support for the following features:

- ▶ IBM z17 Base machine definition
- ▶ PCI function adapter for zHyperLink (HYL)
- ▶ PCI function Network Express adapter (CX6)
- ▶ New hardware (announced with Driver 61)
- ▶ IOCP Dynamic I/O support for stand-alone CF, Linux on Z and z/TPF, running on IBM z16 and IBM z17 CPCs.

**IOCP required level for IBM z17 servers:**

- ▶ z/OS V2.R4 uses IOCP FMID HIO1104
- ▶ z/OS V2.R5 uses IOCP FMID HIO1105
- ▶ z/OS V3R1 uses IOCP FMID HIO1106

For more information, see the following publications:

- ▶ *Stand-Alone Input/Output Configuration Program User's Guide*, SB10-7186
- ▶ *Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7183

## Dynamic Partition Manager V6.0

Dynamic Partition Manager V6.0 is available for managing IBM z17 that are running Linux. DPM 6.0 is available with HMC Driver Level 61 (HMC Version 2.17.0). IOCP does not need to configure a server that is running in DPM mode.

For more information, see *IBM Dynamic Partition Manager (DPM) Guide*, SB10-7188

## 7.4.2 Base CPC features and functions

In this section, we describe the features and functions of Base CPC.

### **z/OS**

#### **HiperDispatch**

The **HIPERDISPATCH=YES/NO** parameter in the IEAOPTxx member of SYS1.PARMLIB and on the **SET OPT=xx** command controls whether HiperDispatch is enabled or disabled for a z/OS image. It can be changed dynamically, without an IPL or any outage.

In z/OS, the IEAOPTxx keyword **HIPERDISPATCH** defaults to YES when it is running on an IBM z17, IBM z16, or IBM z15 system.

The use of SMT on IBM z17 systems requires that HiperDispatch is enabled on the operating system. For more information, see “Simultaneous multithreading” on page 288.

Also, any LPAR that is running with more than 64 logical processors is required to operate in HiperDispatch Management Mode.



The following rules control this environment:

- ▶ If an LPAR is defined at IPL with more than 64 logical processors, the LPAR automatically operates in HiperDispatch Management Mode, regardless of the HIPERDISPATCH= specification.
- ▶ If logical processors are added to an LPAR that has 64 or fewer logical processors and the extra logical processors raise the number of logical processors to more than 64, the LPAR automatically operates in HiperDispatch Management Mode, regardless of the HIPERDISPATCH=YES/NO specification. That is, even if the LPAR has the HIPERDISPATCH=NO specification, that LPAR is converted to operate in HiperDispatch Management Mode.
- ▶ An LPAR with more than 64 logical processors that are running in HiperDispatch Management Mode cannot be reverted to run in non-HiperDispatch Management Mode.

HiperDispatch on IBM z17 systems uses chip and CPC drawer configuration to improve the cache access performance. It optimizes the system PU allocation with Chip/cluster/drawer cache structure on IBM Z servers.

The base support for IBM z17 is provided by PTFs that are identified by:

`IBM.device.server.IBM z17-9175.requiredservice`

PR/SM on IBM z17 servers preferentially assigns memory for a system in one CPC drawer that is striped across the clusters of that drawer to take advantage of the lower latency memory access in a drawer. Also, PR/SM tries to consolidate storage onto drawers with the most processor entitlement.

With HiperDispatch enabled, PR/SM seeks to assign logical processors of a partition to the smallest number of PU chips within a drawer in cooperation with operating system to optimize shared cache usage.

PR/SM automatically keeps a partition's memory and logical processors on the same CPC drawer where possible. This arrangement looks simple for a partition, but it is a complex optimization for multiple logical partitions because some must be split among processors drawers.

All IBM z17 processor types can be dynamically reassigned except IFPs.

To use HiperDispatch effectively, WLM goal adjustment might be required. Review the WLM policies and goals and update them as necessary.

WLM policies can be changed without turning off HiperDispatch. A health check is provided to verify whether HiperDispatch is enabled on a system image.

### ***z/VM V7R3 and V7R4***

z/VM also uses the HiperDispatch facility for improved processor efficiency by better use of the processor cache to take advantage of the cache-rich PU chip, node, and drawer design of the IBM z17 system.

### ***CPU polarization support in Linux on IBM Z***

You can optimize the operation of a vertical SMP environment by adjusting the SMP factor based on the workload demands. For more information about CPU polarization support in Linux on IBM Z, see this [IBM Documentation web page](#).

### ***z/TPF***

z/TPF on IBM z17 can use more processors immediately without reactivating the LPAR or IPLing the z/TPF system.



When z/TPF is running in a shared processor configuration, the achieved MIPS is higher when z/TPF uses a minimum set of processors.

In low-use periods, z/TPF minimizes the processor footprint by compressing TPF workload onto a minimal set of I-streams (engines), which reduces the effect on other LPARs and allows the entire CPC to operate more efficiently.

As a consequence, z/OS and z/VM experience less contention from the z/TPF system when the z/TPF system is operating at periods of low demand.

The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270.

## **zIIP support**

zIIPs do not change the model capacity identifier of IBM z17 servers. IBM software product license charges that are based on the model capacity identifier are not affected by the addition of zIIPs.

No changes to applications are required to use zIIPs. They can be used by the following applications:

- ▶ Db2 V8 and later for z/OS data serving for applications that use data Distributed Relational Database Architecture (DRDA) over TCP/IP, such as data serving, data warehousing, and selected utilities.
- ▶ z/OS XML services.
- ▶ z/OS CIM Server.
- ▶ z/OS Communications Server for network encryption (Internet Protocol Security [IPsec]) and for large messages that are sent by HiperSockets.
- ▶ IBM GBS Scalable Architecture for Financial Reporting.
- ▶ IBM z/OS Global Mirror (formerly XRC) and System Data Mover.
- ▶ IBM z/OS Container Extensions.
- ▶ IBM OMEGAMON® XE on z/OS, OMEGAMON XE on Db2 Performance Expert, and Db2 Performance Monitor.
- ▶ Any Java application that uses the current IBM SDK.
- ▶ Java IBM Semeru Runtime offloading enablement for DLC models that use Integrated Accelerator for AI.
- ▶ WebSphere Application Server V5R1 and later, and products that are based on it, such as WebSphere Portal, WebSphere Enterprise Service Bus (WebSphere ESB), and WebSphere Business Integration (WBI) for z/OS.
- ▶ CICS/TS V2R3 and later.
- ▶ Db2 UDB for z/OS Version 8 and later.
- ▶ IMS Version 8 and later.
- ▶ zIIP Assisted HiperSockets for large messages.
- ▶ z/OSMF (z/OS Management Facility).
- ▶ IBM z/OS Platform for Apache Spark.
- ▶ IBM Watson® Machine Learning for z/OS.
- ▶ z/OS System Recovery Boost.
- ▶ Approved third-party vendor products.



The use of a zIIP is transparent to application programs. The supported operating systems are listed in Table 7-3 on page 269.

On IBM z17 servers, the zIIP processor is designed to run in SMT mode, with up to two threads per processor. This function is designed to help improve throughput for zIIP workloads and provide appropriate performance measurement, capacity planning, and SMF accounting data. zIIP support is available on all currently supported z/OS versions.

Use the **PROJECTCPU** option of the IEAOPTxx parmlib member to help determine whether zIIPs can be beneficial to the installation. Setting PROJECTCPU=YES directs z/OS to record the amount of eligible work for zIIPs in SMF record type 72 subtype 3.

The field APPL% IIPCP of the Workload Activity Report listing by WLM service class indicates the percentage of a processor that is zIIP eligible. Because of the zIIP's lower price as compared to a CP, even a utilization as low as 10% can provide cost benefits.

### Transactional Execution<sup>9</sup> (Limited support)

Transactional Execution (TX) is known in academia and industry as *hardware transactional memory*. Transactional execution is implemented on IBM Z servers.

This feature enables software to indicate to the hardware the beginning and end of a group of instructions that must be treated in an atomic way. All of their results occur or none occur, in true transactional style. The execution is optimistic.

The hardware provides a memory area to record the original contents of affected registers and memory as instruction execution occurs. If the transactional execution group is canceled or must be rolled back, the hardware transactional memory is used to reset the values. Software can implement a fallback capability.

This capability increases the software's efficiency by providing a way to avoid locks (lock elision). This advantage is of special importance for speculative code generation and highly parallelized applications.

TX is used by IBM Java virtual machine (JVM) and might be used by other software. The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270.

### System Recovery Boost

System Recovery Boost is a feature that was implemented on the IBM z15 system. The feature provides more temporary processor capacity and delivers substantially faster system shutdown and restart and critical system operations, such as stand-alone dump, short duration Recovery process boost for sysplex events, and fast catch-up of an accumulated backlog of critical workload after a planned or unplanned event.

Added or increased IBM software costs are *not* incurred by using System Recovery Boost.

**Attention:** IBM z17 provides no additional System Recovery Boost enhancements in comparison to the z16 model.

With the IBM z17 system, more Recovery process boost scenarios are supported that allow the customer to define some boost granularity (see Table 7-14 on page 288). For more information, see *IBM Z System Recovery Boost*, [REDP-5563](#).

<sup>9</sup> Statement of Direction: In a future IBM Z hardware system family, the transactional execution and constrained transactional execution facility will no longer be supported. Users of the facility on current servers should always check the facility indications before use.



Table 7-14 Operating system support for System Recovery Boost

Boost type <sup>a</sup>	z/OS V3R1	z/OS V2R5	z/OS V2R4	z/VM V7R4	z/VM V7R3	z/TPF V1R1	VSE <sup>n</sup> V6R3 <sup>b</sup>	Linux on IBM Z
Subcapacity Boost for IPL, Shutdown <sup>c</sup>	Y	Y	Y <sup>d</sup>		Y	Y <sup>d</sup>	Y <sup>d</sup>	N
zIIP boost <sup>e</sup> for IPL, shutdown, and dump events	Y	Y	Y <sup>d</sup>		N	N	N	N
Subcapacity and zIIP Boost for process recovery boost in a sysplex <sup>f</sup> .	Y	Y	Y <sup>d</sup>		N	N	N	N
Recovery process boosts (IBM z17 <sup>g</sup> ).	Y	Y <sup>d</sup>	N		N	N	N	N

a. Boost must be enabled for LPARs to opt in.

b. VSEn V6.3.1 - 21st Century Software

c. Subcapacity Boost also is available for stand-alone memory dump on z/OS and z/VSE.

d. With Fixes.

e. Allows CP work to be dispatched on zIIPs. zIIP processor capacity boost is available only if the customer has at least one active processor that is characterized as zIIP. For IBM z16 A01 and IBM z15 T01 only, more zIIPs can be used if obtained through eBOD (temporary zIIP boost records).

f. Process recovery boosts support subcapacity CPs speed boost and entitled (purchased) customer zIIPs only; zIIPs that are provided by FC 9930 and FC 6802 cannot be used for process recovery boosts.

g. System Recovery Boost (SRB) Upgrade record is not offered on z17.

### Automation

The client's automation product can be used to automate and control the following System Recovery Boost activities:

- ▶ To activate and deactivate the eBod temporary capacity record to provide more physical zIIPs for an IPL or Shutdown Boost.
- ▶ To dynamically modify LPAR weights, as might be needed to modify sharing physical zIIP capacity during a Boost period.
- ▶ To drive the invocation of the PROC that indicates the beginning of a shutdown process (and the start of the shut-down Boost).
- ▶ To take advantage of new composite hardware API reconfiguration actions.
- ▶ To control the level of parallelism that is present in the workload at startup (for example, starting middleware regions) and shutdown (for example, performing an orderly shutdown of middleware).

### Simultaneous multithreading

SMT is the hardware capability to process up to two simultaneous threads in a single core, which shares the resources of the core, such as cache, translation lookaside buffer (TLB), and execution resources. This sharing improves system capacity and efficiency by reducing processor delays, which increases the overall throughput of the system.

SMT<sup>10</sup> is supported for zIIPs and IFLs.

**Note:** For zIIPs and IFLs, SMT must be enabled on z/OS, z/VM, or Linux on IBM Z instances. An operating system with SMT support can be configured to dispatch work to a thread on a zIIP (for eligible workloads in z/OS) or an IFL (for z/VM) core in single-thread or SMT mode.

The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270.

<sup>10</sup> SMT is also enabled (not user configurable) by default for SAPs.



An operating system that uses SMT controls each core and is responsible for maximizing its throughput and meeting workload goals with the smallest number of cores. In z/OS, consider HiperDispatch cache optimization when you must choose the two threads to be dispatched in the same processor.

HiperDispatch attempts to dispatch guest virtual CPUs on the same logical processor on which they ran. PR/SM attempts to dispatch a vertical low logical processor in the same physical processor. If that process is not possible, it attempts to dispatch it in the same node, or then the same CPC drawer where it was dispatched before to maximize cache reuse.

From the perspective of an application, SMT is transparent and no changes are required in the application for it to run in an SMT environment, as shown in Figure 7-1.

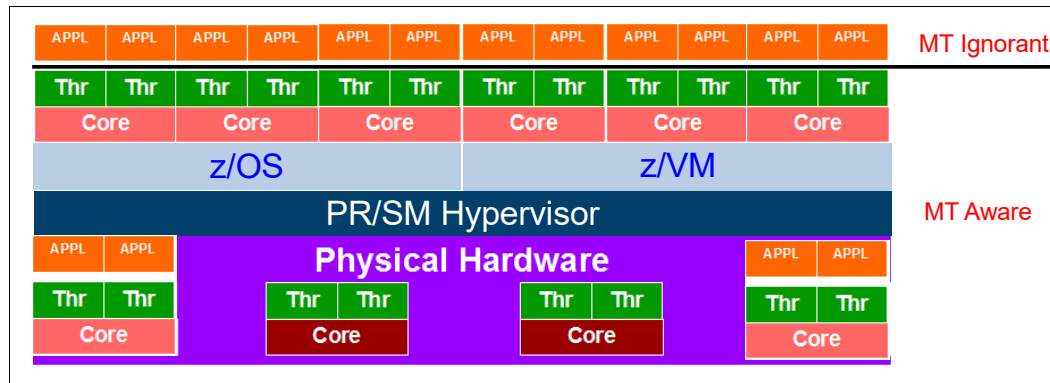


Figure 7-1 Simultaneous multithreading

## z/OS

The use of SMT on z/OS requires enabling HiperDispatch, and defining the processor view (**PROCVIEW**) control statement in the LOADxx parmlib member and the **MT\_ZIIP\_MODE** parameter in the IEAOPTxx parmlib member.

The **PROCVIEW** statement is defined for the life of IPL, and can have the following values:

- ▶ **CORE**: This value specifies that z/OS should configure a processor view of core, in which a core can include one or more threads. The number of threads is limited by IBM z17 to two threads. If the underlying hardware does not support SMT, a core is limited to one thread.
- ▶ **CPU**: This value is the default. It specifies that z/OS should configure a traditional processor view of CPU and not use SMT.
- ▶ **CORE,CPU\_OK**: This value specifies that z/OS should configure a processor view of core (as with the **CORE** value) but the CPU parameter is accepted as an alias for applicable commands.

When **PROCVIEW CORE** or **CORE,CPU\_OK** are specified in z/OS that is running on an IBM z17, HiperDispatch is forced to run as enabled, and you cannot disable HiperDispatch. The **PROCVIEW** statement cannot be changed dynamically; therefore, you must re-IPL after changing it to make the new setting effective.

The **MT\_ZIIP\_MODE** parameter in the IEAOPTxx controls zIIP SMT mode. It can be 1 (the default), where only one thread can be running in a core, or 2, where up two threads can be running in a core. If **PROCVIEW CPU** is specified, the **MT\_ZIIP\_MODE** is always 1. Otherwise, the use of SMT to dispatch two threads in a single zIIP logical processor (**MT\_ZIIP\_MODE=2**) can be changed dynamically by using the **SET OPT=xx** setting in the IEAOPTxx parmlib. Changing the MT mode for all cores can take some time to complete.



**PROCVIEW CORE** requires **DISPLAY M=CORE** and **CONFIG CORE** to display (Example 7-1) the core states and configure an entire core. With the introduction of Multi-Threading support for SAPs, a maximum of 88 logical SAPs can be used. RMF is updated to support this change by implementing page break support in the I/O Queuing Activity report that is generated by the RMF Post processor.

*Example 7-1 Sample D M=CORE Output*

---

```

D M=CORE
IEE174I 09.58.49 DISPLAY M 569
CORE STATUS: HD=Y   MT=2   MT_MODE: CP=1   zIIP=2
ID    ST   ID RANGE  VP   ISCM  CPU  THREAD STATUS
0000  +    0000-0001  H    FC00  +N
0001  +    0002-0003  H    0000  +N
0002  +    0004-0005  H    0000  +N
0003  +    0006-0007  H    0000  +N
0004  +    0008-0009  H    0000  +N
0005  +    000A-000B  M    0000  +N
0006  +    000C-000D  LP   0000  +N
0007  +    000E-000F  LP   0000  +N
0008  +    0010-0011  LP   0000  +N
0009  +    0012-0013  LP   0000  +N
000A  +I    0014-0015  H    0200  ++
000B  +I    0016-0017  H    0200  ++
000C  +I    0018-0019  H    0200  ++
000D  +I    001A-001B  H    0200  ++
000E  +I    001C-001D  M    0200  ++
000F  +I    001E-001F  M    0200  ++
0010  +I    0020-0021  LP   0200  ++
0011  N     0022-0023
0012  N     0024-0025
0013  N     0026-0027
0014  N     0028-0029
0015  N     002A-002B
0016  N     002C-002D
0017  N     002E-002F
0018  N     0030-0031
0019  N     0032-0033
001A  N     0034-0035
001B  NI    0036-0037
001C  NI    0038-0039
001D  NI    003A-003B
001E  NI    003C-003D
001F  NI    003E-003F

CPC ND = 009175.ME1.IBM.02.0000000310A9
CPC SI = 9175.710.IBM.02.00000000000310A9
      Model: ME1
CPC ID = 00
CPC NAME = XXXXXX
LP NAME = XXXXXXXX   LP ID = 1
CSS ID = 0
MIF ID = 1

+ ONLINE   - OFFLINE   N NOT AVAILABLE   / MIXED STATE
W WLM-MANAGED

```



I INTEGRATED INFORMATION PROCESSOR (zIIP)  
 CPC ND CENTRAL PROCESSING COMPLEX NODE DESCRIPTOR  
 CPC SI SYSTEM INFORMATION FROM STSI INSTRUCTION  
 CPC ID CENTRAL PROCESSING COMPLEX IDENTIFIER  
 CPC NAME CENTRAL PROCESSING COMPLEX NAME  
 LP NAME LOGICAL PARTITION NAME  
 LP ID LOGICAL PARTITION IDENTIFIER  
 CSS ID CHANNEL SUBSYSTEM IDENTIFIER  
 MIF ID MULTIPLE IMAGE FACILITY IMAGE IDENTIFIER

---

### ***z/VM V7R4 and V7R3***

The use of SMT in z/VM is enabled by using the **MULTITHREADING** statement in the system configuration file. Multithreading is enabled only if z/VM is configured to run with the HiperDispatch vertical polarization mode enabled and with the dispatcher work distribution mode set to reshuffle.

The default in z/VM is multithreading disabled. Dynamic SMT enables dynamically varying the active threads per core. The number of active threads per core can be changed dynamically without a system outage and potential capacity gains going from no SMT to SMT-2 (one to two threads per core) can be achieved dynamically.

z/VM V7R4 and V7R3 support up to 40 multithreaded cores (80 threads) for IFLs, and each thread is treated as an independent processor. z/VM dispatches virtual IFLs on the IFL logical processor so that the same or different guests can share a core. Each core has a single dispatch vector, and z/VM attempts to place virtual sibling IFLs on the same dispatch vector to maximize cache reuses.

z/VM guests have no awareness of SMT, and cannot use it directly. z/VM SMT use does not include guest support for multithreading. The value of this support for guests is that the first-level z/VM host of the guests can achieve higher throughput from the multi-threaded IFL cores.

### ***Linux on IBM Z and the KVM hypervisor***

The Linux kernel features [SMT functions](#) that were developed by the Linux on IBM Z development team. SMT is supported on LPAR only (not as a second-level guest).

The following *minimum* releases of Linux on IBM Z distributions are supported on IBM z16 (native SMT support):

- ▶ SUSE:
  - SLES 16
  - SLES 15 SP6 with service
  - SUSE SLES 12 SP5 with service
- ▶ Red Hat:
  - Red Hat RHEL 9.4
  - Red Hat RHEL 8.1 with service
  - Red Hat RHEL 7.9 with service
- ▶ Ubuntu:
  - Ubuntu 24.04 LTS
  - Ubuntu 22.04 LTS
  - Ubuntu 20.04.1 LTS with service

The KVM hypervisor is supported on the same Linux on IBM Z distributions in this list.



For more information about the most current support, see the Linux on IBM Z Tested platforms [website](#).

### Single-instruction multiple-data

The SIMD feature introduces a new set of instructions to enable parallel computing that can accelerate code with string, character, integer, and floating point data types. The SIMD instructions allow many operands to be processed with a single complex instruction.

IBM z17 is equipped with a set of instructions to improve the performance of complex mathematical models and analytic workloads through vector processing and complex instructions, which can process numerous data with a single instruction. This set of instructions, which is known as SIMD, enables more consolidation of analytic workloads and business transactions on IBM Z servers.

SIMD on IBM z17 has support for enhanced math libraries that provide performance improvements for analytical workloads by processing more information with a single CPU instruction.

The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270. Operating System support includes the following features<sup>11</sup>:

- ▶ Enablement of vector registers.
- ▶ A math library with an optimized and tuned math function (Mathematical Acceleration Subsystem [MASS]) that can be used in place of some of the C standard math functions. It includes a SIMD vectorized and non-vectorized version.
- ▶ A specialized math library, which is known as Automatically Tuned Linear Algebra Software (ATLAS), that is optimized for the hardware.
- ▶ IBM Language Environment® for C runtime function enablement for ATLAS.
- ▶ DBX to support the disassembly of the new vector instructions, and to display and set vector registers.
- ▶ XML SS exploitation to use new vector processing instructions to improve performance.

MASS and ATLAS can reduce the time and effort for middleware and application developers. IBM provides compiler built-in functions for SIMD that software applications can use as needed, such as for using string instructions.

The followings compilers include built-in functions for SIMD:

- ▶ IBM Java
- ▶ XL C/C++
- ▶ Enterprise COBOL
- ▶ Enterprise PL/I

Code must be developed to use the SIMD functions. Applications with SIMD instructions abend if they run on a lower hardware level system that do not support SIMD. Some mathematical function replacement can be done without code changes by including the scalar MASS library before the standard math library.

Because the MASS and standard math library include different accuracies, assess the accuracy of the functions in the context of the user application before deciding whether to use the MASS and ATLAS libraries.

The SIMD functions can be disabled in z/OS partitions at IPL time by using the **MACHMIG** parameter in the LOADxx member. To disable SIMD code, use the MACHMIG VEF

---

<sup>11</sup> The features that are listed here might not be available on all operating systems that are listed in the tables.



hardware-based vector facility. If you do not specify a **MACHMIG** statement, which is the default, the system not limited in its use of the Vector Facility for z/Architecture (SIMD).

## New IBM z17 machine instructions

By default (unless locally customized), the assembler uses the **OPTABLE(UNI)** universal operation code table. This defines the mnemonics for instructions up to the latest supported z/Architecture level. APAR PH62834 provides exploitation support for HLASM 1.6 (all z/OS releases), adding the new mnemonics for IBM z17.

**Important:** These mnemonics may collide with the names of Assembler macro instructions you have.

It is safer to assemble using an **OPTABLE** option which matches the current highest target hardware level. If you code Assembly Language macros, you should compare the list of new instructions to the names of your Assembler macros.

An Tool will be available with the PTF availability from IBM Support website. If a conflict is identified, then either:

1. Rename your affected macros
2. Specify a separate assembler **OPCODE** table – **PARM=,ASMAOPT**, or **'\*\*PROCESS OPTABLE'** in the source

Use a coding technique that permits both use of a new instruction and a macro with the same name in an assembly such as HLASM's mnemonic tag (:MAC :ASM).

See the APAR documentation for a link to the new mnemonics and what to do in the case of clashes.

**Attention:** We normally generally recommend that rather than using the default of **OPTABLE(UNI)** which will immediately pick up any new instructions, it is safer to assemble using an **OPTABLE** option which matches the current highest target hardware level. This firstly ensures that any accidental attempt to use a newer instruction will get an error, and secondly means that if support for a new hardware level is added by maintenance, this will not impact existing programs at all, so it will only be necessary to check for conflicting macros at the point when new hardware is to be exploited by using a new **OPTABLE** level.

## Hardware decimal floating point

Industry support for decimal floating point is growing, with IBM leading the open standard definition. Examples of support for the draft standard IEEE 754r include Java BigDecimal, C#, XML, C/C++, GCC, COBOL, and other key software vendors, such as Microsoft and SAP.

IBM z17 uses COBOL optimization for Hexadecimal Floating Point (HFP) <--> Binary Coded Decimal (BCD) conversion, and Numeric Editing, and Zoned Decimal operations, introduced with IBM z16.

The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270. For more information, see 7.5.8, “z/OS XL C/C++ considerations” on page 342.

## Out-of-order execution

Out-of-order (OOO) execution yields significant performance benefits for compute-intensive applications by reordering instruction execution, which allows later (newer) instructions to be run ahead of a stalled instruction, and reordering storage accesses and parallel storage accesses. OOO maintains good performance growth for traditional applications.



The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270. For more information, see “3.4.3, “Out-of-Order execution” on page 86.

### **z/OS DFSORT exploitation of IBM Integrated Accelerator for Z Sort**

The Integrated Accelerator for Z Sort is an on chip hardware feature that is available on IBM z15 and newer servers. It is driven by the SORT LISTS (SORTL) instruction.

z/OS DFSORT takes advantage of Z Sort, and provides users with significant performance boosts for their sort workloads. With z/OS DFSORT's Z Sort algorithm, clients can see batch sort job elapsed time improvements of up to 20 - 30% (depending on record size) and CPU time improvements of up to 40% compared to older IBM Z systems w/o integrated SORT accelerator.

The function is used on z/OS V3R1, z/OS V2R5, and enabled on z/OS V2R4 with PTFs for APAR PH03207.

The sort jobs must meet certain eligibility criteria. For the full list and other considerations, see the DFSORT User Guide for PH03207.

### **CPU Measurement Facility with new counter**

For every server generation, the hardware adds, removes, and moves extended counters in the CPU Measurement Facility (CPU/MF). For IBM z17 extended counters will begin with HISCTR\_KEXT8\_ \* in the HISYCTRS mapping (see z/OS MVS Data Areas Volume 1). z/OS Hardware Instrumentation Services (HIS) will collect IBM z17 extended counters for performance problem diagnosis. HIS must be setup and started to collect this information, on each LPAR. HIS writes then the extended counters in SMF 113.

If changed extended counter sets are being used, then any programs relying upon those counters are impacted, and must be updated accordingly.

If unchanged extended counter sets are using, no changes are necessary. This new support is available for z/OS V2.4 and higher.

**Restriction:** z/OS does not collect CPU MF data when running as a guest of z/VM.

Gathering CPU MF Counters is an common industry “best practice” approach.

Capture on pre-IBM z17 server to determine your current LSPR workload. Capture again on IBM z17 server afterwards.

Having both values this can allow you to validate your achieved IBM z17 processor performance, and provide insights for new features and functions.

**Tip:** Additional information are available on the Washington Systems Center website: <https://www.ibm.com/support/pages/node/6354583>

The Exploitation support will be available via APAR.

The supported operating systems are listed in Table 7-3 on page 269.

For more information about this function, see this [IBM Support web page](#).

For more information about the CPU Measurement Facility, see this [IBM Support web page](#).



For more information, see “12.2, “IBM z17 Large System Performance Reference ratio” on page 492.

### Hardened security between BCPii and the HMC/SE

z/OS BCPii is being enhanced to support server-based authorization with JSON Web Tokens, JWT. This enhancement allows the application to issue operations previously not available, including asynchronous notification support, interaction with Dynamic Partition Manager CECs, User Management operations, and more.

This process will correlate a z/OS user ID that issued the BCPii request with a Hardware Management Consoles (HMC) user ID. The correlated HMC user ID will then be used to determine if the ‘user’ is authorized to issue the request.

Another V2 API, HWIREST2, will be introduced which only supports the JWT authorization scheme.

**Restriction:** This new function only supported with C and HLASM interfaces. There is no REXX interface available. Also this support is only available on z/OS V3.1 and higher.

HWIREST2 has the advantage of enabling z/OS BCPii to register an ENF 68 listener on the user’s behalf. HWIREST, the initial V2 API, makes the user application responsible for registering an ENF 68 listener exit. HWIREST supports both FACILITY class and JWT authorization scheme for classical CEC non-HMC target operations. Supports C, HLASM, and REXX.

Local SE and any target SE or HMC, must be at IBM z17. HMC managing the local SE must be at IBM z17.

**Restriction:** z/VM support is not applicable

The Exploitation support will be available via APAR OA65929.

### Workload Classification Pricing

The new Workload Classification Pricing enables classification of programs and collection metrics to allow for price differentiation on z/OS. Possible Examples of workload classes might be:

- ▶ Any General Purpose business workloads
- ▶ Any zIIP-eligible work
- ▶ AI inferencing work
- ▶ zCX work

Workload instrumentation data can be collected with the new IEASYSxx WORKINST parameter and has the following values:

<b>SYSTEM</b>	is the default, and will collect data if the system supports it.
<b>YES</b>	Data collection enabled
<b>NO</b>	Data collection disabled

The SCRT has changed and SMF 70.1 records will have a new data section which report on the workload class usage.



**Note:** The support is available on z/OS V2.5 and higher via APAR OA66812, OA65240, OA65242, OA66596, and OA66937. z/VM PTFs are required to enable workload classification for z/OS as a guest.

## Sustainability and Power consumption reporting

New CEC-level and LPAR-level power consumption information are provided for IBM z17.

**CEC-level**                      total power, unassigned resources power, infrastructure power.

**LPAR-level**                    CPU power, memory power, I/O power.

The Power consumption information is available to be gathered every 5 seconds. However, your SMF interval will control how often it is reported. The SMF 70.1 records has been enhanced in the CPU control section to report on the power consumption counts which are aggregated over the SMF interval. This support is Intended to assist with current and future regulations and demands to report power metrics at a more granular level.

IBM intends reporting products to make use of service class/reporting class resource consumption data in SMF 72.3 to provide workload-level granularity in reporting out the CPU, I/O, and memory power.

**Note:** The Exploitation support is available on z/OS V3.1 and higher via APAR OA63265 and OA66018.

## Replacement for Capacity Records

The new support for Replacement Capacity Records for Tailored Fit Pricing for IBM Z Hardware (TFP-HW). z/OS WLM provides updates to support the model-replacement capacity values. In addition to existing values now new information included such as the model, model-permanent, and model-temporary capacity identifier and ratings.

The z/OS System command D M,CPU has been modified in order to support IBM z17 replacement capacity.

The z/OS Data Gatherer stores new metrics in SMF 70.1 and Monitor III measurement table ERBPCDB. Also z/OS SCRT processes enhanced SMF 70.1 for TFP-HW.

**Note:** This enhancement is available for V2.5 and higher and requires exploitation support via APAR OA66402, OA66054, OA63265, and OA66938.

## z/OS SMF Enhancements for sustainability and power consumption

SMF 70 records were enhanced to indicate the amount of electrical power consumption on a CEC-level and LPAR-level.

**CEC-level**                      total power, unassigned resources power, infrastructure power.

**LPAR-level**                    CPU power, memory power, I/O power.

Power consumption information is available to be gathered every 5 seconds. However, your SMF interval will control how often it is reported. The SMF 70.1 records were expanded in the CPU control section in order to report on the power consumption counts which are aggregated over the SMF interval.

The support is intended to assist with current and future regulations and demands to report power metrics at a more granular level.



IBM intends reporting products to make use of service class/reporting class resource consumption data (in SMF 72.3) to provide workload-level granularity in reporting out the CPU, I/O, and memory power.

**Restriction:** This support is only available on V3.1 and higher and requires APAR OA63265 and OA66018.

## Large page support

In addition to the 1-MB large pages, 4-KB pages, and page frames, IBM z17 servers support pageable 1-MB large pages, large pages that are 2 GB, and large page frames.

Starting with z/OS V2.R5 onwards allows 2 GB LFAREA to exceed the 4 TB limit, up to 16 TB.

**Attention:** All online real storage over 4 TB is part of the 2 GB LFAREA, in addition to what was specified in LFAREA. That means only 4TB are available for 4-KB and 1-MB frames.

Real memory is available for 2 GB pages only. Applications that make use of 2 GB frames should be reviewed to use more frames if applicable (for example, Java and Db2).

## Calculation Example

Lets assume we have 8 TB defined to the LPAR. Our IEASYSxx member contains the following LFAREA definition: LFAREA=(1M=1024, 2G=512)

That means we are requesting:

Table 7-15 Storage allocation example

Storage frame type	Requested	Assigned
1-MB	1-GB	1-GB
2-GB	1-TB	5-TB = 1TB + (8TB - 4TB)

On z/OS V2.5 and higher, no PTFs required.

**Note:** Any machine having a single LPAR consuming more than the amount of memory plugged within a drawer will inherently reduce the performance of the LPAR due to the cross-drawer communication.

This would be 16 TB for IBM z17, 10TB for z15/z16, and 8TB for z14.

The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270.

## ICSF Enhancements

The Clear key HMAC acceleration via CP Assist for Cryptographic Functions (CPACF) has been implemented. ICSF changes clear key HMAC requests internally to use the new CPACF instruction when running on IBM z17.

**Attention:** No new cryptographic algorithms are added.

**Note:** Available on z/OS V2.5 and higher via APAR OA66518.



## Deep learning library (zDNN) IBM Z Integrated Accelerator for AI (zAIU)

The specialized-function-assist instructions are intended to provide performance improvements for specific operations in software libraries, utilities, or operating system services. These facilities and instructions may be replaced or removed in the future, therefore a software library or operating system function should be used instead of directly accessing the instructions.

zDNN is an IBM standard C library that provides APIs to facilitate access to the zAIU, which is an on-chip deep learning inference accelerator.

zDNN provides software functions that enable an exploiter to:

1. Transform tensor memory layout from standard layouts (i.e., NHW C) to the non-standard layout required by the AIU.
2. Convert tensor element data type from standard data types to the AIU required DFloat16 format.
3. Call deep learning primitives supported by the AIU.

zDNN will be available on both z/OS and Linux on Z distributions. Linux on Z will also facilitate zCX exploitation of AIU by z/OS clients using supported Linux AI frameworks.

For Software fallback, all fall back scenario would be NN model executing ops on AIU back to non-AIU HW (standard CPU ops). Exploiting applications should provide software-based implementation of DL primitives. Most z/OS exploitation will be via IBM Deep Learning Compiler, which will generate NN model optimized library which can switch between AIU and normal CPU based models.

**Note:** See IBM Z Deep Neural Network Library Programming Guide and Reference

On APARs/releases for technology that leverages the AI Accelerator (via zDNN):

1. Watson Machine Learning for z/OS has an ONNX model feature.
2. SQL Data insights for Db2 13.

The ONNX model feature would not require any changes from a data scientists perspective (i.e., no model changes). Action on the model deployment may be needed (i.e., to trigger a model recompile that targets the AIU). SQL Data Insights is a new feature, and requires explicit exploitation steps. These are brand new types of built-in SQL functions that will drive AI under the covers to uncover hidden relationships in Db2 data.

zDNN is that deep learning library. zAIU has complex data layout requirements for the tensor to enhance performance characteristics or operations. zDNN formats the tensor appropriately on behalf of the caller, using an optimized approach.

NNPA is the instruction that drives the zAIU.

zDNN library provides a set of APIs that an exploiter will use to drive the desired request. zDNN is available on both z/OS and Linux on Z. Including support for Linux on Z is important as acceleration can be enabled in frameworks for z/OS via zCX.

**Important:** zDNN is supported on IBM z16, but there are new APIs and IBM z17-aware optimizations which require IBM z17.

IBM z17 has updated Neural Network Processor Assist (NNPA) instructions.



**Restriction:** V2.5 and higher is required and via APAR OA66863. When using z/VM live guest relocation, ensure the relocation group members to be at least z16 in order to use the IBM z17 functionality.

## Open XL C/C++ for z/OS exploitation

The IBM z17 support will be added in the next release of the Open XL C/C++ compiler:

- ▶ New compiler sub-options to target IBM z17 instructions: ``-march=arch15`` and ``-march=z17``.
- ▶ Binaries produced with these sub-options for IBM z17 can only execute on IBM z17 and higher.
- ▶ Use ``-march=arch14`` or ``-march=z16`` or below, if you intend to execute on pre-IBM z17.
- ▶ Includes supports and optimization to exploit some of these new IBM z17 instructions.
- ▶ Extensions to the existing C/C++ vector programming support making use of the Vector-Enhancements Facility 3 in IBM z17. Including support of new scalar integer ``__int128`` data type.

As in the past, the z/OS XL C/C++ compiler, included as a priced feature of z/OS, will not be updated with the support mentioned above.

- ▶ XL C/C++ supports up to z15 HW instructions with ARCH(13).
- ▶ Programs compiled with XL C/C++ will run on IBM z17.

**Tip:** Watch for availability of the next release of Open XL C/C++ compiler at this website:

<https://www.ibm.com/support/pages/ibm-open-xl-cc-and-xl-cc-zos-documentation-library>

## AI accelerator exploitation

With the IBM z17 Integrated Accelerator for AI, customers can benefit from the acceleration of AI operations, such as fraud detection, customer behavior predictions, and streamlining of supply chain operations; all in real time. AI accelerators are designed to deliver AI inference in real time, at large scale and rate, with no transaction left behind so that customers can instantly derive the valuable insights from their data.

The AI capability is applied directly into the running transaction, shifting the traditional paradigm of applying AI to the transactions that were completed. This innovative technology can be used for intelligent IT workloads placement algorithms and contribute to better overall system performance. The co-processor is driven by the new Neural Networks Processing Assist (NNPA) instruction.

NNPA is a new nonprivileged Complex Instruction Set Computer (CISC) memory-to-memory instruction that operates on tensor objects that are in client application program memory. AI functions and macros are abstracted by way of NNPA.

## Virtual Flash Memory

IBM Virtual Flash Memory (FC 0644) offers up to 6.0 TB of memory for IBM z17 ME1. VFM is provided for improved application availability and to handle paging workload spikes.

IBM Virtual Flash Memory is designed to help improve availability and handling of paging workload spikes when running z/OS. With this support, z/OS is designed to help improve system availability and responsiveness by using VFM across transitional workload events,



such as market openings, and diagnostic data collection. z/OS also helps improve processor performance by supporting middleware use of pageable large (1 MB) pages.

VFM can help organizations meet their most demanding service level agreements and compete more effectively. VFM is easily configurable, and provides rapid time to value.

The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270.

## Guarded Storage Facility

Also known as *less-pausing garbage collection*, Guarded Storage Facility (GSF) is an architecture that was introduced with IBM z14 to enable enterprise scale Java applications to run without periodic pause for garbage collection on larger heaps.

### z/OS

GSF support allows an area of storage to be identified such that an Exit routine assumes control if a reference is made to that storage. GSF is managed by new instructions that define Guarded Storage Controls and system code to maintain that control information across undispatch and redispatch.

Enabling a less-pausing approach improves Java garbage collection. Function is provided on IBM z14 and subsequent servers that are running z/OS V2.R4 and later with APAR OA51643 installed. The **MACHMIG** statement in **LOADxx** of **SYS1.PARMLIB** disables the function.

### z/VM

GSF is designed to improve the performance of garbage-collection processing by various languages, in particular Java.

The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270.

## Instruction Execution Protection

Instruction Execution Protection (IEP) is a hardware function that enables software, such as Language Environment, to mark specific memory regions (for example, a heap or stack), as nonexecutable to improve the security of programs that are running on IBM Z servers against stack-overflow or similar attacks.

Through enhanced hardware features (based on DAT table entry bit) and specific software requests to obtain memory areas as nonexecutable, areas of memory can be protected from unauthorized execution. A Protection Exception occurs if an attempt is made to fetch an instruction from an address in such an element or if an address in such an element is the target of an execute-type instruction.

### z/OS

To use IEP, Real Storage Manager (RSM) is enhanced to request nonexecutable memory allocation. Use new keyword **EXECUTABLE=YES|NO** on **STORAGE OBTAIN** or **IARV64** to indicate whether memory that is to be used contains executable code. Recovery Termination Manager (RTM) writes LOGREC record of any program-check that results from IEP.

IEP support is for z/OS V2.R4 and later running on IBM z17 with APARs OA51030 and OA51643 installed.

### z/VM

Guest exploitation support for the Instruction Execution Protection Facility is provided with APAR VM65986.



The supported operating systems are listed in Table 7-3 on page 269 and Table 7-4 on page 270.

### Secure Boot

Secure Boot verification ensures that the Linux distribution kernel comes from an official provider and was not compromised. If the signature of the distribution cannot be verified, the process of booting the operating system is stopped.

The Secure Boot feature requires operating system support.

### IBM Integrated Accelerator for zEnterprise Data Compression

The IBM Integrated Accelerator for zEnterprise Data Compression (zEDC) is implemented as on-chip data compression accelerator; that is, Nest Compression Accelerator (NXU) and supports Deflate/gzip/zlib algorithms.

For more information, see Chapter B, “IBM Integrated Accelerator for zEnterprise Data Compression” on page 525.

Each PU chip includes one on-chip compression unit, which is designed to replace the zEnterprise Data Compression (zEDC) Express PCIe feature that is available on IBM z14 and earlier.

**Note:** The zEDC Express feature that is available on older systems is *not* carried forward to IBM z17.

The IBM Integrated Accelerator for zEDC maintains software compatibility with zEDC Express use cases. For more information, see [Integrated Accelerator for zEnterprise Data Compression](#).

The z/OS zEDC capability is a software-priced feature that is designed to support compression capable hardware. With IBM z17, the zEDC feature is implemented in the on-chip compression accelerator unit, but the software (z/OS) component is required to maintain the same functions as previous PCIe-based zEDC features.

All data interchange with zEDC compressed data remains compatible as IBM z17 and zEDC capable machines coexist (accessing same data). Data that is compressed and written with zEDC can be read and decompressed by IBM z17 well into the future.

The on-chip compression unit has the following operating modes:

- ▶ Synchronous execution in Problem State, where user application starts instruction in its virtual address space. This mode provides low latency and high-bandwidth compression and decompression operations. It does not require any special hypervisor support, which removes the virtualization layer (sharing the zEDC Express PCIe adapter among LPARs requires virtualization support).
- ▶ Asynchronous optimization for Large Operations under z/OS. The authorized application (for example, BSAM/QSAM) issues I/O for asynchronous execution and SAP (PU) starts instruction (synchronously) on behalf of application. The on-chip accelerator enables load balancing of high compression loads and low latency and high bandwidth compared to zEDC Express, while maintaining current user experience on compression.

Function support for the IBM Integrated Accelerator for zEDC is listed in Table 7-3 on page 269 and Table 7-4 on page 270.



For more information, see Chapter B, “IBM Integrated Accelerator for zEnterprise Data Compression” on page 525.

### 7.4.3 Coupling and clustering features and functions

In this section, we describe the coupling and cluster features.

#### Coupling facility and CFCC considerations

Coupling facility (CF) connectivity to an IBM z17 is supported on the IBM z16, IBM z15, or another IBM z17. The CFCC levels that are currently supported on IBM Z are listed in Table 7-16.

Table 7-16 IBM Z CFCC code-levels

IBM Z server	Code level
IBM z17 ME1	CFCC Level 26
IBM z16 A01	CFCC Level 25
IBM z15 T01 and T02	CFCC Level 24

#### Coupling Express3 25Gb LR CHPID type CL6

IBM z17 has a new channel path type CL6. It's the Coupling Express3 25Gb LR, intended to provide significantly higher throughput for coupling links at distance.

z/OS releases starting from V2.4 onwards are eligible to support:

1. XCF/XES helps provide the infrastructure from z/OS to support the new link type including:
  - CL6 related information in the Accounting and Measurement data area (IXLYAMDA)
  - updated displays and IPCS reports.
2. Data Gatherer enhances the Monitor III Coupling Facility data gatherer for channel path details on this new channel path type (CL6).
  - Retrieves channel path details from IXLYAMDA and stores information in SMF 74.4 records and in the Monitor III coupling facility information table (ERBCFIG3)
3. IOS has been updated to support the CL6 Adapter
4. HCD and IOCP has been updated to support the CL6 Adapter

**Important:** This new long distance coupling link adapter (CL6) can only be connected to CL6 links. CL5 can be used to communicate with pre-IBM z17 servers. CL5 can be used as fallback from CL6, if necessary.

**Note:** The exploitation support is available via APAR OA64478, OA64591, OA64362, OA64114.

#### CFCC Level 26

CFCC Level 26 is delivered on IBM z17 servers with driver level 61 and introduces the following enhancements:



Scalability improvements for customers with many CF Images in a Sysplex:

- ▶ Support for up to 32 processors per CF image
- ▶ Support for 4096 CF structures in CFCC and vector firmware
- ▶ Nondisruptive push-based system-managed copy process for CF Duplexing
  - Non-disruptive async System Managed copy process for lock structures using “push” between duplex partners
- ▶ Non-disruptive system-managed copy process for lock structures
  - allows lock structures to be system-managed duplexed, re-duplexed, or reconfigured via a system-managed rebuild, more quickly and in a way that is not disruptive to the customer’s ongoing data-sharing workload
- ▶ Deprecate CF Flash Memory Exploitation and Dedicated CPs Exploitation

### Non-disruptive structure copy

System-Managed CF Structure Duplexing provides a desirable redundancy and failover recovery mechanism for CF structure data, providing CF structure resiliency and availability, but it has two major drawbacks:

- ▶ The synchronous protocol used for synchronizing the two copies of the data (the system-managed duplexing protocol) is very expensive, especially when the two CFs have distance between them, making the sysplex overhead cost of duplexing unacceptable to many clients
- ▶ Establishing duplexing (or re-establishing it after a failover, or performing a system-managed rebuild for structure relocation/reconfiguration purposes), requires a quiesce of the structure while the structure contents are copied over to the new secondary copy, resulting in a transient disruption to the customer workload during the copy process

In z13 GA2, we delivered Asynchronous CF Duplexing for CF lock structures (with DB2 exploitation), which addresses the first problem, but for DB2 lock structures only (with DB2 V12/V13 exploitation only)

- ▶ Successful – it provided simplex-like service times for a duplexed lock structure.
- ▶ Need a more general solution for all of the critical data-sharing lock structures.
- ▶ No other exploiters have supported this Asynchronous Duplexing mechanism (IMS IRLM, VSAM RLS, GRS)

Nothing we have delivered yet addresses the second problem of establishing duplexing (or doing a system-managed rebuild) in a non-disruptive way.

The CF and z/OS will make use of the existing PLSO “push” command for lock structures (currently used in the Asynchronous Duplexing process) in a novel way to allow duplexing to be established (or re-established after a failure), or a system-managed rebuild process to be performed, in a non-disruptive way.

- ▶ z/OS creates an empty secondary copy of the structure and binds it to the existing primary in an asynchronous duplexing relationship
  - All subsequent commands that update the primary structure instance will have those updates “pushed” to the secondary structure instance
  - At this point, the secondary copy is marked as not usable for failover purposes
- ▶ z/OS then invokes a new long-running CF structure copy process to “push” all current contents of the lock structure from primary to secondary; these pushes can interleave and overlap with pushes generated by ongoing mainline locking commands, fully ordered and serialized by normal primary structure latching and sequence number generation processes
- ▶ When the long-running CF structure copy process has completed all copy activity, the structure transitions into the desired final state, which can either be simplex mode, synchronous SM duplexing mode, or asynchronous duplexing mode



- In the case of duplexing, z/OS now marks the secondary copy as usable for failover purposes
- ▶ It's expected that the new PLSO-based copy process be faster than existing software-based SM copy processes, but even if it isn't, the non-disruptiveness of the copy process is a major advantage
  - Avoids disruption of the client's data-sharing workload while starting/restarting CF Duplexing or SM rebuilding a structure (a long-standing client pain point)
- ▶ APAR OA65820 (XCF/XES) – z/OS 2.5 and higher

## CFCC Level 25

CFCC Level 25 is delivered on IBM z16 servers with driver level 51 and introduces the following enhancements:

- ▶ Scalability Improvements
 

Processing and dispatching enhancements that result in meaningful scaling of effective throughput up to the limit of 16 ICF processors.
- ▶ Request latency/performance improvements
 

CFCC and coupling link firmware and hardware improvements to reduce link latency.
- ▶ Elimination of VSAM RLS orphaned cast-out lock problems and improved VSAM RLS Structure Full recovery processing.
 

Addresses reoccurring problems that are encountered by installations running VSAM RLS data sharing.

Retry Buffer support that is used on list and lock commands is extended to nonidempotent cache commands and optimized lock commands.

The new support also allows connectors to lock structures to specify a percentage of record data entries to be reserved. These reserved entries are off limits to normal requests to the coupling facility and can be used only if a new keyword is used on lock requests that generate record data entries.
- ▶ Cache residency time metrics
 

The CF calculates in microseconds by way of a moving weighted average the elapsed time a data area or directory entry resides in a storage class before it is reclaimed. XES returns this information on an IXLCACHE REQUEST=READSTGSTATS and IXLMG STRNAME=strname,STGCLASS=stgclass request.
- ▶ DYNDISP=ONIOFF is deprecated
 

For CFCC Level 25, DYNDISP=THIN is the only available behavior for shared-engine CF dispatching.

Specifying OFF or ON in CF commands and the CF configuration file is preserved for compatibility, but a warning message is issued to indicate that these options are no longer supported, and that DYNDISP=THIN behavior is to be used.

Before you begin the migration process, install the compatibility and coexistence PTFs. A planned outage is required when you upgrade the CF or CF LPAR to CFCC Level 25.

## CFCC Level 24

CFCC Level 24 is delivered on IBM z16 servers with driver level 41. CFCC Level 24 introduced the following enhancements:

- ▶ CFCC Fair Latch Manager



This enhancement to the internals of the Coupling Facility (CFCC) dispatcher provides CF work management efficiency and processor scalability improvements. It also improves the “fairness” of arbitration for internal CF resource latches across tasks

► **CFCC Message Path Resiliency**

CF Message Paths use a z/OS-provided system identifier (SYID) to uniquely identify which z/OS system image (and instance of that system image) is sending requests over a message path to the CF. With IBM z15, we are providing a new resiliency mechanism that transparently recovers for this “missing” message path deactivate (if and when that deactivation ever occurs).

During path initialization, the CF provides more information to z/OS about every message path that appears active, including the SYID for the path. Whenever z/OS interrogates the state of the message paths to the CF, z/OS checks this SYID information for currency and correctness, and if incorrect, gather diagnostic information and reactivates the path to correct the problem.

► **CF monopolization avoidance**

z/OS takes advantage of current CF support in CFLEVEL 24 (IBM z15 T01/T02) to deliver improved z/OS support for handling CF monopolization.

With IBM z15 T01/T02, the CF dispatcher monitors in real-time the number of CF tasks that have a command assigned to them for a specific structure on a structure-by-structure basis.

When the number of CF tasks that is used by any structure exceeds a model-dependent CF threshold, and a global threshold on the number of active tasks also is exceeded, the structure is considered to be “monopolizing” the CF. z/OS is informed of this monopolization.

New support in z/OS observes the monopolization state for a structure, and starts to selectively queue and throttle incoming requests to the CF on a structure-specific basis. Other requests for other “non-monopolizing” structures and workloads are unaffected.

z/OS dynamically manages the queue of requests for the “monopolizing” structures to limit the number of active CF requests (parallelism) to them, and monitors the CF’s monopolization state information so as to observe the structure becoming “non-monopolized” again, so that request processing can eventually revert back to a non-throttled mode of operation.

The overall goal of z/OS anti-monopolization support is to protect the ability of ALL well-behaved structures and workloads to access the CF, and get their requests processed in the CF in a timely fashion. At the same time, it implements queuing and throttling mechanisms in z/OS to hold back the specific abusive workloads that are causing problems for other workloads.

z/OS XCF/XES use of APAR support is required to provide this function.

► **CFCC Change Shared-Engine *CF Default* to **DYNDISP=THIN****

Coupling Facility images can run with shared or dedicated processors. Shared processor CFs can operate with different Dynamic Dispatching (DYNDISP) models:

- **DYNDISP=OFF**: LPAR timeslicing controls the CF processor.
- **DYNDISP=ON**: An optimization over pure LPAR timeslicing, in which the CFCC code manages timer interrupts to share processors more efficiently.
- **DYNDISP=THIN**: An interrupt-driven model in which the CF processor is dispatched in response to a set of events that generate Thin Interrupts.



Thin Interrupt support was available since zEC12/zBC12,. It is proven to be efficient and well-performing in numerous different test and customer shared-engine coupling facility configurations.

Therefore, IBM z15 made **DYNDISP=THIN** the *default mode* of operation for coupling facility images that use shared processors.

For more information about CFCC code levels, see [the Parallel Sysplex page](#) of the IBM IT infrastructure website.

For more information about the latest CFCC code levels, see [the current exception letter](#) that is published on Resource Link website (login is required).

CF structure sizing changes are expected when upgrading from a previous CFCC Level to CFCC Level 26. In fact, CFLEVEL 26 can have more noticeable CF structure size increases associated with it, especially for smaller structures, because of task-related memory increases that are associated with the increased number of CF tasks in CFLEVEL 26.

Review the CF LPAR size by using the [CFSizer tool](#).

Alternatively, the batch [SIZER utility](#) also can be used for re-sizing your CF structures as needed. Make sure to update CFRM Policy INITISIZE or SIZE values as needed.

## Coupling links support

Integrated Coupling Adapter Short Reach 2.0 (ICA SR2.0) and Coupling Express3 Long Reach (CE3 LR) coupling link options provide high-speed connectivity at short and longer distances over fiber optic interconnections.

For more information, see 4.6.4, “Parallel Sysplex connectivity” on page 199.

### ***Integrated Coupling Adapter***

PCIe Gen4 coupling fan-out, which is also known as Integrated Coupling Adapter Short Reach (ICA SR2.0), supports a maximum distance of 150 meters (492 feet) and is defined as CHPID type CS6 in IOCP. ICA SR2.0 is fully compatible with ICA SR and ICA SR1.1.

### ***Coupling Express3 Long Reach***

The CE3 LR link provides point-to-point coupling connectivity at distances of 10 km (6.21 miles) unrepeated using either 10Gb optics as CHPID type CL5, or 25Gb optics as CHPID type CL6 (25G) in IOCP. The supported operating systems are listed in Table 7-5 on page 272.

Coupling Express LR and Coupling Express2 LR will not be available on IBM z17, neither as carry-forward or as new build

This new CE3 LR adapter will support 2 varieties of optics:

- ▶ 10Gb (as today)  
When configured to use 10Gb optics, the adapter will remain compatible with existing CE LR and CE2 LR (CL5) links on previous machines, and will be represented as a CL5 coupling link type:  
  
CL5 only connects to CL5; Use for connections back to previous machines using CL5, or for connections to other IBM z17 machines still using CL5
- ▶ A new 25Gb option for higher bandwidth and higher capacity/potential throughput.  
When configured to use 25Gb optics, the adapter will be incompatible with existing CE LR and CE2 LR (CL5) links on previous machines and can only connect to another 25Gb CE3 LR adapter. Because of this incompatibility, one must use a new CL6 coupling link type for



this CL6 only connects to CL6; Use ONLY for connections to other IBM z17 CPCs using CL6.

Other than the higher bandwidth and new link type, CL6 looks and behaves much the same as CL5, except that in the HCD definitions:

Also, because of its higher bandwidth, CL6 operates in a higher path selection “preference tier” than CL5; other things being equal, selection of CL6 CHPIDs is preferred over CL5

**Important:** For using CE3 LR, APAR OA64478 (DG), OA64362 (XCF/XES), OA64591 (IOS), OA65190 (IOCP), OA64114 (HCD) are required – z/OS 2.4 and higher.

## Virtual Flash Memory use by CFCC

IBM z17 VFM use in coupling facility has been discontinued.

## CFCC Coupling Thin Interrupts (required for IBM z17)

The Coupling Thin Interrupts improves the performance of a CF partition and the dispatching of z/OS LPARs that are awaiting the arrival of returned asynchronous CF requests when used in a shared engine environment.

For more information, see “**Coupling Thin Interrupts**” on page 107. The supported operating systems are listed in Table 7-5 on page 272.

## Asynchronous CF Duplexing for lock structures

Asynchronous CF Duplexing enhancement is a general-purpose interface for any CF Lock structure user. It enables secondary structure updates to be performed asynchronously from primary CF updates. It offers performance advantages for duplexing lock structures and avoids the need for synchronous communication delays during the processing of every duplexed update operation.

Asynchronous CF Duplexing for lock structures requires the following software support:

- ▶ z/OS V3R1
- ▶ z/OS V2R5, V2R4
- ▶ z/VM V7R4, V7R3
- ▶ Db2 V12 with PTFs for APAR PI66689
- ▶ IRLM V2.R3 with PTFs for APAR PI68378

The supported operating systems are listed in Table 7-5 on page 272.

## Asynchronous cross-invalidate for CF cache structures

Asynchronous cross-invalidate (XI) for CF cache structures enables improved efficiency in CF data sharing by adopting a more transactional behavior for cross-invalidate (XI) processing. This processing is used to maintain coherency and consistency of data managers' local buffer pools across the sysplex.

Instead of performing XI signals synchronously on every cache update request that causes them, data managers can “opt in” for the CF to perform these XIs asynchronously (and then synchronize them with the CF at or before the transaction is completed). Data integrity is maintained if all XI signals complete by the time transaction locks are released.

The feature enables faster completion of cache update CF requests, especially with the cross-site distance that is involved. It also provides improved cache structure service times and coupling efficiency. It requires specific data manager use or participation, which is not transparent to the data manager. No SMF data changes were made for CF monitoring and reporting.



The following requirements must be met:

- ▶ CFCC Level 24 or higher
- ▶ z/OS V2.R5 or V2.R4
- ▶ PTFs on every exploiting system in the sysplex: Fixes for APAR OA54688 - Exploitation support z/OS V2.R3
- ▶ Db2 V12 with PTFs for exploitation

### **z/VM Dynamic I/O support for ICA CHPIDs**

z/VM dynamic I/O configuration support allows you to add, delete, and modify the definitions of channel paths, control units, and I/O devices to the server and z/VM without shutting down the system.

This function refers exclusively to the z/VM dynamic I/O support of ICA coupling links. Support is available for the CS5 CHPID type in the z/VM dynamic commands, including the **change channel path** dynamic I/O command.

Specifying and changing the system name when entering and leaving configuration mode are also supported. The supported operating systems are listed in Table 7-5 on page 272.

## **7.4.4 Storage connectivity-related features and functions**

In this section, we describe the storage connectivity-related features and functions.

### **zHyperlink Express**

IBM z14 introduced IBM zHyperLink Express as a brand new IBM Z input/output (I/O) channel link technology since FICON. The zHyperLink Express 1.1 feature is available with new IBM z16 systems and is designed to help bring data close to processing power, increase the scalability of transaction processing, and lower I/O latency.

zHyperLink Express is designed for up to 5x lower latency than High-Performance FICON for IBM Z (zHPF) by directly connecting the IBM Z central processor complex (CPC) to the I/O Bay of the DS8000 (DS8880 or later). This short distance (up to 150 m [492.1 feet]), direct connection is intended to speed Db2 for z/OS transaction processing and improve active log throughput.

The improved performance of zHyperLink Express allows the Processing Unit (PU) to make a synchronous request for the data that is in the DS8000 cache. This process eliminates the undispatch of the running request, the queuing delays to resume the request, and the PU cache disruption.

Support for zHyperLink Writes can accelerate Db2 log writes to help deliver superior service levels by processing high-volume Db2 transactions at speed. IBM zHyperLink Express requires compatible levels of DS8000/F hardware, firmware R8.5.1 or later, and Db2 12 with PTFs.

The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

### **FICON Express32-4P**

FICON Express32-4P four port feature is available on IBM z17. It has four PCHID/CHPIDs (ports) and supports a link data rate of 32 gigabits per second (Gbps) and auto negotiation to 16 and 8 Gbps for synergy with switches, directors, and storage devices. With support for native FICON, High-Performance FICON for Z (zHPF), and Fibre Channel Protocol (FCP),



the IBM z17 server enables you to position your SAN for even higher performance, which helps you to prepare for an end-to-end 16 Gbps infrastructure to meet the lower latency and increased bandwidth demands of your applications.

### **FICON Express32S**

FICON Express32S (available for IBM z17 servers) supports a link data rate of 32 gigabits per second (Gbps) and auto negotiation to 16 and 8 Gbps for synergy with switches, directors, and storage devices. With support for native FICON, High-Performance FICON for Z (zHPF), and Fibre Channel Protocol (FCP), the IBM z17 server enables you to position your SAN for even higher performance, which helps you to prepare for an end-to-end 16 Gbps infrastructure to meet the lower latency and increased bandwidth demands of your applications.

The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

### ***IBM Fibre Channel Endpoint Security***

IBM z17 Model ME1 supports IBM Fibre Channel Endpoint Security feature (FC 1146). FC 1146 provides FC/FCP link encryption and endpoint authentication. This optional priced feature requires the following components:

- ▶ FICON Express 32-4P (FCs0387 and 0388), FICON Express32S (FCs 0461 and 0462), and FICON Express16SA (carry forward FCs 0436 and 0437), for both link encryption and endpoint authentication
- ▶ Select DS8000 storage
- ▶ Supporting infrastructure: IBM Guardium Security Key Lifecycle Manager (GSKLM)
- ▶ CPACF enablement (FC 3863)

For more information, see this [announcement letter](#).

### **FICON Express16SA**

FICON Express16SA (carry forward to IBM z16 server) supports a link data rate of 16 gigabits per second (Gbps) and auto negotiation to 8 Gbps for synergy with switches, directors, and storage devices. With support for native FICON, High-Performance FICON for IBM Z (zHPF), and Fibre Channel Protocol (FCP), the IBM z17 server enables you to position your SAN for even higher performance, which helps you to prepare for an end-to-end 16 Gbps infrastructure to meet the lower latency and increased bandwidth demands of your applications.

The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

### **Extended distance FICON**

An enhancement to the industry-standard FICON architecture (FC-SB-3) helps avoid degradation of performance at extended distances by implementing a new protocol for persistent IU pacing. Extended distance FICON is transparent to operating systems and applies to all FICON Express32-4P, FICON Express32S, and FICON Express16SA, features that carry native FICON traffic (CHPID type FC).

To use this enhancement, the control unit must support the new IU pacing protocol. IBM System Storage DS8000 series supports extended distance FICON for IBM Z environments. The channel defaults to current pacing values when it operates with control units that cannot use extended distance FICON.



## High-performance FICON

High-performance FICON (zHPF) was first provided on IBM System z10®, and is a FICON architecture for protocol simplification and efficiency. It reduces the number of information units (IUs) that are processed. Enhancements were made to the z/Architecture and the FICON interface architecture to provide optimizations for online transaction processing (OLTP) workloads.

zHPF is available on the following servers:

- ▶ IBM z17
- ▶ IBM z16
- ▶ IBM z15

As of this writing, the FICON Express32-4P, FICON Express32S, and FICON Express16SA, (CHPID type FC) support the FICON protocol and the zHPF protocol in the server LIC.

When used by the FICON channel, the z/OS operating system, and the DS8000 control unit or other subsystems, the FICON channel processor usage can be reduced and performance improved. Suitable levels of Licensed Internal Code (LIC) are required.

Also, the changes to the architectures provide end-to-end system enhancements to improve reliability, availability, and serviceability (RAS).

zHPF is compatible with the following standards:

- ▶ Fibre Channel Framing and Signaling standard (FC-FS)
- ▶ Fibre Channel Switch Fabric and Switch Control Requirements (FC-SW)
- ▶ Fibre Channel Single-Byte-4 (FC-SB-4) standards

For example, the zHPF channel programs can be used by the z/OS OLTP I/O workloads, Db2, VSAM, the partitioned data set extended (PDSE), and the z/OS file system (zFS).

At the zHPF announcement, zHPF supported the transfer of small blocks of fixed size data (4 K) from a single track. This capability was extended, first to 64 KB, and then to multi-track operations. The 64 KB data transfer limit on multi-track operations was removed by z196. This improvement allows the channel to fully use the bandwidth of FICON channels, which results in higher throughputs and lower response times.

The multi-track operations extension applies to the FICON Express32-4P, FICON Express32S, and FICON Express16SA, when configured as CHPID type FC and connecting to z/OS. zHPF requires matching support by the DS8000 series. Otherwise, the extended multi-track support is transparent to the control unit.

zHPF is enhanced to allow all large write operations (greater than 64 KB) at distances up to 100 km (62.13 miles) to be run in a single round trip to the control unit. This process does not elongate the I/O service time for these write operations at extended distances. This enhancement to zHPF removes a key inhibitor for customers that are adopting zHPF over extended distances, especially when the IBM HyperSwap capability of z/OS is used.

From the z/OS perspective, the FICON architecture is called *command mode* and the zHPF architecture is called *transport mode*. During link initialization, the channel node and the control unit node indicate whether they support zHPF.

**Requirement:** All FICON channel path identifiers (CHPIDs) that are defined to the same LCU must support zHPF. The inclusion of any non-compliant zHPF features in the path group causes the entire path group to support command mode only.



The mode that is used for an I/O operation depends on the control unit that supports zHPF and its settings in the z/OS operating system. For z/OS use, a parameter is available in the IECIOSxx member of SYS1.PARMLIB (**ZHPF=YES or NO**) and in the **SETIOS** system command to control whether zHPF is enabled or disabled. The default is ZHPF=NO.

Support also is added for the **D IOS,ZHPF** system command to indicate whether zHPF is enabled, disabled, or not supported on the server.

Similar to the existing FICON channel architecture, the application or access method provides the channel program (CCWs). How zHPF (transport mode) manages channel program operations is different from the CCW operation for the existing FICON architecture (command mode).

While in command mode, each CCW is sent to the control unit for execution. In transport mode, multiple channel commands are packaged together and sent over the link to the control unit in a single control block. Fewer processors are used compared to the existing FICON architecture. Specific complex CCW chains are not supported by zHPF.

The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

For more information about FICON channel performance, see the performance technical papers that are available [at the IBM Z I/O connectivity page](#) of the IBM IT infrastructure website.

### **Modified Indirect Data Address Word facility**

The Modified Indirect Data Address Word (MIDAW) facility improves FICON performance. It provides a more efficient channel command word (CCW)/indirect data address word (IDAW) structure for specific categories of data-chaining I/O operations.

The MIDAW facility is a system architecture and software feature that is designed to improve FICON performance. This facility was first made available on IBM System z9® servers, and is used by the Media Manager in z/OS.

The MIDAW facility provides a more efficient CCW/IDAW structure for certain categories of data-chaining I/O operations. MIDAW can improve FICON performance for extended format data sets. Nonextended data sets also can benefit from MIDAW.

MIDAW can improve channel use and I/O response time. It also reduces FICON channel connect time, director ports, and control unit processor usage.

IBM laboratory tests indicate that applications that use extended format (EF) data sets, such as Db2, or long chains of small blocks can gain significant performance benefits by using the MIDAW facility.

MIDAW is supported on FICON channels that are configured as CHPID type FC. The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

### ***MIDAW technical description***

An IDAW is used to specify data addresses for I/O operations in a virtual environment.<sup>12</sup> The IDAW design allows the first IDAW in a list to point to any address within a page. Subsequent IDAWs in the same list must point to the first byte in a page. Also, IDAWs (except the first and last IDAW) in a list must manage complete 2 K or 4 K units of data.

---

<sup>12</sup> Exceptions are made to this statement, and many details are omitted in this description. In this section, we assume that you can merge this brief description with an understanding of I/O operations in a virtual memory environment.



Figure 7-2 shows a single CCW that controls the transfer of data that spans noncontiguous 4 K frames in main storage. When the IDAW flag is set, the data address in the CCW points to a list of words (IDAWs). Each IDAW contains an address that designates a data area within real storage.

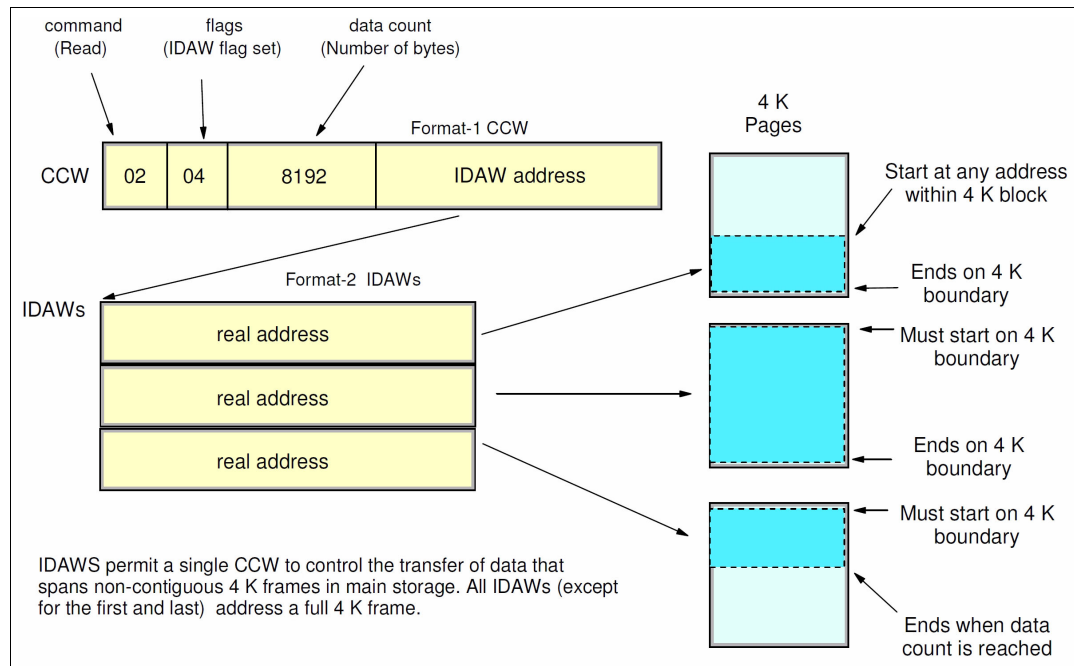


Figure 7-2 IDAW usage

The number of required IDAWs for a CCW is determined by the following factors:

- ▶ IDAW format as specified in the operation request block (ORB)
- ▶ Count field of the CCW
- ▶ Data address in the initial IDAW

For example, three IDAWS are required when the following events occur:

- ▶ The ORB specifies format-2 IDAWs with 4 KB blocks.
- ▶ The CCW count field specifies 8 KB.
- ▶ The first IDAW designates a location in the middle of a 4 KB block.

CCWs with data chaining can be used to process I/O data blocks that have a more complex internal structure, in which portions of the data block are directed into separate buffer areas. This process is sometimes known as *scatter-read* or *scatter-write*. However, as technology evolves and link speed increases, data chaining techniques become less efficient because of switch fabrics, control unit processing and exchanges, and other issues.



The MIDAW facility is a method of gathering and scattering data from and into discontinuous storage locations during an I/O operation. The MIDAW format is shown in Figure 7-3. It is 16 bytes long and aligned on a quadword.

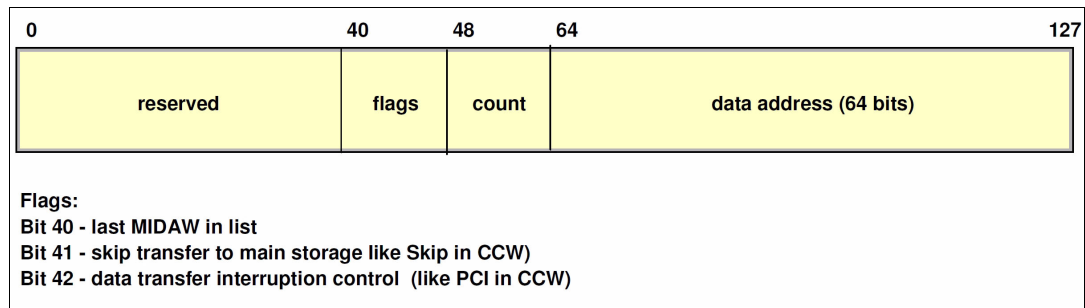


Figure 7-3 MIDAW format

An example of MIDAW usage is shown in Figure 7-4.

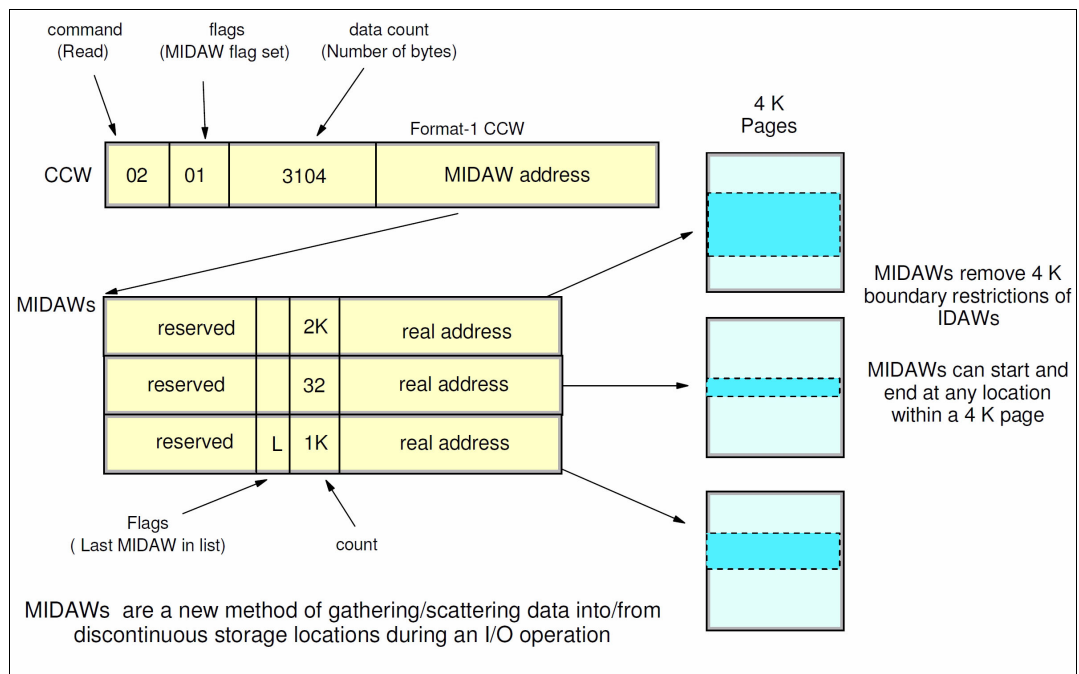


Figure 7-4 MIDAW usage

The use of MIDAWs is indicated by the MIDAW bit in the CCW. If this bit is set, the skip flag cannot be set in the CCW. The skip flag in the MIDAW can be used instead. The data count in the CCW must equal the sum of the data counts in the MIDAWs. The CCW operation ends when the CCW count goes to zero or the last MIDAW (with the last flag) ends.

The combination of the address and count in a MIDAW cannot cross a page boundary. Therefore, the largest possible count is 4 K. The maximum data count of all the MIDAWs in a list cannot exceed 64 K, which is the maximum count of the associated CCW.

The scatter-read or scatter-write effect of the MIDAWs makes it possible to efficiently send small control blocks that are embedded in a disk record to separate buffers from those that are used for larger data areas within the record. MIDAW operations are on a single I/O block, in the manner of data chaining. Do not confuse this operation with CCW command chaining.



**Extended format data sets**

z/OS extended format (EF) data sets use internal structures (often not visible to the application program) that require a scatter-read (or scatter-write) operation. Therefore, CCW data chaining is required, which produces less than optimal I/O performance. Because the most significant performance benefit of MIDAWs is achieved with EF data sets, a brief review of the EF data sets is included here.

VSAM and non-VSAM (DSORG=PS) sets can be defined as EF data sets. For non-VSAM data sets, a 32-byte suffix is appended to the end of every physical record (that is, block) on disk. VSAM appends the suffix to the end of every control interval (CI), which normally corresponds to a physical record.

A 32 K CI is split into two records to span tracks. This suffix is used to improve data reliability, and facilitates other functions that are described next. For example, if the DCB BLKSIZE or VSAM CI size is equal to 8192, the block on storage consists of 8224 bytes. The control unit does not distinguish between suffixes and user data. The suffix is transparent to the access method and database.

In addition to reliability, EF data sets enable the following functions:

- ▶ DFSMS striping
- ▶ Access method compression
- ▶ Extended addressability (EA)

EA is useful for creating large Db2 partitions (larger than 4 GB). Striping can be used to increase sequential throughput, or to spread random I/Os across multiple logical volumes. DFSMS striping is useful for the use of multiple channels in parallel for one data set. The Db2 logs are often striped to optimize the performance of Db2 sequential inserts.

Processing an I/O operation to an EF data set normally requires at least two CCWs with data chaining. One CCW is used for the 32-byte suffix of the EF data set. With MIDAW, the extra CCW for the EF data set suffix is eliminated.

MIDAWs benefit EF and non-EF data sets. For example, to read 12 4 K records from a non-EF data set on a 3390 track, Media Manager chains together 12 CCWs by using data chaining. To read 12 4 K records from an EF data set, 24 CCWs are chained (two CCWs per 4 K record). By using Media Manager track-level command operations and MIDAWs, an entire track can be transferred by using a single CCW.

**Performance benefits**

z/OS Media Manager includes I/O channel program support for implementing EF data sets, and automatically uses MIDAWs when appropriate. Most disk I/Os in the system are generated by using Media Manager.

Users of the Executing Fixed Channel Programs in Real Storage (EXCPVR) instruction can construct channel programs that contain MIDAWs. However, doing so requires that they construct an IOBE with the IOBEMIDA bit set. Users of the EXCP instruction cannot construct channel programs that contain MIDAWs.

The MIDAW facility removes the 4 K boundary restrictions of IDAWs and for EF data sets, which reduces the number of CCWs. Decreasing the number of CCWs helps to reduce the FICON channel processor use. Media Manager and MIDAWs do not cause the bits to move any faster across the FICON link. However, they reduce the number of frames and sequences that flow across the link and use the channel resources more efficiently.



The performance of a specific workload can vary based on the conditions and hardware configuration of the environment. IBM laboratory tests found that Db2 gains significant performance benefits by using the MIDAW facility in the following areas:

- ▶ Table scans
- ▶ Logging
- ▶ Utilities
- ▶ Use of DFSMS striping for Db2 data sets

Media Manager with the MIDAW facility can provide significant performance benefits when used in combination applications that use EF data sets (such as Db2) or long chains of small blocks.

For more information about FICON and MIDAW, see the following resources:

- ▶ The [I/O Connectivity](#) page of the IBM IT infrastructure website includes information about FICON channel performance
- ▶ *DS8000 Performance Monitoring and Tuning*, [SG24-8013](#)

## ICKDSF

Device Support Facilities, ICKDSF, Release 17 is required on all systems that share disk subsystems with an IBM z17 processor.

ICKDSF supports a modified format of the CPU information field that contains a two-digit LPAR identifier. ICKDSF uses the CPU information field instead of CCW reserve/release for concurrent media maintenance. It prevents multiple systems from running ICKDSF on the same volume, and at the same time allows user applications to run while ICKDSF is processing. To prevent data corruption, ICKDSF must determine all sharing systems that might run ICKDSF. Therefore, this support is required for IBM z17.

**Remember:** The need for ICKDSF Release 17 also applies to systems that are not part of the same sysplex, or are running an operating system other than z/OS, such as z/VM.

## z/OS Discovery and Auto-Configuration

z/OS Discovery and Auto Configuration (zDAC) is designed to automatically run several I/O configuration definition tasks for new and changed disk and tape controllers that are connected to a switch or director, when attached to a FICON channel.

The zDAC function is integrated into the hardware configuration definition (HCD). Clients can define a policy that can include preferences for availability and bandwidth that include parallel access volume (PAV) definitions, control unit numbers, and device number ranges. When new controllers are added to an I/O configuration or changes are made to existing controllers, the system discovers them and proposes configuration changes that are based on that policy.

zDAC provides real-time discovery for the FICON fabric, subsystem, and I/O device resource changes from z/OS. By exploring the discovered control units for defined logical control units (LCUs) and devices, zDAC compares the discovered controller information with the current system configuration. It then determines delta changes to the configuration for a proposed configuration.

All added or changed logical control units and devices are added into the proposed configuration. They are assigned proposed control unit and device numbers, and channel paths that are based on the defined policy. zDAC uses channel path chosen algorithms to minimize single points of failure. The zDAC proposed configurations are created as work I/O definition files (IODFs) that can be converted to production IODFs and activated.



zDAC is designed to run discovery for all systems in a sysplex that support the function. Therefore, zDAC helps to simplify I/O configuration on IBM z17 systems that run z/OS, and reduces complexity and setup time.

zDAC applies to all FICON features that are supported on IBM z17 when configured as CHPID type FC. The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

### **Platform and name server registration in FICON channel**

The FICON Express32-4P, FICON Express32S, and FICON Express16SA features support platform and name server registration to the fabric for CHPID types FC and FCP.

Information about the channels that are connected to a fabric (if registered) allows other nodes or storage area network (SAN) managers to query the name server to determine what is connected to the fabric.

The following attributes are registered for the IBM z17 systems:

- ▶ Platform information
- ▶ Channel information
- ▶ Worldwide port name (WWPN)
- ▶ Port type (N\_Port\_ID)
- ▶ FC-4 types that are supported
- ▶ Classes of service that are supported by the channel

The platform and name server registration service are defined in the Fibre Channel Generic Services 4 (FC-GS-4) standard.

### **The 63.75-K subchannels**

Servers before IBM z9 EC reserved 1024 subchannels for internal system use, out of a maximum of 64 K subchannels. Starting with IBM z9 EC, the number of reserved subchannels was reduced to 256, which increased the number of subchannels that are available. Reserved subchannels exist in subchannel set 0 only. One subchannel is reserved in each of subchannel sets 1, 2, and 3.

The informal name, 63.75-K subchannels, represents 65280 subchannels, as shown in the following equation:

$$63 \times 1024 + 0.75 \times 1024 = 65280$$

This equation is applicable for subchannel set 0. For subchannel sets 1, 2 and 3, the available subchannels are derived by using the following equation:

$$(64 \times 1024) - 1 = 65535$$

The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

### **Multiple subchannel sets**

First introduced in IBM z9 EC, multiple subchannel sets (MSS) provide a mechanism for addressing more than 63.75 K I/O devices and aliases for FICON (CHPID types FC) on the IBM z17, IBM z16, IBM z15, IBM z14, IBM z13, IBM z13s, IBM zEC12, and IBM zBC12. IBM z196 introduced the third subchannel set (SS2). With IBM z13, one more subchannel set (SS3) was introduced, which expands the alias addressing by 64 K more I/O devices.



Current z/VM versions MSS support for mirrored direct access storage device (DASD) provides a subset of host support for the MSS facility to allow the use of an alternative subchannel set for Peer-to-Peer Remote Copy (PPRC) secondary volumes.

The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273. For more information about channel subsystem, see Chapter 5, “Central processor complex channel subsystem” on page 209.

### ***Subchannel sets***

IBM z17 ME1 supports four subchannel sets (SS0, SS1, SS2, and SS3).

Subchannel sets SS1, SS2, and SS3 can be used for disk alias devices of primary and secondary devices, and as Metro Mirror secondary devices. This set helps facilitate storage growth and complements other functions, such as extended address volume (EAV) and Hyper Parallel Access Volumes (HyperPAV).

See Table 7-6 on page 272 and Table 7-7 on page 273 for list of supported operating systems.

### **IPL from an alternative subchannel set**

IBM z17 supports IPL from subchannel set 1 (SS1), subchannel set 2 (SS2), or subchannel set 3 (SS3), in addition to subchannel set 0.

See Table 7-6 on page 272 and Table 7-7 on page 273 for list of supported operating systems. For more information, see “IPL from an alternative subchannel set” on page 317.

### **32 K subchannels**

To help facilitate growth and continue to enable server consolidation, the IBM z17 supports up to 32 K subchannels per FICON Express32-4P, FICON Express32S, and FICON Express16SA channels (CHPID). More devices can be defined per FICON channel, which includes primary, secondary, and alias devices. The maximum number of subchannels across all device types that are addressable within an LPAR remains at 63.75 K for subchannel set 0 and 64 K (64 X 1024)-1 for subchannel sets 1, 2, and 3.

This support is available to the IBM z17, IBM z16, IBM z15, z14, IBM z13, and IBM z13s servers, and in IBM z17 it applies to the FICON Express32-4P, FICON Express32S, and FICON Express16SA features (defined as CHPID type FC).

The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

### **Request node identification data**

The request node identification data (RNID) function for native FICON CHPID type FC allows isolation of cabling-detected errors. The supported operating systems are listed in Table 7-6 on page 272.

### **FICON link incident reporting**

FICON link incident reporting allows an operating system image (without operator intervention) to register link incident reports. The supported operating systems are listed in Table 7-6 on page 272.



## Health Check for FICON Dynamic routing

Starting with IBM z13, the channel microcode was changed to support FICON dynamic routing. Although change is required in z/OS to support dynamic routing, I/O errors can occur if the FICON switches are configured for dynamic routing despite the missing support in the processor or storage controllers. Therefore, a health check is provided that interrogates the switch to determine whether dynamic routing is enabled in the switch fabric.

No action is required on z/OS to enable the health check; it is automatically enabled at IPL and reacts to changes that might cause problems. The health check can be disabled by using the **PARMLIB** or **SDSF** modify commands.

The supported operating systems are listed in Table 7-6 on page 272. For more information about FICON Dynamic Routing (FIDR), see Chapter 4, “Central processor complex I/O structure” on page 169.

## Global resource serialization FICON CTC toleration

For some configurations that depend on ESCON CTC definitions, global resource serialization (GRS) FICON CTC toleration that is provided with APAR OA38230 is essential, especially after ESCON channel support was removed from IBM Z starting with IBM zEC12.

The supported operating systems are listed in Table 7-6 on page 272.

## Increased performance for the FCP protocol

The FCP LIC is modified to help increase I/O operations per second for small and large block sizes, and to support 32-Gbps link speeds.

For more information about FCP channel performance, see [the performance technical papers that are available](#) at the IBM Z I/O connectivity page of the IBM IT infrastructure website.

The FCP protocol is supported by z/VM, z/VSE, and Linux on IBM Z. The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

## T10-DIF support

American National Standards Institute (ANSI) T10 Data Integrity Field (DIF) standard is supported on IBM Z for SCSI end-to-end data protection on fixed block (FB) LUN volumes. IBM Z provides added end-to-end data protection between the operating system and the DS8870 unit. This support adds protection information that consists of Cyclic Redundancy Checking (CRC), Logical Block Address (LBA), and host application tags to each sector of FB data on a logical volume.

IBM Z support applies to FCP channels only. The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

## N\_Port ID Virtualization

N\_Port ID Virtualization (NPIV) allows multiple system images (in LPARs or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. First introduced with z9 EC, this feature can be used with supported FICON features on IBM z17 servers. The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

## Worldwide port name tool

Part of the IBM z17 system installation is the planning of the SAN environment. IBM includes a stand-alone tool to assist with this planning before the installation.



The capabilities of the WWPN are extended to calculate and show WWPNs for virtual and physical ports ahead of system installation.

The tool assigns WWPNs to each virtual FCP channel or port by using the same WWPN assignment algorithms that a system uses when assigning WWPNs for channels that use NPIV. Therefore, the SAN can be set up in advance, which allows operations to proceed much faster after the server is installed. In addition, the SAN configuration can be retained instead of altered by assigning the WWPN to physical FCP ports when a FICON feature is replaced.

The WWPN tool takes a .csv file that contains the FCP-specific I/O device definitions and creates the WWPN assignments that are required to set up the SAN. A binary configuration file that can be imported later by the system is also created. The .csv file can be created manually or exported from the HCD/HCM. The supported operating systems are listed in Table 7-6 on page 272 and Table 7-7 on page 273.

The WWPN tool is applicable to all FICON channels that are defined as CHPID type FCP (for communication with SCSI devices) on IBM z17. It is available [for download at the Resource Link](#) at the following website (log in is required).

**Note:** An optional feature can be ordered for WWPN persistency before shipment to keep the same I/O serial number on the new CPC. Current information must be provided during the ordering process.

## 7.4.5 Networking features and functions

In this section, we describe the networking features and functions that are supported on IBM z17.

### IBM z17 integrated I/O architecture

IBM z17 introduces a new I/O accelerator implemented on the processor chip for use with new FICON and Networking cards. This is the Data Processing Unit I/O Accelerator and is a distinct core implemented on the PU chip with access to L2 and L3 cache. The DPU handles much of the channel firmware that used to be performed on an I/O adapter and should therefore be faster and have improved latency.

The DPU I/O Complex moves functionality from the ASIC on the I/O adapters and in-boards it into Assist Processors on the PU chip.

The DPU design aims to build an I/O subsystem with similar or better qualities of services than the existing I/O subsystem. It is designed to deliver value by improved performance and power efficiency. It also delivers value to our customers with decreased channel latencies. For more information see: 5.4, “IBM z17 Data Processing Unit (DPU)” on page 219

The goals of this implementation are to deliver IBM Z platform efficiencies:

1. Improve peak I/O start rates and reduce latencies
2. Provide focused per port recovery for the most common types of failures
3. Improve recurring networking costs for customers by providing integrated RoCE SMC-R and OSA support, allowing single port serviceability for all DPU managed I/O adapters
4. Reduce dependency on the PCI support partition by providing physical function support for PCIe Native use cases

Data Processing Unit (DPU) supported protocols:

1. Legacy mode FICON



2. HPF (High Performance FICON)
3. FCP (SCSI over fiber channel)
4. OSA (Open System Adapter)
5. OSA-ICC (Open System Adapter – Integrated Console Controller)
6. Physical function support for Native Ethernet exploitation

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

### **Measurements**

IBM z17 has an entirely new I/O hardware and architecture model for both storage and networking. The new design moves processor and memory closer, which transforms I/O operations to allow workloads to grow and scale. Adds new channel measurement characteristics and channel utilization counts for new Channel Measurement Groups (CMG) 4 and 5. The channel utilization counts and channel measurement characteristics for the new CMGs are provided in Channel Path Measurement Block (IRACPMB) which is populated by the Channel Path Measurement Facility (CPMF) on IBM zNEXT hardware. Also z/OS WLM provides support in the Channel Path Measurement Block (IRACPMB).

z/OS Data Gatherer extends its data collection for Channel Measurement Groups 4 and 5 for SMF 73 records and Monitor III channel data table (ERBCPDG3).

**Note:** The support is available for z/OS V2.5 or higher. The Exploitation support is available via APAR OA66014 and OA66054.

**Restriction:** The Channel Path Measurement Facility is not available under z/VM, no channel measurement characteristics and utilization data will be retrieved.

### **Shared Memory Communication - Direct Memory Access**

First introduced with IBM z13 servers, the Shared Memory Communication - Direct Memory Access (SMC-D) feature maintains the socket-API transparency aspect of SMC-R so that applications that use TCPI/IP communications can benefit immediately without requiring application software to undergo IP topology changes.

Similar to SMC-R, this protocol uses shared memory architectural concepts that eliminate TCP/IP processing in the data path, yet preserve TCP/IP Qualities of Service for connection management purposes.

Support in select Linux on IBM Z distributions is now provided for Shared Memory Communications over Direct Memory Access (SMC-D). For more information, see [this Linux on IBM Z Blogspot web page](#).

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.



## Shared Memory Communications over RDMA Version 2

Shared Memory Communications v2 is available in z/OS V2R4 (with PTFs), z/OS V2R5, and z/OS V3R1.

Because the initial version of SMC was limited to TCP/IP connections over the same layer 2 network, it was not routable across multiple IP subnets. The associated TCP/IP connection was limited to hosts within a single IP subnet that required the hosts to have direct access to the same physical layer 2 network (that is, the same Ethernet LAN over a single VLAN ID). The scope of eligible TCP/IP connections for SMC was limited to and defined by the single IP subnet.

SMC over RDMA Version 2 (SMC-R v2) provides support for SMC over multiple IP subnets for both SMC-D and SMC-R and is referred to as SMC-D v2 and SMC-R v2. SMC v2 requires updates to the underlying network technology. SMC-D v2 requires ISM v2 and SMC-R v2 requires RoCE v2.

The SMC-R v2 protocol is downward compatible, which allows SMC-R v2 hosts to continue to communicate with SMC-R v1 previous hosts.

Although SMC-R v2 changes the SMC connection protocol to enable multiple IP subnet support, SMC-R v2 does not change how user TCP socket data is transferred, which preserves the benefits of SMC to TCP workloads.

TCP/IP connections that require IPsec are not eligible for any form of SMC.

## HiperSockets Completion Queue

The HiperSockets Completion Queue function is implemented on IBM z17, IBM z16, and IBM z15. This function is designed to allow HiperSockets to transfer data synchronously (if possible) and asynchronously, if necessary. Therefore, it combines ultra-low latency with more tolerance for traffic peaks. HiperSockets Completion Queue can be especially helpful in burst situations.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## HiperSockets Virtual Switch Bridge

The HiperSockets Virtual Switch Bridge is supported on IBM Z servers. With the HiperSockets Virtual Switch Bridge, z/VM virtual switch is enhanced to transparently bridge a guest virtual machine network connection on a HiperSockets LAN segment. This bridge allows a single HiperSockets guest virtual machine network connection to also directly communicate with the following components:

- ▶ Other guest virtual machines on the virtual switch
- ▶ External network hosts through the virtual switch OSA UPLINK port

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## HiperSockets Multiple Write Facility

The HiperSockets Multiple Write Facility allows the streaming of bulk data over a HiperSockets link between two LPARs. Multiple output buffers are supported on a single Signal Adapter (SIGA) write instruction. The key advantage of this enhancement is that it allows the receiving LPAR to process a much larger amount of data per I/O interrupt. This process is transparent to the operating system in the receiving partition. HiperSockets



Multiple Write Facility with fewer I/O interrupts is designed to reduce processor use of the sending and receiving partitions.

Support for this function is required by the sending operating system. For more information, see “HiperSockets” on page 198. The supported operating systems are listed in Table 7-8 on page 275.

## **HiperSockets support of IPv6**

IPv6 is a key element in the future of networking. The IPv6 support for HiperSockets allows compatible implementations between external networks and internal HiperSockets networks. The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## **HiperSockets Layer 2 support**

For flexible and efficient data transfer for IP and non-IP workloads, the HiperSockets internal networks on IBM Z servers can support two transport modes: Layer 2 (Link Layer) and the current Layer 3 (Network or IP Layer). Traffic can be Internet Protocol (IP) Version 4 or Version 6 (IPv4, IPv6) or non-IP (AppleTalk, DECnet, IPX, NetBIOS, or SNA).

HiperSockets devices are protocol-independent and Layer 3-independent. Each HiperSockets device features its own Layer 2 Media Access Control (MAC) address. This MAC address allows the use of applications that depend on the existence of Layer 2 addresses, such as Dynamic Host Configuration Protocol (DHCP) servers and firewalls.

Layer 2 support can help facilitate server consolidation. Complexity can be reduced, network configuration is simplified and intuitive, and LAN administrators can configure and maintain the mainframe environment the same way as they do a non-mainframe environment.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## **HiperSockets network traffic analyzer for Linux on IBM Z**

HiperSockets network traffic analyzer (HS NTA) provides support for tracing Layer2 and Layer3 HiperSockets network traffic in Linux on IBM Z. This support allows Linux on IBM Z to control the trace for the internal virtual LAN to capture the records into host memory and storage (file systems).

Linux on IBM Z tools can be used to format, edit, and process the trace records for analysis by system programmers and network administrators.

## **Network Express features**

Network Express allows RoCE and OSA networking features to converge into a single network feature. Reducing cost for physical networking resources (drawer I/O slots, adapters, ports, cables, switch port). One single port on the new adapter can simultaneously have two „personalities“, managed independently, although service action on the physical card/port requires both entities to be varied offline.

<b>OSH</b>	using Enhanced QDIO for OSA (Ethernet) – new OSA networking CHPID type (TCP/IP)
<b>NETH</b>	using native PCI for RoCE (SMC-R - Shared memory communications over RDMA)
<b>NETD</b>	Linux on IBM Z native support (all protocols TCP/IP, RDMA protocols)



The IBM z17 SMC-Rv2 support requires OSH and NETH CHPID types must be converged on the same PCHID/port, with matching interface statements. The Network Express supports Enhanced QDIO (EQDIO) architecture, allowing z/OS Communications Server to interact with hardware using optimized operations, to allow growing I/O rates.

EQDIO builds the foundation for the introduction of advanced Ethernet and networking capabilities, which support IBM Z Hybrid Cloud Enterprise users. The z/OS Communications Server SMF records are updated in order to support the new ports. z/OS IOS supports the new OSA networking CHPID type, OSH (OSA Hybrid – Network Express for Ethernet).

**Note:** The support is available for z/OS V2.5 or higher. The Exploitation support is available via APAR OA63265, PH56528, OA64896, PH54596.

Each port can be configured to provide support for a single host protocol (EQDIO or native PCIe) or combination of host protocols. Adapters can be configured with both ports either as OSH/NETH or as NETD.

**Important:** When used as guest of a z/VM that supports IBM z17, there are two options for network connectivity:

1. Dedicate a device on an OSH CHPID to the z/OS guest. The z/OS configuration will operate the device as an OSH-type device.
2. Connect the z/OS guest to a z/VM virtual switch. The z/OS configuration will continue operate the virtual NIC as an OSD-type device. The virtual switch uplink will provide physical connectivity via a device on an OSH CHPID.  
A z/VM VSwitch supporting Network Express OSH does not currently support z/OS guests exploiting an EQDIO uplink port. In the interim, clients will be required to use either a guest-attached OSH device or existing functionality available with OSA-Express7S adapters.

The new Converged Multi-Function Network Adapter is a two port feature supporting either 10Gb or 25Gb optics and has and one PCHID/CHPID. Both ports must carry the same speed optics. Single mode (LR/LX) or multimode (SR/SX) fiber with Small form factor pluggable (SFP+) optics using LC Duplex connector. Networking Express does NOT auto-negotiate to a slower speed.

## OSA-Express7S 1.2 25 GbE LR and SR features

OSA-Express7S 1.2 features are an Ethernet technology refresh introduced with IBM z16.

OSA-Express7S 1.2 25 GbE SR (FC 0459) and OSA-Express7S 1.2 25 GbE LR (FC 0460) are installed in the PCIe+ I/O Drawer and have 25 GbE physical port. New with the generation is the Long Reach version, which uses single mode fiber and can be point to point connected to a distance of up to 10 km (6.2 miles). The features connect to a 25 GbE switch and do not support auto-negotiation to a different speed.

Consider the following points regarding operating system support:

- ▶ z/OS V2R2 and V2R3 require fixes for the following APARs: OA55256 (IBM VTAM®) and PI95703 (TCP/IP).
- ▶ V7R1 requires PTF for APAR PI99085.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.



## **OSA-Express7S 1.2 10 GbE LR and SR features**

OSA-Express7S 1.2 features are an Ethernet technology refresh introduced with IBM z16.

OSA-Express7S 1.2 10 GbE SR (FC 0457) and OSA-Express7S 1.2 10 GbE LR (FC 0456) are installed in the PCIe+ I/O Drawer and have 10 GbE physical port. The features connect to a 10 GbE switch and do not support auto-negotiation to a different speed.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## **OSA-Express7S 1.2 GbE LX and SX features**

OSA-Express7S 1.2 features are an Ethernet technology refresh that was introduced with IBM z16.

OSA-Express7S 1.2 GbE SX (FC 0455) and OSA-Express7S 1.2 10 GbE LX (FC 0454) are installed in the PCIe+ I/O Drawer and have two GbE physical ports. The features connect to a GbE switch and do not support auto-negotiation to a different speed.

Each adapter can be configured in the following modes:

- ▶ QDIO, with CHPID types OSD
- ▶ Local 3270 emulation, including OSA-ICC, with CHPID type OSC

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## **OSA-Express7S 25 Gigabit Ethernet SR (carry forward to IBM z17)**

OSA-Express7S 25 GbE SR1.1 (FC 0449) and OSA-Express7S 25 GbE (FC 0429) are installed in the PCIe I/O drawer and have one 25 GbE physical port and requires 25 GbE optics and Ethernet switch 25 GbE support (negotiation down to 10 GbE is not supported).

Consider the following points regarding operating system support:

- ▶ z/OS V2R1, V2R2, and V2R3 require fixes for the following APARs: OA55256 (VTAM) and PI95703 (TCP/IP).
- ▶ z/VM V6R4 and V7R1 require PTF for APAR PI99085.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## **OSA-Express7S 10-Gigabit Ethernet LR and SR (carry forward)**

OSA-Express7S 10-Gigabit Ethernet features are installed in the PCIe I/O drawer, which is supported by the 16 GBps PCIe Gen3 host bus. The performance characteristics are comparable to the OSA-Express6S features and they also retain the same form factor and port granularity.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## **OSA-Express6S 10-Gigabit Ethernet LR and SR (carry forward)**

OSA-Express7S 10-Gigabit Ethernet features are installed in the PCIe I/O drawer, which is supported by the 16 GBps PCIe Gen3 host bus. The performance characteristics are comparable to the OSA-Express5S features. They also retain the same form factor and port granularity. OSA-Express6S features were introduced with IBM z14, can be carried forward to an IBM z15 (T01 and T02), and ordered with a new IBM z16 A01



The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

### **OSA-Express7S Gigabit Ethernet LX and SX (carry forward)**

IBM z16 introduces an Ethernet technology refresh with OSA-Express7S Gigabit Ethernet features to be installed in the PCIe I/O drawer, which is supported by the 16 Gbps PCIe Gen3 host bus. The performance characteristics are comparable to the OSA-Express6S features and they also retain the same form factor and port granularity.

Each adapter can be configured in the following modes:

- ▶ QDIO, with CHPID types OSD
- ▶ Local 3270 emulation, including OSA-ICC, with CHPID type OSC

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

With the OSA-ICC function, 3270 emulation for console session connections is integrated through a port on the OSA-Express7S GbE.

OSA-ICC can be configured on a PCHID-by-PCHID basis, and is supported at any of the feature settings. Each port can support up to 120 console session connections.

To improve security of console operations and to provide a secure, validated connectivity, OSA-ICC supports Transport Layer Security/Secure Sockets Layer (TLS/SSL) with Certificate Authentication starting with IBM z13 GA2 (Driver level 27).

**Note:** OSA-ICC supports up to 48 *secure* sessions per CHPID (the overall maximum of 120 connections is unchanged).

### **OSA-ICC Enhancements**

With HMC 2.14.1 and newer the following enhancements are available:

- ▶ The IPv6 communications protocol is supported by OSA-ICC 3270 so that clients can comply with regulations that require all computer purchases to support IPv6.
- ▶ TLS negotiation levels (the supported TLS protocol levels) for the OSA-ICC 3270 client connection can now be specified:
  - TLS 1.0 OSA-ICC 3270 server permits TLS 1.0, TLS 1.1, and TLS 1.2 client connections.
  - TLS 1.1 OSA-ICC 3270 server permits TLS 1.1 and TLS 1.2 client connections.
  - TLS 1.2 OSA-ICC 3270 server permits only TLS 1.2 client connections.
- ▶ Separate and unique OSA-ICC 3270 certificates are supported (for each PCHID) for the benefit of customers who host workloads across multiple business units or data centers where cross-site coordination is required. Customers can avoid interruption of all the TLS connections at the same time when having to renew expired certificates. OSA-ICC also continues to support a single certificate for all OSA-ICC PCHIDs in the system.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

### **Checksum offload for in QDIO mode (CHPID type OSD)**

Checksum offload provides the capability of calculating the Transmission Control Protocol (TCP), User Datagram Protocol (UDP), and IP header checksum. Checksum verifies the



accuracy of files. By moving the checksum calculations to a Gigabit or 1000BASE-T Ethernet feature, host processor cycles are reduced and performance is improved.

Checksum offload provides checksum offload for several types of traffic and is supported by the following features when configured as CHPID type OSD (QDIO mode only):

- ▶ OSA-Express7S 1.2, OSA\_Express7S 1.1, and OSA-Express7S 25 GbE
- ▶ OSA-Express7S and OSA-Express7S 1.2 10 GbE
- ▶ OSA-Express7S and OSA-Express7S 1.2 GbE
- ▶ OSA-Express7S and OSA-Express7S 1.2 1000BASE-T Ethernet

When checksum is off-loaded, the OSA-Express feature runs the checksum calculations for Internet Protocol version 4 (IPv4) and Internet Protocol version 6 (IPv6) packets. The checksum offload function applies to packets that go to or come from the LAN.

When multiple IP stacks share an OSA-Express, and an IP stack sends a packet to a next hop address that is owned by another IP stack that is sharing the OSA-Express, OSA-Express sends the IP packet directly to the other IP stack. The packet does not have to be placed out on the LAN, which is termed LPAR-to-LPAR traffic. Checksum offload is enhanced to support the LPAR-to-LPAR traffic, which was not originally available.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

### Querying and displaying an OSA configuration

OSA-Express3 introduced the capability for the operating system to query and display directly the current OSA configuration information (similar to OSA/SF). z/OS uses this OSA capability by introducing the TCP/IP operator command **display OSAINFO**. z/VM provides this function with the **NETSTAT OSAINFO TCP/IP** command.

The use of **display OSAINFO** (z/OS) or **NETSTAT OSAINFO** (z/VM) allows the operator to monitor and verify the current OSA configuration and helps improve the overall management, serviceability, and usability of OSA-Express cards.

These commands apply to CHPID type OSD. The supported operating systems are listed in Table 7-8 on page 275.

### QDIO data connection isolation for z/VM

The QDIO data connection isolation function provides a higher level of security when sharing an OSA connection in z/VM environments that use VSWITCH. The VSWITCH is a virtual network device that provides switching between OSA connections and the connected guest systems.

QDIO data connection isolation allows disabling internal routing for each QDIO connected. It also provides a means for creating security zones and preventing network traffic between the zones.

QDIO data connection isolation is supported by all OSA-Express features on IBM z16. The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

### QDIO interface isolation for z/OS

Some environments require strict controls for routing data traffic between servers or nodes. In specific cases, the LPAR-to-LPAR capability of a shared OSA connection can prevent such controls from being enforced. With interface isolation, internal routing can be controlled on an LPAR basis. When interface isolation is enabled, the OSA discards any packets that are destined for a z/OS LPAR that is registered in the OSA Address Table (OAT) as isolated.



QDIO interface isolation is supported on all OSA-Express features on IBM z16. The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

### **QDIO optimized latency mode**

QDIO optimized latency mode (OLM) can help improve performance for applications that feature a critical requirement to minimize response times for inbound and outbound data.

OLM optimizes the interrupt processing in the following manner:

- ▶ For inbound processing, the TCP/IP stack looks more frequently for available data to process. This process ensures that any new data is read from the OSA-Express features without needing more program-controlled interrupts (PCIs).
- ▶ For outbound processing, the OSA-Express cards also look more frequently for available data to process from the TCP/IP stack. Therefore, the process does not require a Signal Adapter (SIGA) instruction to determine whether more data is available.

The supported operating systems are listed in Table 7-8 on page 275.

### **QDIO Diagnostic Synchronization**

QDIO Diagnostic Synchronization enables system programmers and network administrators to coordinate and simultaneously capture software and hardware traces. It allows z/OS to signal OSA-Express features (by using a diagnostic assist function) to stop traces and capture the current trace records.

QDIO Diagnostic Synchronization is supported by the OSA-Express features on IBM z16 when in QDIO mode (CHPID type OSD). The supported operating systems are listed in Table 7-8 on page 275.

### **Inbound workload queuing (IWQ) for OSA**

Inbound workload queuing (IWQ) creates multiple input queues and allows OSA to differentiate workloads “off the wire.” It then assigns work to a specific input queue (per device) to z/OS.

Each input queue is a unique type of workload, and has unique service and processing requirements. The IWQ function allows z/OS to preassign the appropriate processing resources for each input queue. This approach allows multiple concurrent z/OS processing threads to process each unique input queue (workload), which avoids traditional resource contention.

IWQ reduces the conventional z/OS processing that is required to identify and separate unique workloads. This advantage results in improved overall system performance and scalability.

A primary objective of IWQ is to provide improved performance for business-critical interactive workloads by reducing contention that is created by other types of workloads. In a heavily mixed workload environment, this “off the wire” network traffic separation is provided by OSA-Express7S 1.2, OSA-Express7S, and OSA-Express6S<sup>13</sup> features that are defined as CHPID type OSD.

---

<sup>13</sup> Only OSA-Express6S and OSA-Express7S cards are supported on IBM z16 as carry forward.



OSA IWQ is shown in Figure 7-5.

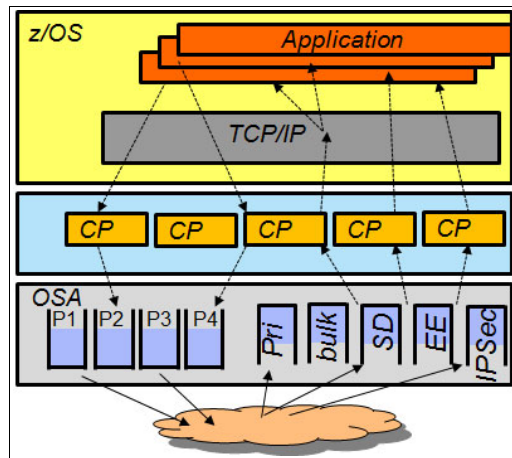


Figure 7-5 OSA inbound workload queuing

The following types of z/OS workloads are identified and assigned to unique input queues:

- ▶ **z/OS Sysplex Distributor traffic**  
Network traffic that is associated with a distributed virtual Internet Protocol address (VIPA) is assigned to a unique input queue. This configuration allows the Sysplex Distributor traffic to be immediately distributed to the target host.
- ▶ **z/OS bulk data traffic**  
Network traffic that is dynamically associated with a streaming (bulk data) TCP connection is assigned to a unique input queue. This configuration allows the bulk data processing to be assigned the suitable resources and isolated from critical interactive workloads.
- ▶ **EE (Enterprise Extender / SNA traffic)**  
IWQ for the OSA-Express features is enhanced to differentiate and separate inbound Enterprise Extender traffic to a dedicated input queue.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## GARP VLAN Registration Protocol

All OSA-Express features support VLAN prioritization, which is a component of the IEEE 802.1 standard. GARP VLAN Registration Protocol (GVRP) support allows an OSA-Express port to register or unregister its VLAN IDs with a GVRP-capable switch and dynamically update its table as the VLANs change. This process simplifies the network administration and management of VLANs because manually entering VLAN IDs at the switch is no longer necessary.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

## Link aggregation support for z/VM

Link aggregation (IEEE 802.3ad) that is controlled by the z/VM Virtual Switch (VSWITCH) allows the dedication of an OSA-Express port to the z/VM operating system. The port must be participating in an aggregated group that is configured in Layer 2 mode.

Link aggregation (trunking) combines multiple physical OSA-Express Network Express ports into a single logical link. This configuration increases throughput, and provides



nondisruptive failover if a port becomes unavailable. The target links for aggregation must be of the same type.

Link aggregation is applicable to CHPID type OSD (QDIO) and to OSH (EQDIO). The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

### **Multi-VSwitch Link Aggregation**

Multi-VSwitch Link Aggregation support allows a port group of OSA-Express features to span multiple virtual switches within a single z/VM system or between multiple z/VM systems. Sharing a Link Aggregation Port Group (LAG) with multiple virtual switches increases optimization and use of the OSA-Express or Network Express features when handling larger traffic loads.

Higher adapter use protects customer investments, which is increasingly important as 10 GbE deployments become more prevalent. The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

### **Large send for IPv6 packets**

Large send for IPv6 packets improves performance by offloading outbound TCP segmentation processing from the host to an OSA-Express feature by using a more efficient memory transfer into it.

Large send support for IPv6 packets applies to the OSA-Express7S 1.2, OSA-Express7S, and OSA-Express6S<sup>13</sup> features (CHPID type OSD) on IBM z16, IBM z15, and IBM z14.

OSA-Express6S added TCP checksum on large send, which reduces the cost (CPU time) of error detection for large send.

The supported operating systems are listed in Table 7-8 on page 275 and Table 7-9 on page 276.

### **OSA Dynamic LAN idle**

The OSA Dynamic LAN idle parameter helps reduce latency and improve performance by dynamically adjusting the inbound blocking algorithm. System administrators can authorize the TCP/IP stack to enable a dynamic setting that previously was static.

The blocking algorithm is modified based on the following application requirements:

- ▶ For latency-sensitive applications, the blocking algorithm is modified considering latency.
- ▶ For streaming (throughput-sensitive) applications, the blocking algorithm is adjusted to maximize throughput.

In all cases, the TCP/IP stack determines the best setting that is based on the current system and environmental conditions, such as inbound workload volume, processor use, and traffic patterns. It can then dynamically update the settings.

Supported OSA-Express features adapt to the changes, which avoids thrashing and frequent updates to the OAT. Based on the TCP/IP settings, OSA holds the packets before presenting them to the host. A dynamic setting is designed to avoid or minimize host interrupts.

OSA Dynamic LAN idle is supported by all OSA-Express features on IBM z16 when in QDIO mode (CHPID type OSD). The supported operating systems are listed in Table 7-8 on page 275.



### OSA Layer 3 virtual MAC for z/OS environments

To help simplify the infrastructure and facilitate load balancing when an LPAR is sharing an OSA MAC address with another LPAR, each operating system instance can have its own unique logical or virtual MAC (VMAC) address. All IP addresses that are associated with a TCP/IP stack are accessible by using their own VMAC address instead of sharing the MAC address of an OSA port. This situation also applies to Layer 3 mode and to an OSA port spanned among channel subsystems.

OSA Layer 3 VMAC is supported by all OSA-Express features on IBM z16 when in QDIO mode (CHPID type OSD). The supported operating systems are listed in Table 7-8 on page 275.

### Network Traffic Analyzer

IBM Z servers offer systems programmers and network administrators the ability to more easily solve network problems despite high traffic. With the OSA-Express Network Traffic Analyzer and QDIO Diagnostic Synchronization on the server, you can capture trace and trap data. This data can then be forwarded to z/OS tools for easier problem determination and resolution.

The Network Traffic Analyzer is supported by all OSA-Express features on IBM z16 when in QDIO mode (CHPID type OSD). The supported operating systems are listed in Table 7-8 on page 275.

## 7.4.6 Cryptography Features and Functions Support

IBM z17 provides the following major groups of cryptographic functions:

- ▶ Synchronous cryptographic functions, which are provided by CPACF
- ▶ Asynchronous cryptographic functions, which are provided by the Crypto Express8S feature

The minimum software support levels are described in the following sections. Review the current PSP buckets to ensure that the latest support levels are known and included as part of the implementation plan.

### CP Assist for Cryptographic Function

Central Processor Assist for Cryptographic Function (CPACF), which is standard<sup>14</sup> on every IBM z17 core, now supports pervasive encryption. Simple policy controls allow business to enable encryption to protect data in mission-critical databases without stopping the database or re-creating database objects.

Database administrators can use z/OS Dataset Encryption, z/OS Coupling Facility Encryption, z/VM encrypted hypervisor paging, and z/TPF transparent database encryption, which use the performance enhancements in the hardware.

CPACF supports the following features in IBM z17:

- ▶ Processor Activity Instrumentation to count cryptographic operations
- ▶ Advanced Encryption Standard (AES, symmetric encryption)
- ▶ Data Encryption Standard (DES, symmetric encryption)
- ▶ Secure Hash Algorithm (SHA, hashing)
- ▶ SHAKE Algorithms

---

<sup>14</sup> CPACF hardware is implemented on each IBM z15 core. CPACF functions are enabled with FC 3863.



- ▶ True Random Number Generation (TRNG)
- ▶ Improved GCM (Galois Counter Mode) encryption (enabled by a single hardware instruction)

In addition, the IBM z17, IBM z16, and IBM z15 cores implement a Modulo Arithmetic unit in support of Elliptic Curve Cryptography.

CPACF is used by several IBM software product offerings for z/OS, such as IBM WebSphere Application Server for z/OS. For more information, see 6.4, “CP Assist for Cryptographic Functions” on page 231.

The supported operating systems are listed in Table 7-10 on page 279 and Table 7-11 on page 280.

### **Crypto Express8S**

Crypto Express8S includes a single- or dual- HSM adapter (single or dual IBM 4770 PCIe Cryptographic Co-processor [PCIeCC]) and complies with the following Physical Security Standards:

- ▶ FIPS 140-3 level 4
- ▶ Common Criteria EP11 EAL4+
- ▶ Payment Card Industry (PCI) HSM
- ▶ German Banking Industry Commission (GBIC, formerly DK)
- ▶ AusPayNet (APN)

Support of Crypto Express8S functions varies by operating system and release and by the way that the card is configured as a coprocessor or an accelerator. The supported operating systems are listed in Table 7-10 on page 279 and Table 7-11 on page 280.

### **Crypto Express7S (carry forward on IBM z17)**

Introduced with IBM z15, Crypto Express7S includes a single- or dual-port adapter (single or dual IBM 4769 PCIe Cryptographic Co-processor [PCIeCC]) and complies with the following Physical Security Standards:

- ▶ FIPS 140-2 level 4
- ▶ Common Criteria EP11 EAL4
- ▶ Payment Card Industry (PCI) HSM
- ▶ German Banking Industry Commission (GBIC, formerly DK)

The supported operating systems are listed in Table 7-10 on page 279 and Table 7-11 on page 280.

### **Web deliverables**

For more information about web-deliverable code on z/OS, see [the z/OS downloads website](#).

For Linux on IBM Z, support is delivered through IBM and the distribution partners. For more information, see [Linux on IBM Z on the IBM Developer website](#).

### **z/OS Integrated Cryptographic Service Facility**

To achieve security in a distributed computing environment, a combination of elements must work together. A security policy should be based on an appraisal of the value of data and the potential threats to that data. This security policy provides the foundation for a secure environment.



IBM categorized the following security functions according to International Organization for Standardization (ISO) standard 7498-2:

- ▶ Identification and authentication: Includes the ability to identify users to the system and provide proof that they are who they claim to be.
- ▶ Access control: Determines which users can access which resources.
- ▶ Data confidentiality: Protects an organization's sensitive data from being disclosed to unauthorized persons.
- ▶ Data integrity: Ensures that data is in its original form and that nothing altered it.
- ▶ Security management: Administers, controls, and reviews a business security policy.
- ▶ Nonrepudiation: Assures that the suitable individual sent the message.

Only cryptographic services can provide the data confidentiality and the identity authentication that is required to protect business commerce on the internet<sup>15</sup>.

Integrated Cryptographic Service Facility (ICSF) is a base component of z/OS. It is designed to transparently use the available cryptographic functions (whether CPACF or Crypto Express) to balance the workload and help address the bandwidth requirements of the applications.

ICSF support for IBM z17 is provided with PTFs, not as previously was the case, through Web deliverables.

Supported levels of ICSF automatically detect what hardware cryptographic capabilities are available where it is running. Then, it enables functions accordingly. No toleration of new hardware is necessary because it is "just there". ICSF maintenance is necessary if you want to use new capabilities.

Use of new functions is available for:

- ▶ z/OS - ICSF Web deliverable 19 (HCR77D1), unless otherwise noted. WD19 supports z/OS V2R4, V2R5, V3R1.  
WD 20 supports z/OS V2R5 (base, which is HCR77D2)
- ▶ z/VM Version 7.3 or newer for guest use

New function exploitation includes:

- ▶ CCA and EP11 CEX8 Coprocessor support
- ▶ CCA and EP11 Quantum Safe Algorithms (Kyber & Dilithium 8,7)
- ▶ EP11 mechanism for data reencryption and new ECC curve support
- ▶ Fully homomorphic encryption
- ▶ Usage counters that count classes of crypto operations (to meet audit requirements)

When new Quantum Safe Algorithms are used and a KDS is shared in a sysplex, ensure that all ICSF PTFs are installed on all systems.

For more information about ICSF versions and FMID cross-references, see this [IBM Support web page](#).

---

<sup>15</sup> Quoted from z/OS V2.R5 publications.



## RMF Support for Crypto Express

RMF enhances the Monitor I Crypto Activity data gatherer to recognize and use performance data for the new Crypto Express8S (CEX8), Crypto Express7S (CEX7), and CryptoExpress6S (CEX6) cards. RMF supports all valid card configurations on IBM z16 and provides CEX7 and CEX6 crypto activity data in the SMF type 70 subtype 2 records and RMF Postprocessor Crypto Activity Report.

Reporting can be done at an LPAR/domain level to provide more granular reports for capacity planning and diagnosing problems. This feature requires fix for APAR OA54952.

The supported operating systems are listed in Table 7-10 on page 279.

## z/OS Data Set Encryption

Aligned with IBM Z Pervasive Encryption initiative, IBM provides application-transparent, policy-controlled dataset encryption in IBM z/OS.

Policy-driven z/OS Data Set Encryption enables users to perform the following tasks:

- ▶ De-couple encryption from data classification; encrypt data automatically independent of labor-intensive data classification work.
- ▶ Encrypt data immediately and efficiently at the time that it is written.
- ▶ Reduce risks that are associated with mis-classified or undiscovered sensitive data.
- ▶ Help protect digital assets automatically.
- ▶ Achieve application transparent encryption.

IBM Db2 for z/OS and IBM Information Management System (IMS) intend to use z/OS Data Set Encryption.

With z/OS, Data Set Encryption DFSMS enhances data security with support for data set level encryption by using DFSMS access methods. This function is designed to give users the ability to encrypt their data sets without changing their application programs.

DFSMS users can identify which data sets require encryption by using JCL, Data Class, or the RACF data set profile. Data set level encryption can allow the data to remain encrypted during functions, such as backup and restore, migration and recall, and replication.

z/OS Data Set Encryption requires CP Assist for Cryptographic Functions (CPACF).

Considering the significant enhancements that were introduced with z14, the encryption mode of XTS is used by access method encryption to obtain the best performance possible. It is not recommended to enable z/OS data set encryption until all sharing systems, fallback, backup, and DR systems support encryption.

In addition to applying PTFs enabling the support, ICSF configuration is required. The supported operating systems are listed in Table 7-10 on page 279.

## Quantum-safe encryption with IBM z17 and Crypto Express8S

Quantum-safe cryptography strengthens the portfolio of pervasive encryption capabilities on IBM z16, which allows customers to encrypt the data with a quantum-safe cryptographic algorithm (AES with 256-bit keys) as is the case with prior IBM Z buy using quantum-safe algorithms for internal system protection of encryption keys.

The following Quantum-safe enhancements were introduced with IBM z16 to accomplish this encryption:



- ▶ Key generation
- ▶ Hybrid key exchange schemes
- ▶ Dual digital signature schemes

### **Crypto Analytics Tool for IBM Z**

The IBM Crypto Analytics Tool (CAT) for IBM Z is an analytics solution that collects data on your z/OS cryptographic infrastructure, presents reports, and analyzes if any vulnerabilities exist. CAT collects cryptographic information from across the enterprise and provides reports to help users better manage the crypto infrastructure and ensure it follows best practices. The use of CAT can help you deal with managing complex cryptography resources across your organization.

### **z/VM encrypted hypervisor paging (encrypted paging support)**

With the PTF for APAR VM65993, z/VM V6.4 provides support for encrypted paging in support of the IBM z16 pervasive encryption philosophy of encrypting all data that is in flight and at rest. Ciphering occurs as data moves between active memory and a paging volume that is owned by z/VM.

Included in this support is the ability to dynamically control whether a running z/VM system is encrypting this data. This support protects guest paging data from administrators or users with access to volumes. Enabled with AES encryption, z/VM Encrypted Paging includes low overhead by using CPACF.

The supported operating systems are listed in Table 7-10 on page 279.

### **z/TPF transparent database encryption**

Shipped in August 2016, z/TPF at-rest Data Encryption provides following features and benefits:

- ▶ Automatic encryption of at-rest data by using AES CBC (128 or 256).
- ▶ No application changes required.
- ▶ Database level encryption by using highly efficient CPACF.
- ▶ Inclusion of data on disk and cached in memory.
- ▶ Ability to include data integrity checking (optionally by using SHA-256) to detect accidental or malicious data corruption.
- ▶ Tools to migrate a database from unencrypted to encrypted state or change the encryption key/algorithm for a specific DB while transactions are flowing (no database downtime).

### **Pervasive encryption for Linux on IBM Z**

Pervasive encryption for Linux on IBM Z combines the full power of Linux with IBM z16 capabilities by using the support of the following features:

- ▶ Kernel Crypto: IBM z16 CPACF
- ▶ LUKS dm-crypt Protected-Key CPACF
- ▶ Libica and openssl: IBM z16 CPACF and acceleration of RSA handshakes by using SIMD
- ▶ Secure Service Container: High security virtual appliance deployment infrastructure

### ***Protection of data at-rest***

By using the integration of industry-unique hardware accelerated CPACF encryption into the standard Linux components, users can achieve optimized encryption transparently to prevent raw key material from being visible to operating systems and applications.



Because of the potential costs and overhead, most of the organizations avoid the use of host-based network encryption today. By using enhanced CPACF and SIMD on IBM z16, TLS and IPsec can use hardware performance gains while benefiting from transparent enablement. Reduced cost of encryption enables broad use of network encryption.

## **IBM Z Security and Compliance Center**

The IBM Z Security and Compliance Center is a modern, browser-based application that provides compliance capability mapping, fact collection, and validation. Designed for use with minimal technical skills, this solution can automate evidence collection of compliant-relevant facts from IBM Z platforms, including the new CPACF usage counters that demonstrate crypto algorithm strength and key protection.

The IBM Z Security and Compliance Center enables clients to:

- ▶ Generate detailed reports to enable executives, administrators, and auditors to understand compliance metrics with ease.
- ▶ Track compliance drift over time with dashboard visualizations that include historical compliance information.
- ▶ Use compliance evidence generation facilities from IBM Z software stack (for example, z/OS, z/OS Middleware, z/OS Compliance Integration Manager, Oracle on Linux on IBM Z, and PostgreSQL on Linux on IBM Z).
- ▶ Provide an interactive view of the compliance posture and details around the severity of control deviations from regulations, such as PCI-DSS v3.2.1, NIST SP800-53, and CIS Benchmarks.

### ***z/OS support for compliance***

z/OS was enhanced to enable the collection of compliance data from IBM z16 CPACF counters and several z/OS products and components.

A new z/OSMF Compliance fact collection REST API sends an ENF86 signal to all systems. Participating products and components collect and write compliance data to new SMF 1154 records that are associated with its unique subtype. These new SMF 1154 records can be integrated into solutions, such as the IBM z16 IBM Z Security and Compliance Center.

This support requires PTFs for z/OS 2.4 and z/OS 2.5. The PTFs are identified by a fix category that is designated specifically for Compliance data collection support named IBM.Function.Compliance.DataCollection. For more information about how to use this fix category to identify and install the specific PTFs that enable compliance data collection, see “IBM Fix Category Values and Descriptions”.

For more information about z/OS collection sources and enablement, see the following resources:

- ▶ Software Announcement 222-005, IBM Z Security and Compliance Center.
- ▶ Software Announcement 222-092, CICS Transaction Server for z/OS 6.1.
- ▶ Software Announcement 222-003, Db2 13 for z/OS powered by AI innovations provides industry scalability, business resiliency and intelligence.

### ***Linux support for compliance***

Linux on IBM Z supports the collection of compliance data from the Linux environment.

The following prerequisite operating system versions are supported for the collection of compliance data:

- ▶ Red Hat Enterprise Linux 8.0 (RHEL) on IBM Z, or later



- ▶ SUSE Linux Enterprise Server (SLES) on IBM Z 15
- ▶ Ubuntu Server LTS for IBM Z 22.04

The following optional software is available for collecting compliance data:

- ▶ Oracle 19c
- ▶ PostgreSQL 13.x, 14.x

## 7.5 z/OS migration considerations

Except for base processor support, z/OS releases do not require any of the functions that are introduced with the IBM z17. Minimal toleration support that is needed depends on z/OS release.

Although IBM z17 servers do not require any “functional” software, it is recommended to install all IBM z17 service before upgrading to the new server. The support matrix for z/OS releases and the IBM Z servers that support them are listed in Table 7-17, where “X” indicates that the hardware model is supported.

Table 7-17 z/OS support summary

z/OS Release	IBM z15 <sup>a</sup>	IBM z16	IBM z17	End of Service	Extended Defect Support <sup>b</sup>
V2R4	Y	Y	Y	09/2024	09/2027 <sup>b</sup>
V2R5	Y	Y	Y	09/2026	09/2029 <sup>b</sup>
V3R1	Y	Y	Y	09/2028	09/2031 <sup>b</sup>

a. Server is withdrawn from marketing.

b. The IBM Software Support Services provides the ability for customers to purchase extended defect support service for z/OS.

### 7.5.1 General guidelines

The IBM z17 introduces the latest IBM Z technology. Although support is provided by z/OS starting with z/OS V2.R2, the capabilities and use of IBM z17 depends on the z/OS release. Optional web deliverables<sup>16</sup> are needed for some functions on some releases.

**New:** ICSF support for IBM z17 is provided with PTFs, not Web deliverables

In general, consider the following guidelines:

- ▶ Do not change software releases and hardware at the same time.
- ▶ At a minimum, apply maintenance from the following FIXCAT to all systems that participate in a sysplex with IBM z17, regardless of whether the systems is to be migrated to the current hardware:

IBM.Device.Server.IBM z17-9175.RequiredService

<sup>16</sup> For example, the use of Crypto Express7S requires the Cryptographic Support for z/OS V2R2 - z/OS V2R3 web deliverable.



- ▶ Keep members of the sysplex at the same software level, except during brief migration periods.
- ▶ Upgrade Coupling Facility LPARs to current levels (review all structure sizes by using the CFSIZER tool *before* the CF is upgraded).
- ▶ Review any restrictions and migration considerations before creating an upgrade plan.
- ▶ Acknowledge that some hardware features cannot be ordered or carried forward for an upgrade from an earlier server to IBM z17 and plan accordingly.
- ▶ Determine the changes in IOCP, HCD, and HCM to support defining IBM z17 configuration and the new features and functions it introduces.
- ▶ Ensure that none of the new z/Architecture Machine Instructions (mnemonics) that were introduced with IBM z17 are colliding with the names of Assembler macro instructions you use<sup>17</sup>.
- ▶ Check the use of **MACHMIG** statements in **LOADxx PARMLIB** commands.
- ▶ Contact software vendors to inform them of new machine model and request new license keys, if applicable.
- ▶ Review the z/OS Upgrade Workflow for z/OSMF that is provided as a ++APAR for z/OS V2R4 and higher<sup>18</sup>. This Workflow also is available in the IBM Documentation library.

## 7.5.2 Hardware Fix Categories

Base support includes fixes that are required to run z/OS on the IBM z16 server. They are identified by:

IBM.Device.Server.IBM z17-9175.RequiredService

The use of many functions covers fixes that are required to use the capabilities of the IBM z16 servers. They are identified by:

IBM.Device.Server.IBM z17-9175.Exploitation

Recommended service is identified by:

IBM.Device.Server.IBM z17-9175.RecommendedService

**Attention:** Starting with z17, PSP Bucket is no longer available.

Consider the following other Fix Categories of Interest:

- ▶ Fixes that are required to use the Server Time Protocol function:  
IBM.Function.ServerTimeProtocol
- ▶ Fixes that are required to use the High-Performance FICON function:  
IBM.Function.zHighPerformanceFICON
- ▶ Fixes that are required for IBM Z System Recovery Boost:  
IBM.Function.SystemRecoveryBoost
- ▶ PTFs that allow previous levels of ICSF to coexist with the latest Cryptographic Support for z/OS V2R4 - z/OS V3R1 (HCR77E1) web deliverable:  
IBM.Coexistence.ICSF.z/OS\_V2R4-V3R1-HCR77E1

<sup>17</sup> For more information, see the [Tool to Compare IBM z16 Instruction Mnemonics with Macro Libraries](#) IBM Technote.

<sup>18</sup> At General Availability a PTF will be provided and it will be marked with FIXCAT:  
IBM.Device.Server.z17-9175.Required Service



Use the SMP/E **REPORT MISSINGFIX** command to determine whether any FIXCAT APARs exist that are applicable and are not yet installed, and whether any SYSMODs are available to satisfy the missing FIXCAT APARs.

Before any action to install required service can take place you should update your SMP/E HOLDDATA to the most current level. In Example 7-2 we show you an example in how to update your HOLDDATA information.

---

*Example 7-2 RECEIVE HOLDDATA sample JCL*

---

```
//T445REC JOB (CP00,KE10,LK),'LUTZ KUEHNER',MSGCLASS=T,
//      NOTIFY=&SYSUID,TIME=1440,CLASS=1
//SMPREC EXEC PGM=GIMSMP
//SMPCSI DD DSN=<Your SMP/E Global CSI>,DISP=SHR
//SMPOUT DD SYSOUT=*
//BPXPRINT DD SYSOUT=*
//SYSPRINT DD SYSOUT=*
//SMPNTS DD PATHDISP=KEEP,PATH='/SERVICE/tmp/'
//SMPJHOME DD PATHDISP=KEEP,PATH='/usr/lpp/java/J8.0_64/'
//SMPCPATH DD PATHDISP=KEEP,PATH='/usr/lpp/smp/classes/'
//SMPCNTL DD *
SET BOUNDARY (GLOBAL) .
RECEIVE SYSMODS HOLDDATA DELETEPKG
        ORDER(ORDERSERVER(ORDSRVR)
        CLIENT(MYCLIENT)
        CONTENT(ALL
                ) FORTGTZONES(<Your Target Zone>
                ).
/*
//ORDSRVR DD *
<ORDERSERVER
    url="<Your IBM Service URL>"
    keyring="<Your Keyring>"
    certificate="<Your SMPE Client Certificate>"
</ORDERSERVER>
/*
//MYCLIENT DD *
<CLIENT
    debug="yes"
    downloadmethod="https"
    javahome="/usr/lpp/java/J8.0_64/"
    javadebugoptions="-Xmx128m -verbose">
<HTTPPROXY
    host="<Your local Internet Proxy>" port="<Your outbound proxy Port>">
</HTTPPROXY>
</CLIENT>
/*
//
```

---

In the workflow, all PFT FIXCATs are clearly documented with steps to run **SMP/E REPORT MISSINGFIX** command you see in Example 7-3 to assist in knowing if prepared for IBM z17.

---

*Example 7-3 Example of MISSINGFIX*

---

```
//T445RFX JOB (CP00,KE10,LK),'LUTZ KUEHNER',MSGCLASS=T,
```



```
//      NOTIFY=&SYSUID,TIME=1440,CLASS=1
//*****
//SMPREC  EXEC PGM=GIMSMP
//SMPCSI  DD  DSN=<Your SMP/E Global CSI>,DISP=SHR
//SMPOUT  DD  SYSOUT=*
//SMPRPT  DD  SYSOUT=*
//SMPHRPT DD  SYSOUT=*
//SYSPRINT DD SYSOUT=*
//SMPCNTL DD  *
  SET BOUNDARY (GLOBAL) .
  REPORT MISSINGFIX ZONES(<Your Target Zone>)
    FIXCAT(IBM.Device.Server.z17* ).
/*
//
```

---

If no further maintenance is required for your product in order to support z17 hardware you will see an SMP/E output like we show in Example 7-4.

*Example 7-4 SMP/E no further actions needed*

```
GIM69228I      NO FIXCAT HOLDDATA MATCHING THE SPECIFIED FIXCAT AND FORFMID VALUES
                WAS FOUND.
```

---

If any additional maintenance is required you will see an detailed SMPRPT who shows all required software that is currently not installed. In Example 7-5 an additional SMPPUNCH output is available who can be used in another SMP/E JCL to apply all missing PTF's that are found in the MISSINGFIX report.

*Example 7-5 sample SMP/E APPLY*

```
SET BDY(MVST) .
  APPLY  CHECK
        SELECT(
/* IBM.Function.z17-9175.RequiredService */
        UA64554
        )
        BYPASS(HOLDSYSTEM)
        GROUPEXTEND.
```

---

For more information about IBM Fix Category Values and Descriptions, see this [IBM Support web page](#).

### 7.5.3 z/OS V3.R1

IBM z/OS, Version 3 Release 1, was made generally available on 29 September 2023. This release delivers innovation through an agile, optimized, and resilient platform that helps companies build applications and services that are based on a highly scalable and secure infrastructure. This infrastructure provides the performance and availability for on-premises or provisioned as-a-service workloads. Table 7-18 provides a list of APARs for z/OS.

*Table 7-18 APARs for z17 exploitation support*

z17 Function	z/OS Exploitation Support via APARs
New z/Architecture and IBM zNEXT machine instructions	PH62834



z17 Function	z/OS Exploitation Support via APARs
CPU Measurement Facility new extended counters	<<APAR Not yet defined, please check again>
z/OS BCPII and HMC/SE hardened security	OA65929
Workload-level sustainability and power consumption reporting	OA63265 and OA66018
Workload Classification Pricing	OA66812, OA65240, OA65242, OA66596, and OA66937
Replacement Capacity Records	OA66402, OA66054, OA63265 and OA66938
IBM z17 integrated I/O architecture	OA66014 and OA66054
25Gb Long Distance Coupling Link Adapter (CL6)	OA64478, OA64591, OA64362 and OA64114
CF Level 26: Non-disruptive system-managed copy process for lock structures	OA65820
Integrated Cryptographic Support Facility	OA66518
zDNN for IBM Z Integrated Accelerator for AI	OA66863
Open XL C/C++ exploitation	via Web Deliverable
System Recovery Boost enhancements	None

#### 7.5.4 z/OS V2.R5

IBM z/OS, Version 2 Release 5 was announced 27 July 2021. One of the highlights of this release is the support of 16 TB real memory per z/OS image. This support allows new workloads that require more storage than is available.

For more information about this release, see this [announcement letter](#).

#### 7.5.5 z/OS V2.R4

IBM z/OS, Version 2 Release 4, was made generally available on 30 September 2019. This release delivers innovation through an agile, optimized, and resilient platform that helps companies build applications and services that are based on a highly scalable and secure infrastructure. This infrastructure provides the performance and availability for on-premises or provisioned as-a-service workloads.

z/OS V2.R4 delivers many capabilities, including the following examples:

- ▶ IBM z/OS Container Extensions (zCX), which enables the ability to run almost any Linux on IBM Z Docker container in z/OS alongside existing z/OS applications and data without a separate provisioned Linux server.
- ▶ Easier integration of z/OS into private and multi-cloud environments with improvements that deliver a more robust, easy to use, and highly available implementation that uses IBM Cloud Provisioning and Management for z/OS, IBM z/OS Cloud Broker, and IBM Cloud Storage Access for z/OS Data.



- ▶ Enhancements that continue to simplify and modernize the z/OS environment for a better user experience and improved productivity by reducing the level of IBM Z specific skills that are required to maintain z/OS.
- ▶ Ongoing industry-wide simplification improvements to help companies install and configure software by using a common and modern method. These installation improvements range from the packaging of software through the configuration so that faster time to value can be realized throughout the enterprise.
- ▶ IBM Open Data Analytics for z/OS provides enhancements to simplify data analysis by combining open source run times and libraries with analysis of z/OS data at its source,
- ▶ Enhancements to security and data protection on the system with support for new industry cryptography and continued enhancements driving pervasive encryption through the ability to encrypt data without application changes. A new RACF capability improves management of access and privileges.
- ▶ The use of IBM z16 capabilities

System Recovery Boost reduces the time that z/OS is offline when the operating system is offline for any reason. The use of IBM System Recovery Boost expedites planned operating system shutdown processing, operating system Initial Program Load (IPL), middleware and workload restart and recovery, and the client workload execution that follows.

It enables businesses to return their systems to work faster, not just from catastrophes, but after all kinds of disruptions (planned and unplanned). Another aspect of System Recovery Boost is to expedite and streamline the execution of GDPS recovery scripts that perform reconfiguration actions during various planned and unplanned operational scenarios.

### 7.5.6 Remote dynamic activation of I/O configurations for stand-alone Coupling Facilities, Linux on Z and z/TPF

The remote activation of dynamic changes avoid the need for disruptive hardware/firmware actions (Power-on Reset or IML) to be taken to instantiate those configuration changes, reducing, or completely eliminating the client workload impact that would otherwise have resulted from taking these disruptive actions.

IBM z16 provide a supported capability to drive these hardware-only I/O configuration changes from a driving z/OS HCD instance to a remote target CPC which is a Coupling facility of hosts IBM Linux on Z and z/TPF.

This new support is applicable only when both driving CPC and the target CPC are z16s with the required firmware support (Bundle S24 or higher) and when the driving system's z/OS level is 2.3 or higher with APA OA65559.

### 7.5.7 Coupling links

IBM z17 servers support only active participation in the same Parallel Sysplex with IBM z16 and IBM z15. Configurations with z/OS on one of these servers can add an IBM z17 Server to their Sysplex for a z/OS or a Coupling Facility image.

Configurations with a Coupling Facility on one of these servers can add an IBM z17 Server to their Sysplex for a z/OS or a Coupling Facility image. IBM z17 does not support participating in a Parallel Sysplex with System IBM z14/IBM z14 ZR1 and earlier systems.

Each system can use, or not use, internal coupling links, CE LR links, or ICA SR coupling links independently of what other systems are using.



Coupling connectivity is available only when other systems also support the same type of coupling. For more information about supported coupling link technologies on IBM z16, see 4.6.4, “Parallel Sysplex connectivity” on page 199, and the [Coupling Facility Configuration Options](#) white paper.

## 7.5.8 z/OS XL C/C++ considerations

IBM z/OS V3.R1 XL C/C++ is an optional feature of z/OS that continues to ship with IBM z17, however, z/OS XL C/C++ compiler, will not be updated with the support for new z17 ARCH. XL C/C++ supports up to z15 HW instructions with ARCH(13) and programs compiled with XL C/C++ will run on IBM z17

For more information about the **ARCHITECTURE**, **TUNE**, and **VECTOR** compiler options, see [z/OS XL C/C++ User's Guide, SC14-7307-40](#).

z/OS XL C/C++ Web deliverables are available at no charge to z/OS XL C/C++ customers:

- ▶ Based on open-source LLVM infrastructure; supports up to date C++ language standards
- ▶ 64-bit, UNIX System Services only

**Statement of Direction:** IBM will continue to adopt the LLVM and Clang compiler infrastructure in future C/C++ offerings on IBM Z<sup>a</sup>.

- a. Any statements regarding IBM's future direction, intent, or product plans are subject to change or withdrawal without notice.

## 7.6 z/VM migration considerations

IBM z17 supports z/VM 7.4 and z/VM 7.3. z/VM is moving to continuous delivery model. For more information, see [this web page](#).

### 7.6.1 IBM z/VM 7.4

IBM z/VM 7.4 includes some new features. Some of the Support will be available as PTFs concurrently at IBM zNext general availability. That includes PTFs for IOCP, HCD, and HLASM.

- ▶ Enable Guest Exploitation for the following new facilities. Vector-Enhancements Facility 3 will have new instructions intended to provide performance improvements
- ▶ Vector-Packed-Decimal-Enhancement 3 is intended to provide performance improvements of COBOL programs when compiled using the NUMCHECK option to detect and convert data.
- ▶ Workload-Instrumentation provides a means of classifying and sampling workloads to enhance the z/OS pricing model.
- ▶ Message-Security-Assist Extensions provides enhancements which allow use of XTS and HMAC algorithms and all for generation of XTS and HMAC encryption keys while using AES algorithms
- ▶ Reduced support for TX results non-constrained transactions in unconditionally aborting with a CC1 set and no TDB stored.
- ▶ Perform Lock Operation (PLO) provides operations for managing locks in storage to replace capabilities previously provided by the constrained-transactional-execution facility (CTX)



- ▶ Concurrent-Functions provides new instructions intended to replace use of TX for software serialization.
- ▶ Network Express Adapter EQDIO Support within the z/VM VSwitch allows customers to configure the VSwitch to take advantage of lower latency and higher bandwidths provided by networking EQDIO devices within their datacenter. The VSwitch EQDIO exploitation includes QDIO to EQDIO translation allowing guests which do not support EQDIO to directly take advantage of this networking support.<sup>19</sup>
- ▶ Power Consumption metrics provided within z/VM monitor provides enhancements to the z/VM monitor to include entire CPC and LPAR specific level power consumption information like you see in Figure 7-6. This information includes power metrics for CPU, I/O, and Memory usage. Consumers of z/VM monitor, such as the z/VM performance datapump, can be enhanced to calculate/approximate guest level apportionment.



Figure 7-6 Sample Power Apportionment Data

- ▶ CPU-Measurement Facility provides CPU-MF specific support for IBM z17
- ▶ Data Processing Unit Next Generation I/O accelerator Instrumentation provided within z/VM monitor to be able to collect instrumentation data within z/VM monitor for the new IBM z17 I/O complex
- ▶ Dynamic I/O support and guest exploitation for the following
  - 25G LR for Long Distance Coupling CHPID type CL6

**Note:** Dynamic I/O support only; No guest exploitation

- Network Express Adapter CHPID type OSH and PCI Function Types NETH and NETD
- AI Accelerator Adapter PCI Function Type PAIA

The new Guest Exploitation support for the following new features will be available with z17:

<sup>19</sup> A z/VM VSwitch supporting Network Express OSH does not currently support z/OS guests exploiting an EQDIO uplink port. In the interim, clients will be required to use either a guest-attached OSH device or existing functionality available with OSA-Express7S adapters.

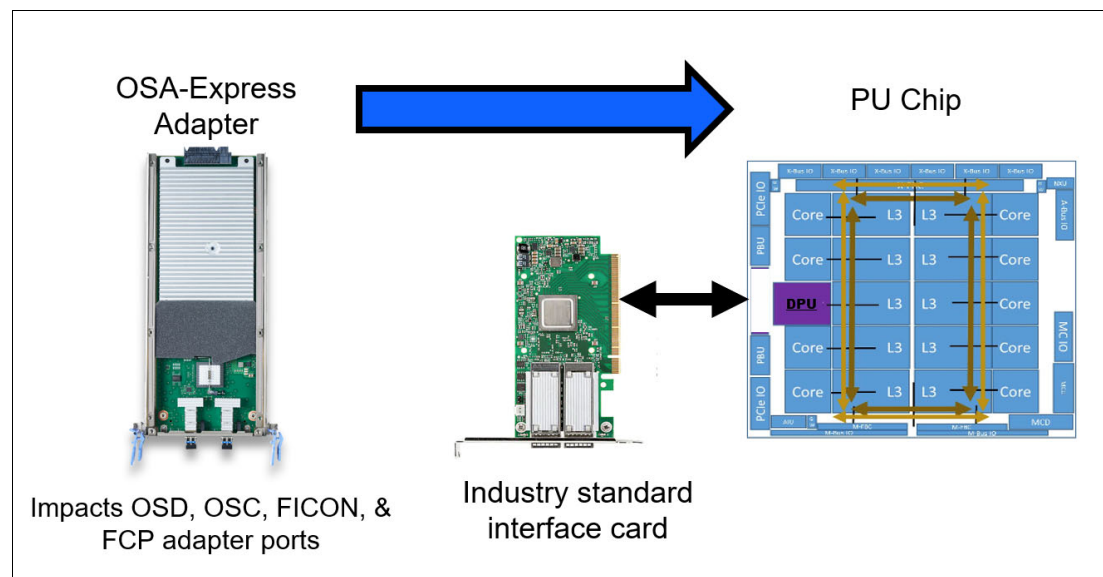


- ▶ RoCE Network Express Adapter Hybrid (NETH) & Direct (NETD) Virtual Function support allows guests to directly exploit RoCE functionality of the Network Express adapter (additional details later in presentation)
- ▶ Networking Express Adapter EQDIO OSA Hybrid (OSH) CHPID support allows guests to directly exploit OSH functionality of the Network Express adapter, and allows guests to exploit OSD simulated devices via the z/VM VSwitch to a real OSH device.
- ▶ AI Accelerator Adapter allow guests to take advantage of and exploit the capabilities of the enhanced AI Accelerator Adapter

## Transformation thru I/O Infrastructure Modernization

The I/O subsystem of IBM Z is built on a solid architectural basis that has provided a long and rich history of I/O (storage and networking) function, throughput and bandwidth improvements, while maintaining application backward-compatibility.

Changes to the current physical packaging and, in the case of networking, architecture are needed. In Figure 7-7 we show how the I/O Functionality has been moved to the z17 DPU in order to accelerate I/O Operations.



*Figure 7-7 Moving I/O Functionality*

The entire System Z I/Os are moved to the I/O abstraction layer (firmware) located within an adapter-based RISC Processor to a Z Core. That:

- Eliminates the need for proprietary adapter hardware.
- Locates the adapter closer to memory and the nest (cores).

The next generation of abstraction layer for networking called Enhanced QDIO which has been designed to:

- Reduce protocol message exchanges.
- Cache collision reduction.
- Increased memory efficiency.
- Increased bandwidth potential.
- Scalability for Z Virtualization



## Enhanced OSA Architecture (EQDIO)

The Queued-Direct-I/O (QDIO), the Z architecture used to drive OSA channels, was developed 20+ years ago. IBM has provided ability for common I/O device driver to be maintained in the operating systems, even as new underlying hardware adapters have been upgraded. IBM has addressed the limitations in traditional architecture/implementation and simplify the interface. EQDIO support is available for the z/VM Hypervisor itself as also for the z/VM TCPIP Stack.

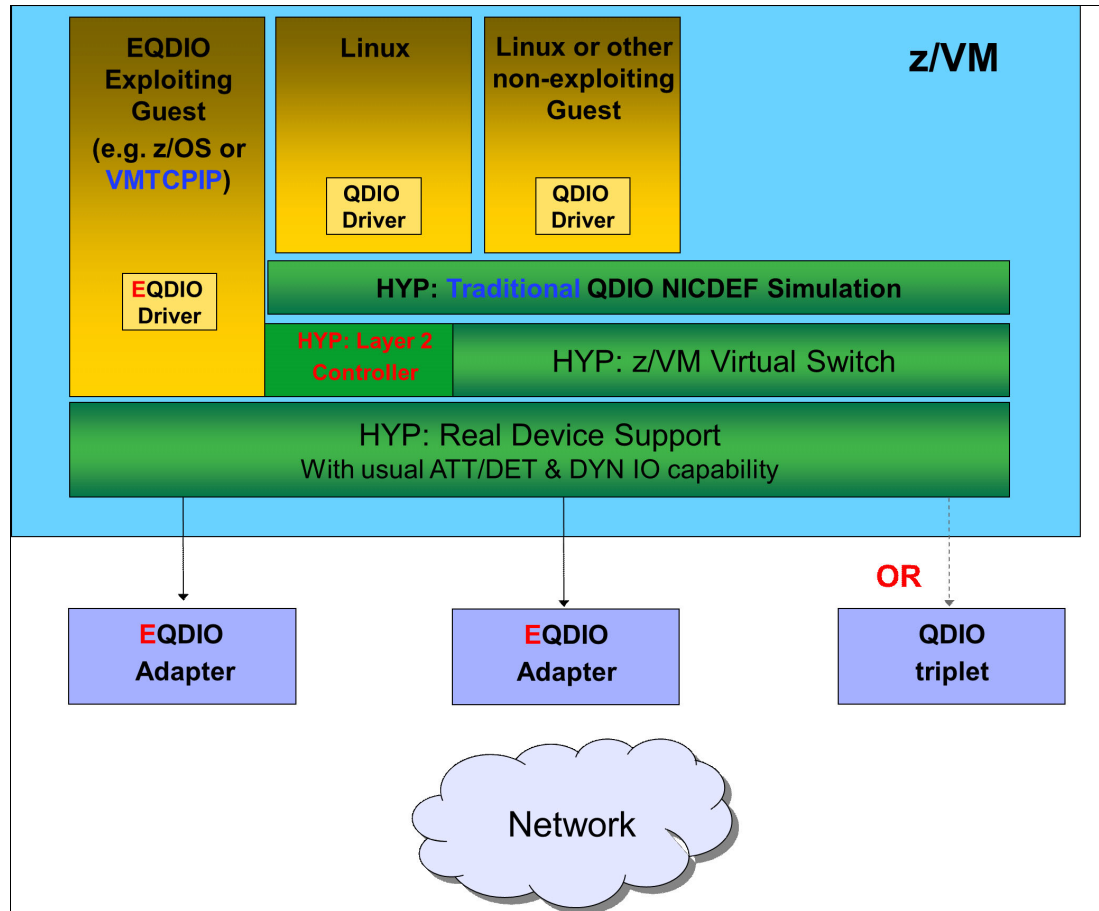


Figure 7-8 EQDIO Overview

## Hardware provided Converged Multi-Function Network Adapter

Converged support for multiple networking protocols – provide ability to run existing functions on single physical adapter, also known as Appliance.

**OSD+** Enhanced QDIO protocol (EQDIO) capability (aka OSH)

**RoCE** RDMA, SMC-R & TCP/IP capabilities (aka NETH)

The Adapter are now multi-function networking adapters and will have 2 Ports per I/O Slot. Initial z17 support will be available for 10GbE and 25GbE transmission speed. Each port on card is a unique CHPID. Multiple protocols can be shared on the same physical port. Each port can be configured to provide a single function or combination of functions.

## z/VM Dedicated Device Support

This new support, you see in Figure 7-9 on page 346, allows customer to attach one or more OSH EQDIO Devices to a virtual machine. z/VM itself must get between the program and the adapter in order to do virtual to real address translation and the virtual Memory Page



Management, also known as Pin and Unpin SBs. Furthermore z/VM maintains an set of shadow queues in memory who are only accessible by z/VM and the adapter. z/VM's responsibility is to keep the guest and shadow queues synchronized.

**Note:** The Queue synchronization is performed by intercepting the EQDIO Millicode Instructions.

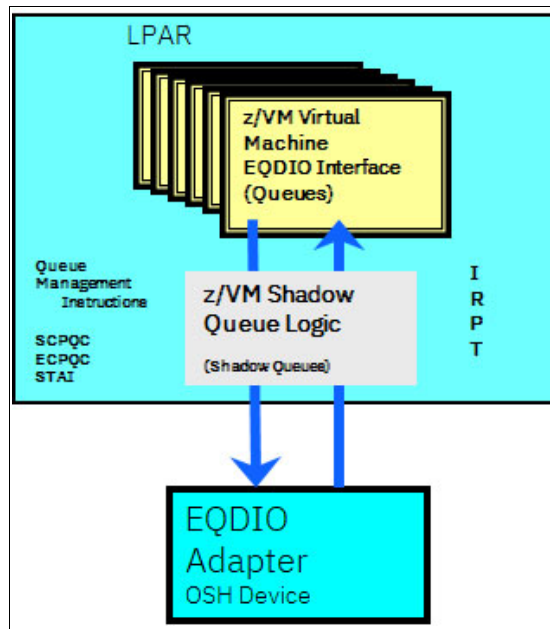


Figure 7-9 z/VM dedicated Device Support

## EQDIO VSWITCH Support

The VM VSWITCH supporting traditional QDIO-based simulated LAN to real EQDIO adapters.

**Note:** VM VSWITCH provides a customer the ability to define its own virtual network to interconnect multiple virtual machines using a simulated device to a physical network.

The DEFINE NIC TYPE QDIO creates an network adapter working either on Layer 2 (Ethernet) or Layer 3 (IP).

**Attention:** New EQDIO supports Layer 2 mode only. Using Layer 3 requires traditional QDIO uplink.

The VSWITCH configuration externals remain mostly unchanged with some existing parms not applicable to EQDIO.

For migration of VSWITCH to EQDIO is transparent, no configuration changes are necessary. Previous used RDEV statements can be used, the z/VM hypervisor will figure out how to use them with EQDIO. For Linux Guests also LGR capability is available.



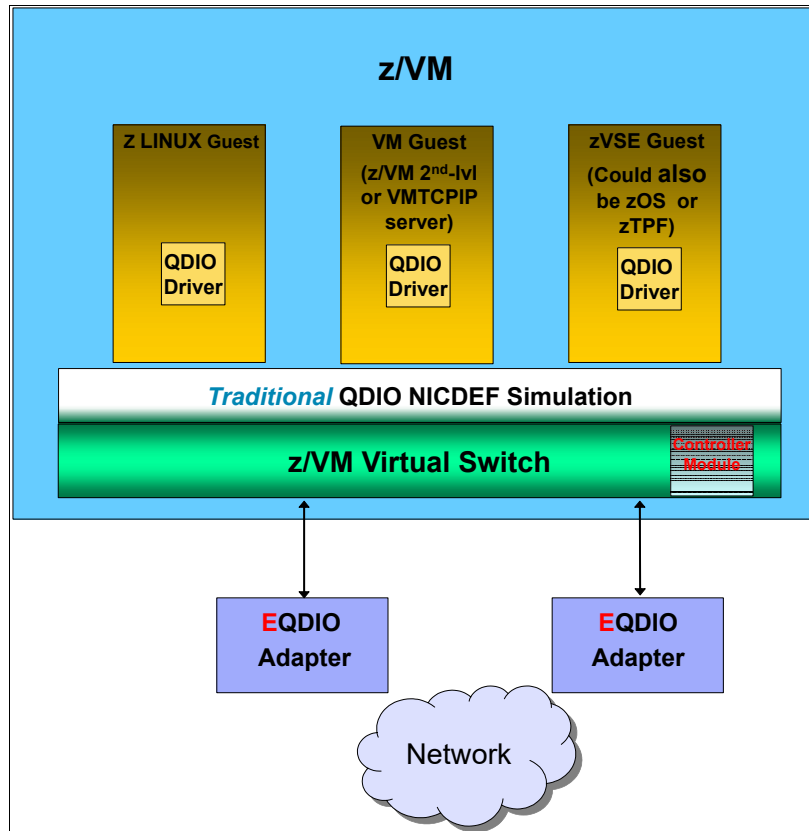


Figure 7-10 EQDIO VSWITCH Support

## 7.6.2 IBM z/VM 7.3

IBM z/VM 7.3 includes the following features:

- ▶ 8-Member SSI: increases the maximum size of a Single System Image (SSI) cluster from four members to eight, which enables customers to grow their SSI clusters to allow for more workload. It also provides more flexibility to use live guest relocation (LGR) for nondisruptive upgrades and workload balancing.
- ▶ New Architecture Level Set of IBM z14 and IBM z14 ZR1.
- ▶ All new functions are made available in z/VM 7.2 throughout the continuous delivery process.

## 7.6.3 Capacity

For the capacity of any z/VM logical partition (LPAR) and any z/VM guest, you might want to adjust the number to accommodate the PU capacity of IBM z17 servers in terms of the number of Integrated Facility for Linux (IFL) processors and central processors (CPs), real or virtual.

## 7.7 VSE<sup>n</sup> migration considerations

As described in 7.2.4, “VSE<sup>n</sup>” on page 266, IBM z17 supports only VSE<sup>n</sup> V6.3.1



Consider the following general guidelines when you are migrating VSE<sup>n</sup> environment to IBM z17 servers:

► Collect reference information before migration

This information includes baseline data that reflects the status of, for example, performance data, CPU use of reference workload, I/O activity, and elapsed times.

This information is required to size IBM z17 and is the only way to compare workload characteristics after migration.

**Note:** For more information, see: <https://www.21stcenturysoftware.com/vsen/>

► Apply all required maintenance for IBM z17

## 7.8 Software licensing

The IBM z17 software portfolio includes operating system software (that is, z/OS, z/VM, z/VSE, and z/TPF) and middleware that runs on these operating systems. The portfolio also includes middleware for Linux on IBM Z environments.

This section provides an overview of the following IBM Z software licensing options that are available for IBM z17 software, including MLC, zIPLA, subcapacity, sysplex, and Taylor Fit Pricing:

► Monthly license charge (MLC)

MLC is a recurring charge that is applied monthly. It includes the right to use the product and provides access to IBM product support during the support period. Select an MLC pricing metric that is based on your goals and environment.

The selected metric is used to price MLC products, such as z/OS, z/TPF, z/VSE, middleware, compilers, and selected systems management tools and utilities:

– Key MLC metrics and offerings

MLC metrics include various offerings. The metrics and pricing schemes that are available on IBM z15, IBM z16, and IBM z17 are listed in Table 7-19.

Table 7-19 MLC metrics offerings

Key MLC metric	Sub-capacity	Sysplex aggregation	Contract number
Advanced Workload License Charges (AWLC) <sup>a</sup>	Y	Y	Z125-8538
Country Multiplex License Charges (CMLC) <sup>b</sup>	Y	n/a	Z126-6965
Flat Workload License Charges (FWLC) <sup>c,a</sup>	see footnotes A/C	see footnotes A/C	see footnotes A/C
IBM Z New Application License Charges (zNALC) <sup>d</sup>	Y	Y	Z125-7454
Parallel Sysplex License Charges (PSLC) <sup>e</sup>	n/a	Y	Z125-5205
Midrange Workload License Charges (MWLC)	Y	n/a	Z125-7452

a. AWLC and FWLC are available only on IBM z17 ME1, IBM z16, or IBM z15 when that machine is participating in a qualified Parallel Sysplex environment.



- b. The Country Multiplex offering was withdrawn as of 1 January 2021. For existing CMP customers, machines that are eligible to be included in a multiplex cannot be older than two generations *before* the most recently available server.
- c. Metric available with AWLC or CMLC only.
- d. This metric is eligible for subcapacity charges or for aggregation in a qualified Parallel Sysplex environment.
- e. PSLCs are available only on IBM z16 A02 or IBM z16 AGZ, or IBM z15 T02, when that machine is participating in a qualified Parallel Sysplex environment.

► **zIPLA licensing**

International Program License Agreement (IPLA) programs include a one-time charge (OTC) and an optional annual maintenance charge, called Subscription and Support. This annual charge includes access to IBM technical support and enables you to obtain version upgrades at no charge for products that generally fall under the zIPLA such as application development tools, CICS tools, data management tools, WebSphere for IBM Z products, Linux on IBM Z middleware and z/VM.

The following pricing metrics apply to IBM Z IPLA products:

- Value Unit pricing applies to most IPLA products that run on z/OS. Value Unit pricing is typically based on a number of MSUs and allows for lower cost of incremental growth.
- z/VM and specific z/VM middleware include pricing that is based on the number of engines. Engine-based Value Unit pricing allows for a lower cost of incremental growth with more engine-based licenses that are purchased.
- Most Linux middleware also is priced based on the number of engines. The number of engines is converted into Processor Value Units under the IBM Passport Advantage® terms and conditions.

For more information, see this [web page](#).

► **Subcapacity licensing**

Subcapacity licensing includes software charges for specific IBM products that are based on the use capacity of the logical partitions (LPARs) on which the product runs.

Subcapacity licensing removes the dependency between the software charges and CPC (hardware) installed capacity.

The subcapacity licensed products are charged monthly based on the highest observed 4-hour rolling average use of the logical partitions in which the product runs.

The 4-hour rolling average use of the logical partition can be limited by a defined capacity value on the image profile of the partition. This value activates the soft capping function of PR/SM, which limits the 4-hour rolling average partition use to the defined capacity value. Soft capping controls the maximum 4-hour rolling average usage (the last 4-hour average value at every 5-minute interval), but does not limit the maximum instantaneous partition use.

You can also use an LPAR group capacity limit, which sets soft capping by PR/SM for a group of logical partitions that are running z/OS. Only the 4-hour rolling average use of the LPAR group is tracked, which allows use peaks above the group capacity value.

► **Sysplex licensing**

Sysplex licensing allows monthly software licenses to be aggregated across a qualified Parallel Sysplex. To be eligible for Sysplex pricing aggregation, the customer environment must meet hardware, software, operation, and verification criteria to be considered “actively coupled”. For more information about Sysplex licensing, see this [web page](#).

► **Taylor Fit Software Consumption**

Taylor Fit Software Consumption Solution is a cloud-like, usage-based licensing model. Usage is measured based on MSUs that are used, which removes the need for manual or



automated capping. It also allows customers to configure their systems to support optimal response times and service level agreements.

Tailored Fit Pricing (TFP) requires z/OS V2.4, or later, with the applicable PTFs applied.

The requirements for TFP vary with the solution. The specific requirements for a solution must be met before IBM can accept and process subcapacity reports in which TFP solutions are reported. For more information about TFP, see this [web page](#).

## Technology Transition Offerings with IBM z17

Complementing the announcement of the IBM z16 server, IBM introduced the following Technology Transition Offerings (TTOs):

- ▶ Technology Update Pricing for the IBM z17.
- ▶ New and revised Transition Charges for Sysplexes or Multiplexes TTOs for actively coupled Parallel Sysplexes (z/OS), Loosely Coupled Complexes (z/TPF), and Multiplexes (z/OS and z/TPF).

Technology Update Pricing for the IBM z17 extends the software price and performance that is provided by AWLC for IBM z17 servers. The new and revised Transition Charges for Sysplexes offerings provide a transition to Technology Update Pricing for the IBM z17 for customers who have not fully migrated to IBM z17 servers. This transition ensures that aggregation benefits are maintained and phases in the benefits of Technology Update Pricing for the IBM z17 pricing as customers migrate.

When an IBM z17 server is in an actively coupled Parallel Sysplex or a Loosely Coupled Complex, you might choose aggregated Advanced Workload License Charges (AWLC) pricing or aggregated Parallel Sysplex License Charges (PSLC) pricing (subject to all applicable terms and conditions).

When an IBM z17 server is part of a Multiplex under Country Multiplex Pricing (CMP) terms, Country Multiplex License Charges (CMLC), Multiplex zNALC (MzNALC), and Flat Workload License Charges (FWLC) are the only pricing metrics that are available (subject to all applicable terms and conditions).

When an IBM z17 server is running z/VSE, you can choose Mid-Range Workload License Charges (MWLC), which are subject to all applicable terms and conditions.

For more information about AWLC, CMLC, MzNALC, PSLC, MWLC, or the Technology Update Pricing and Transition Charges for Sysplexes or Multiplexes TTO offerings, see the [IBM Z Software Pricing page](#) of the IBM IT infrastructure website.

## 7.9 References

For more information about planning, see the home pages for the following operating systems:

- ▶ [z/OS](#)
- ▶ [z/VM](#)
- ▶ [z/VSE](#)
- ▶ [z/TPF](#)
- ▶ [Linux on IBM Z](#)
- ▶ [KVM for IBM Z](#)



## 8



# System upgrades

This chapter provides an overview of the IBM Z server upgrade process and how, in many cases, customers can manage capacity upgrades by using online tools and automation. The chapter also includes a detailed description of capacity on demand (CoD) offerings available on the IBM z17.

IBM z17 servers support dynamic provisioning features to give clients exceptional flexibility and control over system capacity and costs.

This chapter includes the following topics:

- ▶ 8.5, “Permanent upgrade by using the CIU facility” on page 372
- ▶ 8.2, “Permanent and Temporary Upgrades” on page 354
- ▶ 8.3, “Concurrent upgrades” on page 360
- ▶ 8.4, “Miscellaneous equipment specification upgrades” on page 366
- ▶ 8.5, “Permanent upgrade by using the CIU facility” on page 372
- ▶ 8.6, “On/Off Capacity on Demand” on page 376
- ▶ 8.7, “z/OS Capacity Provisioning” on page 382
- ▶ 8.8, “System Recovery Boost” on page 387
- ▶ 8.10, “Flexible Capacity for Cyber Resiliency” on page 388
- ▶ 8.11, “Capacity Backup (CBU)” on page 390
- ▶ 8.12, “Planning for nondisruptive upgrades” on page 394
- ▶ 8.13, “Summary of Capacity on-Demand offerings” on page 399



## 8.1 Introduction

A key resource for managing client IBM Z servers is the [IBM Resource Link website](#). Once registered, a client can view product information by clicking **Resource Link** → **Client Initiated Upgrade Information**, and selecting **Education**. Select your particular product from the list of available systems.

The scalability of IBM z17 servers includes the following benefits:

- ▶ Enabling new business opportunities
- ▶ Support for dynamic capacity growth and cloud environments
- ▶ Risk management of volatile, high-growth, and high-volume applications
- ▶ Enabling 24 x 7 application availability
- ▶ Enabling capacity growth during lockdown periods
- ▶ Enabling planned downtime without availability effects

## 8.2 Permanent and Temporary Upgrades

The terminology for CoD and the types of upgrades for an IBM z17 are described in this section.

### 8.2.1 Overview

Upgrades can be categorized as described in this section.

#### Permanent versus temporary upgrades

Deciding whether to perform a permanent or temporary upgrade depends on the situation. For example, a growing workload might require more memory, I/O cards, or processor capacity. However, to handle a peak workload or to temporarily replace a system that is down during a disaster or data center maintenance, might require only a temporary upgrade.

IBM z17 servers offer the following solutions:

- ▶ Permanent upgrades
  - Miscellaneous equipment specification (MES)  
An MES upgrade might involve adding physical hardware or installing Licensed Internal Code Configuration Control (LICCC). In both cases, the hardware installation is performed by IBM personnel.
  - Customer Initiated Upgrade (CIU)  
The use of the CIU facility for a system requires that the online CoD buying feature (FC 9900) is installed on the system and the relevant CIU contract agreements are in place. The CIU facility supports only LICCC upgrades.

For more information, see 8.2.4, “Permanent upgrades” on page 358.

**Tip:** An MES provides system upgrades that can result in more enabled processors, a different central processor (CP) capacity level, more processor drawers, memory, PCIe+ I/O drawers, and I/O features (physical upgrade). Extra planning tasks are required for nondisruptive logical upgrades. An MES is ordered through your IBM representative and installed by IBM service support representatives (IBM SSRs).



► Temporary upgrades

All temporary upgrades are LICCC-based. The one billable capacity offering is On/Off Capacity on Demand (CoD), which can be used for short-term capacity requirements and are pre-paid or post-paid.

The replacement capacity offering available is the Capacity Backup (CBU), and the Flexible Capacity for Cyber Resiliency.

**Note:** Capacity for Planned Event is not available on IBM z17.

System Recovery Boost zIIP capacity is a pre-paid offering that is available on IBM z16 A01 and IBM z17 ME1. It is intended to provide temporary zIIP capacity to be used to boost CPU performance for boost events. For more information, see *Introducing IBM Z System Recovery Boost*, [REDP-5563](#).

Flexible Capacity for Cyber Resiliency is new type of temporary record that was introduced with IBM z16. This record holds the Flexible Capacity Entitlements for IBM z17 machines across two or more sites.

## 8.2.2 CoD for IBM z17 systems-related terminology

The most frequently used terms that are related to CoD for IBM z17 systems are listed in Table 8-1.

Table 8-1 CoD terminology

Term	Description
Activated capacity	Capacity that is purchased and activated. Purchased capacity can be greater than the activated capacity.
Billable capacity	Capacity that helps handle workload peaks (expected or unexpected). The one billable offering that is available is On/Off CoD.
Capacity	Hardware resources (processor and memory) that can process the workload can be added to the system through various capacity offerings.
Capacity Backup (CBU)	This capacity allows you to place model capacity or specialty engines in a backup system. CBU is used in an unforeseen loss of system capacity because of an emergency or for Disaster Recovery testing.
Capacity for Planned Event (CPE) <sup>a</sup>	Used when temporary replacement capacity is needed for a short-term event. CPE activates processor capacity temporarily to facilitate moving systems between data centers, upgrades, and other routine management tasks. CPE is an offering of CoD.
Capacity levels	Can be full capacity or sub-capacity. For an IBM z17 ME1 system, capacity levels for the CP engine are 7, 6, 5, and 4.
Capacity setting	<p>Derived from the capacity level and the number of processors.</p> <p>For the IBM z17 ME1 system, the capacity levels are 7nn, 6yy, 5yy, and 4xx, where xx, yy, or nn indicates the number of active CPs.</p> <p>The number of processors can have the following ranges:</p> <ul style="list-style-type: none"> <li>► 0 - 43 for capacity levels 4xx. An all IFL or an all ICF system has a capacity level of 400.</li> <li>► 1 - 43 for capacity levels 5yy and 6yy.</li> <li>► 1 - 99 in decimal and A0 - K8, where A0 represents 100 and K8 represents 208, for capacity level 7nn.</li> </ul>



Term	Description
Customer Initiated Upgrade (CIU)	A web-based facility where you can request processor and memory upgrades by using the IBM Resource Link and the system's Remote Support Facility (RSF) connection.
Capacity on Demand (CoD)	The ability of a system to increase or decrease its performance capacity as needed to meet fluctuations in demand.
Capacity Provisioning Manager (CPM)	As a component of z/OS Capacity Provisioning, CPM monitors business-critical workloads that are running z/OS on IBM z17.
Customer profile	This information is on Resource Link and contains client and system information. A customer profile can contain information about systems that are related to their IBM customer numbers.
Flexible Capacity for Cyber Resiliency	Available on IBM z17 ME1 servers, the optional Flexible Capacity Record is an orderable feature that entitles a customer to active MIPS flexibility for all engine types between IBM z17 servers across two or more sites. It allows capacity swaps for an extended term.
Full capacity CP feature	For IBM z17 servers, feature (CP7) provides full capacity. Capacity settings 7nn are full capacity settings with the ranges of 1 - 99 in decimal and A0 - K0, where A0 represents 100 and K8 represents 208, for capacity level 7nn.
High-water mark	Capacity that is purchased and owned by the client.
Installed record	The LICCC record is downloaded, staged to the Support Element (SE), and is installed on the central processor complex (CPC). A maximum of eight different records can be concurrently installed.
Model capacity identifier (MCI)	Shows the current active capacity on the system, including all replacement and billable capacity. For IBM z17 ME1 servers, the model capacity identifier is in the form of 4xx, 5yy, 6yy, or 7nn, where xx, yy, or nn indicates the number of active CPs: <ul style="list-style-type: none"> <li>▶ xx can have a range of 00 - 43. An all IFL or an all ICF system has a capacity level of 400.</li> <li>▶ yy can have a range of 01 - 43.</li> <li>▶ 1 - 99 in decimal and A0 - K8, where A0 represents 100 and K8 represents 208, for capacity level 7nn.</li> </ul>
Model Permanent Capacity Identifier (MPCI)	Keeps information about the capacity settings that are active before any temporary capacity is activated.
Model Temporary Capacity Identifier (MTCI)	Reflects the permanent capacity with billable capacity only, without replacement capacity. If no billable temporary capacity is active, MTCI equals the MPCI.
On/Off Capacity on Demand	Represents a function that allows spare capacity in a CPC to be made available to increase the total capacity of a CPC. For example, On/Off CoD can be used to acquire capacity for handling a workload peak.
Permanent capacity	The capacity that a client purchases and activates. This amount might be less capacity than the total capacity purchased.
Permanent upgrade	LICC that is licensed by IBM to enable the activation of applicable computing resources, such as processors or memory, for a specific CIU-eligible system on a permanent basis.
Purchased capacity	Capacity that is delivered to and owned by the client. It can be higher than the permanent capacity.
Permanent/Temporary entitlement record	The internal representation of a temporary (TER) or permanent (PER) capacity upgrade that is processed by the CIU facility. An <i>entitlement record</i> contains the encrypted representation of the upgrade configuration with the associated time limit conditions.



Term	Description
Replacement capacity	A temporary capacity that is used for situations in which processing capacity in other parts of the enterprise is lost. This loss can be a planned event or an unexpected disaster. The two replacement offerings available are Capacity for Planned Events and Capacity Backup.
Resource Link	The <a href="#">IBM Resource Link website</a> a technical support website that provides a comprehensive set of tools and resources (log in required).
Secondary approval	An option that is selected by the client that requires second approver control for each CoD order. When a secondary approval is required, the request is sent for approval or cancellation to the Resource Link secondary user ID.
Staged record	The point when a record that represents a temporary or permanent capacity upgrade is retrieved and loaded on the SE disk.
Subcapacity	For IBM z17 ME1 servers, CP features (CP4, CP5, and CP6) provide reduced capacity relative to the full capacity CP feature (CP7).
System Recovery Boost Upgrade record	Available on IBM z17 ME1 servers, the optional System Recovery Boost Upgrade is an orderable feature that provides more capacity for a limited time to enable speeding up shutdown, restart, and catchup processing for a limited event duration.
Temporary capacity	An optional capacity that is added to the current system capacity for a limited amount of time. It can be capacity that is owned or not owned by the client.
Vital product data (VPD)	Information that uniquely defines system, hardware, software, and microcode elements of a processing system.

a. Capacity for Planned Event (CPE) is not supported on IBM z17.

### 8.2.3 Concurrent and nondisruptive upgrades

Depending on the effect on the system and application availability, upgrades can be classified in the following manner:

- Concurrent

In general, *concurrency* addresses the continuity of operations of the *hardware* during an upgrade; for example, whether a system (hardware) must be turned off during the upgrade. For more information, see 8.3, “Concurrent upgrades” on page 360.

- Non concurrent

This type of upgrade requires turning off the hardware that is being upgraded. Examples include memory upgrades to an IBM z17 ME1 max 43.

- Nondisruptive

*Nondisruptive* upgrades do not require the software or operating system to be restarted for the upgrade to take effect.

- Disruptive

An upgrade is considered *disruptive* when resources that are modified or added to an operating system image require that the operating system be restarted to configure the newly added resources.

A Concurrent upgrade might be disruptive to operating systems or programs that do not support the upgrades while being nondisruptive to others. For more information, see 8.12, “Planning for nondisruptive upgrades” on page 394.



## 8.2.4 Permanent upgrades

Permanent upgrades can be obtained by using the following processes:

- ▶ Ordered through an IBM marketing representative
- ▶ Initiated by the client with the CIU on the IBM Resource Link

**Tip:** The use of the CIU facility for a system requires that the online CoD buying feature (FC 9900) is installed on the system. The CIU facility is enabled through the permanent upgrade authorization feature code (FC 9898).

### Permanent upgrades that are ordered through an IBM representative

Through a permanent upgrade, you can accomplish the following tasks:

- ▶ Add:
  - Processor drawers
  - Peripheral Component Interconnect Express (PCIe) drawers and features
  - Model capacity
  - Specialty engines
  - Memory
  - I/O channels
  - Crypto Express cards
- ▶ Activate unassigned model capacity or IFLs, ICFs, or zIIPs
- ▶ Deactivate activated model capacity or IFLs, ICFs, or zIIPs
- ▶ Change specialty engines (recharacterization)

**Considerations:** Most of the MESs can be concurrently applied without disrupting the workload. For more information, see 8.3, “Concurrent upgrades” on page 360. However, specific MES changes might be disruptive, such as adding PCIe IO drawers.

Memory upgrades that require dual inline memory module (DIMM) changes can be made non disruptively if multiple CPC drawers are available and the flexible memory option is used.

### Permanent upgrades by using CIU on the IBM Resource Link

Ordering the following permanent upgrades by using the CIU application through Resource Link allows you to add capacity to fit within your hardware:

- ▶ Add:
  - Model capacity
  - Specialty engines
  - Memory
- ▶ Activate unassigned model capacity or IFLs, ICFs, or zIIPs
- ▶ Deactivate activated model capacity or IFLs, ICFs, or zIIPs



## 8.2.5 Temporary upgrades

IBM z17 ME1 offers the following types of temporary upgrades:

- ▶ On/Off CoD

This offering allows you to temporarily add capacity or specialty engines to cover seasonal activities, period-end requirements, peaks in workload, or application testing. This temporary upgrade can be ordered by using the CIU application through Resource Link only.

**Prepaid On/Off CoD tokens:** Beginning with IBM z16, new prepaid On/Off CoD tokens that are purchased do not carry forward to future systems.

- ▶ CBU

This offering allows you to replace model capacity or specialty engines in a backup system that is used in an unforeseen loss of system capacity because of a disaster.

- ▶ System Recovery Boost record (FC 6802)

This offering allows you to add up to 20 zIIPs for use with the System Recovery Boost facility. System Recovery Boost provides temporary extra capacity for CP workloads to allow rapid shutdown, restart, and recovery of eligible systems. System Recovery Boost records are prepaid, licensed for 1 - 5 years, and can be renewed at any time.

- ▶ Flexible Capacity Record

This offering allows you to move CPU capacity between machines across two or more sites. Capacity can be moved between sites a maximum of 12 times per year for a maximum of 12 months per move.

Consider the following points:

- ▶ CBU, and System Recovery Records can be ordered by using the CIU application through Resource Link or by contacting your IBM marketing representative.
- ▶ Flexible Capacity can be ordered by contacting your IBM representative.
- ▶ Temporary upgrade capacity changes might be billable or a replacement.

### Billable capacity

To handle a peak workload, you can activate up to double the purchased capacity of any processor unit (PU) type temporarily. You are charged daily.

This billable capacity offering is On/Off CoD.

### Replacement capacity

When processing capacity is lost in part of an enterprise, replacement capacity can be activated. It allows you to activate any PU type up to your authorized limit.

The following replacement capacity offerings are available:

- ▶ Capacity Backup
- ▶ Flexible Capacity for Cyber Resiliency



## 8.3 Concurrent upgrades

Concurrent upgrades on IBM z17 servers can provide more capacity with no system outage. In most cases, a concurrent upgrade can be nondisruptive to the operating system with planning and operating system support.

This capability is based on the flexibility of the design and structure, which allows concurrent hardware installation and Licensed Internal Control Code (LICC) configuration changes.

The sub-capacity models allow more configuration granularity within the family. The added granularity is available for models that are configured with up to 43 CPs, and provides 129 extra capacity settings. Sub-capacity models provide for CP capacity increase in two dimensions that can be used together to deliver configuration granularity. The first dimension is adding CPs to the configuration. The second is changing the capacity setting of the CPs currently installed to a higher model capacity identifier.

IBM z17 allows the concurrent and nondisruptive addition of processors to a running logical partition (LPAR). As a result, you can have a flexible infrastructure to which you can add capacity. This function is supported by z/OS, z/VM, and z/VSE. This addition is made by using one of the following methods:

- ▶ With planning ahead for the future need of extra processors. Reserved processors can be specified in the LPAR's profile. When the extra processors are installed, the number of active processors for that LPAR can be increased without the need for a partition reactivation and initial program load (IPL).
- ▶ Another (easier) way is to enable the dynamic addition of processors through the z/OS LOADxx member. Set the **DYNCPADD** parameter in member LOADxx to **ENABLE**.

### 8.3.1 PU Capacity feature upgrades

IBM z17 Model ME1 has a machine type 9175 and a model capacity identifier.

The 9175 is available in the following CPC drawer configurations:

- ▶ Feature Max43 (one CPC Drawer installed) can have a maximum of 43 PUs for client characterization
- ▶ Feature Max90 (two CPC Drawers) can have a maximum of 90 client PUs
- ▶ Feature Max136 (three CPC Drawers) can have a maximum of 136 client PUs
- ▶ Feature Max183 (four CPC Drawers) can have a maximum of 183 client PUs
- ▶ Feature Max208 (four CPC Drawers) can have a maximum of 208 client PUs
- ▶ Model capacity identifiers 4xx, 5yy, 6yy, or 7nn

The *xx* is a range of 00 - 43<sup>1</sup>, *yy* is a range of 01 - 43, and *nn* is a range of 01 - 99, A0 - K0, where A0 represents the decimal number 100, which combines the character A with decimal 0 and where K0 represents the decimal number 200. It is obtained by continuing the hexadecimal counting to the following values:

- F = 15
- G = 16
- H = 17
- I = 18
- J = 19
- K = 20

<sup>1</sup> The IBM z17 zero CP MCI is 400. This setting applies to an all-IFL or all-ICF system.



- Adding the decimal digit 8 to make 208
- An IBM z17 ME1 with 208 client usable processors is an IBM z17 - 7K8. The model capacity identifier describes how many CPs are characterized (*xx*, *yy*, or *nn*) and the capacity setting (4, 5, 6, or 7) of the CPs.

A hardware configuration upgrade always requires more physical hardware (processor drawers, PCIe+ I/O drawers, or both). A system upgrade can change the system model or the MCI.

Consider the following points regarding model upgrades:

- ▶ LICCC upgrade:
  - Can add memory or Virtual Flash Memory (VFM) up to the amount that is physically installed
  - Can change the model capacity identifier, the capacity setting, or both
- ▶ Hardware installation upgrade:
  - Can change the CPC drawer feature by adding one or more drawers
  - Can change the model capacity identifier, the capacity setting, or both
  - Can add physical memory, PCIe+ I/O drawers, and other hardware features

The model capacity identifier can be concurrently changed. Concurrent upgrades can be performed for permanent and temporary upgrades.

**Tip:** A CPC drawer feature upgrade can be performed concurrently only for a Max43 or a Max902 machine if feature codes 2933 or 2934 were ordered with the base machine.

### LICCC upgrades (MES ordered)

The LICCC provides for system upgrades without hardware changes by activating extra (physically installed) unused capacity. Concurrent upgrades through LICCC can be performed for the following resources:

- ▶ Processors, such as CPs, ICFs, z Integrated Information Processors (zIIPs), IFLs, and SAPs, if unused PUs are available on the installed processor drawers, or if the model capacity identifier for the CPs can be increased.
- ▶ Memory, when unused capacity is available on the installed memory cards. The Flexible memory option is available to give you better control over future memory upgrades. For more information, see 2.5.7, “Flexible Memory Option” on page 53.

### Concurrent hardware installation upgrades (MES ordered)

Configuration upgrades can be concurrent when installing the following resources:

- ▶ Processor drawers (which contain processors, memory, and fan-outs). Up to two processor drawers can be added concurrently on an IBM z17 ME1 Max43 if feature codes 2933 and 2934 were ordered with the initial configuration.
- ▶ PCIe+ Gen3 fan-outs.
- ▶ I/O cards, when slots are still available on the installed PCIe+ I/O drawers.
- ▶ PCIe+ I/O drawers.

The concurrent I/O upgrade capability can be better used if a future target configuration is considered during the initial configuration.



### Concurrent PU conversions (MES ordered)

IBM z17 ME1 supports concurrent conversion between all PU types, to provide flexibility and meet changing business requirements.

**Important:** The LICCC-based PU conversions require that at least one PU (CP, ICF, or IFL), remains unchanged. Otherwise, the conversion is disruptive. The PU conversion generates a LICCC that can be installed concurrently in two steps:

1. Remove the assigned PU from the configuration.
2. Activate the newly available PU as the new PU type.

LPARs also might have to free the PUs to be converted. The operating systems must include support to configure processors offline or online so that the PU conversion can be done non disruptively.

**Considerations:** Client planning and operator action are required to use concurrent PU conversion. Consider the following points about PU conversion:

- ▶ It is disruptive if *all* current PUs are converted to different types.
- ▶ It might require individual LPAR outages if dedicated PUs are converted.

Unassigned CP capacity is recorded by a model capacity identifier. CP feature conversions change (increase or decrease) the model capacity identifier.

## 8.3.2 Customer Initiated Upgrade facility

The CIU facility is an IBM online system through which you can order, download, and install permanent and temporary upgrades for IBM Z servers. Access to and use of the CIU facility requires a contract between the client and IBM through for which the terms and conditions for use of the CIU facility are accepted.

The CIU facility is controlled through the permanent upgrade authorization FC 9898. A prerequisite to FC 9898 is the online CoD buying feature code (FC 9900). Although FC 9898 can be installed on your IBM z17 servers at any time, often it is added when ordering an IBM z17.

After you place an order through the CIU facility, you receive a notice that the order is ready for download. You can then download and apply the upgrade by using functions that are available through the Hardware Management Console (HMC), along with the RSF. After all of the prerequisites are met, the entire process (from ordering to activation of the upgrade) is performed by the customer, and does not require any onsite presence of IBM SSRs.

### CIU prerequisites

The CIU facility supports LICCC upgrades only. It does not support I/O upgrades. All other capacity that is required for an upgrade must be previously installed. Extra processor drawers or I/O cards cannot be installed as part of an order that is placed through the CIU facility. The sum of CPs, unassigned CPs, ICFs, zIIPs, IFLs, and unassigned IFLs cannot exceed the client PU count of the installed processor drawers. The total number of zIIPs can be twice the number of purchased CPs.

### CIU registration and contract for CIU

To use the CIU facility, a customer must be registered and the system must be set up. After you complete the CIU registration, access to the CIU application is available through the [IBM Resource Link website](#).



As part of the setup, provide one resource link ID for configuring and placing CIU orders and, if required, a second ID as an approver. The IDs are then set up for access to the CIU support. The CIU facility allows upgrades to be ordered and delivered much faster than through the regular MES process.

To order and activate the upgrade, log on to the [IBM Resource Link website](#), and start the CIU application to upgrade a system for processors or memory. You can request a client order approval to conform to your operational policies. You also can allow the definition of more IDs to be authorized to access the CIU. More IDs can be authorized to enter or approve CIU orders, or only view orders.

## Permanent upgrades

Permanent upgrades can be ordered by using the CIU facility. Through the CIU facility, you can generate online permanent upgrade orders to concurrently add processors (CPs, ICFs, zIIPs, and IFLs), and memory, or change the model capacity identifier. You can do so up to the limits of the installed processor drawers on a system.

## Temporary upgrades

The IBM z17 ME1 base model describes permanent and dormant capacity by using the capacity marker and the number of PU features that are installed on the system. Up to eight temporary offerings can be present. Each offering includes its own policies and controls, and each can be activated or deactivated independently in any sequence and combination. Although multiple offerings can be active at any time, only *one* On/Off CoD offering can be active at any time if enough resources are available to fulfill the offering specifications.

Temporary upgrades are represented in the system by a *record*. All temporary upgrade records are on the SE hard disk drive (HDD). The records can be downloaded from the RSF or installed from portable media. At the time of activation, you can control everything locally.

The provisioning architecture is shown in Figure 8-1.

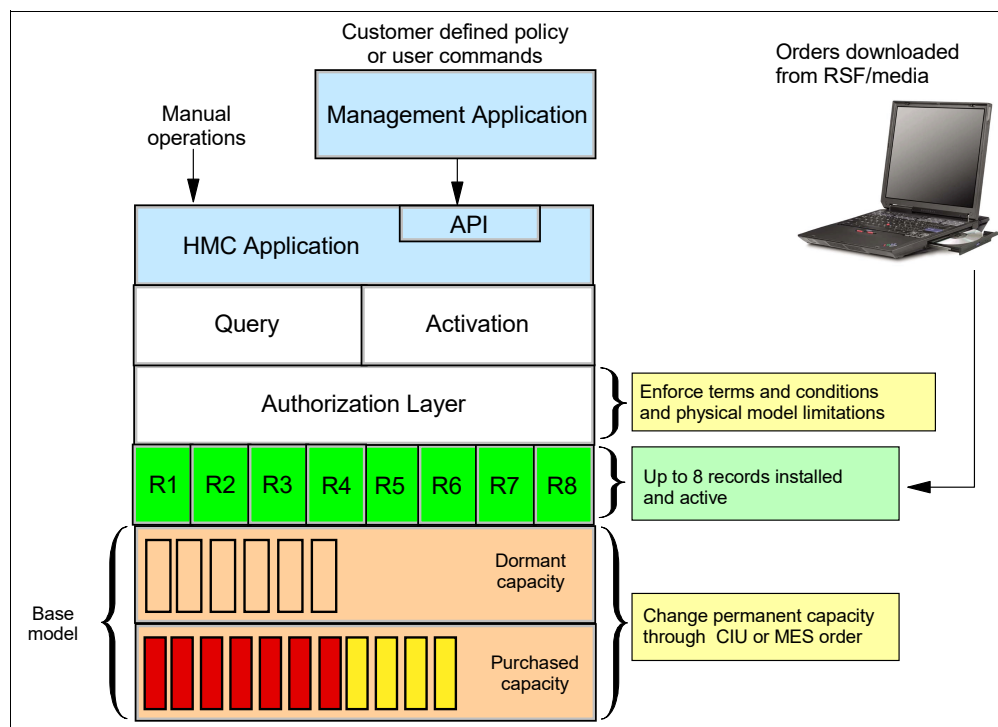


Figure 8-1 Provisioning architecture



The authorization layer enables administrative control over the temporary offerings. The activation and deactivation can be driven manually or under the control of an application through a documented application programming interface (API)<sup>2</sup>.

By using the API approach, you can customize at activation time the resources that are necessary to respond to the current situation up to the maximum that is specified in the order record. If the situation changes, you can add or remove resources without having to return to the base configuration. This process eliminates the need for temporary upgrade specifications for all possible scenarios.

This approach also enables you to update and replenish temporary upgrades, even in situations where the upgrades are active. Likewise, depending on the configuration, permanent upgrades can be performed while temporary upgrades are active. Examples of the activation sequence of multiple temporary upgrades are shown in Figure 8-2.

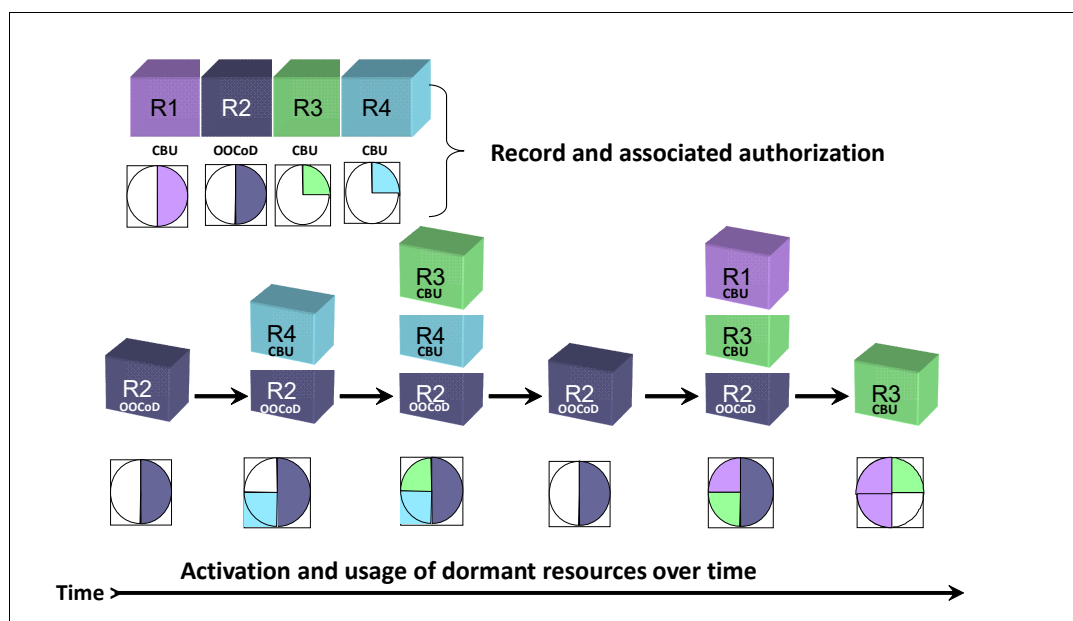


Figure 8-2 Example of temporary upgrade activation sequence

As shown in Figure 8-2, if R2, R3, and R1 are active at the same time, only parts of R1 can be activated because not enough resources are available to fulfill all of R1. When R2 is deactivated, the remaining parts of R1 can be activated as shown.

Temporary capacity can be billable as On/Off CoD, or replacement capacity as CBU, Flexible Capacity, or System Recovery Boost. Consider the following points:

- On/Off CoD is a function that enables *concurrent* and *temporary* capacity growth of the system.

On/Off CoD can be used for client peak workload requirements, for any length of time, and includes a daily hardware and maintenance charge. The software charges can vary according to the license agreement for the individual products. For more information, contact your IBM Software Group representative.

<sup>2</sup> API details can be found in *z/OS MVS Programming: Callable Services for High-Level Languages*, SA23-1377



On/Off CoD can concurrently add processors (CPs, ICFs, zIIPs, and IFLs), increase the model capacity identifier, or both. It can do so up to the limit of the installed processor drawers of a system. It is restricted to twice the installed capacity. On/Off CoD requires a contractual agreement between you and IBM.

You decide whether to pre-pay or post-pay On/Off CoD. Capacity tokens that are inside the records are used to control activation time and resources.

- CBU is a concurrent and temporary activation of more CPs, ICFs, zIIPs, and IFLs; or an increase of the model capacity identifier; or both.

**Note:** CBU cannot be used for peak workload management in any form.

On/Off CoD is the correct method to use for workload management. A CBU activation can last up to 90 days when a disaster or recovery situation occurs.

CBU features are optional, and require unused capacity to be available on installed processor drawers of the backup system. They can be available as unused PUs, an increase in the model capacity identifier, or both.

A CBU contract must be in place before the special code that enables this capability can be loaded on the system. The standard CBU contract provides for five 10-day tests (the *CBU test activation*) and one 90-day activation over a five-year period. For more information, contact your IBM representative.

You can run production workload on a CBU upgrade during a CBU test. At least an *equivalent* amount of production capacity must be shut down during the CBU test. If you signed CBU contracts, you also must sign an Amendment (US form #Z125-8145) with IBM to allow you to run production workload on a CBU upgrade during your CBU tests. More 10-day tests can be purchased with the CBU record.

- The System Recovery Boost Upgrade allows a concurrent activation of extra zIIPs.

The System Recovery Boost Upgrade record offering can be used to provide extra zIIP capacity that can be used by the System Recovery Boost facility. You might want to consider the use of this offering if your server is a full capacity model (7nn) and can benefit from more CP capacity (running on zIIPs) for system shutdown and restart. The capacity is delivered as zIIPs that can perform CP work during the boost periods for an LPAR.

A System Recovery Boost Record contract must be in place before the special code that enables this capability can be loaded on the system. The standard contract provides for one 6-hour activation for the specific purpose of System Recovery Boost only. For more information, contact your IBM representative.

Activation of System Recovery Boost Record does not change the MCI of your system.



### 8.3.3 Concurrent upgrade functions summary

The possible concurrent upgrades combinations are listed in Table 8-2.

Table 8-2 Concurrent upgrade summary

Type	Name	Upgrade	Process
Permanent	MES	CPs, ICFs, zIIPs, IFLs, processor drawer, memory, and I/Os	Installed by IBM SSRs
	Online permanent upgrade	CPs, ICFs, zIIPs, IFLs, and memory	Performed through the CIU facility
Temporary	On/Off CoD	CPs, ICFs, zIIPs, and IFLs	Performed through the On/Off CoD facility
	CBU	CPs, ICFs, zIIPs, and IFLs	Activated through model conversion
	System Recovery Boost Record	zIIPs	Activated through model conversion
	Flexible Capacity Record	CPs, ICFs, zIIPs, and IFLs	Activated through model conversion

## 8.4 Miscellaneous equipment specification upgrades

MES upgrades enable concurrent and permanent capacity growth. MES upgrades allow the concurrent adding of processors (CPs, ICFs, zIIPs, and IFLs), memory capacity, and I/O ports. For sub-capacity models, MES upgrades allow the concurrent adjustment of the number of processors and the capacity level.

The MES upgrade can be performed by using LICCC only, installing more processor drawers, adding PCIe+ I/O drawers, adding I/O<sup>3</sup> features, or using the following combinations:

- ▶ MES upgrades for processors are done by any of the following methods:
  - LICCC assigning and activating unassigned PUs up to the limit of the installed processor drawers.
  - LICCC to adjust the number and types of PUs to change the capacity setting, or both.
  - Installing more processor drawers and LICCC assigning and activating unassigned PUs on the installed processor drawers.
- ▶ MES upgrades for memory are done by one of the following methods:
  - By using LICCC to activate more memory capacity up to the limit of the memory cards on the currently installed processor drawers. Flexible memory features enable you to implement better control over future memory upgrades. For more information about the memory features, see 2.5.7, “Flexible Memory Option” on page 53.
  - Installing more processor drawers and the use of LICCC to activate more memory capacity on installed processor drawers.
  - By using the CPC Enhanced Drawer Availability (EDA), where possible, on multi-drawer systems to add or change the memory cards.

<sup>3</sup> Other adapter types, such as zHyperlink, Coupling Express LR, and Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE), also can be added to the PCIe+ I/O drawers through an MES.



- MES upgrades for I/O are done by installing I/O features and supporting infrastructure (if required) on PCIe drawers that are installed, or installing PCIe drawers to hold the new cards.

An MES upgrade requires IBM SSRs for the installation. In most cases, the time that is required for installing the LICCC and completing the upgrade is short, depending on how up to date the machine microcode levels are.

To better use the MES upgrade function, carefully plan the initial configuration to allow a concurrent upgrade to a target configuration. The availability of PCIe+ I/O drawers improves the flexibility to perform unplanned I/O configuration changes concurrently.

The Store System Information (STSI) instruction gives more useful and detailed information about the base configuration and temporary upgrades.

The model and model capacity identifiers that are returned by the STSI instruction are updated to coincide with the upgrade. For more information, see “Store System Information instruction” on page 396.

**Upgrades:** An MES provides the physical upgrade, which results in more enabled processors, different capacity settings for the CPs, and more memory, I/O ports, I/O adapters, and I/O drawers. Extra planning tasks are required for nondisruptive logical upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 398.

### 8.4.1 MES upgrade for processors

An MES upgrade for processors can concurrently add CPs, ICFs, zIIPs, and IFLs to an IBM z17 by assigning available PUs on the processor drawers through LICCC. Depending on the quantity of the extra processors in the upgrade, more processor drawers might be required, and can be concurrently installed before the LICCC is enabled if plan-ahead features are available. With the sub-capacity models, capacity can be provided by adding CPs, changing the capacity identifier on the current CPs, or both.

**Limits:** The sum of CPs, inactive CPs, ICFs, unassigned ICFs, zIIPs, unassigned zIIPs, IFLs, and unassigned IFLs, cannot exceed the maximum limit of PUs available for client use.

An example of an MES upgrade for processors (with two upgrade steps) is shown in Figure 8-3 on page 368



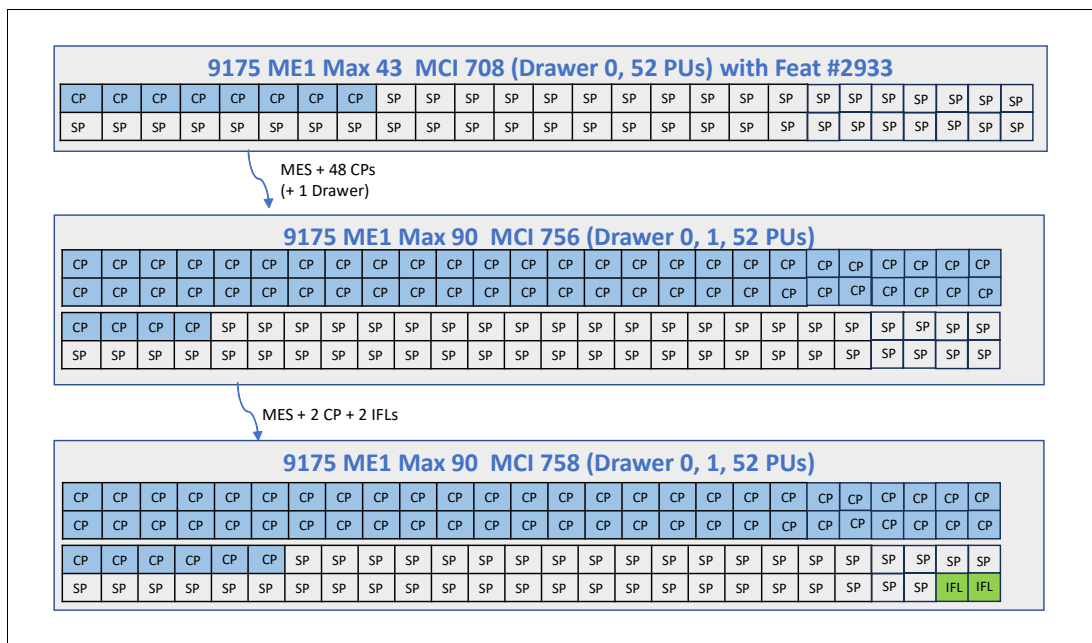


Figure 8-3 MESs for Processor examples

An IBM z17 model ME1 Max43 (one processor drawer), model capacity identifier 708 (eight CPs), is concurrently upgraded to a model ME1 Max90 (two processor drawers), with MCI 756 (56 CPs). The model upgrade requires adding a processor drawer and assigning and activating 48 PUs as CPs. Then, model Max90, MCI 756, is concurrently upgraded to a capacity identifier 758 (58 CPs) with two IFLs. This process is done by assigning and activating four more unassigned PUs (two as CP and two as IFLs). If needed, LPARs can be created concurrently to use the newly added processors.

The example that is shown in Figure 8-3 shows how the addition of PUs as CPs and IFLs and the addition of a processor drawer works. The addition of a processor drawer to an IBM z17 Max43 upgrades the machine to Max90.

After the second CPC drawer addition, CPC drawer 0 has 52 configurable PUs and CPC drawer 1 has 52 configurable PUs, which allows 90 PUs to be characterized on the new Max90 configuration.

**Consideration:** All available processors on a server (including reserved processors) can be defined to an LPAR. However, do not define more processors to an LPAR than the target operating system supports.



The number of processors that are supported by various operating systems releases are listed in Table 8-3.

*Table 8-3 Number of processors that are supported by operating system*

Operating system	Number of processors that are supported
z/OS V2R4 and later	200 PUs per z/OS LPAR in non-SMT mode and 128 PUs per z/OS LPAR in SMT mode. For both, the PU total is the sum of CPs and zIIPs.
z/VM V7R3 and later	80 (or 40 in SMT mode).
z/VSE <sup>n</sup> V6.3	z/VSE Turbo Dispatcher can use up to 4CPs, and tolerates up to 10-way LPARs.
z/TPF	86 CPs.
Linux on IBM z17	The IBM z17 limit is 200 CPs although Linux <sup>a</sup> supports 256 cores without SMT and 128 cores with SMT (256 threads).

a. Supported Linux on IBM Z distributions (for more information, see 322).

Software charges, which are based on the total capacity of the system on which the software is installed, are adjusted to the new capacity after the MES upgrade.

Software products that use Workload License Charges (WLC) or Taylor Fit Pricing (TFP) might not be affected by the system upgrade. Their charges are based on partition usage, not on the system total capacity. For more information about WLC, see 7.8, “Software licensing” on page 348.

## 8.4.2 MES upgrades for memory

MES upgrades for memory can concurrently add memory in the following ways:

- ▶ Through LICCC, which enables more capacity up to the limit of the installed DIMM memory cards.
- ▶ Concurrently installing CPC drawers and LICCC-enabling memory capacity on the new CPC drawers.

The Flexible Memory Feature is available to allow better control over future memory upgrades. For more information about flexible memory features, see 2.5.7, “Flexible Memory Option” on page 53.

If the IBM z17 is a multiple processor drawer configuration, you can use the EDA feature to remove a processor drawer and add DIMM memory cards. It also can be used to upgrade the installed memory cards to a larger capacity size. You can then use LICCC to enable the extra memory.

With suitable planning, memory can be added non disruptively to z/OS partitions and z/VM partitions. If necessary, new LPARs can be created non disruptively to use the newly added memory.

**Concurrency:** Upgrades that require DIMM changes can be concurrent by using the EDA feature. Planning is required to see whether this option is a viable for your configuration. The use of the flexible memory option ensures that EDA can work with the least disruption.



The one-processor drawer feature Max43 requires a minimum of 1024 GB addressable memory. The client addressable storage in this case is 340 GB. Memory can be upgraded up to 16 TB of memory. An upgrade changes the DIMM sizes and adding DIMMs in all available slots in the processor drawer. Total available memory for customer use is the result of total installed memory minus 884 GB allocated to HSA.

You also can add memory by *concurrently* adding a second processor drawer with sufficient memory into the configuration and then, using LICCC to enable that memory. Changing DIMMs in a single CPC drawer system is disruptive.

An LPAR can dynamically take advantage of a memory upgrade if reserved storage is defined to that LPAR. The reserved storage is defined to the LPAR as part of the image profile.

Reserved memory can be configured online to the LPAR by using the LPAR dynamic storage reconfiguration (DSR) function. DSR allows a z/OS operating system image and z/VM partitions to add reserved storage to their configuration if any unused storage exists.

The nondisruptive addition of storage to a z/OS and z/VM partition requires the correct operating system parameters to be set. If reserved storage is not defined to the LPAR, the LPAR must be deactivated, the image profile changed, and the LPAR reactivated. This process allows the extra storage resources to be available to the operating system image.

### 8.4.3 MES upgrades for I/O

MES upgrades for I/O can concurrently add I/O features by using one of the following methods:

- ▶ Installing I/O features on an installed PCIe+ I/O drawer
- ▶ Adding a PCIe+ I/O drawer to hold the new I/O features

For more information about PCIe+ I/O drawers, see 4.2, “I/O system overview” on page 172.

The number of PCIe+ I/O drawers that can be present in an IBM z17 depends on how many CPC drawers are present. It also depends on whether the CPC drawer reserve features are present.

The number of drawers for IBM z17 configuration options is listed in Table 8-4.

**Note:** The maximum number of I/O drawers in the table is reduced by one for each CPC drawer reserve feature that is present.

Table 8-4 PCIe+ I/O drawer limit for the IBM z17 systems

Number of frames	Max43	90	Max136	Max183/Max208
1	3	2	1	-
2	6	7	6	4
3	-	12	11	9
4	-	-	-	12

Depending on the number of I/O features, the configurator determines the number of PCIe+ I/O drawers that is required.



To better use the MES for I/O capability, carefully plan the initial configuration to allow concurrent upgrades up to the target configuration.

If a PCIe+ I/O drawer is added to an IBM z17 and original features must be physically moved to another PCIe+ I/O drawer, original card moves are disruptive.

z/VSE<sup>n</sup>, z/TPF, and Linux on Z do *not* provide dynamic I/O configuration support. Although installing the new hardware is done concurrently, defining the new hardware to these operating systems requires an IPL.

**Tip:** IBM z17 ME1 features a hardware system area (HSA) of 884 GB HSA is *not* part of the client-purchased memory.

## 8.4.4 Feature on Demand

Only one Feature on Demand (FoD) LICCC record is installed or staged at any time in the system. Its contents can be viewed in the Manage window, as shown in Figure 8-4.

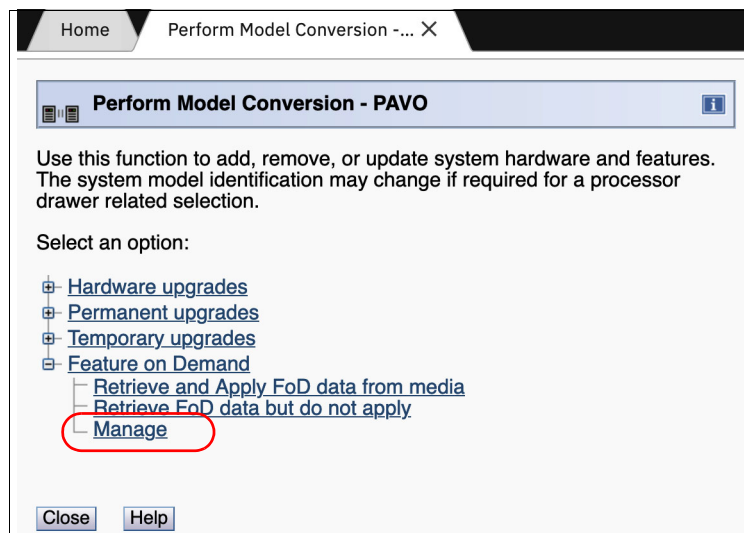


Figure 8-4 Features on-Demand

A staged record can be removed without installing it. A FoD record can be installed only completely; no selective feature or partial record installation is available. The features that are installed are merged with the CPC LICCC after activation.

A FoD record can be installed only once. If it is removed, a new FoD record is needed to reinstall. A remove action cannot be undone.

## 8.4.5 Summary of plan-ahead feature

The flexible memory plan-ahead feature is available for IBM z17 servers. No feature code is associated with flexible memory. The purpose of flexible memory is to enable enhanced processor drawer availability. If a processor drawer must be serviced, the flexible memory is activated to accommodate storing the CPC drawer that is taken offline. After the repair action, the memory is taken offline again and is made unavailable for use.



**Tip:** Accurate planning and the definition of the target configuration allows you to maximize the value of these plan-ahead features.

## 8.5 Permanent upgrade by using the CIU facility

By using the CIU facility (through [IBM Resource Link](https://www.ibm.com/support/resourcelink)), you can start a permanent upgrade for CPs, ICFs, zIIPs, IFLs, or memory. When performed through the CIU facility, you add the resources without IBM personnel present at your location. You can also unassign previously purchased CPs, zIIPs, ICFs and IFL processors through the CIU facility.

Adding permanent upgrades to a system through the CIU facility requires that the permanent upgrade enablement feature (FC 9898) is installed on the system. A permanent upgrade might change the system model capacity identifier (4xx, 5yy, 6yy, or 7nn) if more CPs are requested, or if the capacity identifier is changed as part of the permanent upgrade. If necessary, more LPARs can be created concurrently to use the newly added processors.

**Consideration:** A permanent upgrade of processors can provide a concurrent upgrade, which results in more enabled processors that are available to a system configuration. More planning and tasks are required for *nondisruptive* logical upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 398.

Maintenance charges are automatically adjusted as a result of a permanent upgrade.

Software charges that are based on the total capacity of the system on which the software is installed are adjusted to the new capacity after the permanent upgrade is installed. Software products that use WLC or customers with TFP might not be affected by the system upgrade because their charges are based on LPAR usage rather than system total capacity.

For more information about WLC, see 7.8, “Software licensing” on page 348.

The CIU facility process on IBM Resource Link is shown in Figure 8-5.

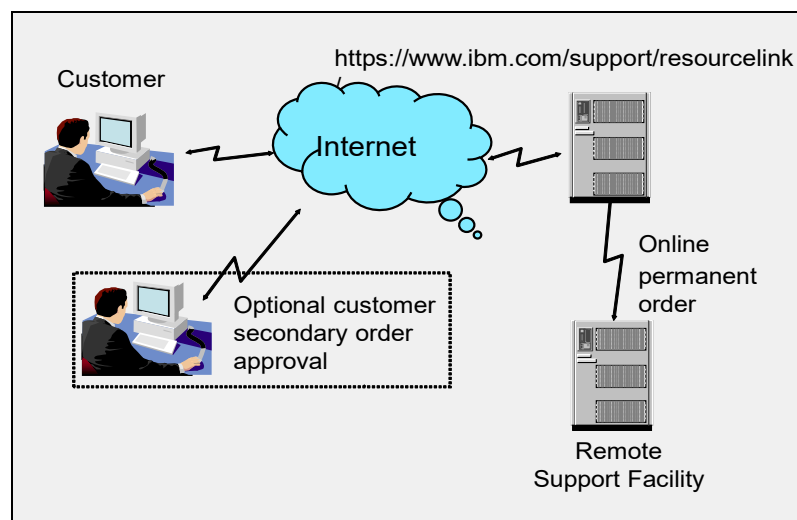


Figure 8-5 Permanent upgrade order example



The following sample sequence shows how to start an order on IBM Resource Link:

1. Sign on to Resource Link.
2. Select **Customer Initiated Upgrade** from the main Resource Link page. Client and system information that is associated with the user ID are displayed.
3. Select the system to receive the upgrade. The current configuration (PU allocation and memory) is shown for the selected system.
4. Select **Order Permanent Upgrade**. Resource Link limits the options to those options that are valid or possible for the selected configuration (system).
5. After the target configuration is verified by the system, accept or cancel the order. An order is created and verified against the pre-established agreement.
6. Accept or reject the price that is quoted. A secondary order approval is optional. Upon confirmation, the order is processed. The LICCC for the upgrade is available within hours.

The order activation process for a permanent upgrade is shown in Figure 8-6. When the LICCC is passed to the Remote Support Facility, you are notified through an e-mail that the upgrade is ready to be downloaded.

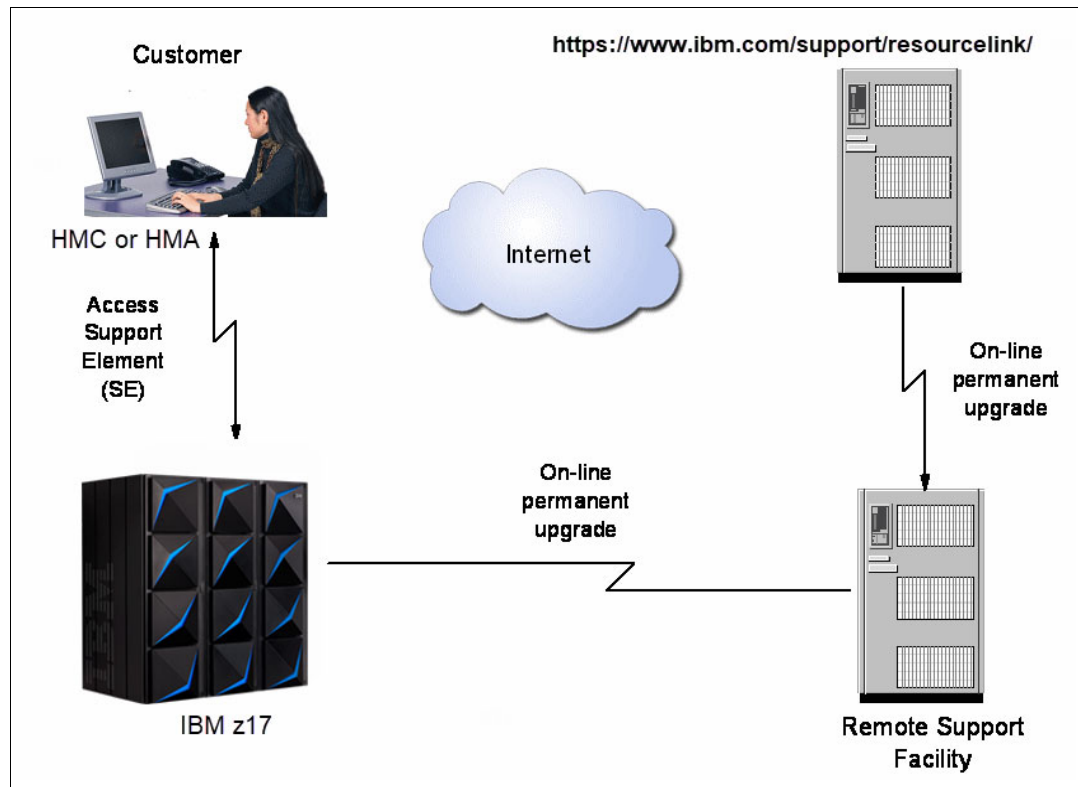


Figure 8-6 CIU-eligible order activation example

### 8.5.1 Ordering

IBM Resource Link provides the interface that enables you to order a concurrent upgrade for a system. You can create, cancel, or view the order, and view the history of orders that were placed through this interface.

Configuration rules enforce that only valid configurations are generated within the limits of the individual system. Warning messages are issued if you select invalid upgrade options. The process allows only one permanent CIU-eligible order for each system to be placed at a time.



For more information, see the [IBM Resource Link website](#) (log in required).

The initial view of the Machine profile on Resource Link is shown in Figure 8-7.

**Machine profile**  
8561 - [REDACTED]

Current configuration	
Model Capacity:	729 (29 CPs)
ICF:	3
zIIP:	11
IFL:	28
SAP:	8
Memory:	4864
Unassigned IFLs:	0

Current configuration as of 26 Feb 2022 18:49:18

**Machine summary**  
Type, model, serial:  
8561 - T01 - [REDACTED]  
System name:  
AFPS06SE

**Customer summary**  
Company name:  
[REDACTED]  
Customer number:  
[REDACTED]  
GEO, country:  
Americas - zDutchy of Merwyn

**Ordering options**  
[→ Order permanent upgrade](#)  
[→ Order On/Off CoD record](#)  
[→ Order On/Off CoD test record](#)  
[→ Order On/Off CoD record with prepaid upgrades](#)  
[→ Order On/Off CoD record with spending limits](#)  
[→ Order administrative On/Off CoD test record](#)  
[→ Order Capacity Backup \(CBU\) record](#)  
[→ Order Capacity for Planned Events \(CPE\) record](#)  
[→ Order System Recovery Boost Upgrade record](#)  
[Display upgrade matrix](#)

**About ordering**  
 Authorization to create orders  
 User ID: [brunofarrugia@fr.ibm.com](#) and 3 more  
 Authorization to approve orders

**Ordering options**  
 CIU permanent: Enabled  
 On/Off CoD: Enabled

**To update profile**  
[Upload VPD](#)  
[Upload upgrade billing XML data](#)  
[Disable machine profile...](#)

Figure 8-7 Machine profile window

The number of CPs, ICFs, zIIPs, IFLs, SAPs, memory size, and unassigned IFLs on the current configuration are displayed on the left side of the page.

Resource Link retrieves and stores relevant data that is associated with the processor configuration, such as the number of CPs and installed memory cards. It allows you to select only those upgrade options that are deemed valid by the order process. It also allows upgrades only within the bounds of the currently installed hardware.

## 8.5.2 Retrieval and activation

After an order is placed and processed, the suitable upgrade record is passed to the IBM support system for download.

When the order is available for download, you receive an e-mail that contains an activation number. You can then retrieve the order by using the Perform Model Conversion task from the SE, or through the Single Object Operation to the SE from an HMC.



In the Perform Model Conversion window, select **Permanent upgrades** to start the process, as shown in Figure 8-8.

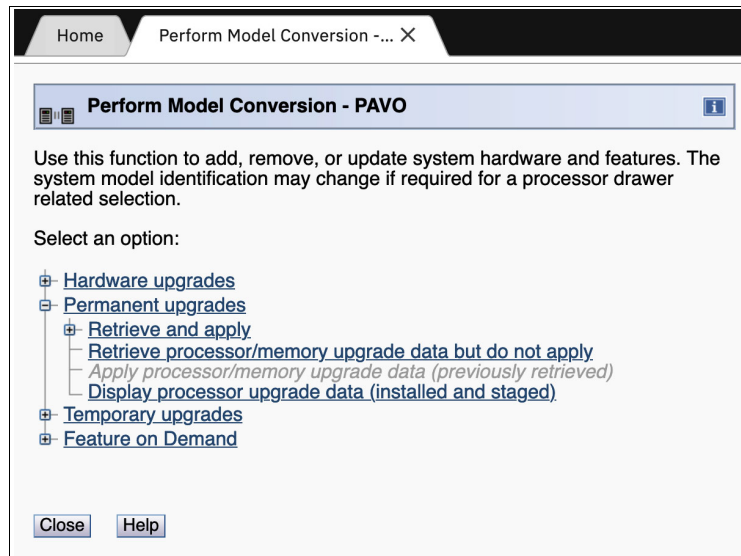


Figure 8-8 IBM z17 Perform Model Conversion window

The window provides several possible options. If you select the **Retrieve and apply** data option, you are prompted to enter the order activation number to start the permanent upgrade, as shown in Figure 8-9.

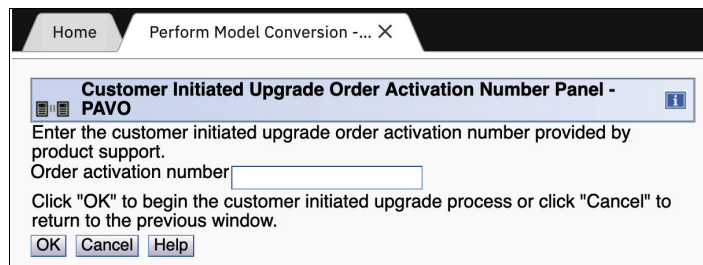


Figure 8-9 Customer Initiated Upgrade Order Activation Number window



## 8.6 On/Off Capacity on Demand

On/Off CoD allows you to temporarily enable additional PUs and unassigned PUs that are available within the current hardware model. You can also use it to change capacity settings for CPs to help meet your peak workload requirements.

**Note:** Details about On/Off Capacity on Demand are taken from the Capacity on Demand User's Guide (SC28-7025-01). Please make sure to download the latest copy on IBM Documentation. Go to <https://www.ibm.com/docs/en/systems-hardware>, select IBM Z or IBM LinuxONE, then select your configuration, and click Library Overview on the navigation bar.

### 8.6.1 Overview

Before implementing any temporary capacity upgrades using On/Off CoD, plan in advance to determine what configurations you might need based on workload projections. This is important because, when properly planned, you only need to order one On/Off CoD record; and this record should be able to handle any possible configurations you want to activate.

When you order an On/Off CoD record, you can prepay for the upgrade or post-pay for the upgrade.

- ▶ When ordering a post-paid On/Off CoD record without spending limits, you select your upgrade configuration. There is no cost incurred when you order or install this type of record. You pay for what you activate during the activation time. You are charged on a 24-hour basis.
- ▶ When ordering a prepaid On/Off CoD record, you can select one or more configurations and identify the duration of each configuration. Then Resource Link calculates the total number of tokens you will need. As resources are used, the tokens are decremented.
- ▶ When ordering a post-paid On/Off CoD record with spending limits, you can select your upgrade configuration and identify your maximum spending limit. Then, Resource Link calculates the number of tokens that will not allow you to exceed that limit. As the resources are used, the tokens are decremented.

For CP engines, a token represents an amount of processing capacity resulting in one MSU of software cost for one day (an MSU day). For specialty engines, a token represents the activation of one engine of that type for one day (a processor day).

Ensure that you enable your system well in advance of needing to place an order.

On/Off CoD allows you to temporarily turn on unowned PUs, unassigned CPs (or unassigned CP capacity), and unassigned IFLs, zIIPs and ICFs available within the current model with the following limitations:

- ▶ Temporary model capacity with CPs and capacity level equal to or greater than the active model capacity, up to 100% of the purchased capacity (active permanent capacity plus unassigned permanent capacity)
- ▶ As many temporary IFLs up to the total of purchased IFLs (permanently active IFLs plus unassigned IFLs)



- ▶ As many additional specialty engines of each type up to the total purchased specialty engines of each type.

**Note:** On/Off CoD requires that the Online CoD Buying feature (FC 9900) is installed on the system that you want to upgrade.

The temporary addition of memory and I/O ports or adapters is *not* supported.

An On/Off CoD upgrade cannot change the system capacity feature. The addition of processor drawers is *not* supported. However, the activation of an On/Off CoD upgrade can increase the model capacity identifier (*4nn*, *5nn*, *6nn*, or *7nn*).

## 8.6.2 On/Off CoD testing

The On/Off Capacity on Demand offering provides two types of tests:

- ▶ an On/Off CoD test
- ▶ an administrative On/Off CoD test

An **On/Off CoD test record** allows you to validate that the retrieve, install, activate, and deactivate On/Off CoD capacity upgrade process performs non disruptively. Authorized users can train to activate an On/Off CoD record, test an LPAR configuration and verify you can change between CP activation levels.

Each On/Off CoD registered machine is entitled to **one** free On/Off CoD test. No IBM charges are assessed for the test, including charges that are associated with temporary hardware capacity, IBM software, and IBM maintenance.

This test can have a maximum duration of 24 hours, which commences upon the activation of any capacity resource that is contained in the On/Off CoD record. Activation levels of capacity can change during the 24-hour test period. The On/Off CoD test deactivates at the end of the 24-hour test period.

An **administrative On/Off CoD test** record allows you to test the Capacity on Demand process for training and API testing without incurring hardware or software charges. An administrative On/Off CoD test does not activate any additional capacity. The capacity level is fixed at 0%.

## 8.6.3 Ordering

The enablement process for each Capacity on Demand offering begins when you order the associated enablement feature code and sign the associated IBM contract document(s), and for online buying capability, completes when you receive an e-mail from Resource Link notifying you that your machine is enabled for ordering upgrade records.

Before you order an On/Off CoD record, consider the following:

- ▶ For a single On/Off CoD record,— The maximum upgrade for CP capacity is 100% of the current purchased capacity. Current purchased capacity (also referred to as the “high water mark” or HWM) includes owned and active permanent capacity and owned and unassigned permanent capacity. Capacity is computed based on processing capacity



gained by adding the engines. It is based off the published LSPR values for the configuration.

- The maximum upgrade allowed for specialty engines is doubling the number of engines.

For example, for an increase in model capacity, if you have a 711, you can activate up to a 725. If you have a 711 with 2 unassigned engines (713 purchased), you would be able to activate up to a 730. For an increase in specialty engines, if you have 6 ICFs, you can add up to 6 more ICFs.

It is recommended that when you order a post-paid On/Off CoD record, you order the maximum capacity and maximum number of specialty engines.

**Note:** Resource Link will not allow you to order beyond the maximum.

Although it is recommended that you order the maximum capacity and number of specialty engines when you order a On/Off CoD record, there may be reasons when you do not want to maximize. For example, you may:

- Not want all engines available for use.
- Want to prevent certain types of upgrades.
- Want to reactivate just the unassigned capacity (order 0%).

**Note:** Even though Resource Link displays the high water mark model when you specify 0% when ordering, a 0% On/Off CoD record on a downgraded machine allows you to activate any supported On/Off CoD upgrades to unassigned model capacities between the active permanent configuration and your high water mark.

By default, an On/Off CoD record is initially available up to 180 days, starting on the date you place your order. After the 180 days, the record will expire unless you “replenish” the record. Replenish allows you to use an existing configuration to either increase your capacity, add specialty engines, or extend the expiration date rather than ordering a new On/Off CoD record.

You can order a replenishment record to manually extend the expiration date or you can enable the automatic renewal function to automatically extend the expiration date of installed records. With the automatic renewal function, a replenishment record is automatically generated 90 days before the record expires. The expiration date on the newly generated replenishment record is set to 180 days from the date the record was automatically generated, which extends the expiration date 90 days from the previous expiration date.

The automatic renewal function is available on post-paid On/Off CoD records. Automatic renewal requires a Remote Support Facility (RSF) connection.

If you apply a permanent upgrade, by default, any active On/Off CoD resources of the same type are converted to permanent upgrades. If all On/Off CoD resources are consumed by the permanent upgrade, the On/Off CoD record remains active with zero resources allocated. Therefore, after the permanent upgrade is complete, you should deactivate (or Undo) the On/Off CoD record.

If your business process requires you to have a purchase order before placing an order, make sure you have the purchase order number ready before placing your order.



On/Off CoD can be ordered as prepaid or postpaid. A prepaid On/Off CoD offering record contains resource descriptions, MSUs, specialty engines, and tokens that describe the total capacity that can be used. For CP capacity, the token contains MSU-days. For specialty engines, the token contains specialty engine-days.

When resources on a prepaid offering are activated, they must have enough capacity tokens to allow the activation for an entire billing window, which is 24 hours. The resources remain active until you deactivate them or until one resource uses all of its capacity tokens. Then, all activated resources from the record are deactivated.

A postpaid On/Off CoD offering record contains resource descriptions, MSUs, specialty engines, and can contain capacity tokens that denote MSU-days and specialty engine-days.

When resources in a postpaid offering record *without* capacity tokens are activated, those resources remain active until they are deactivated, or until the offering record expires. The record normally expires 180 days after its installation.

When resources in a postpaid offering record *with* capacity tokens are activated, those resources must include enough capacity tokens to allow the activation for an entire billing window (24 hours). The resources remain active until they are deactivated until all of the resource tokens are used, or until the record expires. The record usually expires 180 days after its installation. If one capacity token type is used, resources from the entire record are deactivated.

For example, for an IBM z17 with capacity identifier 502 (two CPs), a capacity upgrade through On/Off CoD can be delivered in the following ways:

- ▶ Add CPs of the same capacity setting. With this option, the model capacity identifier can be changed to a 503, which adds another CP to make it a three-way CP. It also can be changed to a 504, which adds two CPs and makes it a four-way CP.
- ▶ Change to a different capacity level of the current CPs and change the model capacity identifier to a 602 or 702. The capacity level of the CPs is increased, but no other CPs are added. The 502 also can be temporarily upgraded to a 603, which increases the capacity level and adds a processor. The capacity setting 439 does not have an upgrade path through On/Off CoD because you cannot reduce the number of CPs and a 539 is more than twice the capacity.

Use the Large System Performance Reference (LSPR) information to evaluate the capacity requirements according to your workload type. For more information about LSPR data for current IBM processors, see [this web page](#).

The On/Off CoD hardware capacity is charged on a 24-hour basis. A grace period is granted at the end of the On/Off CoD day. This grace period allows up to an hour after the 24-hour billing period to change the On/Off CoD configuration for the next 24-hour billing period or deactivate the current On/Off CoD configuration. The times when the capacity is activated and deactivated are maintained in the IBM z17 and returned to the IBM support systems.

If On/Off capacity is active, On/Off capacity can be added without having to return the system to its original capacity. If the capacity is increased multiple times within a 24-hour period, the charges apply to the highest amount of capacity active in that period.

If capacity is added from an active record that contains capacity tokens, the system checks whether the resource has enough capacity to be active for an entire billing window (24 hours). If that criteria is not met, no extra resources are activated from the record.

If necessary, more LPARs can be activated concurrently to use the newly added processor resources.



**Consideration:** On/Off CoD provides a concurrent hardware upgrade that results in more capacity being made available to a system configuration. Extra planning tasks are required for nondisruptive upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 398.

To participate in this offering, you must accept contractual terms for purchasing capacity through Resource Link, establish a profile, and install an On/Off CoD enablement feature on the system. Later, you can concurrently install temporary capacity up to the limits in On/Off CoD and use it for up to 180 days.

Monitoring occurs through the system call-home facility. An invoice is generated if the capacity is enabled during the calendar month. You are billed for the use of temporary capacity until the system is returned to the original configuration. Remove the enablement code if the On/Off CoD support is no longer needed.

Resource Link provides the interface to order a dynamic upgrade for a specific system. You can create, cancel, and view the order. Configuration rules are enforced, and only valid configurations are generated based on the configuration of the individual system. After you complete the prerequisites, orders for the On/Off CoD can be placed. The order process uses the CIU facility on Resource Link.

Memory and channels are not supported on On/Off CoD.

An example of an On/Off CoD order on the Resource Link web page is shown in Figure 8-10.

IBM Systems > z Systems > Resource Link > Customer Initiated Upgrade > Machine profiles > Machine 2964 - 8DA87 >

## Order On/Off CoD record

Step 2 of 2: Review and submit your order

Review the range of upgrades you selected on the previous page. The On/Off CoD record you are about to order will be configured to support activating any configurations within the range.

(\*) indicates accepting the [Terms and Conditions of this order](#) is required to submit it. Mark the check box to indicate acceptance.

Expiration date:	11 Oct 2015	Renew automatically:	Yes
<b>Model capacity:</b>	0% more model capacity	<b>Daily hardware prices</b>	<b>Daily maintenance prices (estimated)<sup>1</sup></b>
<b>ICF:</b>	2 more ICF engines	\$0.00	\$12.00
<b>zIIP:</b>	2 more zIIP engines	\$0.00	\$12.00
<b>IFL:</b>	0 more IFL engines		
<b>SAP:</b>	0 more SAP engines		

**Machine summary**

Type: 2964 N63  
Model: 735  
Serial number: 8DA87

**Current configuration**

Model capacity (CPs): 35  
ICF: 8  
zIIP: 12  
IFL: 8  
SAP: 12  
Available engines: 0

**Supported upgrades**

[Show upgrades](#)  
[Show upgrade prices](#)

**Description:** +0% model capacity, +2 ICF, +2 zIIP, +0 IFL, +0 SAP

**Notes:**

1. Reflects current established prices for the selected machine. Prices are subject to change; the actual prices in effect at the time of use will apply.
2. Daily prices for ICF, zIIP, IFL, and SAP upgrades are **per engine**.
3. The IFL upgrade daily hardware price includes per IFL for the management enablement level in effect for this machine.

Figure 8-10 On/Off CoD order example

The example order that is shown in Figure 8-10 is an On/Off CoD order for 0% more CP capacity (system is at capacity level 7), and for two more ICFs and two more zIIPs. The maximum number of CPs, ICFs, zIIPs, and IFLs is limited by the current number of available unused PUs of the installed processor drawers. The maximum number of SAPs is determined by the model number and the number of available PUs on the already installed processor drawers.



To finalize the order, you must accept Terms and Conditions for the order, as shown in Figure 8-11.

Terms of Order

You have requested an On/Off Capacity on Demand, or Temporary Capacity upgrade. Your enterprise has previously accepted the Temporary Capacity terms, restated here. In the event there is a conflict between the terms shown on this website and the terms specified in your contract with IBM, the terms of such contract prevail:

1) upon download and installation of this Temporary Capacity Upgrade, IBM grants you only a temporary license to use the LIC enabling such Temporary Capacity Upgrade. You may use such Temporary Capacity Upgrade only on the TC Eligible Machine for which such LIC is provided, and only to the extent of the authorization identified via the CIU Facility.

☒ I accept the Terms and Conditions of this order\*

Submit

Figure 8-11 CIU order Terms and Conditions

## 8.6.4 Activation and deactivation

Before an ordered On/Off CoD record can be activated, it has to be retrieved and installed. To retrieve a temporary upgrade record, log onto the HMC in system programmer mode, find **Perform Model Conversion** in the **Configuration** task list and click on **Temporary upgrades** and **Retrieve**. The record is now placed in a staging area so it can be installed at a later time.

To install the record, go to the Staged Records tab, select the record and click Install. The installed Records page opens showing the newly installed record. The On/Off CoD record is now ready to be activated.

A temporary upgrade record can be activated using any of the following methods:

- ▶ Manually, using the Support Element.
- ▶ Setting up scheduled operations
- ▶ Using APIs
- ▶ z/OS Capacity Provisioning

Please refer to the Capacity on Demand User's Guide for details about these activation methods.

**Deactivating** is the process of removing temporary processors or decreasing temporary model capacity. Deactivation can be performed manually or automatically upon expiration.

When you are finished using all or part of a capacity upgrade, you can take action to remove processors or decrease model capacity using the Support Element. You can only remove activated resources for the specific offering. You cannot remove dedicated engines or the last of the engine type.



If you do not manually deactivate the added capacity, the activated resources are automatically deactivated at expiration time (including any grace period). You will receive daily warning messages (hardware messages) starting five days in advance of the expiration. Once a temporary record enters the grace period, the only customer option is to deactivate all resources from this record. You cannot change the activation level by increasing or decreasing partial resources. If you attempt to partially increase or decrease resources, you will receive an error indicating the temporary record has expired.

The Capacity on Demand User's Guide explains in depth how to deactivate temporary records and any considerations before deactivating.

### 8.6.5 Discontinuing and removing Capacity on Demand features

Certain events, for example selling your machine to another party or returning your machine to IBM or another leasing company require that you discontinue your use of one or more of the Capacity on Demand (CoD) features on a machine and remove those features from the machine. This may require you to deactivate CoD records or delete staged or installed CoD records.

Please refer to the Capacity on Demand Users's Guide to which actions are needed for your event.

When you no longer need your CIU machine profile you may also disable the profile on Resource Link. Disabling a machine profile does not delete it. A disabled machine profile remains listed on the CIU All machine profiles page on Resource Link so you can review its order history, billing history, or other information if necessary.

## 8.7 z/OS Capacity Provisioning

This section describes how z/OS Capacity Provisioning can help you manage the addition of capacity to a server to handle workload peaks.

z/OS Capacity Provisioning is delivered as part of the z/OS MVS Base Control Program (BCP).

Capacity Provisioning includes the following components:

- ▶ Capacity Provisioning Manager (Provisioning Manager)
- ▶ Capacity Provisioning Management Console (available in the IBM z/OS Management Facility)
- ▶ Sample data sets and files

The Provisioning Manager monitors the workload on a set of z/OS systems and organizes the provisioning of extra capacity to these systems when required. You define the systems to be observed in a domain configuration file.



The details of extra capacity and the rules for its provisioning are stored in a policy file. These two files are created and maintained through the Capacity Provisioning Management Console (CPMC).

The operational flow of Capacity Provisioning is shown in Figure 8-12.

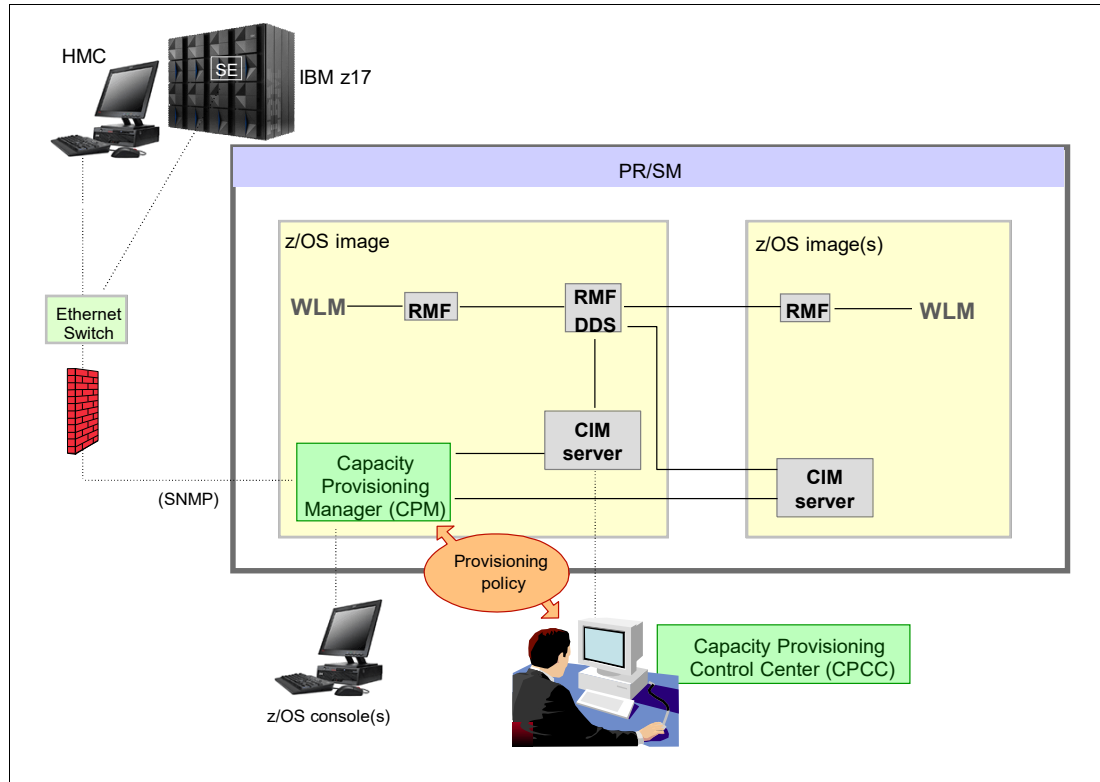


Figure 8-12 The capacity provisioning process and infrastructure

The z/OS WLM manages the workload by goals and business importance on each z/OS system. WLM metrics are available through existing interfaces, and are reported through IBM Resource Measurement Facility (RMF) Monitor III, with one RMF gatherer for each z/OS system.

Sysplex-wide data aggregation and propagation occur in the RMF Distributed Data Server (DDS). The RMF Common Information Model (CIM) providers and associated CIM models publish the RMF Monitor III data.

CPM retrieves critical metrics from one or more z/OS systems' CIM structures and protocols. CPM communicates to local and remote SEs and HMCs by using the Simple Network Management Protocol (SNMP).

CPM can see the resources in the individual offering records and the capacity tokens. When CPM activates resources, a check is run to determine whether enough capacity tokens remain for the specified resource to be activated for at least 24 hours. If insufficient tokens remain, no resource from the On/Off CoD record is activated.



If a capacity token is used during an activation that is driven by the CPM, the corresponding On/Off CoD record is deactivated prematurely by the system. This process occurs even if the CPM activates this record, or parts of it. However, you receive warning messages five days before a capacity token is fully used.

The five days are based on the assumption that the consumption is constant for the five days. You must put operational procedures in place to handle these situations. You can deactivate the record manually, allow it to occur automatically, or replenish the specified capacity token by using the Resource Link application.

The Capacity Provisioning Management Console (CPMC) is a console that administrators use to work with provisioning policies and domain configurations and to monitor the status of a Provisioning Manager. The management console is implemented by the Capacity Provisioning task in the IBM z/OS Management Facility (z/OSMF). z/OSMF provides a framework for managing various aspects of a z/OS system through a web browser interface.

## Capacity Provisioning Domain

The provisioning infrastructure is managed by the CPM through the Capacity Provisioning Domain (CPD), which is controlled by the Capacity Provisioning Policy (CPP). The CPD is shown in Figure 8-13.

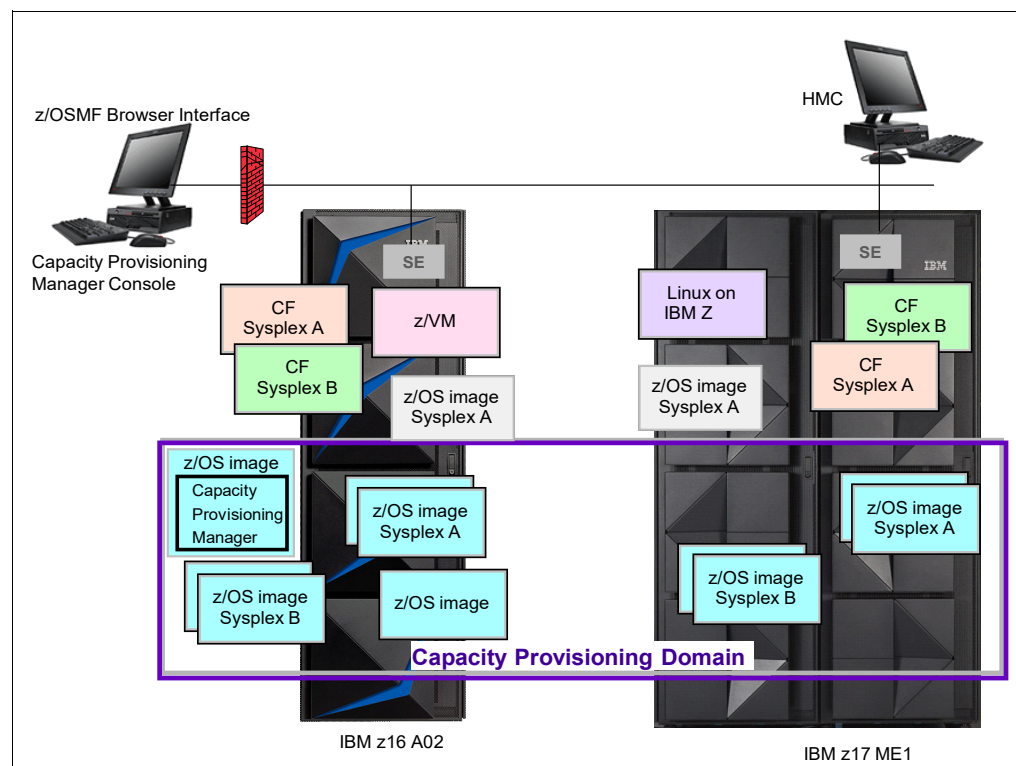


Figure 8-13 Capacity Provisioning Domain

The CPD configuration defines the CPCs and z/OS systems that are controlled by an instance of the CPM. One or more CPCs, sysplexes, and z/OS systems can be defined into a domain. Although sysplexes and CPCs do not have to be contained in a domain, they must not belong to more than one domain.

Each domain has one active capacity provisioning policy.

CPM operates in the following modes, which allows four different levels of automation:



► Manual

Use this command-driven mode when no CPM policy is active.

► Analysis

In analysis mode, CPM processes capacity-provisioning policies and informs the operator when a provisioning or de-provisioning action is required according to policy criteria.

Also, the operator determines whether to ignore the information or to manually upgrade or downgrade the system by using the HMC, SE, or available CPM commands.

► Confirmation

In this mode, CPM processes capacity provisioning policies and interrogates the installed temporary offering records. Every action that is proposed by the CPM must be confirmed by the operator.

► Autonomic

This mode is similar to the confirmation mode, but no operator confirmation is required.

Several reports are available in all modes that contain information about the workload, provisioning status, and the rationale for provisioning guidelines. User interfaces are provided through the z/OS console and the CPMC application.

The provisioning policy defines the circumstances under which more capacity can be provisioned (when, which, and how). The criteria features the following elements:

- A time condition is when provisioning is allowed:
  - Start time indicates when provisioning can begin.
  - Deadline indicates that provisioning of more capacity is no longer allowed.
  - End time indicates that deactivation of capacity must begin.
- A workload condition is which work qualifies for provisioning. It can have the following parameters:
  - The z/OS systems that can run eligible work.
  - The importance filter indicates eligible service class periods, which are identified by WLM importance.
  - Performance Index (PI) criteria:
    - Activation threshold: PI of service class periods must exceed the activation threshold for a specified duration before the work is considered to be suffering.
    - Deactivation threshold: PI of service class periods must fall below the deactivation threshold for a specified duration before the work is considered to no longer be suffering.
  - Included service classes are eligible service class periods.
  - Excluded service classes are service class periods that must not be considered.

**Tip:** If no workload condition is specified, the full capacity that is described in the policy is activated and deactivated at the start and end times that are specified in the policy.

- Provisioning scope is how much more capacity can be activated and is expressed in MSUs.

The number of zLIPs must be one specification per CPC that is part of the CPD and are specified in MSUs.



The maximum provisioning scope is the maximum extra capacity that can be activated for all the rules in the CPD.

In the specified time interval, the provisioning rule is that up to the defined extra capacity can be activated if the specified workload is behind its objective.

The rules and conditions are named and stored in the Capacity Provisioning Policy.

For more information about z/OS Capacity Provisioning functions, see *z/OS MVS Capacity Provisioning User's Guide*, SC33-8299.

## Planning considerations for using automatic provisioning

Although only one On/Off CoD offering can be active at any one time, several On/Off CoD offerings can be present on the system. Changing from one offering to another requires stopping the active offering before the inactive one can be activated. This operation decreases the current capacity during the change.

The provisioning management routines can interrogate the installed offerings, their content, and the status of the content of the offering. To avoid the decrease in capacity, create only one On/Off CoD offering on the system by specifying the maximum allowable capacity. Then, when an activation is needed, the CPM can activate a subset of the contents of the offering sufficient to satisfy the demand. If more capacity is needed later, the Provisioning Manager can activate more capacity up to the maximum allowed increase.

Multiple offering records can be pre-staged on the SE HDD. Changing the content of the offerings (if necessary) also is possible.

**Remember:** CPM controls capacity tokens for the On/Off CoD records. In a situation where a capacity token is used, the system deactivates the corresponding offering record. Therefore, you must prepare routines for catching the warning messages about capacity tokens being used, and have administrative procedures in place for such a situation.

The messages from the system begin five days before a capacity token is fully used. To avoid capacity records being deactivated in this situation, replenish the necessary capacity tokens before they are used.

The CPM operates based on Workload Manager (WLM) indications, and the construct that is used is the Performance Index (PI) of a service class period. It is important to select service class periods that are suitable for the business application that needs more capacity.

For example, the application in question might be running through several service class periods, where the first period is the important one. The application might be defined as importance level 2 or 3, but might depend on other work that is running with importance level 1. Therefore, it is important to consider which workloads to control and which service class periods to specify.



## 8.8 System Recovery Boost

System Recovery Boost (SRB) is a feature introduced with the IBM z15, and enhanced with the IBM z16, that provides capabilities to reduce the time it takes to shut down and restart (IPL) a system, by providing additional processor capacity and throughput for the boosted image.

In addition to boosting shutdown and IPL, System Recovery Boost can provide short-term acceleration for specific system and sysplex recovery and diagnostic capture events in z/OS, including, with the IBM z16, SVC dumps, HyperSwap configuration load, and middleware region startup.

IBM z17 provides no new System Recovery Boost enhancements, meaning, all IBM z17 System Recovery Boost enhancements available on z16, are the same enhancements which are available on IBM z17, with one exception: the System Recovery Boost Upgrade is sunset on IBM z17.

There are three classes of boost: IPL (startup) boost, recovery process boost, and shutdown boost. Each class has different capabilities.

1. IPL boost is enabled by default and delivers extra processor capacity after an IPL to get you back up and running faster.
2. Recovery process boost provides increased short-duration processor capacity for the acceleration of some sysplex recovery situations. Starting with the IBM z16, additional recovery events can be boosted. This includes Standalone Dumps, CF Structure Recovery, SVC Dump and other recovery events
3. Shutdown boost enables a faster shutdown by delivering extra processor capacity upon indication that a shutdown is in progress.

The increased capacity can be provided in one or more of the following ways:

### **Speed boost:**

Speed boost is a capability of SRB that improves the recovery time of exploiting operating systems when running on a subcapacity CPC. If you are running on a subcapacity CPC, then while System Recovery Boost is active, z/OS will request that the CPC firmware increase the CP speed of the image to full capacity model speed for the duration of the boost. After the boost ends, the image will return to the subcapacity model speed.

Speed boost applies only to the image being boosted; all other images not being boosted

### **zIIP boost**

If your system has z Integrated Information Processors (zIIPs), then zIIP boost can improve z/OS recovery time, assuming zIIP capacity is available to the image.

z/OS is the only operating system that can exploit the zIIP boost capability, as it's the only OS that can natively exploit zIIPs. While zIIP boost is active, z/OS will make most non-zIIP eligible work zIIP eligible, thus allowing most work to run on zIIPs if there isn't sufficient CP capacity available. This provides additional capacity and parallelism to accelerate processing during the boost periods. IBM refers to this as blurring the CPs and zIIPs together.



On z15 and z16 IBM also offered a third way to activate additional capacity: the priced feature System Recovery Boost Upgrade. However, the System Recovery Boost Upgrade is sunset on IBM z17:

- ▶ New IBM z17 machines will not be able to order the SRB Upgrade Record
- ▶ Current IBM z15 or z16 machines with the SRB Upgrade Record configured will not be able to carry forward the record when upgrading to IBM z17
- ▶ The SRB Upgrade Record will not be altered on IBM z15 or z16. Current z15 and z16 clients will not be impacted. Currently installed records may be extended, or new records added
- ▶ Base SRB functionality will not be impacted

Please refer to the Whitepaper [“System Recovery Boost for the IBM z15 and z16”](#) by Kevin McKenzie for more details and setup.

**Note:** System Recovery Boost Upgrade was introduced with IBM z15 and carry forward with IBM z16.  
System Recovery Boost Upgrade *is not supported* with the IBM z17.

## 8.9 Capacity for Planned Event (CPE)

The Capacity Planned Event feature (6833) can no longer be ordered for a new IBM z17; however, if installed on the base system, the record is brought forward during an upgrade into IBM z17 (by way of the support element Save/Restore process). Also, the CPE record cannot be replenished through e-configuration or IBM Resource Link.

## 8.10 Flexible Capacity for Cyber Resiliency

IBM Z Flexible Capacity for Cyber Resiliency is a Capacity on Demand offering available on IBM z16 and z17 servers. It enables you to shift capacity between participating IBM z16 and z17 machines and use the target configuration for up to one year.

IBM Z Flexible Capacity for Cyber Resiliency supports a broad range of use case scenarios: DR and DR Testing, compliance, facility maintenance, and pro-active avoidance.

The offering has to be differentiated between two editions:

- ▶ IBM Z Flexible Capacity for Cyber Resiliency Enterprise Edition (EE) and
- ▶ IBM Z Flexible Capacity for Cyber Resiliency Limited Terms Edition (LiTe)

Flex Cap EE	Flex Cap LiTe
12 annual capacity changes per serial number	4 annual capacity changes per serial number



Flex Cap EE	Flex Cap LiTe
Stay-out period for a maximum of 12 months	Stay-out period for a maximum of 1 month (real DR: 90 days)
Inter-Site and Intra-site	Inter-site only (no Intra-site)
License: Term (maximum of 5 years) or perpetual	License: Term (maximum of 5 years) or perpetual

Stay-out period refers to the time a swapped capacity can stay on the backup server.

Inter-site means capacity can only be swapped between servers in different datacenters - Intra-site refers to servers in the same datacenter.

Capacity shifts can be done under full customer control without IBM intervention and can be fully automated by using IBM GDPS automation tools.

Flexible Capacity for Cyber Resiliency supports a broad set of scenarios and can be combined with other IBM On-Demand offerings.

- ▶ Flexible Capacity Authorization (#9933)
- ▶ Flexible Capacity Record (#0376)
- ▶ Billing feature codes (FC 0317 - 0322, and FC 0378 - 0386)

Flexible Capacity for Cyber Resiliency can be ordered by contacting your IBM hardware sales representative. The offering requires that an order is placed against each serial number (SN) that is involved in capacity transfer with one record per SN.

Installation and setup: The new Flexible Capacity Record is installed and set up on each participating IBM Z server.

Consider the following points:

- ▶ On the IBM z17 or IBM z16 source system, the permanent capacity is unassigned to the base level.
- ▶ The new Flexible Capacity Record is installed and activated on the IBM z17 source system to restore capacity back to the purchased level.
- ▶ On the IBM z16 or IBM z17 target systems, the new Flexible Capacity Record enables clients to bring the capacity up to the level of the production system when activated. The Flexible Capacity Record remains inactive until capacity is shifted from the base system to the target systems.
- ▶ After deactivating the Flexible Capacity Record on the base system, the capacity active through Flexible Capacity Transfer records on the target systems must not exceed the capacity active on the base system before the swap.

Please note that the above setup description is only one example: Flex Cap impresses with its flexibility and many, also complex configurations, including multiple z System machines in several data centers are feasible.

For more information and implementation examples, see Appendix C, “Tailored Fit Pricing and IBM Z Flexible Capacity for Cyber Resiliency” on page 531.

Refer to the Redpaper “*IBM Z Flexible Capacity for Cyber Resiliency*,” [REDP-5702](#).



## 8.11 Capacity Backup (CBU)

CBU provides reserved emergency backup processor capacity for unplanned situations in which capacity is lost in another part of your enterprise. It allows you to recover by adding the reserved capacity on a designated IBM Z server.

CBU is the quick, temporary activation of PUs:

- ▶ For up to 90 contiguous days, for a loss of processing capacity as a result of an emergency or disaster recovery situation.
- ▶ For 10 days, for testing your disaster recovery procedures or running the production workload. This option requires that IBM Z workload capacity that is equivalent to the CBU upgrade capacity is shut down or otherwise made unusable during the CBU test.

**Important:** CBU is for disaster and recovery purposes only. It *cannot* be used for peak workload management or for a planned event.

### 8.11.1 Ordering

The CBU process allows for CBU to activate CPs, ICFs, zIIPs, IFLs, and SAPs. To use the CBU process, a CBU enablement feature (FC 9910) must be ordered and installed. You must order the quantity and type of PU that you require by using the following feature codes:

- ▶ FC 6805: More CBU test activations
- ▶ FC 6817: Total CBU years ordered
- ▶ FC 6818: CBU records that are ordered
- ▶ FC 6820: Single CBU CP-year
- ▶ FC 6821: 25 CBU CP-year
- ▶ FC 6822: Single CBU IFL-year
- ▶ FC 6823: 25 CBU IFL-year
- ▶ FC 6824: Single CBU ICF-year
- ▶ FC 6825: 25 CBU ICF-year
- ▶ FC 6828: Single CBU zIIP-year
- ▶ FC 6829: 25 CBU zIIP-year
- ▶ FC 6830: Single CBU SAP-year
- ▶ FC 6831: 25 CBU SAP-year
- ▶ FC 6832: CBU replenishment

The CBU entitlement record (FC 6818) contains an expiration date that is established at the time of the order. This date depends on the quantity of CBU years (FC 6817). You can extend your CBU entitlements through the purchase of more CBU years.

The number of FC 6817 per instance of FC 6818 remains limited to five. Fractional years are rounded up to the nearest whole integer when calculating this limit.

If two years and eight months exist before the expiration date at the time of the order, the expiration date can be extended by no more than two years. One test activation is provided for each CBU year that is added to the CBU entitlement record.

FC 6805 allows for ordering more tests in increments of one. The maximum number of tests that is allowed is 15 for each FC 6818.

The processors that can be activated by CBU come from the available unassigned PUs on any installed processor drawer. A maximum of 208 CBU CP features can be ordered on a



z17. The number of features that can be *activated* is limited by the number of unused PUs on the system; for example:

- ▶ An IBM z17 Max43 with Capacity Model Identifier 401 can activate up to 43 CBU features. These CBU features can be used to change the capacity setting of the CPs, and to activate unused PUs.
- ▶ An IBM z17 Max90 with 15 CPs, 4 IFLs, and 1 ICF has 70 unused PUs available. It can *activate* up to 70 CBU features.

The ordering system allows for over-configuration in the order. You can order up to 208 CBU features, regardless of the current configuration. However, at activation, only the capacity that is installed can be activated. At activation, you can decide to activate only a subset of the CBU features that are ordered for the system.

Sub-capacity makes a difference in the way that the CBU features are completed. On the full-capacity models, the CBU features indicate the amount of extra capacity that is needed. If the amount of necessary CBU capacity is equal to four CPs, the CBU configuration is four CBU CPs.

The sub-capacity models feature multiple capacity settings of 4xx, 5yy, or 6yy. The standard models use the capacity setting 7nn. To change the capacity setting, the number of CBU CPs must be equal to or greater than the number of CPs in the base configuration.

For example, if the base configuration is a two-way 402, two CBU feature codes are required to provide a CBU configuration of a four-way of the same capacity setting (404). If the desired CBU target configuration is a four way 504, going from model capacity identifier 402 to a 504 requires four CBU feature codes (2 CBU features to change from 402 to 502 plus 2 CBU features to go from 502 to 504).

If the capacity setting of the CPs is changed, more CBU features are required, not more physical PUs. Therefore, your CBU contract requires more CBU features when the capacity setting of the CPs is changed.

CBU can add CPs through LICCC only, and the IBM z17 ME1 must have the correct number of installed processor drawers to allow the required upgrade. CBU can change the model capacity identifier to a *higher* value than the base setting (4xx, 5yy, or 6yy), but the CBU feature cannot *decrease* the capacity setting.

A CBU contract must be in place before the special code that enables this capability can be installed on the system. CBU features can be added to an IBM z17 nondisruptively. For each system enabled for CBU, the authorization to use CBU is available for 1 - 5 years.

The alternative configuration is activated *temporarily*, and provides more capacity than the system's original, *permanent* configuration. At activation time, determine the capacity that you require for that situation. You can decide to activate only a subset of the capacity that is specified in the CBU contract.

The base system configuration must have sufficient memory and channels to accommodate the potential requirements of the large CBU target system. Ensure that all required functions and resources are available on the backup systems. These functions include CF LEVELs for coupling facility partitions, memory, and cryptographic functions, and connectivity capabilities.



When the emergency is over (or the CBU test is complete), the system must be returned to its original configuration. The CBU features can be deactivated at any time before the expiration date. Failure to deactivate the CBU feature before the expiration date can cause the system to downgrade resources gracefully to the original configuration. The system does not deactivate dedicated engines, or the last of in-use shared engines.

**Planning:** CBU for processors provides a concurrent upgrade. This upgrade can result in more enabled processors, changed capacity settings that are available to a system configuration, or both. You can activate a subset of the CBU features that are ordered for the system. Therefore, more planning and tasks are required for *nondisruptive* logical upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 398.

For more information, see the *Capacity on Demand User's Guide*, SC28-7058-00.

### 8.11.2 CBU activation and deactivation

The activation and deactivation of the CBU function is your responsibility and does not require the onsite presence of IBM SSRs. The CBU function is activated or deactivated concurrently from the HMC by using the API. On the SE, CBU is activated by using the Perform Model Conversion task or through the API. The API enables task automation.

#### CBU activation

CBU is activated from:

- ▶ The SE by using the HMC and SSO to the SE
- ▶ By using the Perform Model Conversion task
- ▶ Through automation by using the API on the SE or the HMC

During a real disaster, use the Activate CBU option to activate the 90-day period.

#### Image upgrades

After CBU activation, the IBM z17 can have more capacity, more active PUs, or both. The extra resources go into the resource pools and are available to the LPARs. If the LPARs must increase their share of the resources, the LPAR weight can be changed or the number of logical processors can be concurrently increased by configuring reserved processors online. The operating system must concurrently configure more processors online. If necessary, more LPARs can be created to use the newly added capacity.

#### CBU deactivation

To deactivate the CBU, the extra resources must be released from the LPARs by the operating systems. In some cases, this process involves varying the resources offline. In other cases, it can mean shutting down operating systems or deactivating LPARs. After the resources are released, the same facility on the HMC/SE is used to turn off CBU. To deactivate CBU, select the **Undo temporary upgrade** option from the Perform Model Conversion task on the SE.



## CBU testing

Test CBUs are provided as part of the CBU contract. CBU is activated from the SE by using the Perform Model Conversion task. Select the test option to start a 10-day test period. A standard contract allows one test per CBU year. However, you can order more tests in increments of one up to a maximum of 15 for each CBU order.

**Tip:** The CBU test activation is done the same way as the real activation; that is, by using the same SE Perform a Model Conversion window and selecting the **Temporary upgrades** option. The HMC windows were changed to avoid accidental real CBU activations by setting the test activation as the default option.

The test CBU must be deactivated in the same way as the regular CBU. Failure to deactivate the CBU feature before the expiration date can cause the system to degrade gracefully back to its original configuration. The system does not deactivate dedicated engines or the last in-use shared engine.

## CBU example

An example of a CBU operation is shown in Figure 8-14. The permanent configuration is a 504, and a record contains seven CP CBU features. During an activation, multiple target configurations are available. With 7 CP CBU features, you can add up to 7CPs within the same MCI, which allows the activation of a 506, 507, through to a 511 (the blue path).

Alternatively, 4 CP CBU features can be used to change the MCI (in the example from a 504 to a 704) and then add the remaining 3 CP CBU features to upgrade to a 707 (the red path).

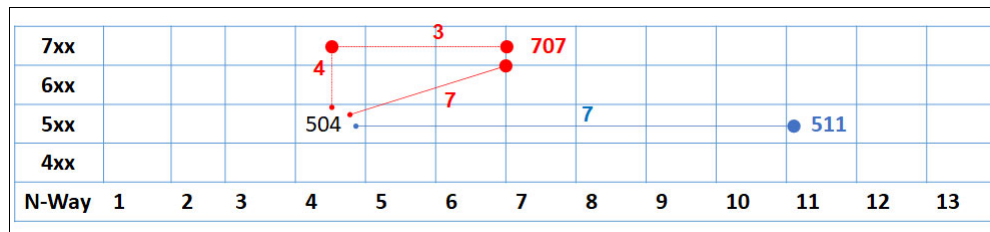


Figure 8-14 CBU example

## 8.11.3 Automatic CBU enablement for GDPS

The IBM Geographically Dispersed Parallel Sysplex (GDPS) enables automatic management of the PUs that are provided by the CBU feature during a system or site failure. Upon detection of a site failure or planned disaster test, GDPS concurrently adds CPs to the systems in the take-over site to restore processing power for mission-critical production workloads. GDPS automation runs the following tasks:

- ▶ The analysis that is required to determine the scope of the failure. This process minimizes operator intervention and the potential for errors.
- ▶ Automates authentication and activation of the reserved CPs.
- ▶ Automatically restarts the critical applications after reserved CP activation.
- ▶ Reduces the outage time to restart critical workloads from several hours to minutes.

The GDPS service is for z/OS only, or for z/OS in combination with Linux on Z.



## 8.12 Planning for nondisruptive upgrades

Continuous availability is an important requirement for customers, and planned outages are no longer acceptable. Although Parallel Sysplex clustering technology is the best continuous availability solution for z/OS environments, nondisruptive upgrades within a single system can avoid system outages and cover non-z/OS operating systems.

IBM z17 allows *concurrent* upgrades, which means that dynamically adding capacity to the system is possible. If the operating system images that run on the upgraded system do not require disruptive tasks to use the new capacity, the upgrade is also *nondisruptive*. This process avoids power-on resets (POR), LPAR deactivation, and IPLs.

If the concurrent upgrade is intended to satisfy an *image* upgrade to an LPAR, the operating system that is running in this partition must concurrently configure more capacity online. z/OS operating systems include this capability. z/VM can concurrently configure new processors and I/O devices online, and memory can be dynamically added to z/VM partitions.

If the concurrent upgrade is intended to satisfy the need for more operating system images, more LPARs can be created *concurrently* on the IBM z17 system. These LPARs include all resources that are needed. These extra LPARs can be activated concurrently.

These enhanced configuration options are available through the HSA, which is an IBM reserved area in system memory.

In general, Linux operating systems cannot add more resources concurrently. However, Linux and other types of virtual machines that run under z/VM can benefit from the z/VM capability to nondisruptively configure more resources online (processors and I/O).

With z/VM, Linux guests can manipulate their logical processors by using the Linux CPU hotplug daemon. The daemon can start and stop logical processors that are based on the Linux *load average* value. The daemon is available in Linux SLES 10 SP2 and later, and in Red Hat Enterprise Linux (RHEL) V5R4 and up.

### 8.12.1 Components

The following components can be added, depending on the considerations as described in this section:

- ▶ PUs
- ▶ Memory
- ▶ I/O
- ▶ Cryptographic adapters
- ▶ Special features

#### PUs

CPs, ICFs, zIIPs, and IFLs, can be added concurrently to an IBM z17 if unassigned PUs are available on any installed processor drawer.

zIIP Processors require at least one PU characterized as CP, and the total number of zIIPs is one minus the IBM z17 model. For instance, an IBM z17 Max90 can have up to 89 zIIPs, considering one required CP. The IBM z17 allows the concurrent addition of a second and third processor drawer if the CPC reserve features are installed.

If necessary, more LPARs can be created concurrently to use the newly added processors.



The Coupling Facility Control Code (CFCC) also can configure more processors online to coupling facility LPARs by using the CFCC image operations window.

## Memory

Memory can be added concurrently up to the physical installed memory limit. More processor drawers can be installed concurrently, which allows further memory upgrades by LICCC, and enables memory capacity on the new processor drawers.

By using the previously defined reserved memory, z/OS operating system images, and z/VM partitions, you can dynamically configure more memory online. This process allows nondisruptive memory upgrades. Linux on Z supports Dynamic Storage Reconfiguration.

## I/O

I/O features can be added concurrently if all the required infrastructure (I/O slots and PCIe Fan-outs) is present in the configuration. PCIe+ I/O drawers can be added concurrently without planning if free space is available in one of the frames and the configuration permits.

Dynamic I/O configuration changes are supported by specific operating systems (z/OS and z/VM), which allows for nondisruptive I/O upgrades. Dynamic I/O reconfiguration on a stand-alone coupling facility system also is possible by using the Dynamic I/O activation for stand-alone CF CPCs features.

## Cryptographic adapters

Crypto Express8S features can be added concurrently if all the required infrastructure is in the configuration.

## Special features

Special features, such as zHyperlink, Coupling Express3 LR, and RoCE features can be added concurrently if all infrastructure is available in the configuration.

## 8.12.2 Concurrent upgrade considerations

By using an MES upgrade or an Capacity on Demand offering, an IBM Z server can be upgraded concurrently from one model to another (temporarily or permanently).

Enabling and using the extra processor capacity is not apparent to most applications. However, specific programs depend on processor model-related information, such as ISV products. Consider the effect on the software that is running on an IBM z17 when you perform any of these configuration upgrades.

## Processor identification

The following instructions are used to obtain processor information:

- ▶ Store System Information (STSI) instruction

The STSI instruction can be used to obtain information about the current execution environment and any processing level that is below the current environment. It can be used to obtain processor model and model capacity identifier information from the basic machine configuration form of the system information block (SYSIB). It supports concurrent upgrades and is the recommended way to request processor information.

- ▶ Store CPU ID (STIDP) instruction

STIDP returns information that identifies the execution environment, system serial number, and machine type.



**Note:** To ensure unique identification of the configuration of the issuing CPU, use the STSI instruction specifying basic machine configuration (SYSIB 1.1.1).

## Store System Information instruction

The format of the basic machine configuration SYSIB that is returned by the STSI instruction is shown in Figure 8-15. The STSI instruction returns the model capacity identifier for the permanent configuration and the model capacity identifier for any temporary capacity. This data is key to the functioning of CoD offerings.

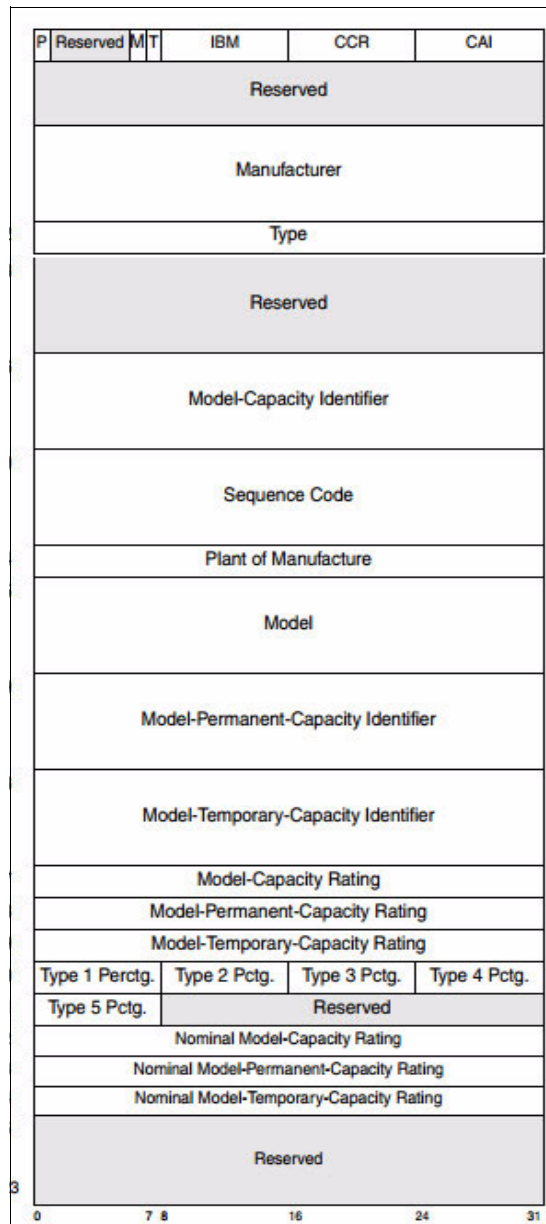


Figure 8-15 Format of system-information block (SYSIB)

The model capacity identifier contains the base capacity, On/Off CoD, and CBU. The Model Permanent Capacity Identifier and the Model Permanent Capacity Rating contain the base capacity of the system. The Model Temporary Capacity Identifier and Model Temporary Capacity Rating contain the base capacity and On/Off CoD.



For more information about the STSI instruction, see *z/Architecture Principles of Operation*, SA22-7832.

### Store CPU ID instruction

The Store CPU ID (STIDP) instruction returns information about the processor type, serial number, and LPAR identifier, as shown in Figure 8-16.

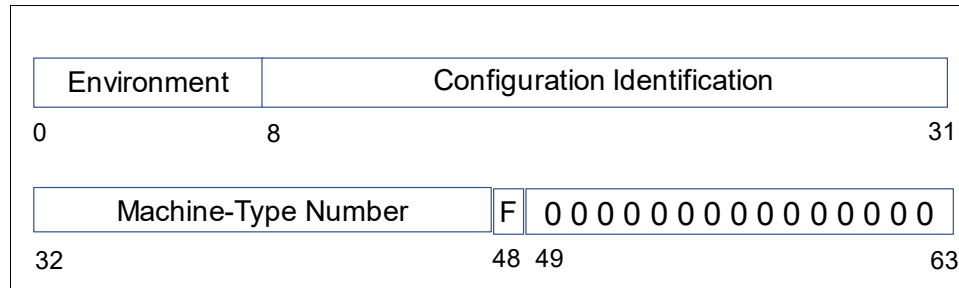


Figure 8-16 STIDP Information

Consider the following points:

- ▶ Bits 0 - 7:
  - For a program that is run by an IBM machine in a level-1 configuration (basic machine mode), or for a program being run by a level-2 configuration (in a logical partition), the environment field contains 00 hex.
  - For a program that is run natively by the System z® Personal-Development Tool, the environment field contains C1 hex or D3 hex.
  - For a program that is run by a level-3 configuration (a virtual machine, such as a z/VM guest), the environment field contains FF hex.
- ▶ Bit positions 8 - 31
 

Contains six hexadecimal digits. The right-most of these digits can represent the machine's serial number.
- ▶ Bit positions 32 - 47
 

Contains an unsigned packed-decimal number that identifies the machine type of the CPU.
- ▶ Bit position 48
 

Specifies the format of the first two hexadecimal digits of the configuration-identification field.
- ▶ Bit positions 49 - 63 are reserved and stored as zeros.

For more information about the STIDP instruction, see *z/Architecture Principles of Operation*, SA22-7832.

### Planning for nondisruptive upgrades

Online permanent upgrades and other Capacity on Demand offerings can be used to upgrade an IBM Z server concurrently. However, specific situations require a disruptive task to enable capacity that was recently added to the system. Some of these situations can be avoided if planning is done. Planning is a key factor for nondisruptive upgrades.

In a multi-site high-availability configuration, another option is the use of Flexible Capacity for Cyber Resiliency to move workload to another site while hardware maintenance is performed.



Disruptive upgrades are performed for the following reasons:

- ▶ LPAR memory upgrades when reserved storage was not previously defined are disruptive to image upgrades. z/OS and z/VM support this function.
- ▶ An I/O upgrade when the operating system cannot use the dynamic I/O configuration function is disruptive to that partition. Linux, z/VSE, and z/TPF do not support dynamic I/O configuration.

You can minimize the need for these outages by carefully planning and reviewing “Guidelines to avoid disruptive upgrades” on page 398.

### **Guidelines to avoid disruptive upgrades**

Based on the reasons for disruptive upgrades (see “Planning for nondisruptive upgrades” on page 397), you can use the following guidelines to avoid or at least minimize these situations, which increases the chances for nondisruptive upgrades:

- ▶ By using an SE function that is called Logical Processor add, which is under Operational Customization tasks, CPs and zIIPs can be added concurrently to a running partition. The CP and zIIP and initial or reserved number of processors can be changed dynamically.
- ▶ The operating system that runs in the targeted LPAR must support the dynamic addition of resources and to configure processors online. The total number of defined and reserved CPs cannot exceed the number of CPs that are supported by the operating system.

z/OS V3R1, V2.R5, and V2.R4, support 200 PUs per z/OS LPAR in non-SMT mode and 128 PUs per z/OS LPAR in SMT mode. For both, the PU total is the sum of CPs and zIIPs. z/VM supports up to 80 processors.

- ▶ Configure reserved storage to LPARs.

Configuring reserved storage for all LPARs before their activation enables them to be nondisruptively upgraded. The operating system that is running in the LPAR must configure memory online. The amount of reserved storage can be greater than the CPC drawer threshold limit, even if no other CPC drawer is installed.

With IBM z17 servers, z/OS 3.1 supports a limit of 16 terabytes (TB) of processor storage per logical partition (LPAR). z/VM V7.R2 and later support 4 TB memory partitions.

- ▶ Consider the flexible memory options.

Use a convenient entry point for memory capacity, and select memory options that allow future upgrades within the memory cards that are installed on the CPC drawers.

For more information about the offerings, see 2.5.7, “Flexible Memory Option” on page 53.

### **Considerations when installing CPC drawers**

During an upgrade, a second and third processor drawer can be installed concurrently if they are planned. Depending on the number of processor drawers in the upgrade and your I/O configuration, a fan-out rebalancing might be needed for availability reasons.

A fourth or fifth processor drawer can be installed at the IBM Manufacturing plant only.



## 8.13 Summary of Capacity on-Demand offerings

The CoD infrastructure and its offerings are based on client requirements for more flexibility, granularity, and better business control over the IBM Z infrastructure, operationally, and financially.

One major client requirement was to eliminate the need for a client authorization connection to the IBM Resource Link system when activating an offering. This requirement is met by the IIBM z15, IBM z16, and IBM z17 servers.

After the offerings are installed on the IBM z17 SE, they can be activated at any time at the customer's discretion. No intervention by IBM or IBM personnel is necessary.

The IBM z17 ME1 can have up to eight temporary upgrade records (On/Off CoD, CBU, System Recovery Boost Upgrade) installed or active at any given time. However, you can only have one On/Off CoD record active at any given time.

The installed offerings can be activated fully or partially, and in any sequence and any combination. The offerings can be controlled manually through command interfaces on the HMC, or programmatically through a number of APIs. IBM applications, ISV programs, and client-written applications can control the use of the offerings.

Resource usage (and therefore, financial exposure) can be controlled by using capacity tokens in the On/Off CoD offering records.

The CPM is an example of an application that uses the CoD APIs to provision On/Off CoD capacity that is based on the requirements of the workload. The CPM cannot control other offerings.

For more information about any of the topics in this chapter, see *Capacity on Demand User's Guide*, SC28-7058.









# Reliability, availability, and serviceability

From the quality perspective, the IBM z17 reliability, availability, and serviceability (RAS) design is driven by a set of high-level program RAS objectives. The IBM Z platform continues to drive toward Continuous Reliable Operation (CRO) at the single footprint level.

**Note:** Throughout this chapter, IBM z17 refers to IBM z17 Model ME1 (Machine Type 9175).

The key objectives, in order of priority, are to ensure data integrity, computational integrity, reduce or eliminate unscheduled outages, reduce scheduled outages, reduce planned outages, and reduce the number of Repair Actions.

RAS can be accomplished with improved concurrent replace, repair, and upgrade functions for processors, memory, drawers, and I/O. RAS also extends to the nondisruptive capability for installing Licensed Internal Code (LIC) updates. In most cases, a capacity upgrade can be concurrent without a system outage. As an extension to the RAS capabilities, environmental controls are implemented in the system to help reduce power consumption and meet cooling requirements.

This chapter includes the following topics:

- ▶ “RAS strategy” on page 402
- ▶ “Technology” on page 402
- ▶ “Structure” on page 407
- ▶ “Reducing complexity” on page 408
- ▶ “Reducing touches” on page 408
- ▶ “IBM z17 availability characteristics” on page 408
- ▶ “IBM z17 RAS functions” on page 412
- ▶ “Enhanced drawer availability” on page 416
- ▶ “Concurrent Drawer Maintenance” on page 424
- ▶ “RAS capability for the HMA and SE” on page 426



## 9.1 RAS strategy

The RAS strategy is to manage change by learning from previous generations and investing in new RAS functions to eliminate or minimize all sources of outages. Enhancements introduced with IBM z16 RAS designs are implemented on the IBM z17 design by introducing new technology, structure, and requirements. Continuous improvements in RAS are associated with new features and functions to ensure that IBM Z servers deliver exceptional value to clients.

The following overriding RAS requirements are principles as shown in Figure 9-1:

- ▶ Include existing (or equivalent) RAS characteristics from previous generations.
- ▶ Learn from current field issues and addressing the deficiencies.
- ▶ Understand the trend in technology reliability (hard and soft) and ensure that the RAS design points are sufficiently robust.
- ▶ Invest in RAS design enhancements (hardware and firmware) that provide IBM Z and Customer valued differentiation.

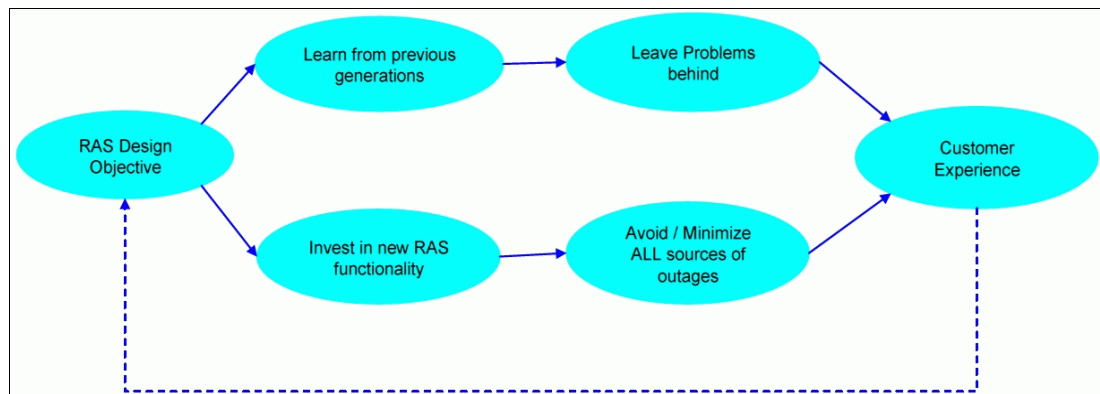


Figure 9-1 Overriding RAS requirements

## 9.2 Technology

This section introduces some of the RAS features that are incorporated in the IBM z17 design.

### 9.2.1 Processor Unit chip

IBM z17 uses the Processor Unit (PU) chip with the following technical changes:

- ▶ A PU chip is implemented by using EUV process with 5 nm FinFET lithography, and has nine cores (eight PUs plus one Data processing Unit, DPU) per chip (design) that are running at 5.5 GHz.
- ▶ Each core has private L1 (instruction and data) caches and a semi-private L2 cache, which are 36 MB. The nine cores and L2 caches on the chip communicate through bidirectional high-speed on-chip ring and with all SMP, I/O, and memory interfaces (see Figure 9-2 on page 403).



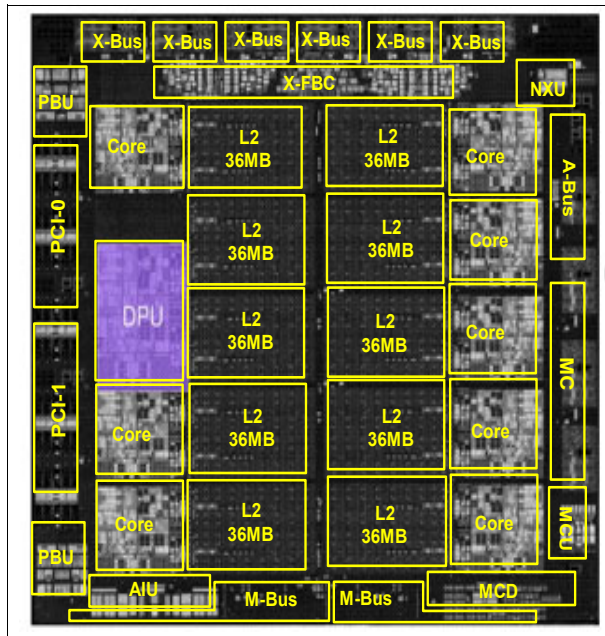


Figure 9-2 PU chip structure

- Two PU chips are packaged on a dual chip module (DCM), as shown in Figure 9-3. PU chip to PU chip communication is performed through the M-Bus, which is a high-speed bus that acts as ring-to-ring extension communication at 160 Gbps data rate, with the following RAS characteristics:
  - ECC on data path and snoop bus
  - Parity on miscellaneous control lines
  - One spare per ~50 data bit bus

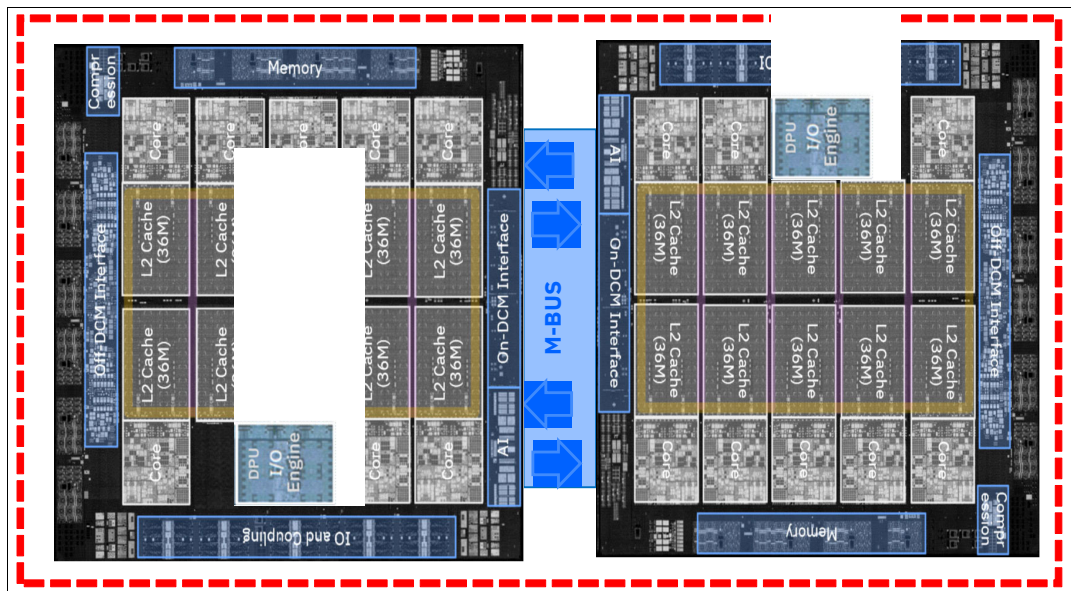


Figure 9-3 IBM z17 Dual Chip Module (M-Bus connects the two chips)

- The processor recovery logic is on the DCM and includes the following RAS characteristics:
  - PU refresh for soft errors



- Hardened RU (recovery unit) with wordline failure detection
- Improved latch interleaving (spacing) for soft error resistance
- Core sparing for hard errors: All servers ship with two spare cores
- Improved long-term thresholding
- Redesigned nest
- Concurrent repair by way of concurrent drawer repair (CDR)
- Concurrent upgrade by way of LICCC and enhanced drawer availability (EDA)
- ▶ L1 and L1 shadow: L1 shadow is new on IBM z17 and includes the following characteristics:
  - Behaves like the L1 (for repairs)
  - Contains changed data
  - All unrecoverable error (UE) checkstop core: UEs are refetched before acting
  - UE impact depends on system state:
    - Before end OP, IPD without storage validity
    - Before SIE sync, System Damaged
- ▶ IBM Z Integrated Accelerator for AI (AIU)
 

Second generation AI Accelerator (AIU) is implemented on IBM z17. AIU behaves like a co-processor to process the synchronous instructions, but AIU is in nest. The core control the AIU by issuing instruction (NNPA):

  - The array macro has row and column repair (MD and ABIST)
  - 1 MB cache with single error correction double error detection (SECDEC) ECC
- ▶ On-chip L2, caches are implemented in dense SRAM. The IBM z16 and IBM z17 removed the physical L3 (on-chip for IBM z15) and L4 (on extra storage controller single chip module - SC SCM) and implemented clustered cache algorithms that provide virtualized L3 (shared victim) and virtual L4 (shared victim) caches (see Figure 9-2 on page 403).
- ▶ The dedicated L2 cache (dense SRAM) is semi-private to the core; that is, each core features an associated 36-MB slice of the L2 cache (semi-private):
  - Virtual L3 on PU Chip (shared victim 360 MB)
  - Virtual L4 on drawer up to 2.88 GB
- ▶ L2 cache and L2 cache recovery:
  - The L2 cache includes expected inline Symbol ECC (RAID4) recovery:
    - 64:1 interleaving (4:1 physical, 16:1 logical)
    - L2 symbol ECC is inherited by virtual L3 and virtual L4
  - The array macro includes row and column repair (module manufacturing and ABIST)
  - Ring can be fenced from L2 for yield and fences core (module manufacturing):
    - If the core checkstops, all 36 MB can be used by the system
    - L2 dedicated and shared split are managed in the LRU logic
  - Dynamic macro sparing: Four spare macros, one spare for each quarter slice
  - L2 1/8 portion is taken offline when a spare is needed, and none is available (named 7/8 recovery): The core must be spared
  - L2 directory is SECDED ECC protected



IBM z17 processor memory and cache structure are shown in Figure 9-4. The physical L3 cache (on chip) and System Controller (SC) SCM (physical L4 cache for IBM z15), which was implemented in EDRAM, were removed and replaced on IBM z17 virtual L3 and L4 cache structure.

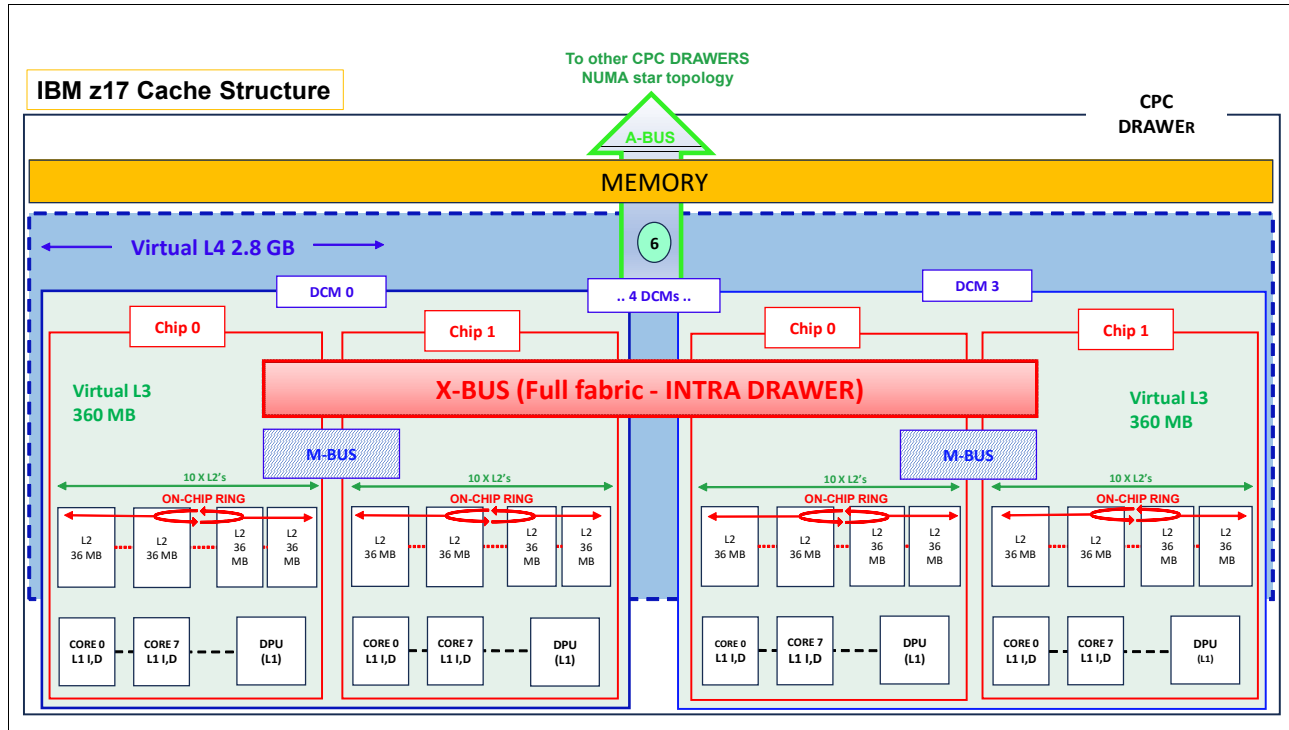


Figure 9-4 IBM z17 CPC Drawer Cache structure

## 9.2.2 Main memory

IBM z17 memory consists of the following features:

- ▶ Up to 48 DIMM are used per drawer in two rows (64 TB system maximum):
  - Organized in 8-channel Reed-Solomon RAIM groups
  - Up to 16 TB per drawer / 64 TB per system
    - 50% reduced RAIM overhead (compared to IBM z15)
    - RAIM protection (similar to IBM z16)
    - Up to six Memory Controllers (same as z16)
  - Up to three chip marks and one channel mark
  - Memory Encryption strengthened to AES256 (AES128 on z16)
  - DDR4 and DDR5 DRAM with on chip power regulation: N+1 voltage regulators
    - Odyssey memory buffer chip enables DDR5
  - Standard Open Memory Interface (OMI); up to six OMI per drawer:
    - CRC/Retry for soft errors
    - Degrade bus lanes from four lanes to two lanes on hard error
  - No waiting for all eight RAIM channels, use first seven to respond:
    - RAIM recovery is used to enhance performance without affecting RAS
    - Processor uses data from the first seven channels to respond



- Refetch data by using all eight RAIM channels if an error occurs
- ▶ Flexible memory:
  - Provides more physical memory DIMMs that are needed to support activation of all customer-purchased memory and HSA on a multiple drawer zNext with one drawer down for:
    - Scheduled concurrent drawer upgrade (for example, memory add)
    - Scheduled concurrent drawer maintenance (N+1 repair)
    - Concurrent repair of an out-of-service CPC drawer “fenced” during Activation (POR)

**Note:** All of these features are available without Flexible Memory, but all customer purchased memory is not available for use in most cases. Some work might need to be shut down or not restarted.

- Offered on Max90, Max136, Max183, and Max208:
  - Flex Memory doubles the installed (configurable) memory for a two-CPC drawer system
  - Flex memory adds 1/2, 1/3, or 1/4 capacity to the ordered memory for a 2-, 3-, or 4-CPC drawer system
- Virtual Flash Memory considerations:
  - Each VFM Feature (FC 0644) takes 512 GB of memory. Up to 12 VFM features can be ordered.
  - If VFM is present, it is included in the Flexible memory calculations.
- ▶ Concurrent service and upgrade by way of Concurrent Drawer Repair (CDR)

### 9.2.3 I/O and service

I/O and service consist of the following features:

- ▶ New I/O features on IBM z16:
  - FICON Express 32G (4 port)
  - Coupling Express3 LR
  - ICA-SR2.0
  - OSA Express7 1.2
  - Network Express
  - Crypto Express 8S
- ▶ The number of PSP, support partitions, for managing native PCIe I/O:
  - Four partitions
  - Reduced effect on MCL updates
  - Better availability
- ▶ Faster Dynamic Memory Relocation engine:
  - Enables faster reallocation of memory that is used for LPAR activations, CDR, and concurrent upgrade
  - Enables also LPAR Optimizations with Dynamic Storage Reconfiguration (DSR)
  - Provides faster, more robust service actions
- ▶ IBM z17 eliminates the classic IBM Z fill and drain tool, FDT, and fluid handling in the field.



## 9.3 Structure

The IBM z17 was designed in a 19-inch frames format. The IBM z17 ME1 can be delivered only as an air-cooled system and fulfills the requirements for ASHRAE A3 class environment.

The IBM z17 ME1 can have up to 12 PCIe I/O drawers delivered with Power Distribution Unit (PDU). The structure of the IBM z17 is designed with the following goals:

- ▶ Enhanced system modularity
- ▶ Standardization to enable rapid integration
- ▶ Platform simplification

Cables are keyed to ensure that correct lengths are plugged. Plug detection ensures correct location, and custom latches ensure retention. Further improvements to the fabric bus include symmetric multiprocessing (SMP-10<sup>1</sup>) cables that connect the drawers.

To improve field-replaceable unit (FRU) isolation, advanced continuity checking techniques are applied to the new SMP-10 cables (Concurrent Cable Repair), which plug into a SMP10 Assembly (Cupid) at the rear top of the CPC drawer with direct connections to the adjacent DCM location.

The thermal RAS design also was improved for the field-replaceable water manifold for PU cooling.

The IBM z17 includes the following characteristics:

- ▶ Processing infrastructure is designed by using drawer technology
- ▶ Keyed cables and plugging detection
- ▶ SMP-10 cables that are used for fabric bus connections
- ▶ Master-master redundant oscillator design in the main memory
- ▶ Point of load cards are separate FRUs
- ▶ Improved serviceability for the water manifold
- ▶ Improved redundant oscillator design
- ▶ Redundant combined Base Management Card (BMC) and Oscillator Cards (OSC) are provided per CPU drawer
- ▶ Redundant N+1 Power Supply Units (PSU) to CPC drawer and PCIe+ I/O drawer

---

<sup>1</sup> IBM z16 uses SMP-9 cables to interconnect the CPC drawers



## 9.4 Reducing complexity

IBM z17 servers continue the IBM z16 enhancements that reduced system RAS complexity.

Independent channel recovery with replay buffers on all interfaces allows recovery of a single DIMM channel, while other channels remain active. Further redundancies are incorporated in I/O pins for clock lines to main memory, which eliminates the loss of memory clocks because of connector (pin) failure. The following RAS enhancements reduce service complexity:

- ▶ Continued use of RAIM ECC
- ▶ RAIM logic was moved on DIMM
- ▶ N+1 on-DIMM voltage regulation
- ▶ Replay buffer for hardware retry on soft errors on the main memory interface
- ▶ Redundant I/O pins for clock lines to main memory
- ▶ Staggered refresh for performance enhancement
- ▶ The new RAIM scheme achieves a higher ratio of data to ECC symbols, while also providing another chip mark

## 9.5 Reducing touches

IBM Z RAS efforts focus on the reduction of unscheduled, scheduled, planned, and unplanned outages. IBM Z technology has a long history of demonstrated RAS improvements. This effort continues with changes that reduce service *touches* on the system.

Firmware was updated to improve filtering and resolution of errors that do not require action. The following RAS enhancements reduce service touches:

- ▶ Improved error resolution to enable filtering
- ▶ Enhanced integrated sparing in processor cores
- ▶ Cache relocates
- ▶ N+1 SEEPROM
- ▶ N+2 POL - (Point of Load)
- ▶ DRAM marking
- ▶ (Dynamic) Spare BUS lanes for PU-PU, PU-MEM, MEM-MEM fabric
- ▶ N+1 Support Element (SE) with N+1 SE power supplies
- ▶ Redundant temperature sensor (one SEEPROM and one temperature sensor per I2C bus)
- ▶ FICON forward error correction
- ▶ A-Bus Lane Sparing
- ▶ OMI (Open Memory Interface) Bus Lane Sparing
- ▶ PU Core Sparing

## 9.6 IBM z17 availability characteristics

The following functions include availability characteristics on IBM z17:



- ▶ Enhanced drawer availability (EDA)

EDA is a *procedure* under which a CPC drawer in a multi-drawer system can be removed and reinstalled during an upgrade or repair action with no effect on the workload.

- ▶ Concurrent memory upgrade or replacement

Memory can be upgraded concurrently by using Licensed Internal Code Configuration Control (LICCC) if physical memory is available on the drawers.

The EDA function can be useful if the physical memory cards must be changed in a multi-drawer configuration (requiring the drawer to be removed).

It requires the availability of more memory resources on other drawers or reducing the need for memory resources during this action. Select the flexible memory option to help ensure that the suitable level of memory is available in a multiple-drawer configuration. This option provides more resources to use EDA when repairing a drawer or memory on a drawer. They also are available when upgrading memory when larger memory cards might be required.

- ▶ Concurrent Driver Maintenance (CDM)

One of the greatest contributors to downtime during planned outages is LIC driver updates that are performed in support of new features and functions. IBM z17 supports the concurrent activation of a selected new driver level. In most cases, the system Support Element driver level will not need a driver upgrade. If the HMA feature is installed, there may be a need to upgrade the HMCs to the new HMC driver level concurrently.

- ▶ Concurrent fan-out addition or replacement

A PCIe+ fan-out card provides the path for data between memory and I/O through PCIe cables. With IBM z17, a hot-pluggable and concurrently upgradeable fan-out card is available. Up to 12 PCIe fan-out cards per CPC drawer are available for IBM z17 servers. An IBM z17 Model ME1 feature Max183 and Max208 hold four CPC drawers and can have 48 PCIe fan out slots.

Internal I/O paths from the CPC drawer fan-out ports to a PCIe I/O drawer are spread across multiple CPC drawers (for features Max90, Max136, Max183, and Max208) and across different nodes within a single CPC drawer Feature Max49. During an outage, a fan-out card that is used for I/O can be repaired concurrently while redundant I/O interconnect ensures that no I/O connectivity is lost.

- ▶ Redundant I/O interconnect

Redundant I/O interconnect helps maintain critical connections to devices. IBM z17 ME1 allows a single drawer in a multi-drawer system to be removed and reinstalled concurrently during an upgrade or repair. Connectivity to the system I/O resources is maintained through a second (alternate) path from a different drawer.

- ▶ Base Management Card (BMC)/Oscillator Cards (OSC)

IBM z17 ME1 features two combined BMC and OSC per CPU drawer. The strategy of redundant clock and switchover stays the same. One primary and one backup is available. If the primary OSC fails, the backup detects the failure, takes over transparently, and continues to provide the clock signal to the CPC.

- ▶ Processor unit (PU) sparing

IBM z17 ME1 has two spare PUs per system to maintain performance levels if an active PU, Internal Coupling Facility (ICF), Integrated Facility for Linux (IFL), IBM Z Integrated Information Processor (zIIP), integrated firmware processor (IFP), or system assist processor (SAP) fails. Transparent sparing for failed processors is supported and sparing is supported across the drawers.

- ▶ Application preservation



This function is used when a PU fails and no spares are available. The state of the failing PU is passed to another active PU, where the operating system uses it to successfully resume the task (in most cases, without customer intervention).

► Cooling change

The IBM z17 air-cooled configuration includes a front-to-rear radiator cooling system. The radiator pumps, blowers, controls, and sensors are N+1 redundant. In normal operation, one active pump supports the system. A second pump is turned on and the original pump is turned off periodically, which improves reliability of the pumps. The replacement of pumps or blowers is concurrent with no effect on performance.

**Note:** Customer chilled water (WCU) is not offered on IBM z17.

► FICON Express 32S and FICON Express 32-4P with Forward Error Correction (FEC)

FICON Express32-4P and 32S features continue to provide a new standard for transmitting data over 32 Gbps links by using 64b/66b encoding. The new standard that is defined by T11.org FC-FS-3 is more efficient than the current 8b/10b encoding.

FICON Express32-4P and 32S channels that are running at 32 Gbps can take advantage of FEC capabilities when connected to devices that support FEC.

FEC allows FICON Express32-4P and 32S channels to operate at higher speeds, over longer distances, with reduced power and higher throughput. They also retain the same reliability and robustness for which FICON channels are traditionally known.

FEC is a technique that is used for controlling errors in data transmission over unreliable or noisy communication channels. When running at 32 Gbps link speeds, customers often see fewer I/O errors, which reduces the potential effect to production workloads from those I/O errors.

Read Diagnostic Parameters (RDP) improve Fault Isolation. After a link error is detected (for example, IFCC, CC3, reset event, or a link incident report), link data that is returned from Read Diagnostic Parameters is used to differentiate between errors that result from failures in the optics versus failures because of dirty or faulty links.

Key metrics can be displayed on the operator console. The results of a display matrix command with the LINKINFO=FIRST parameter, which collects information from each device in the path from the channel to the I/O device (see Figure on page 411):

- Transmit (Tx) and Receive (Rx) optic power levels from the PCHID, Switch Input and Output, and I/O device
- Capable and Operating speed between the devices
- Error counts
- Operating System requires new function APAR OA49089





Figure 9-5 Read Diagnostic Parameters function

The new IBM Z Channel Subsystem Function performs periodic polling from the channel to the end points for the logical paths that are established and reduces the number of useless Repair Actions (RAs).

The RDP data history is used to validate Predictive Failure Algorithms and identify Fibre Channel Links with degrading signal strength before errors start to occur. The new Fibre Channel Extended Link Service (ELS) retrieves signal strength.

#### ► FICON Dynamic Routing

FICON Dynamic Routing (FIDR) enables the use of storage area network (SAN) dynamic routing policies in the fabric. With the IBM z17, FICON channels are no longer restricted to the use of static routing policies for inter-switch links (ISLs) for cascaded FICON directors.

FICON Dynamic Routing dynamically changes the routing between the channel and control unit that is based on the Fibre Channel Exchange ID. Each I/O operation has a unique exchange ID. FIDR supports static SAN routing policies and dynamic routing policies.

FICON Dynamic Routing can help clients reduce costs by providing the following features:

- Share SANs between their FICON and FCP traffic.



- Improve performance because of SAN dynamic routing policies that better use all of the available ISL bandwidth through higher use of the ISLs,
- Simplify management of their SAN fabrics by using static routing policies that assign different ISL routes with each power-on-reset (POR), which makes the SAN fabric performance difficult to predict.

Customers must ensure that all devices in their FICON SAN support FICON Dynamic Routing before they implement this feature.

## 9.7 IBM z17 RAS functions

Hardware RAS function improvements focus on addressing all sources of outages. Sources of outages feature the following classifications:

► **Unscheduled**

This outage occurs because of an unrecoverable malfunction in a hardware component of the system.

► **Scheduled**

This outage is caused by changes or updates that must be done to the system in a timely fashion. A scheduled outage can be caused by a disruptive patch that must be installed, or other changes that must be made to the system.

► **Planned**

This outage is caused by changes or updates that must be done to the system. A planned outage can be caused by a capacity upgrade or a driver upgrade. A planned outage usually is requested by the customer, and often requires planning. The IBM z17 design phase focuses on enhancing planning to simplify or eliminate planned outages.

The difference between scheduled outages and planned outages might not be obvious. The general consensus is that scheduled outages occur sometime soon. The time frame is approximately two weeks.

Planned outages are outages that are planned well in advance and go beyond this approximate two-week time frame. This chapter does not distinguish between scheduled and planned outages.

Preventing unscheduled, scheduled, and planned outages was addressed by the IBM Z system design for many years.

IBM z17 ME1 has a fixed size HSA of 884 GB. This size helps eliminate planning requirements for HSA and provides the flexibility to update dynamically the configuration. You can perform the following tasks dynamically:<sup>2</sup>

► **Add:**

- Logical partition (LPAR)
- Logical channel subsystem (LCSS)
- Subchannel set
- Logical PU to an LPAR
- Cryptographic coprocessor
- Memory
- Physical processor

► **Remove a cryptographic coprocessor**

---

<sup>2</sup> Some planning considerations might exist. For more information, see Chapter 8, “System upgrades” on page 353.



- ▶ Enable I/O connections
- ▶ Swap processor types

By addressing the elimination of planned outages, the following tasks also are possible:

- ▶ Concurrent driver upgrades
- ▶ Concurrent and flexible customer-initiated upgrades

For more information about the flexible upgrades that are started by customers, see 8.3.2, “Customer Initiated Upgrade facility” on page 362.

- ▶ STP management of concurrent CTN Split and Merge
- ▶ Dynamic I/O for stand-alone CF, Linux on Z and z/TPF running on IBM z16 and IBM z17

Dynamic I/O configuration changes can be made to a stand-alone CF, Linux on Z and z/TPF without requiring a disruptive power on reset. A firmware LPAR with a firmware-based appliance version of an HCD instance is used to apply the new I/O configuration changes. The firmware-based LPAR is driven by updates from an HCD instance that is running in a z/OS LPAR on a different IBM z16 or IBM z17 CPC that is connected to the same IBM z17 HMA<sup>3</sup>.

- ▶ System Recovery Boost Stage 3

System Recovery Boost enhancements for IBM z17 allows the possibility of significantly reducing the effect of these disruptions by boosting a set of recovery processes that create significant pain points for users.

These recovery processes include the following examples:

- IBM SAN Volume Controller memory dump boost
- Middleware restart or recycle boost
- HyperSwap configuration load boost

For more information about System Recovery Boost, see *Introducing IBM Z System Recovery Boost*, REDP-5563.

### 9.7.1 Scheduled outages

Concurrent hardware upgrades, parts replacement, driver upgrades, and firmware fixes that are available with IBM z17 all address the elimination of scheduled outages. Also, the following indicators and functions that address scheduled outages are included:

- ▶ Memory data bus lane sparing.  
This feature reduces the number of repair actions for memory.
- ▶ Dual tabbed Memory Clock signals.
- ▶ Triple DRAM chipkill tolerance.
- ▶ Processor drawer power distribution  $N+2$  design.

The CPC drawer uses POL cards in a highly redundant  $N+2$  configuration. POL regulators are daughter cards that contain the voltage regulators for the principle logic voltage boundaries in the IBM z17 CPC drawer. They plug into the CPC drawer system board and are nonconcurrent FRUs for the affected drawer (similar to the memory DIMMs). If you can use EDA, the replacement of POL cards is concurrent for the entire IBM Z server.

- ▶ Redundant ( $N+1$ ) Ethernet switches.
- ▶ Redundant ( $N+2$ ) ambient temperature sensors.

<sup>3</sup> Function also supported by an IBM z16 HMA/HMC upgraded to driver 2.17.0



- Dual inline memory module (DIMM) field-replaceable unit (FRU) indicators.

These indicators imply that a memory module is not error-free and might fail eventually. This indicator gives IBM a warning and provides time to concurrently repair the storage module if the IBM z17 is a multi-drawer system.

The process to repair the storage module is to isolate or “fence off” the drawer, remove the drawer, replace the failing storage module and then, add the drawer. The flexible memory option might be necessary to maintain sufficient capacity while repairing the storage module.

- Single processor core checkstop and sparing.

This indicator shows that a processor core malfunctioned and is *spared*. IBM determines what to do based on the system and the history of that system.

- Point-to-point fabric for symmetric multiprocessing (SMP) RAS design

- SMP-10, introduced with IBM z17, is divided into SMP-10-i, used for internal drawer communication, and SMP-10-e used for external drawer to drawer communication. The SMP-10-e is the A-BUS (point to point between drawers).

- SMP10-10-e assembly, installed at the top rear of the CPC drawer, supports ABUS connectivity of DCMs between drawers by bridging the SMP10-i connection on the DCM to the SMP10-e connection point used to link the drawers. 11 bits per A-Bus

- Concurrent repair of the SMP10 cables.

- One SMP10 cable has nine lanes and It supports time domains reflexometry (TDR) for failure isolation, and Concurrent Cable Repair (CCR)
  - RAS lane Degrade and Call Home Strategy design

Having fewer components that can fail is an advantage. In a multi-drawer system, all of the drawers are connected by point-to-point connections. A drawer can always be added concurrently.

- Air-cooled system: radiator with redundant ( $N+1$ ) pumps.

IBM z17 implements true  $N+1$  redundancy on pumps and blowers for the radiator.

One radiator unit in Frame A and one radiator unit in frame B, which is configuration-dependent. The radiator cooling system can support up to three CPC drawers simultaneously with a redundant design that consists of two pumps and two blowers.

The replacement of a pump or blower causes no performance effect.

- The Gen4 PCIe+ I/O drawer is available for IBM z17. It supports concurrent repair, and all of the Gen4 PCIe+ I/O drawer-supported features can be installed concurrently.
- Memory interface logic to maintain channel synchronization when one channel goes into replay. IBM z17 can isolate recovery to only the failing channel.
- Out-of-band access to DIMM (for background maintenance functions).  
Out-of-band access (by using an I2C interface) allows maintenance (such as logging) without disrupting customer memory accesses.
- Open Memory Interface (OMI) Memory Bus lane sparing.
- Improved DIMM exerciser for testing memory during IML.
- PCIe redrive hub cards plug straight in (no blind mating of connector). Simplified plugging that is more reliable is included.
- ICA SR (short distance) coupling cards plug straight in (no blind mating of the connector). Simplified plugging that is more reliable is included.



- Coupling Express3 LR (CE3 LR) coupling cards plug into the Gen4 PCIe+ I/O drawer, which allows more connections with a faster bandwidth.

## 9.7.2 Unscheduled outages

An *unscheduled outage* occurs because of an unrecoverable malfunction in a hardware component of the system.

The following improvements can minimize unscheduled outages:

- Continued focus on firmware quality

For LIC and hardware design, failures are eliminated through rigorous design rules; design walk-through; peer reviews; element, subsystem, and system simulation; and extensive engineering and manufacturing testing.

- Memory subsystem

Redundant Array of Independent Memory (RAIM) on IBM Z servers is a concept that is similar to the concept of Redundant Array of Independent Disks (RAID). The RAIM design detects and recovers from dynamic random access memory (DRAM), socket, memory channel, or DIMM failures. Memory size now includes RAIM protection and recovery.

Starting with IBM z16, memory channels are organized in 8-card RAIM groups that provide 50% reduced RAIM overhead (compared to IBM z15).

RAIM protection is similar to IBM z16:

- Up to three chip marks and one channel mark
- DDR4 and DDR5 DRAM with on chip power regulation
- N+1 voltage regulators

Memory is implemented by using standard Open Memory Interface (OMI) with up to six OMIs per drawer, has CRC/Retry for soft errors, degrade bus lanes 4->2 on hard error, and no waiting for all eight cards (use first seven to respond).

A precise marking of faulty chips helps ensure timely DIMM replacements. The design of the IBM z17 further improved this chip marking technology. Graduated DRAM marking is available, and channel marking and scrubbing calls for replacement on the third DRAM failure is available.

For more information about the memory system on IBM z17 ME1, see 2.5, “Memory” on page 44.

- Soft-switch firmware

IBM z17 is equipped with the capabilities of soft-switching firmware. Enhanced logic in this function ensures that every affected circuit is powered off during the soft-switching of firmware components. For example, when you are upgrading the microcode of a FICON feature, enhancements are implemented to avoid any unwanted side effects that were detected on previous systems.

## Time synchronization enhancements

- Server Time Protocol (STP) recovery enhancement. IBM z17 updated the clocking structure:
  - System uses a mesosynchronous clocking structure
  - Two redundant oscillator cards in each drawer with dynamic oscillator switchover
  - STP now has external clock reference support, separate Ethernet ports for ETS
  - PTP and NTP (Ethernet cabling) are now directly connected to the CPC (no SE ETS connection needed)



- Moved from Simple NTPv3 to NTPv4, and from PTPv2 to PTPv2.1
- Introduction of ‘mixed-mode’
  - Up to three NTP sources and/or two PTP sources
  - Dynamic reference intervals
- Introduction of security features
  - Network Time Security (NTS)
  - Combine, Select, Falseticker Algorithms
- Oscillator card shares FRU package with drawer’ Base Control Card (BMC)
- Concurrent repair of oscillator or control card
- Enhanced monitoring and reporting
  - Monitoring of PTP/NTP/DNS traffic
  - Audit Log improvements
  - Introduction of Availability Baseline
  - New HW messages, STP alerts, IQYYLOGs
  - PCSIB improvements

A new signal was implemented in IBM z16 hardware for helping with STP recovery. This signal is called *N-Mode power signal*. In a CTN with PTS and BTS being IBM z17 servers, CTN recovery can be started by the new signal. Systems must have dual power with at least one side of the power capable of holding the power (UPS or similar) for five minutes if a utility power failure occurs.

When PCIe-based integrated communication adapter (ICA) Short Reach (SR) links are used, an unambiguous “going away signal” is sent when the server on which the coupling link is running is about to enter a failed (check stopped) state.

When the “going away signal” that is sent by the Current Time Server (CTS) in an STP-only Coordinated Timing Network (CTN) is received by the Backup Time Server (BTS), the BTS can safely take over as the CTS without relying on the previous Offline Signal (OLS) in a two-server CTN, or as the Arbiter in a CTN with three or more servers.

Enhanced Console Assisted Recovery (ECAR) contains recovery algorithms during a failing Primary Time Server (PTS) and uses communication over the HMA/SE network to assist with BTS takeover.

For more information, see Chapter 10, “Hardware Management Console and Support Element” on page 429, and.

Coupling Express3 LR does not support the “going away signal”; however, ECAR can be used to assist with recovery in the following configurations:

- ▶ Design of pervasive infrastructure controls in processor chips in memory ASICs.
- ▶ Improved error checking in the processor recovery unit (RU) to better protect against word line failures in the RU arrays.

## 9.8 Enhanced drawer availability

EDA is a procedure in which a drawer in a multi-drawer system can be removed and reinstalled during an upgrade or repair action. This procedure has no effect on the running workload with suitable planning and depending on your system configuration. A single CPC drawer IBM z17 ME1 (Max43) does *not* support EDA.



The EDA procedure and careful planning help ensure that all the resources are still available to run critical applications in an  $(n-1)$  drawer configuration. This process allows you to avoid planned outages. Consider the flexible memory option to provide more memory resources when you are replacing a drawer.

For more information about flexible memory, see 2.5.7, “Flexible Memory Option” on page 53.

To minimize the effect on current workloads, ensure that sufficient inactive physical resources exist on the remaining drawers to complete a drawer removal. Also, consider deactivating noncritical system images, such as test or development LPARs. After you stop or deactivate these noncritical LPARs and free their resources, you might find sufficient inactive resources to contain critical workloads while completing a drawer replacement.

### 9.8.1 EDA planning considerations

To use the EDA function, configure enough physical memory and engines so that the loss of a single drawer does not result in any degradation to critical workloads during the following occurrences:

- ▶ A degraded restart in the rare event of a drawer failure
- ▶ A drawer replacement for repair or a physical memory upgrade

The following configurations especially enable the use of the EDA function. These IBM z17 features need enough spare capacity so that they can cover the resources of a fenced or isolated drawer. This configuration imposes limits on the following number of the client-owned PUs that can be activated when one drawer within a model is fenced:

- ▶ A maximum of:
  - 43 PUs are configured on the Max43
  - 90 PUs are configured on the Max90
  - 136 PUs are configured on the Max136
  - 183 PUs are configured on the Max183
  - 208 PUs are configured on the Max208
- ▶ No special feature codes are required for PU and model configuration.
- ▶ IBM z17 feature Max43 and Max90 each have 5 SAPs in each drawer. Max136 has three CPC drawers and a total 16 standard SAPs, Max183 has four CPC drawers and a total of 21 SAPs, and Max208 also has four CPC drawers and 24 SAPs.
- ▶ The flexible memory option delivers physical memory so that 100% of the purchased memory increment can be activated even when one drawer is fenced.

The system configuration must have sufficient dormant resources on the remaining drawers in the system for the *evacuation* of the drawer that is to be replaced or upgraded. Dormant resources include the following possibilities:

- ▶ Unused PUs or memory that is not enabled by LICCC
- ▶ Inactive resources that are enabled by LICCC (memory that is not being used by any activated LPARs)
- ▶ Memory that is purchased with the flexible memory option
- ▶ Extra drawers

The I/O connectivity also must support drawer removal. Most of the paths to the I/O feature redundant I/O interconnect support in the I/O infrastructure (drawers) that enable connections through multiple fan-out cards.



If sufficient resources are not present on the remaining drawers, specific noncritical LPARs might need to be deactivated. One or more PUs or storage might need to be configured offline to reach the required level of available resources. Plan to address these possibilities to help reduce operational errors.

**Exception:** Single-drawer systems cannot use the EDA procedure.

Include the planning process as part of the initial installation and any follow-on upgrade that modifies the operating environment. A customer can use the IBM Call Home Connect Cloud reports, tasks on the Support Element (Storage Information, View Hardware Configuration), and CHPID Mapping Tool reports, to determine the number of drawers, active PUs, memory configuration, and channel layout.

If the IBM z17 is installed, click **Prepare for Enhanced Drawer Availability** in the Perform Model Conversion window of the EDA process on the Hardware Management Appliance (HMA). This task helps you determine the resources that are required to support the removal of a drawer with acceptable degradation to the operating system images.

The EDA process determines which resources (including memory, PUs, and I/O paths) are free to allow for the removal of a drawer. You can run this preparation on each drawer to determine which resource changes are necessary. Use the results as input in the planning stage to help identify critical resources.

With this planning information, you can examine the LPAR configuration and workload priorities to determine how resources might be reduced and still allow the drawer to be concurrently removed.

Include the following tasks in the planning process:

- ▶ Review of the IBM z17 configuration to determine the following values:
  - Number of drawers that are installed and the number of PUs enabled. Consider the following points:
    - Use the IBM Call Home Connect Cloud or the HMA to determine the model, number, and types of PUs (CPs, IFLs, ICFs, and zIIPs).
    - Determine the amount of memory (physically installed and LICCC-enabled).
    - Work with your IBM Service Support Representative (IBM SSR) to determine the memory card size in each drawer. The memory card sizes and the number of cards that are installed for each drawer can be viewed from the SE under the CPC configuration task list. Use the View Hardware Configuration option.
  - ICA SR fan-out layouts and ICA to ICA connections.

Use the IBM Call Home Connect Cloud to review the channel configuration. This process is a normal part of the I/O connectivity planning. The alternative paths must be separated as far into the system as possible.
- ▶ Review the system image configurations to determine the resources for each image.
- ▶ Determine the importance and relative priority of each LPAR.
- ▶ Identify the LPAR or workloads and the actions to be taken:
  - Deactivate the entire LPAR.
  - Configure PUs.
  - Reconfigure memory, which might require the use of reconfigurable storage unit (RSU) values.



- Vary off the channels.
- Review the channel layout and determine whether any changes are necessary to address single paths.
- Develop a plan to address the requirements.

When you perform the review, document the resources that can be made available if the EDA is used. The resources on the drawers are allocated during a POR of the system and can change after that process.

Perform a review when changes are made to your IBM z17, such as adding drawers, PUs, memory, or channels. Also, perform a review when workloads are added or removed, or if the HiperDispatch feature was enabled and disabled since the last time you performed a POR.

## 9.8.2 Enhanced Drawer Availability processing

To use the EDA, first ensure that the following conditions are met:

- The used processors (PUs) on the drawer that is removed are freed.
- The used memory on the drawer is freed.
- For all I/O domains that are connected to the drawer, ensure that alternative paths exist. Otherwise, place the I/O paths offline.

For the EDA process, this phase is the preparation phase. It is started from the SE, directly or on the HMA, by using the Single object operation option on the Perform Model Conversion window from the CPC configuration task list, as shown in Figure 9-6.

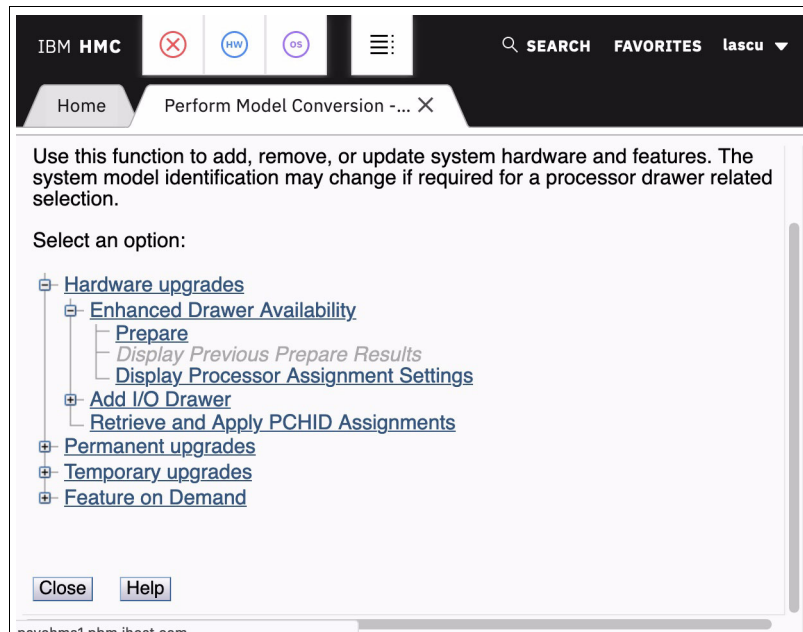


Figure 9-6 Clicking Prepare for Enhanced Drawer Availability option

### Processor availability

Processor resource availability for reallocation or deactivation is affected by the type and quantity of the resources in use, such as the following examples:

- Total number of PUs that are enabled through LICCC



- ▶ PU definitions in the profiles that can be dedicated and dedicated reserved or shared
- ▶ Active LPARs with dedicated resources at the time of the drawer repair or replacement

To maximize the PU availability option, ensure that sufficient inactive physical resources are on the remaining drawers to complete a drawer removal.

### Memory availability

Memory resource availability for reallocation or deactivation depends on the following factors:

- ▶ Physically installed memory
- ▶ Image profile memory allocations
- ▶ Amount of memory that is enabled through LICCC
- ▶ Flexible memory option
- ▶ Virtual Flash Memory if enabled and configured

For more information, see 2.7.2, “Enhanced drawer availability” on page 60.

### Fan out card to I/O connectivity requirements

The optimum approach is to maintain maximum I/O connectivity during drawer removal. The redundant I/O interconnect (RII) function provides for redundant connectivity to all installed I/O domains in the PCIe+ I/O drawers.

### Preparing for Enhanced Drawer Availability

The Prepare Concurrent Drawer replacement option validates that enough dormant resources are available for this operation. If enough resources are not available on the remaining drawers to complete the EDA process, the process identifies those resources. It then guides you through a series of steps to select and free up those resources. The preparation process does not complete until all processors, memory, and I/O conditions are successfully resolved.

**Preparation:** The preparation step does not reallocate any resources. It is used only to record customer choices and produce a configuration file on the SE that is used to run the concurrent drawer replacement operation.

The preparation step can be done in advance. However, if any changes to the configuration occur between the preparation and the physical removal of the drawer, you must rerun the preparation phase.

The process can be run multiple times because it does not move any resources. To view the results of the last preparation operation, click **Display Previous Prepare Enhanced Drawer Availability Results** from the Perform Model Conversion window in the SE.

The preparation step can be run without performing a drawer replacement. You can use it to dynamically adjust the operational configuration for drawer repair or replacement before IBM SSR activity. The Perform Model Conversion window in you click **Prepare for Enhanced Drawer Availability** is shown in Figure 9-6 on page 419.

After you click **Prepare for Enhanced Drawer Availability**, the Enhanced Drawer Availability window opens. Select the drawer that is to be repaired or upgraded; then, select **OK**, as shown in Figure 9-7 on page 421. Only one target drawer can be selected at a time.



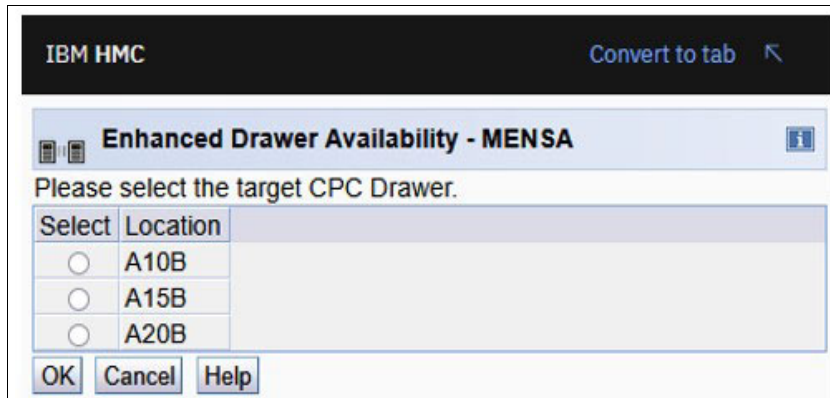


Figure 9-7 Selecting the target drawer

The system verifies the resources that are required for the removal, determines the required actions, and presents the results for review. Depending on the configuration, the task can take from a few seconds to several minutes.

The preparation step determines the readiness of the system for the removal of the targeted drawer. The configured processors and the memory in the selected drawer are evaluated against unused resources that are available across the remaining drawers. The system also analyzes I/O connections that are associated with the removal of the targeted drawer for any single path I/O connectivity.

If insufficient resources are available, the system identifies the conflicts so that you can free other resources.

The following states can result from the preparation step:

- ▶ The system is ready to run the EDA for the targeted drawer with the original configuration.
- ▶ The system is not ready to run the EDA because of conditions that are indicated by the preparation step.
- ▶ The system is ready to run the EDA for the targeted drawer. However, to continue with the process, processors are reassigned from the original configuration.

Review the results of this reassignment relative to your operation and business requirements. The reassignments can be changed on the final window that is presented. However, before making any changes or approving reassignments, ensure that the changes are reviewed and approved by the correct level of support based on your organization's business requirements.

### **Preparation tabs**

The results of the preparation are presented for review in a tabbed format. Each tab indicates conditions that prevent the EDA option from being run. The following tab selections are available:

- ▶ Processors
- ▶ Memory
- ▶ Single I/O
- ▶ Single Domain I/O
- ▶ Single Alternative Path I/O

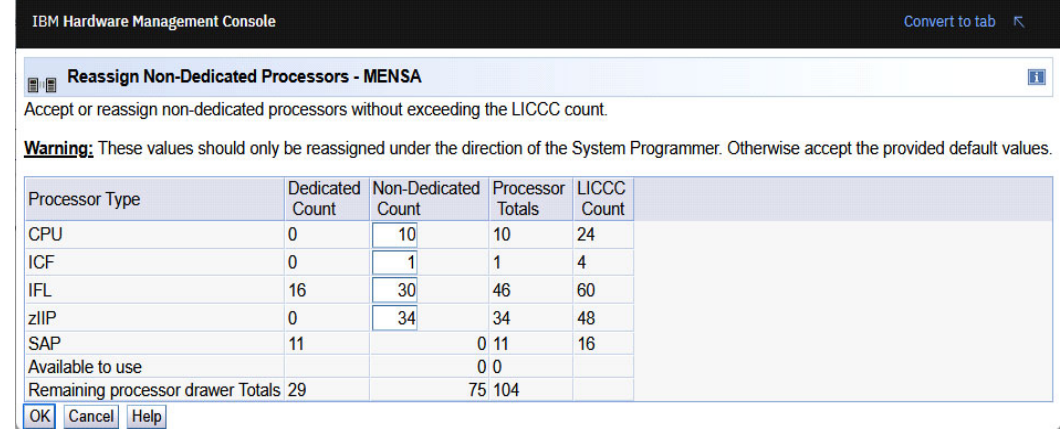
Only the tabs that feature conditions that prevent the drawer from being removed are displayed. Each tab indicates the specific conditions and possible options to correct them.



For example, the preparation identifies single I/O paths that are associated with the removal of the selected drawer. These paths must be varied offline to perform the drawer removal. After you address the condition, rerun the preparation step to ensure that all the required conditions are met.

## Preparing the system to perform enhanced drawer availability

During the preparation, the system determines the PU configuration that is required to remove the drawer. The results and the option to change the assignment on non-dedicated processors are shown in Figure 9-8.



IBM Hardware Management Console Convert to tab

**Reassign Non-Dedicated Processors - MENSA**

Accept or reassign non-dedicated processors without exceeding the LICCC count.

**Warning:** These values should only be reassigned under the direction of the System Programmer. Otherwise accept the provided default values.

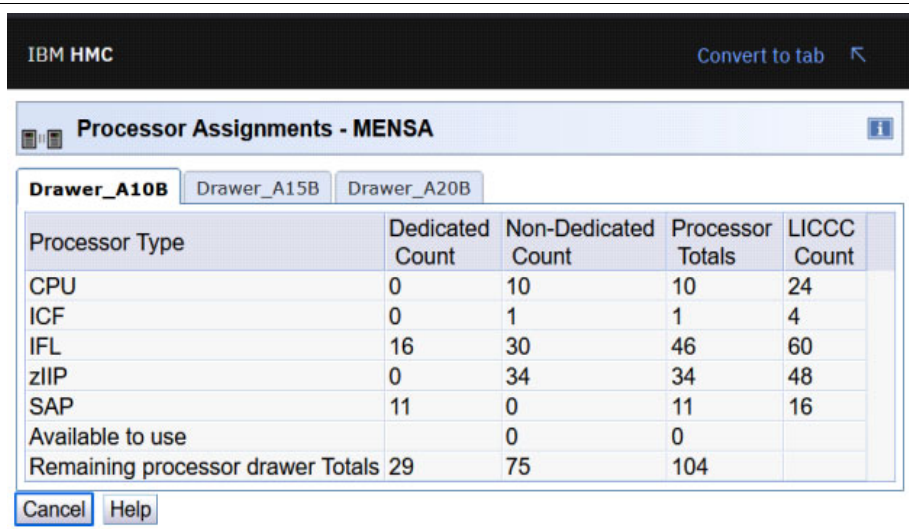
Processor Type	Dedicated Count	Non-Dedicated Count	Processor Totals	LICCC Count
CPU	0	10	10	24
ICF	0	1	1	4
IFL	16	30	46	60
zIIP	0	34	34	48
SAP	11		0 11	16
Available to use			0 0	
Remaining processor drawer Totals	29		75 104	

OK Cancel Help

Figure 9-8 Reassign Non-Dedicated Processors results

**Important:** Consider the results of these changes relative to the operational environment. Understand the potential effect of making such operational changes. Changes to the PU assignment, although technically correct, can result in constraints for critical system images. In specific cases, the solution might be to defer the reassignments to another time that has less effect on the production system images.

After you review the reassignment results and make any necessary adjustments, click **OK** (see Figure 9-9).



IBM HMC Convert to tab

**Processor Assignments - MENSA**

Drawer\_A10B Drawer\_A15B Drawer\_A20B

Processor Type	Dedicated Count	Non-Dedicated Count	Processor Totals	LICCC Count
CPU	0	10	10	24
ICF	0	1	1	4
IFL	16	30	46	60
zIIP	0	34	34	48
SAP	11	0	11	16
Available to use		0	0	
Remaining processor drawer Totals	29	75	104	

Cancel Help

Figure 9-9 Reassign Non-Dedicated Processors, message ACT37294



## Summary of the drawer removal process steps

To remove a drawer, the following resources must be moved to the remaining active drawers:

- ▶ PUs:
  - Enough PUs must be available on the remaining active drawers, including all types of PUs that can be characterized (CPs, IFLs, ICFs, zIIPs, SAPs, and IFP).
- ▶ Memory:
  - Enough installed memory must be available on the remaining active drawers.
- ▶ I/O connectivity:
  - Alternative paths to other drawers must be available on the remaining active drawers, or the I/O path must be taken offline.

By understanding the system configuration and the LPAR allocation for memory, PUs, and I/O, you can make the best decision about how to free the necessary resources to allow for drawer removal.

Complete the following steps to concurrently replace a drawer:

1. Run the preparation task to determine the necessary resources.
2. Review the results.
3. Determine the actions to perform to meet the required conditions for EDA.
4. When you are ready to remove the drawer, free the resources that are indicated in the preparation steps.
5. Repeat the step that is shown in Figure 9-6 on page 419 to ensure that the required conditions are all satisfied.

Upon successful completion, the system is ready for the removal of the drawer.

The preparation process can be run multiple times to ensure that all conditions are met. It does not reallocate any resources; instead, it produces only a report. The resources are not reallocated until the Perform Drawer Removal process is started.

### **Rules during EDA**

During EDA, the following rules are enforced:

- ▶ Processor rules
 

All processors in any remaining drawers are available to be used during EDA. This requirement includes the two spare PUs or any available PU that is non-LICCC.

The EDA process also allows conversion of one PU type to another PU type. One example is converting a zIIP to a CP during the EDA function. The preparation for the concurrent drawer replacement task indicates whether any SAPs must be moved to the remaining drawers.
- ▶ Memory rules
 

All physical memory that is installed in the system, including flexible memory, is available during the EDA function. Any physical installed memory, whether purchased or not, is available to be used by the EDA function.
- ▶ Single I/O rules
 

Alternative paths to other drawers must be available, or the I/O path must be taken offline.

Review the results. The result of the preparation task is a list of resources that must be made available before the drawer replacement can occur.



***Freeing any resources***

At this stage, create a plan to free these resources. The following resources and actions are necessary to free them:

- ▶ Freeing any PUs:
  - Vary off the PUs by using the Perform a Model Conversion window, which reduces the number of PUs in the shared PU pool.
  - Deactivate the LPARs.
- ▶ Freeing memory:
  - Deactivate an LPAR.
  - Vary offline a portion of the reserved (online) memory. For example, in z/OS, run the following command:
 

```
CONFIG_STOR(E=1),<OFFLINE/ONLINE>
```

This command enables a storage element to be taken offline. The size of the storage element depends on the RSU value. In z/OS, the following command configures offline smaller amounts of storage than the amount that was set for the storage element:

```
CONFIG_STOR(nnM),<OFFLINE/ONLINE>
```
  - A combination of both LPAR deactivation and varying memory offline.

**Reserved storage:** If you plan to use the EDA function with z/OS LPARs, set up reserved storage and an RSU value. Use the RSU value to specify the number of storage units that are to be kept free of long-term fixed storage allocations. This configuration allows for storage elements to be varied offline.

## 9.9 Concurrent Driver Maintenance

CDM is one more step toward reducing the necessity for and the duration of a scheduled outage. One of the components to planned outages is LIC Driver updates that are run in support of new features and functions.

When correctly configured, IBM z17 supports concurrently activating a selected new LIC Driver level. Concurrent activation of the selected new LIC Driver level is supported only at specific released sync points. Concurrently activating a selected new LIC Driver level anywhere in the maintenance stream is not possible. Specific LIC updates do not allow a concurrent update or upgrade.

Consider the following key points about CDM:

- ▶ The HMA can query whether a system is ready for a concurrent driver upgrade.
- ▶ Previous firmware updates, which require an initial machine load (IML) of the IBM z16 system to be activated, can block the ability to run a concurrent driver upgrade.
- ▶ A function on the SE allows you or your IBM SSR to define the concurrent driver upgrade sync point to be used for a CDM.
- ▶ The ability to concurrently install and activate a driver can eliminate or reduce a planned outage.
- ▶ IBM z17 allows Concurrent Driver Upgrade (CDU) cloning support to other CPCs for CDU pre-installation and activation.



- ▶ Concurrent crossover from Driver level  $N$  to Driver level  $N+1$ , then to Driver level  $N+2$ , must be done serially. No composite moves are allowed.
- ▶ Disruptive upgrades are permitted at any time, and allow for a composite upgrade (Driver  $N$  to Driver  $N+2$ ).
- ▶ Concurrently backing up to the previous driver level is not possible. The driver level must move forward to driver level  $N+1$  after CDM is started. Unrecoverable errors during an update might require a scheduled outage to recover.

The CDM function does not eliminate the need for planned outages for driver-level upgrades. Upgrades might require a system level or a functional element scheduled outage to activate the new LIC. The following circumstances require a scheduled outage:

- ▶ Specific complex code changes might dictate a disruptive driver upgrade. You are alerted in advance so that you can plan for the following changes:
  - Design data or hardware initialization data fixes
  - CFCC release level change
- ▶ OSA CHPID code changes might require PCHID Vary OFF/ON to activate new code.
- ▶ Crypto code changes might require PCHID Vary OFF/ON to activate new code.

**Note:** zUDX clients should contact their User Defined Extensions (UDX) provider before installing Microcode Change Levels (MCLs). Any changes to Segments 2 and 3 from a previous MCL level might require a change to the customer's UDX. Attempting to install an incompatible UDX at this level results in a Crypto checkstop.

### 9.9.1 Resource Group and native PCIe features MCLs

Microcode fixes, referred to as *individual MCLs* or *packaged in bundles*, might be required to update the Resource Group code and the native PCIe features. Although the goal is to minimize changes or make the update process concurrent, the maintenance updates at times can require the Resource Group or the affected native PCIe to be *toggled* offline and online to implement the updates.

The IBM z17 Coupling Express3 LR FC 0498 is the only native PCIe feature managed by Resource Group code.

Consider the following points for managing native PCIe adapters microcode levels:

- ▶ Updates to the Resource Group require all native PCIe adapters that are installed in that RG to be offline.
- ▶ Updates to the native PCIe adapter require the adapter to be offline. If the adapter is not defined, the MCL session automatically installs the maintenance that is related to the adapter.

The PCIe native adapters are configured with Function IDs (FIDs) and might need to be configured offline when changes to code are needed. To help alleviate the number of adapters (and FIDs) that are affected by the Resource Group code update, IBM z17 have four Resource Groups per system (CPC).

**Note:** Other adapter types, such as FICON Express, Network Express, OSA Express, zHyperLink2.0, and Crypto Express that are installed in the PCIe+ I/O drawer are not affected because they are not managed by the Resource Groups.



The front, rear, and top view of the PCIe+ I/O drawer and the Resource Group assignment by card slot are shown in Figure 9-10. All PCIe+ I/O drawers that are installed in the system feature the same Resource Group assignment.

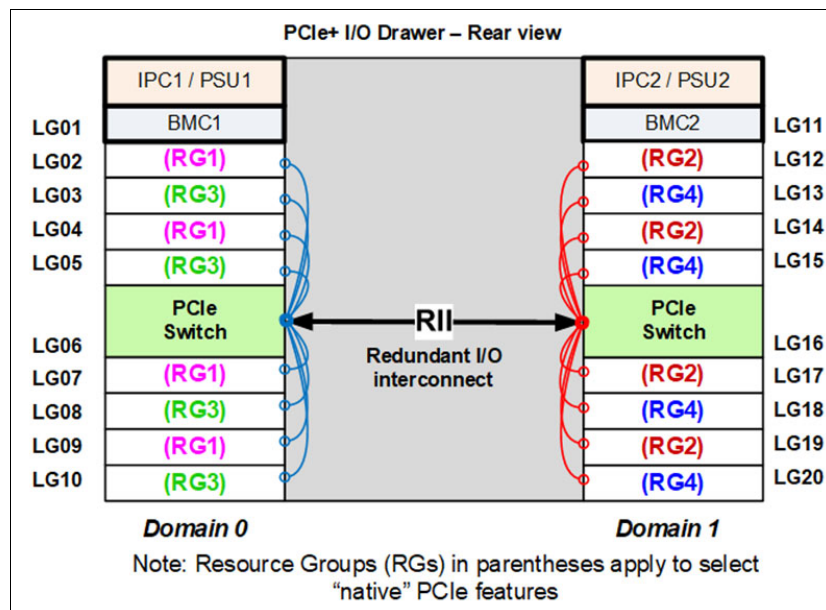


Figure 9-10 Resource Group slot assignment

## 9.10 RAS capability for the HMA and SE

The HMA and the SE include the following RAS capabilities:

- ▶ Back up from HMA and SE

On a scheduled basis, the HMA hard disk drive (HDD) is backed up to the USB flash memory drive (UFD), a defined FTP server, or both.

The primary SE and an alternative SE operating systems are virtualized along with the HMA. Backing up the SEs is still an available option.

For more information, see 10.2, "HMC and SE changes and new features" on page 430.

- ▶ Remote Support Facility (RSF)

The HMA RSF provides the important communication to a centralized IBM support network for hardware problem reporting and service.

For more information, see 10.4, "Remote Support Facility" on page 452.

- ▶ Microcode Change Level (MCL)

Regular installation of MCLs is key for RAS, optimal performance, and new functions. Generally, plan to install MCLs quarterly at a minimum. Review hiper MCLs continuously. You must decide whether to wait for the next scheduled apply session, or schedule one earlier if your risk assessment of the new hiper MCLs warrants.

For more information, see 10.6, "HMC, SE, and CPC microcode" on page 466.

- ▶ SE

IBM z17 servers are provided with two 1U trusted servers inside the IBM Z server A frame: One is always the primary SE and the other is the alternative SE<sup>4</sup>. The primary SE is the active SE. The alternative acts as the backup. Information is mirrored once per day. The



SE servers include N+1 redundant power supplies. The SEs hardware also runs the virtualized HMC code.

For more information, see 10.2.2, “HMC and SE server” on page 444.

---

<sup>4</sup> If HMA feature is installed on the system, special upgrade procedure must be followed to ensure nondisruptive SE upgrade.









# Hardware Management Console and Support Element

The Hardware Management Console (HMC) supports the functions and tasks that are required to manage the IBM Z. When tasks are performed on the HMC, the commands are sent to the Primary Support Element (SE) of the targeted system, which then issues commands to their respective central processor complex (CPC).

This chapter describes the newest and most important elements for the HMC and SE.

**Tip:** The Help function is a good starting point to get more information about all of the functions that can be used by the HMC and SE. The Help feature is available by clicking **Help** from the drop-down menu that appears when you click the **User menu** in the top upper right corner.

For more information, see the [IBM Z Documentation](#), select the applicable server and then **Library Overview**, select **Hardware Management (HMC) Version 2.17.0 help system content** or **Support Element (SE) Version 2.17.0 help system content**.

This chapter includes the following topics:

- ▶ 10.1, “Introduction” on page 430
- ▶ 10.2, “HMC and SE changes and new features” on page 430
- ▶ 10.3, “HMC and SE connectivity” on page 446
- ▶ 10.4, “Remote Support Facility” on page 452
- ▶ 10.5, “HMC and SE capabilities” on page 453
- ▶ 10.6, “HMC,SE, and CPC microcode” on page 466



## 10.1 Introduction

The HMC is a closed system (appliance), which means that no other applications can be installed on it. The HMC runs a set of management applications.

The HMC code runs on the two integrated 1U rack-mounted servers on the top of the IBM z17 A-frame. Stand-alone HMCs (Tower or Rack Mount) can no longer be ordered and are not supported.

HMC Driver 61/Version 2.17.0 is introduced with IBM z17. Driver 61 can be installed on z15 and z16 HMA HMCs.

With IBM z17, the HMA feature, installed on top of the A-frame, shares the integrated 1U server hardware with the SE code. The SE code runs virtualized under the HMC on each of the two integrated 1U rack-mounted servers. One SE is the Primary SE (active), and the other is the Alternative SE (backup). As with the HMCs, the SEs are closed systems (appliances), and other applications cannot be installed. The HMC is used to set up, manage, monitor, and operate one or more IBM Z CPCs. It manages IBM Z hardware and its logical partitions (LPARs), and provides support applications. At least one HMC is required to operate an IBM Z. An HMC can manage multiple IBM Z CPCs.

When tasks are performed at the HMC, the commands are routed to the Primary SE of the IBM Z. The SE then issues those commands to the target CPC.

**Note:** The new Driver level for HMC and SE for IBM z17 is Driver 61. Driver 61 is equivalent to Version 2.17.0.

Since IBM z15, several “traditional” SE-only functions were moved to HMC tasks. On HMC Driver 61/Version 2.17.0, these functions appear as native HMC tasks but run on the SE. These HMC functions run in parallel with Single Object Operations (SOOs), simplifying and streamlining system management. For more information about SOOs, see “Single Object Operations” on page 456.

For more information about the HMC, see also the [IBM Z Hardware Management Console Videos](#).

**Note:** HMC 2.17.0 supports managing IBM z15, IBM z16, and IBM z17 IBM Z server generations (N-2),

## 10.2 HMC and SE changes and new features

The initial release included with IBM z17 is HMC and SE Driver 61/Version 2.17.0. Check the “What’s new” widget (available in the HMC Dashboard) to examine the new features.

For more information about HMC and SE functions, use the HMC and SE console help system or see [IBM Resource Link](#) (login required), select **Library**, the applicable server and then, select **Hardware Management Console Operations Guide** or **Support Element Operations Guide**.

Further, in the HMC Dashboard, under **Helpful Links**, you will find links to Resource Link, videos, APIs, and so on under Helpful links. Under **About this HMC**, you can find information like the installed Bundle level. If the HMC is part of an HMA, this HMC shows if it is currently



running the Primary or Alternate SE. Also, the name of the peer HMC, where the 2nd SE is running, is displayed in Figure 10-1.

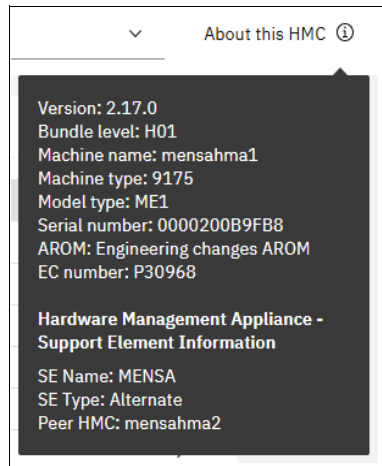


Figure 10-1 Example of “About this HMC” Driver 61/Version 2.17.0 HMC and SE new features

The following support was added with Driver 61/Version 2.17.0:

### Import/Export from Remote Browsing File System

As with Driver 61/Version 2.17.0, only HMAs are supported; most clients will use the HMC via remote connections. For the tasks listed below, Import/Export from Remote Browsing File System is supported:

- ▶ Fibre Channel Endpoint Security
- ▶ Secure Boot Certificate Management
- ▶ Certificate Management
- ▶ System Input/Output Configuration Analyzer
- ▶ Analyze Console Internal Code
- ▶ Change Console Internal Code
- ▶ FCP Configuration
- ▶ Audit log Scheduled Operations
- ▶ Export/Import IOCDS
- ▶ Save/Restore Customizable Console Data
- ▶ Crypto Configuration
- ▶ View/Archive Security Logs
- ▶ Advanced Facilities
- ▶ OSA Advanced Facilities
- ▶ Crypto UDX configuration
- ▶ Transmit Service Data
- ▶ Transmit VPD
- ▶ Import Secure Execution Keys
- ▶ Export/Import Profiles



- ▶ Reassign HMC
- ▶ Manage Firmware Features
- ▶ System I/O Configuration Analyzer
- ▶ Perform Model Conversion

An example of how an export to a File System is shown in Figure 10-2.

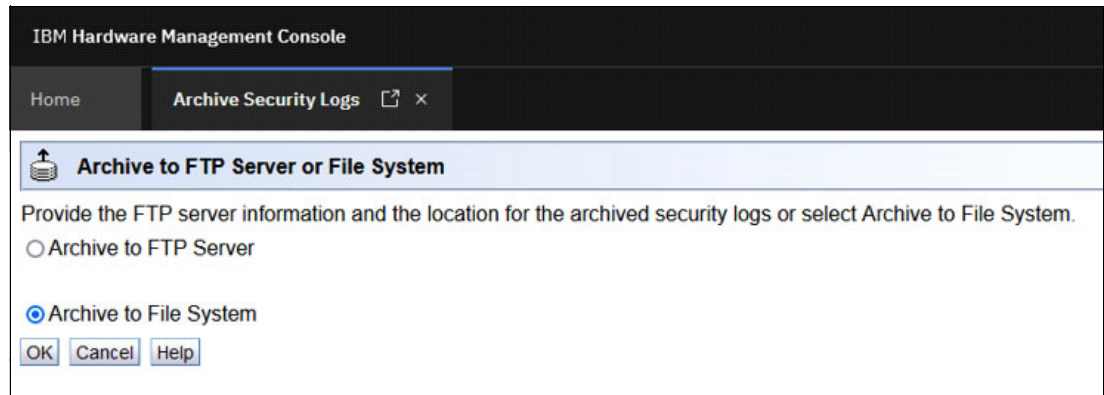


Figure 10-2 Example of export Archive Security Logs to File System

## Single Sign On (SSO)

SSO is a new authentication method for the HMC/SE. It allows users to use credentials from other services and is based on OpenID Connect (OIDC) technology.

The SSO possibilities are implemented for the different options in the HMC User Management. First, you must define your SSO servers in the SSO task, as shown in Figure 10-3 on page 381.

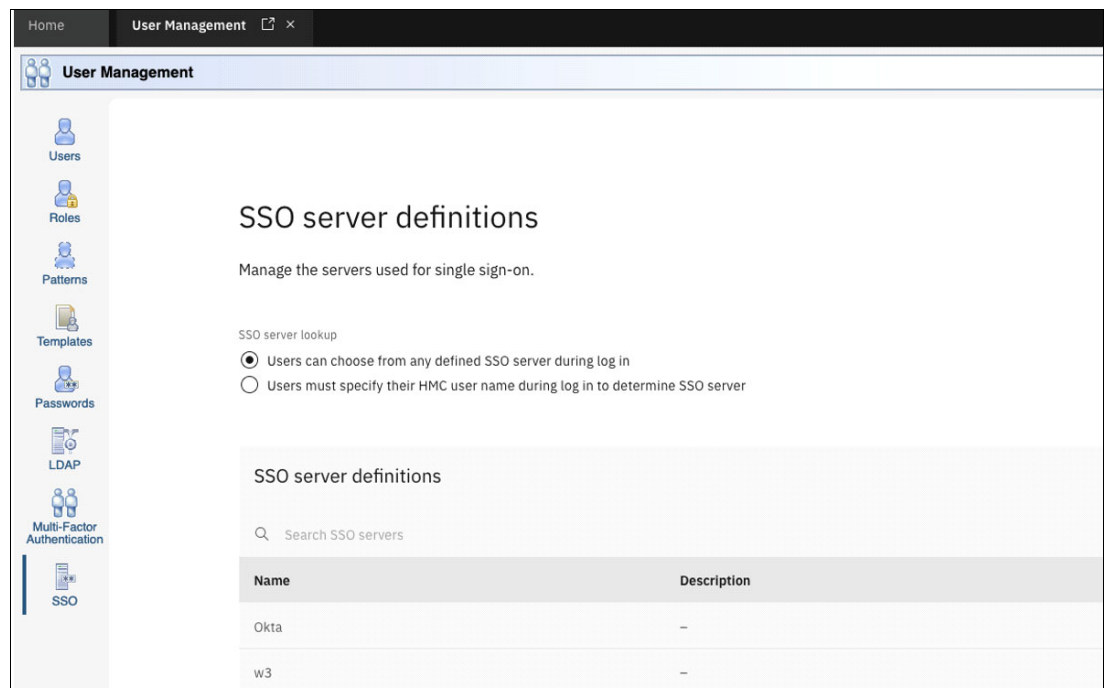


Figure 10-3 Example of SSO server definitions



If you create a new User, Template, or Pattern you can allocate the according SSO servers.

## External Time Source (ETS) enhancements for STP and HMC NTP

The following enhancements are available in the ETS for STP:

- ▶ It is now supported to have 3 Network Time Protocol (NTP) ETS servers
- ▶ Support for Mixed Mode (use of both NTP and PTP ETS servers in parallel)
- ▶ Allow use of certificates for secure Network Time Security (NTS) NTP communication between the Central Processor Complex (CPC) and configured ETS(es)
- ▶ Allow use of advanced monitoring commands for NTP and PTP
- ▶ NTS support for the STP ETS NTP and for the HMC NTP connection
- ▶ For PTP you can choose the Domain number and between Multicast and Unicast.

For more information, see the *IBM Z Server Time Protocol Guide*, [SG24-8480](#).

## OSA-ICC CA Signed Shared Certificates

It is now supported to import CA-signed certificates in shared scope. Once you import a CA-signed certificate in shared scope, the panel will replace a certificate that is shared across OSA PCHIDs with the imported CA-signed certificate. The new certificate will not be applied on the online OSA PCHIDs until they are configured off, and then back on.

## Propagate HMC Certificates to HMC managed CPCs (/theirs SEs)

Certificates imported to the HMC Certificate Management task will also be imported to the IBM z17 SEs. Functions like LDAP, MFA, or FTPS profit from it. For example, locally defined SE users configured for LDAP authentication can now log into the SE with trusted LDAP. Certificates deleted from the HMC Certificate Management task will also be deleted from managed IBM z17 SEs. If multiple HMCs are managing the same SE, then the Certificate is only deleted if no other HMC has this Certificate imported.

If the HMC and SE stop communicating, the Certificates associated with that HMC are removed from the SE until communications are back up.

Figure 10-4 shows an example of the new message if you import an HMC Certificate.

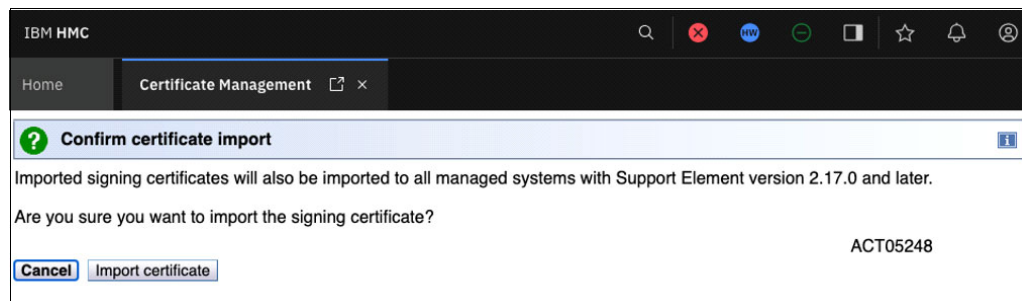


Figure 10-4 Example new message for HMC Certification import

### ▶ New audit log scenarios

When you import or delete a Certificate on an HMC, you will see a corresponding HMC Audit log entry. If a Certificate is imported or deleted on a SE, you will see a corresponding SE Audit log entry.

List of events that have implications on Certifications and corresponding Audit log entries:

- Adding or removing a CPC/SE as Object Definition



- Data Replication of Certificates
- Restore Customizable Console Data
- Restore Upgrade Data

### **Remote Code Load (RCL) enhancements**

Some of the enhancements to RCL with the IBM z17:

- ▶ HMC RCL Alerts - these alerts can be sent via the HMC UI, e-mail, Web Services API, or IBM HMC Mobile notifications, in addition to IBM Monitoring e-mail, which provides information on:
  - RCL Scheduled
  - RCL Running
  - RCL Completed
  - Additional RCL Health checks
    - e.g., condition of Alternate SE switchover capability needed for HMA update
  - RCL Health checks are available on IBM Resource Link before or at the scheduling time of these Health checks
  - Block RCL scheduling rather than RCL scheduling failing after further verification on the HMC
  - Can be used to make the client aware to issues and allow them to address any blocking RCL conditions prior to the need to schedule RCL
  - Autopopulate RCL Backup location
- ▶ Several other RCL infrastructure enhancements were made which should provide:
  - Overall better client scheduling/notification experience
  - Higher quality execution of RCL

For more information see 10.6.1, “Remote Code Load (RCL)” on page 468.

### **HMC Dual Control**

Dual Control is an optional security feature designed to ensure that sensitive actions or operations cannot proceed without approval from a second person. Its purpose is to minimize the negative impacts caused by the inexperience or malicious intent of a single employee, comply with security and compliance regulations, and reduce cost-incurring task errors by requiring the approval of a second employee before allowing an operation to proceed. Dual Control supports Web Services API. More information can be found in the Hardware Management Console Web Services API (Version 2.17.0), SC27-2646 document.

Figure 10-5 on page 435 shows an example of a Dual Control request for deactivating a LPAR.



**IBM Hardware Management Console**

Home User Management x Deactivate - A32:CF1 x

### Create dual control request

This task requires approval before you can run it. Create a dual control request that includes an approval due date and instruction that indicates how you want to proceed after receiving an approval. You can also provide a description of the request, and comments for the approvers.

**Request name** Approval due date

Deactivate - A32:CF1 09/25/2024

**Description (optional)** 57/1024

Deactivation of the CF1 partition per direction from BDV.

**Instructions for running the approved task**

☒ Run immediately ⓘ

☐ Run at a specific date and time ⓘ

☐ Run the task manually ⓘ

**Comment (optional)** 81/1024

Please approve this deactivate request as discussed in planning meeting with BDV.

Help Cancel Submit request

**GUIDANCE**

After you send the request, reviewers are notified and either approve or deny the request. You can track the status of your request through the Dual Control Management task.

If your request is approved, the task is run according to the instruction that you select.

Figure 10-5 Create Dual Control request interface

Key features of Dual Control include:

- ▶ **Role-based task and object enablement**  
Security Administrators can specify which tasks, objects, and users require Dual Control, as well as designate which users are authorized to approve Dual Control requests.
- ▶ **Flexible request options**  
Users can submit Dual Control requests for supported tasks with different run options.
- ▶ **Real-time notifications**  
Users are notified about the status of their Dual Control requests through the HMC, as well as external methods like e-mail, HMC mobile notifications, and APIs, providing near real-time visibility
- ▶ **Request management and tracking**  
Both requestors and approvers can track the status of Dual Control requests via the Dual Control management task.
- ▶ **Approver autonomy**  
Approvers can assign themselves to requests and are given relevant information to make well-informed decisions on whether to approve or reject a Dual Control request

The first release of Dual Control will support the following tasks:

- ▶ Activate
- ▶ Deactivate
- ▶ Stop (DPM)



- ▶ Reset
- ▶ Load
- ▶ Change LPAR Cryptographic controls
- ▶ Perform Model Conversion

Dual Control is available on IBM z15 and IBM z16 systems when connected to an HMC updated to Driver 61/Version 2.17.0. For SE versions below 2.17.0, it remains supported but with certain limitations:

- ▶ Users should remove the Single Object Operations permission from any role with Dual Control enabled on an IBM z15 and IBM z16. These systems' SEs do not natively support dual control, and users could bypass HMC dual control enforcement by performing specific supported actions, such as activating the SE.
- ▶ Dual Control is not supported for Change LPAR Cryptographic Controls.
- ▶ Dual Control is not supported for Perform Model Conversion.

### **TLS Cipher Suite Configuration Improvement**

The IBM z16 added HMC and SE support for TLS 1.3. Before setting the TLS level to 1.3, you must ensure that all services and servers that are connecting by way of TLS to the HMC and SE support TLS 1.3 as well (for example, the remote browser, LDAP Authentication Servers, Web Services API connections, Fibre Channel End Point Security, and FTPS, Single Object Operations).

If minimum TLS version is set to 1.2, TLS 1.3 is attempted first and then falls back to TLS 1.2 if required.

The Customize Console Services task (for HMC and SE), Configure TLS Settings, allows you to set both the minimum TLS protocol version and the enabled TLS cipher suites. This task identifies the protocol versions each cipher suite is compatible with and only shows those that can be used with the chosen minimum TLS protocol version.

Figure 10-6 on page 437 shows the HMC window that allows the selection of the wanted TLS version.



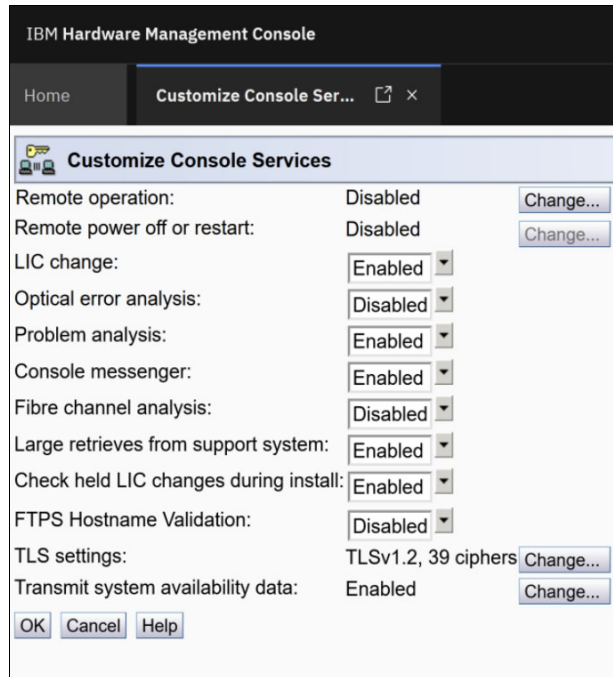


Figure 10-6 Customize Console Services window

A single Customize Console Services sub-task, Configure TLS Settings, is provided to configure the minimum TLS protocol version and the enabled TLS cipher suites.

The task identifies the protocol version(s) each cipher suite can use and only displays those that can potentially be used with the selected minimum TLS protocol version.

Figure 10-7 shows the usable cipher suites.



Figure 10-7 Configure TLS Settings window

In the Audit Log of the HMC and SE you can prove the according settings.

## BCPii enhancements

We enable you to handle HMC/SE UI functions automated via BCPii. Since z15 BCPii v2 is available, BCPii v1 & SNMP are being deprecated. We recommend you move to BCPii v2 if



you have not already. Below, you can find most of the new enhancements with IBM z17 for BCPii v2.

More information can be found in the *Hardware Management Console Web Services API (Version 2.17.0)*, SC27-2646 and 10.5.9, “Automated operations via APIs” on page 464. On the HMC Dashboard -> Helpful links -> APIs is a shortcut to find the appropriate documentation.

### **Summary of API version updates**

In general there are lot of changes / updates in the HMC Web Services API. Not all are mentioned here in this Redbook. Please see the *Hardware Management Console Web Services API Version 2.17.0*, SC27-2646, “Summary of API version updates” for all new functions.

### **Asynchronous Notification z/OS Support for BCPii v2**

Asynchronous Notification z/OS Support for BCPii v2 requires BCPii enhanced security introduced with IBM z17. Access control is dependent on the permissions for the associated HMC user identity. This means the HMC will limit the sent notifications to only those permitted, and z/OS will act as a simple transport for notifications messages.

The SE notifications are limited to those objects which are exposed to BCPii. For example /api/users APIs are not exposed so Property/Inventory Change notifications for users won't be generated.

The HMC notifications are only restricted by permissions of the associated HMC identity.

Registration is made available via a set of API operations only available to the BCPii interface, which can be seen in Table 10-1.

*Table 10-1 API operations to registering for Asynchronous Notifications*

GET	<i>/api/sessions/operations/get-notification-registrations</i>	Gets the existing notification registrations for a session.
POST	<i>/api/sessions/operations/register-for-notifications</i>	Create a new notification registration. Clients can fine tune which notifications they are interested in via event name and object filters.
POST	<i>/api/sessions/operations/update-notifications-registration</i>	Update an existing notification registration for the session.
POST	<i>/api/sessions/operations/delete-notifications-registration</i>	Delete an existing notification registration for the session.

There are no authorization requirements to register for notifications. All permission checking is done when a notification is to be sent. Clients can only list and modify registrations pertaining to their own session. In this context a session is related to the z/OS user ID making the request.

### **BCPii HMC Targeting with Enhanced Security**

BCPii v2 is able to target an HMC with a request. A BCPii request is associated with a real HMC user or template. BCPii passes the z/OS User ID with the request. The z/OS User ID is inside a JSON Web Token (JWT), and the JWT is signed.

When the HMC receives the request, it authenticates the JWT content (validates JWT is not expired and has a valid signature and ensures there is an HMC user mapping for the



associated z/OS User ID). If the JWT is valid, there will be an API sessions established for the HMC identity mapped to the z/OS User ID and use the identity associated with the API session to authorize the request. The HMC needs to be configured accordingly. The following tasks has to be followed: **Customize Console Services -> BCPii authorization -> Change....** Here you define the according users.

**Data Replication** and **Save/Restore Customizable Console Data** has new data type *BCPii Authorisation Data* to replicate, save, and restore.

### **BCPii HWIREST/v2 functional enhancements**

The following new functional BCPii HWIREST/v2 enhancements are introduced (non exhaustive):

- ▶ REST API support for UI task: **I/O Element Management**. Associated API feature: *io-element*.
- ▶ REST API support for UI task: **Object Definition**. Associated API feature: *add-remove-object-definition*
- ▶ REST API support for UI tasks regarding Dual Control. Associated API feature: *dual-control*
- ▶ REST API support for UI tasks regarding **Customize/Delete Activation Profiles**. Associated API feature: *update-profile-apis*
  - Image Activation Profile ‘Data Model’: *hmac-key-enablement / load-timeout*
- ▶ REST API support for UI tasks: **View Code Level**. Associated API feature: *adapter-code-level*
- ▶ REST API support for power consumption. Associated API feature: *power-consumption-monitoring*

### **New HMC user Data replicated to SE**

The long-term goal is to provide the ability to manage all user data in a single place (HMC). Since z15, standard HMC user definitions are replicated to the managed SEs. You can utilize these HMC user definitions to logon locally on the SE. New HMC-defined user patterns and templates are also replicated in the managed SEs.

**Note:** We recommend to do all HMC user data definitions (not just users) only on the HMC. User Management on the SE should no longer be necessary.

Things to consider:

- ▶ Locally defined SE users/patterns/templates take priority over inherited HMC user definitions
- ▶ HMC user definitions do not become managed on the SE and will not show in User Management
- ▶ LDAP authentication is executed through an HMC
- ▶ MFA enabled users are not supported when logging on locally on the SE, only HMC users via Single Object Operations to SE

### **Support a timeout interval for List-directed IPL's**

A List-Directed IPL is an enhanced IPL process with the ability to pass parameters and not just a loadparm. A list-directed IPL is an IPL that is initiated when an IPL command or IPL directory statement specifies one of the following operands:

- ▶ `fcv_vdev`



- ▶ LOADDEV
- ▶ DUMPDEV (DUMPDEV is not a valid operand for IPL directory statements)

In these cases, the IPL command or IPL directory statement operands do not provide all the required information for the IPL.

List-directed IPL parameters (LOADDEV and DUMPDEV parameters) provide the IPL information that the IPL command or IPL directory statement operands don't offer. We have added the ability to customize the timeout value for this IPL. Up until now, it has been a hardcoded value of 60 seconds. Since most list-directed IPL's occur on FCP devices, this value was not always sufficient.

The new timeout value defaults to 300 seconds, but you can now choose from a range of 60 to 600 seconds.

The timeout value is supported through the LOAD and IMAGE profiles and if you are using the Web-Services API (WSAPI).

### Hardware Message SNMP traps include Case Number

Hardware message traps are created for every hardware message created on an HMC or SE. Some of these hardware messages are related to cases. To enable the identification of the messages relating to cases, a new field has been added to hardware message traps for the Case Number associated with the message. (An empty string means no Case Number associated with the message.)

Hardware message delete traps have the same format as hardware message traps; these will also have the new Case Number field.

SNMP traps are defined in SNMP MIB file format; the hardware message trap has the following fields, each field has an SNMP object identifier (OID) and a value.

- ▶ OID for the object the message is associated with, Value: 1 (hardware message)
- ▶ OID for the message type, Value: 1 (hardware message)
- ▶ OID for the message text, Value: Message text string
- ▶ OID for the message refresh indicator, Value: Boolean
- ▶ OID for the message time stamp, Value: Time stamp string
- ▶ OID for the list of CPC images associated with the message, Value: Image list string
- ▶ OID for the name of the object associated with the message, Value: Name of the associated object
- ▶ OID for the case number associated with the message, Value: case number string  
This is a new field that will contain the Case Number if the message is related to a case or the empty string if not related to a case.

If you are already sending SNMP Traps, then you have nothing to implement.

### Remote HMC Application Restart

The **Remote power off or restart** in the **Customize Console Services** option has been enhanced to allow additional granularity. You can now also choose "Restart application." This can be important now as standalone HMCs are no longer supported with IBM z17. You can now choose between the following options:

- ▶ Disabled: Nothing allowed remotely. Prevents selection of other options.
- ▶ Restart application: Remote application restart is permitted.



- ▶ Restart console: Remote restart (aka reboot) is permitted.
- ▶ Power off: Remote power off is allowed.

In Figure 10-8 you can see an example of these settings on the HMC.

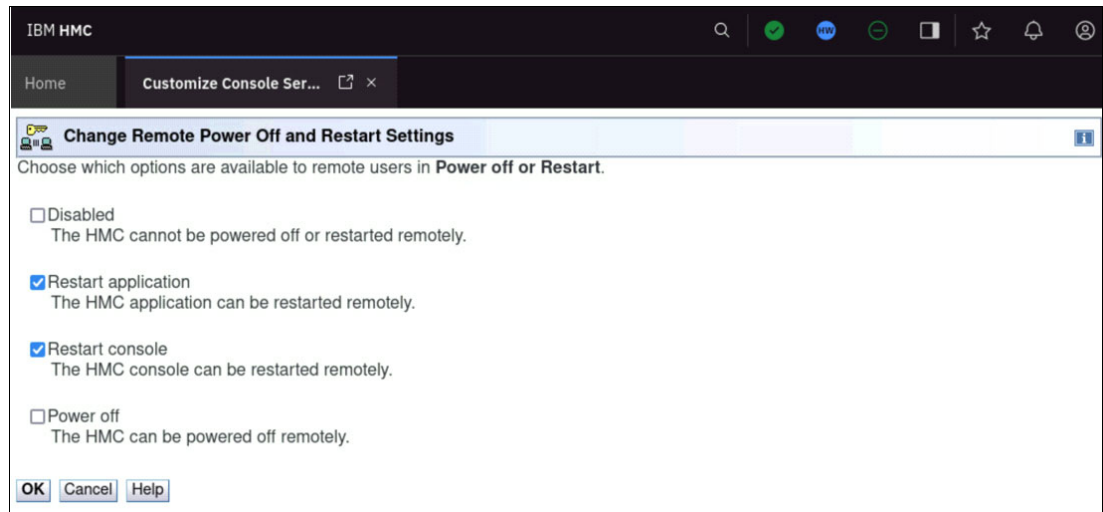


Figure 10-8 HMC Change Remote Power Off and Restart Setting

**Note:** To change these settings you have to login locally on the HMC.

## I/O Element Management

An I/O Element is responsible for management and execution of storage and networking protocol adapters across two I/O drawer domains. The **I/O Element Management** task provides the ability to:

- ▶ View I/O Elements and associated adapters
- ▶ Apply pending firmware updates to an I/O Element

I/O Element Firmware Updates require that all associated adapters are Configured Off/On. This may be a disruptive action to the I/O adapters depending on redundancy. We expect very rare cases in which I/O Element Firmware Updates are required.

Figure 10-9 on page 442 show an example of the task I/O Element Management.



I/O element management

View and manage I/O elements.

Q

Search I/O elements and adapters

Element ID	Status	Mode	Update status	
01	Operating	IO	No updates	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>

Figure 10-9 Example of the task I/O Element Management

Suppose you start an Update Firmware for an I/O Element ID. In that case, all associated Adapters to this I/O Element ID may be taken offline via the Operating System - if you continue, they will be forced offline. A disruptive confirmation panel will be displayed to confirm potentially disruptive changes to online I/O Adapters. The action has to be confirmed via password.

These new and other possible pending conditions can also be checked via **System Information -> Query Additional Actions**.

## HMC Mobile update

There are recent updates on HMC Mobile:

- ▶ Dual Control handling
- ▶ Sustainability metrics (System and Partition Power Consumption)
- ▶ Survey intercept

For more information on this topic see “IBM HMC Mobile” on page 457.

## Report a Problem using HMC Web Services API interface

You can report a problem for the HMC via task **Service Management -> Report a Console Problem**. You can report a problem for the CPC/LPAR via task **Service -> Report a Problem (RaP)**.

Now you can also perform these tasks with the Web Services API interface:

- Report a Console Problem can be invoked with the following URI:
  - `/api/console/operations/report-problem`
- Report a CPC Problem can be invoked with the following URI:
  - `/api/cpcs/{cpc-id}/operations/report-problem`  
Where {cpc-id} is the Object ID of the CPC object.
- Report a Logical Partition Problem can be invoked with the following URI:
  - `/api/logical-partitions/{logical-partition-id}/operations/report-problem`  
Where {logical-partition-id} is the Object ID of the Logical Partition object.
- Report a Partition Problem can be invoked with the following URI:



- `/api/partitions/{partition-id}/operations/report-problem`  
Where {partition-id} is object ID of the Partition object.

## 10.2.1 Hardware Management Appliance

Starting with IBM z15, the two 1U rack-mounted servers on the top of the A-frame provide increased hardware capacity, which allows instances of HMC and SE to run collocated on the same physical server. The SE code runs as a virtual guest of the HMC code.

The SE interface can be accessed from the HMC using the Single Object Operation on the HMC.

The HMA feature FC 0355 consists of the HMC code that is installed on the two 1U rack-mounted servers on the top of the A-frame, which are collocated with the SE code. The servers are configured with processor, memory, storage, and networking resources to support all processing and security requirements for running HMC and SE code.

**Recommendation:** We highly recommend at least one HMA feature per location/data center. However, at a maximum of 2 HMA per location/data center, as otherwise, the management, cabling, and microcode updates can be too much of an effort.

The two HMCs (HMC1 and HMC2; these names can be changed) from manufacturing are configured as independent HMCs. They are *not* Primary or Alternative HMCs. HMC Data Replication can be established, if wanted.

The SE runs as a guest of the HMC. The two SE code instances are clustered for high availability. One SE code runs the Primary SE the other Alternative SE. These SEs perform data mirroring and their role can be switched for maintenance purposes.

Switching the Primary and Alternative SE roles is essential because HMC microcode maintenance can be performed only on the server that runs the Alternative SE as a guest.

**Important:** With the IBM HMA, shutdown or restart of the HMC that includes the Primary SE code as guest also restarts the Primary SE code. An application restart of the HMC is not disruptive to the guest SE code.

If the HMC, which receiving microcode updates, runs the Primary SE guest, an SE switchover must be performed. Figure 10-10 shows the HMA relation to the HMCs and SEs.

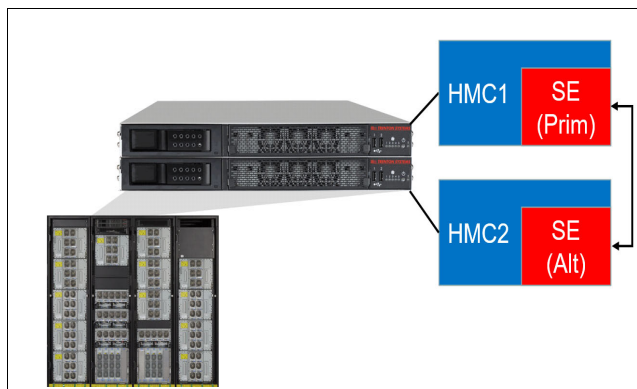


Figure 10-10 HMA: HMCs and SEs



## 10.2.2 HMC and SE server

The two 1U rack-mounted hardware servers that are installed at the top of the IBM z17 A frame (as shown in Figure 10-11) are used for HMA HMC (optional feature) and SE functions.



Figure 10-11 HMC and SE server location (front view)

### HMC and Support Element Keyboard Mouse Monitor

With IBM z17, a Keyboard Mouse Monitor (KMM) device that is in the front of rack A under the 1U servers replaces the previous KMM assembly that was mounted on a swing gate in IBM z14.

The KMM device is shown in Figure 10-12.

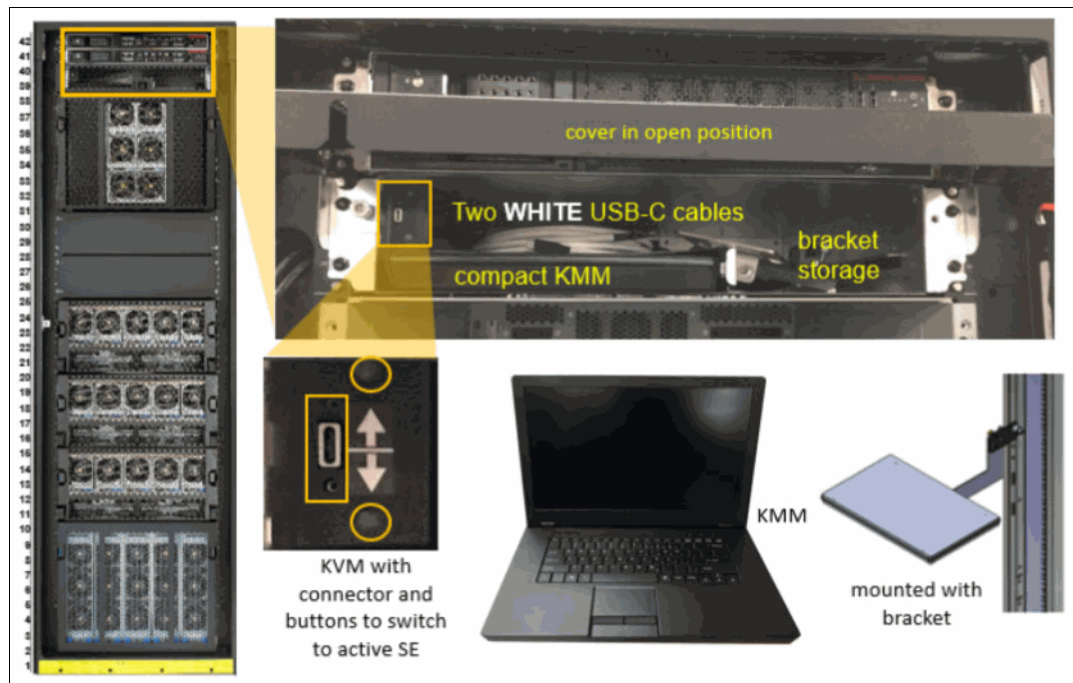


Figure 10-12 HMC / SE KMM device inside A frame



Consider the following points:

- ▶ The device is intended to be used by the IBM System Service Representative (SSR) only. If remote access to the HMC and SE is not possible, the customer can use the KMM as an emergency option. Local SE can be managed by the Virtual Support Element Management task, as shown in Figure 10-13.

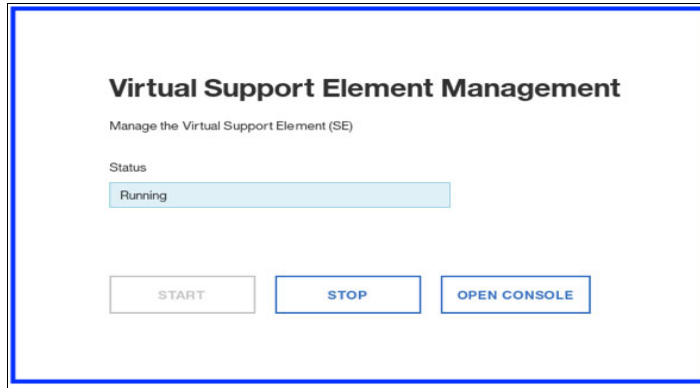


Figure 10-13 Local Virtual Support Element Management

- ▶ One KMM is provided.
- ▶ The KMM is stored at the front of the A-frame, below the two 1U servers. A USB-C cable and mounting bracket are stored with the KMM.
- ▶ The USB cable can be used to plug the device into a KVM switch at the front or the rear of the rack when servicing the system.
- ▶ Switching between servers is done using a button that is on the KMM. The KMM screen also indicates which server is selected.
- ▶ The KMM mounting bracket can be used to mount the device to any frame in the system (front or rear).
- ▶ The KMM can be used on any IBM z17 system (no affinity to the system with which it is shipped exists).

For more information about the KMM and how to attach it to the A frame, see *9175 Installation Manual*, GC28-7050.

### 10.2.3 USB support for HMC and SE

This section describes two methods for service and functional operations for HMC Driver 61/Version 2.17.0.

#### Microcode load

Microcode can be loaded by using the following options:

- ▶ USB
 

If the HMC and SE code is included with a USB drive when a new system is ordered, the load procedures can be done using the USB drive
- ▶ Electronic
 

If USB load is not allowed or if FC 0846 (no physical media options) is ordered, an ISO image is used for a firmware load over a local area network (LAN). The ISO image can be downloaded through zRSF or an FTP (FTPS/SFTP) server accessible by the LAN.



**Important:** The ISO image server *must* be in the same IP subnet with the target system to load the HMC or SE code.

### Operating system load from removable media or server

z/OS, z/VM, and Linux on Z are available via a USB or network distribution. z/TPF does not use the HMC for code load.

## 10.2.4 SE driver and version support with the HMC Driver 61/Version 2.17.0

- ▶ The driver of the HMC and SE is equivalent to a specific HMC and SE version
- ▶ Driver 41 is equivalent to Version 2.15.0
- ▶ Driver 51 is equivalent to Version 2.16.0
- ▶ Driver 61 is equivalent to Version 2.17.0

An HMC with Driver 61/Version 2.17.0 supports N-2 IBM Z server generations. Some functions that are available on Driver 61/Version 2.17.0 and later are supported only when the HMC is connected to an IBM z17 with Driver 61/Version 2.17.0.

The SE drivers and versions that are supported by the HMC Driver 61/Version 2.17.0 are listed in Table 10-2.

*Table 10-2 SEs that are supported by HMC Driver 61/Version 2.17.0*

IBM Z product name	Machine type	SE driver	SE version
IBM z17 ME1	9175	61	2.17.0
IBM z16 A02	3932	51	2.16.0
IBM z16 A01	3931	51	2.16.0
IBM z15 T02	8562	41	2.15.0
IBM z15 T01	8561	41	2.15.0

## 10.3 HMC and SE connectivity

The HMC and SE must be connected with copper Ethernet RJ45 cables through a client-provided Switch infrastructure.

**Security:** The configuration of network components, such as routers or firewalls, is beyond the scope of this document. Whenever the networks are interconnected, security exposures can exist. For more information about HMC security, see *Hardware Management Console Security*, SC28-7061, and *Integrating IBM Remote Support into your Enterprise*, SC28-7060.

For more information about the HMC settings related to access and security, see [IBM Resource Link](#). On that web page, select **Library**, **IBM Z**, **IBM z17** then **Library Overview**.



### 10.3.1 Hardware Management Appliance (HMA) connectivity

On IBM z17, two HMCs are delivered with the HMA FC 0355. The HMA HMC code runs on the two integrated 1U rack-mounted servers on the top of the A-frame. Stand-alone HMCs (Tower or Rack Mount) can no longer be ordered and are not supported on IBM z17.

With IBM z17 and HMA, the SE code runs virtualized on the two integrated HMCs on the two integrated 1U rack-mounted servers on the top of the IBM z17 A-frame. One SE is the Primary SE (active), and the other is the Alternative SE (backup).

The HMC communicates with the SE through a customer-supplied Ethernet switch infrastructure (two switches are recommended for redundancy). With the HMA, each HMC and each SE has its own two physical Ethernet RJ45 ports, as shown in Figure 10-14 on page 396. Each HMC and SE requires at minimum one physical Ethernet connection to the customer-supplied Ethernet Switch. These are a total of four physical copper RJ45 Ethernet cables. For redundancy or, for example, separation of different services, you may connect all eight physical RJ45 ports; then you need eight physical Ethernet copper RJ45 cables. The red lines highlight the Ethernet connectivity for the SEs; the green lines highlight the Ethernet connectivity for the HMA HMCs.

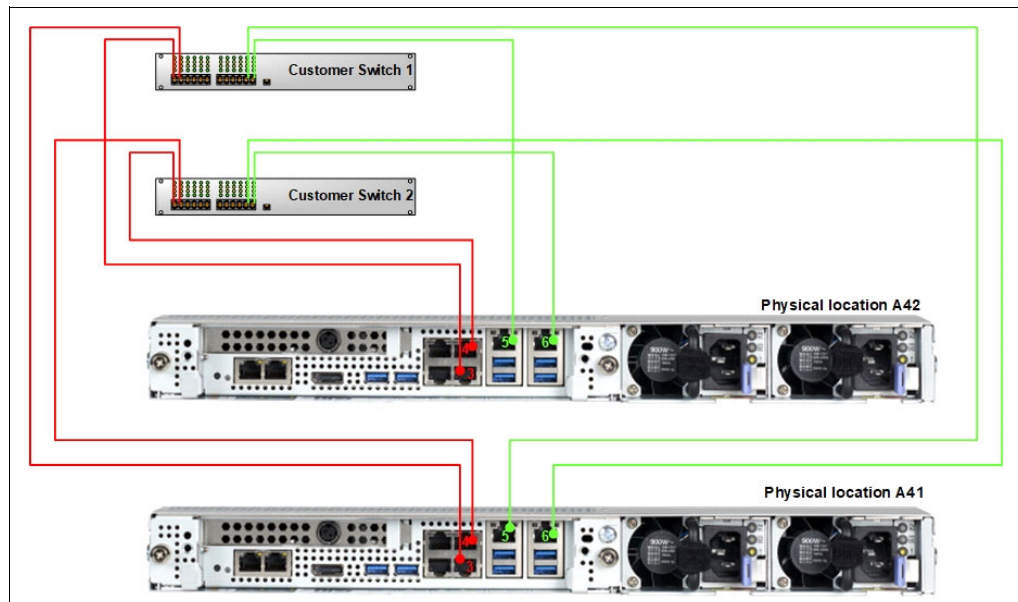


Figure 10-14 HMA connectivity

**Note:** The HMA HMCs *must* be connected to the SEs using a customer-provided switch infrastructure. Direct connection between the HMCs and the SEs is *not* supported.

### 10.3.2 Support Element (SE) connectivity

Each SE has two Ethernet RJ45 ports. An example of these connections made without an HMA for the IBM Z where the SEs are installed is shown in Figure 10-15 on page 397.



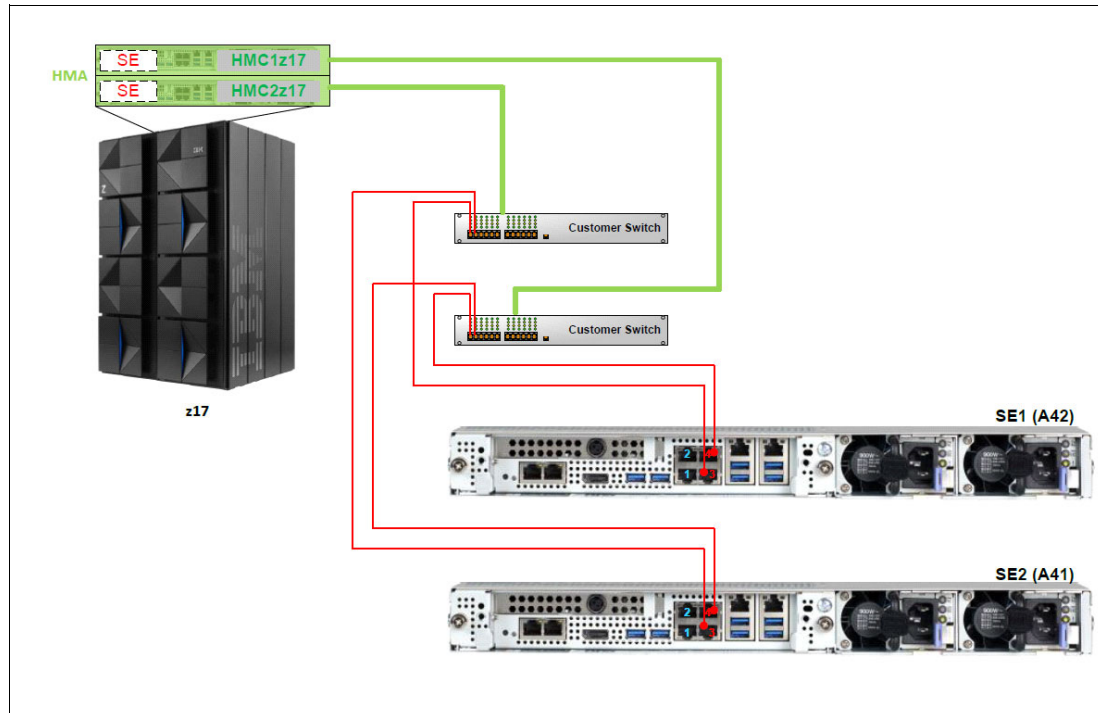


Figure 10-15 HMC / SE connectivity without HMA

The connectivity for multiple IBM Z generations HMA's environments (IBM z17 N-2 only) is shown in Figure 10-16.

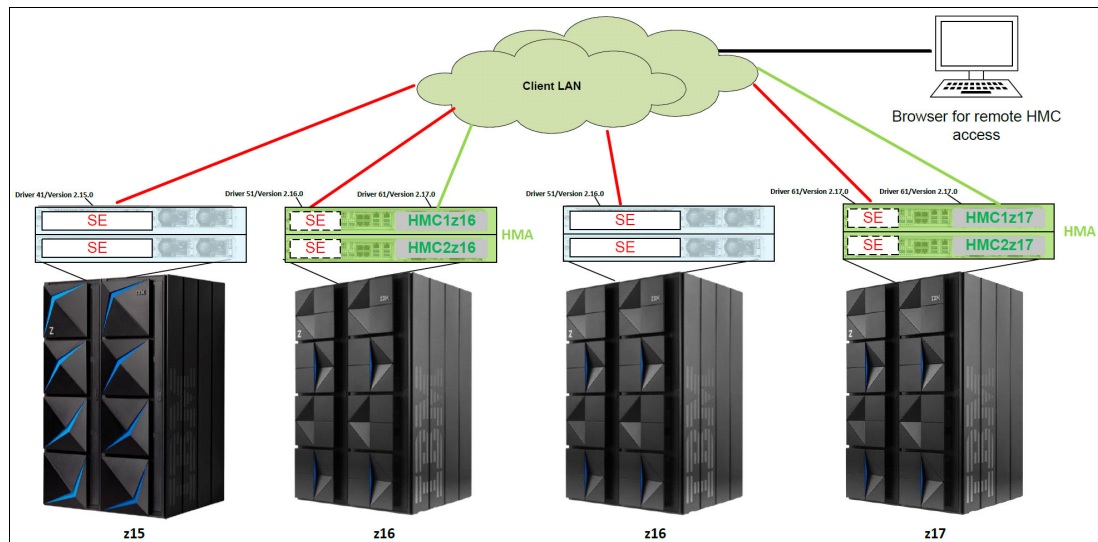


Figure 10-16 IBM z17 HMC/SE connectivity with multiple CPCs

<< replace z17 with correct picture>>

Various methods are available to set up the network. Designing and planning the HMC and SE connectivity is the customers' responsibility, based on the environment's connectivity and security requirements.



### 10.3.3 Network planning for the HMC and SE

Plan the HMC and SE network connectivity carefully for current and future use. Many of the IBM Z capabilities benefit from the various network connectivity options.

The following functions are examples that depend on the HMC connectivity:

- ▶ Lightweight Directory Access Protocol (LDAP) support, which can be used for HMC user authentication
- ▶ Network Time Protocol (NTP)
- ▶ RSF through broadband
- ▶ HMC remote access and HMC Mobile
- ▶ RSA SecurID support
- ▶ MFA with Time-based One Time Password (TOTP)

#### HMC File Transfer support

The HMC and SE support the FTP, FTPS, and SFTP protocols. All three file transfer protocols require login ID and password credentials.

FTPS is based on Secure Sockets Layer (SSL) cryptographic protocol and requires certificates to authenticate the servers. SFTP is based on Secure Shell protocol (SSH) and requires SSH keys to authenticate the servers. Certificates and key pairs are hosted on the IBM z17 HMC.

The following FTP server requirements must be met:

- ▶ Passive data connections are supported
- ▶ A server configuration is available that allows the client to connect on an ephemeral port

The following FTPS server requirements must be met:

- ▶ Operates in “explicit” mode
- ▶ Allows a server to offer secure and unsecured connections
- ▶ Supports passive data connections
- ▶ Supports secure data connections

The SFTP server must support password-based authentication.

The HMC file transfer protocol choices for the backup task are shown in Figure 10-17.

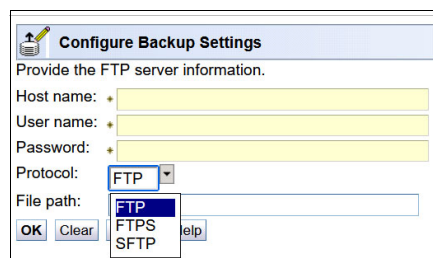


Figure 10-17 FTP protocol choices



## Secure console-to-console communications

The IBM z17 HMC features an industry-standard-based, password-driven cryptography system. The Domain Security Settings are used to provide authentication and high-quality encryption. We recommend that clients use unique Domain Security settings to provide maximum security. These settings provide greater security than anonymous cipher suites, even if the default settings are used.

For more information about HMC networks, see the following resources:

- ▶ The HMC and SE Driver 61/Version 2.17.0 console help system
- ▶ [IBM Resource Link](#). On the web page, select **Library**, the applicable server, and then select **Hardware Management Console Operations Guide** or **Support Element Operations Guide**. *IBM Z 9175 Installation Manual for Physical Planning*, GC28-7049.

## 10.3.4 Hardware considerations

The following hardware considerations are important for the IBM z17:

- ▶ IBM does not provide Ethernet cables with the system
- ▶ IBM does not provide Ethernet switches with the system for HMC and SE communication

### Ethernet switches

The customer provides ethernet switches for HMC and SE connectivity. Existing supported switches can still be used.

Ethernet switches often include the following characteristics:

- ▶ 16 auto-negotiation ports
- ▶ 1000 Mbps data rate
- ▶ Full duplex operation
- ▶ Auto medium-dependent interface crossover (MDIX) on all ports
- ▶ Port status LEDs
- ▶ Copper RJ45 connections

## 10.3.5 TCP/IP Version 6 on the HMC and SE

The HMC and SE can communicate by using IPv4, IPv6, or both.

IPv6 link-local addresses feature the following characteristics:

- ▶ Every IPv6 network interface is assigned a link-local IP address
- ▶ A link-local address is used on a single link (subnet) only and is never routed
- ▶ Two IPv6-capable hosts on a subnet can communicate by using link-local addresses, without having any other IP addresses assigned. If HMC-to-SE IPv6 link-local is working, the SE/CPC appears in **System Management** → **Unmanaged Systems** on the HMC.

## 10.3.6 Assigning TCP/IP addresses to the HMC, SE and ETS

Use the *HMC and SE Network Information Worksheet*, [GC28-7057](#), to plan your HMC, SE, and External Time Reference (ETS) IP configuration.

For ETS configuration please see *IBM Z Server Time Protocol Guide*, [SG24-8480](#)

An HMC can have the following IP configurations:



- ▶ Statically assigned IPv4 or IPv6 addresses
- ▶ Dynamic Host Configuration Protocol (DHCP)-assigned IPv4 or DHCP-assigned IPv6 addressees
- ▶ Auto-configured IPv6:
  - Link-local is assigned to every network interface
  - Router-advertised, which is broadcast from the router, can be combined with a Media Access Control (MAC) address to create a unique address
  - Privacy extensions can be enabled for these addresses to avoid using the MAC address as part of the address to ensure uniqueness.

An SE can have the following IP addresses:

- ▶ Statically assigned IPv4 or statically assigned IPv6
- ▶ Auto-configured IPv6 as link-local or router-advertised

IP addresses on the SE cannot be dynamically assigned through DHCP to ensure repeatable address assignments. DHCP privacy extensions are not used on the SE.

The HMC uses IPv4 and IPv6 multi-casting<sup>1</sup> to automatically discover the SEs. The HMC Network Diagnostic Information task can be used to identify the IP addresses (IPv4 and IPv6) that the HMC uses to communicate to the SEs (of a CPC).

IPv6 addresses are easily identified. A fully qualified IPV6 address features 16 bytes. It is written as eight 16-bit hexadecimal blocks that are separated by colons, as shown in the following example:

```
2001:0db8:0000:0000:0202:b3ff:fe1e:8329
```

Because many IPv6 addresses are not fully qualified, shorthand notation can be used. In shorthand notation, the leading zeros can be omitted, and a series of consecutive zeros can be replaced with a double colon. The address in the previous example also can be written in the following manner:

```
2001:db8::202:b3ff:fe1e:8329
```

If an IPv6 address is assigned to the HMC for remote operations that use a web browser, browse to it by specifying that address. The address must be surrounded with square brackets in the browser's address field, as shown in the following example:

```
https://[fdab:1b89:fc07:1:201:6cff:fe72:ba7c]
```

Your browser must support the use of link-local addresses.

### 10.3.7 HMC multi-factor authentication

MFA is an optional and configurable feature on a per-user, per-template basis. It enhances security by requiring what you know (the first factor) and what you have available, which means that only a person who owns a specific phone number, for example, can log in.

The MFA first factor is the combination of login ID and password; the second factor is TOTP (Time-based One-Time Password) that is sent to your smartphone, desktop, or application (for example, Google Authenticator or IBM Verify). This TOTP is defined in RFC 6238 standard and uses a cryptographic hash function that combines a secret key with the current time to generate a one-time password.

<sup>1</sup> For the customer-supplied switch, multicast must be enabled at the switch level.



The secret key is generated by HMC/SE/TKE while the user is performing the first-factor log-on. The secret key is known only to HMC/SE/TKE and to the user's device. For that reason, it must be protected as much as your first-factor password.

MFA code that was generated as a second factor is time-sensitive. Therefore, it is essential to remember that it must be used as soon as possible after it is generated.

The algorithm within the HMC that is responsible for MFA code generation changes the code every 30 seconds. However, the HMC and SE console accepts current, previous, and next MFA codes to ease the process.

Having HMC, SE, and device clocks synchronized is also important. If the clocks are not synchronized, the MFA log-on attempt fails. Time zone differences are irrelevant because the MFA code algorithm uses UTC.

IBM z15, HMC Driver 41/Version 2.15.0 provided the integration of HMC authentication and z/OS MFA support. Therefore, RSA SecurID authentication is achieved through centralized support from IBM MFA for z/OS, with the MFA policy defined in RACF and the HMC IDs assigned to RACF user IDs. The RSA authentication server verifies the RSA SecurID passcode (from an RSA SecurID Token). This authentication is supported only on HMC, not on SE.

The following support was added with IBM z16 Driver 51/Version 2.16.0:

► Enhanced MFA functions

MFA via Time-based One-time Password (TOTP) or IBM Z MFA (z/OS) and RSA Secure ID is possible on the HMC.

New with Driver 51/Version 2.16.0, the following further MFA possibilities are supported:

- Certificates:
  - Personal Identity Verification (PIV)
  - Common Access Card (CAC)
  - Certificates on USB keys
- Generic Remote Authentication Dia-In User Service (RADIUS) allows for support of all various RADIUS factor types. Involves customer provided RADIUS server.

Also, Driver 51/Version 2.16.0 provides support of IBM Z MFA for Red Hat Enterprise Linux Server or SUSE Linux Enterprise Server that runs on z/VM or native in an LPAR.

## 10.4 Remote Support Facility

The HMC Remote Support Facility (RSF) provides important communication to a centralized IBM Support network for hardware problem reporting and service. The following types of communication are provided:

- Problem reporting and repair data
- Microcode Change Level (MCL) delivery
- Hardware inventory data, which is also known as vital product data (VPD)
- Health and diagnostic data
- On-demand enablement (CoD)

### 10.4.1 Security characteristics

The following security characteristics are in effect:



- ▶ RSF requests are always started from the HMC to IBM. An inbound connection is never started from the IBM Service Support System.
- ▶ All data that is transferred between the HMC and the IBM Service Support System is encrypted with high-grade SSL/Transport Layer Security (TLS) encryption.
- ▶ When starting the SSL/TLS-encrypted connection, the HMC validates the trusted host with the digital signature that is issued for the IBM Service Support System.
- ▶ Data sent to the IBM Service Support System consists of hardware problems and configuration data.

For more information about the benefits of Broadband RSF and the SSL/TLS-secured protocol, as well as a sample configuration for the Broadband RSF connection, see *Integrating IBM Remote Support into your Enterprise*, SC28-7060.

### 10.4.2 RSF connections to IBM and Enhanced IBM Service Support System

- ▶ To ensure the best availability and redundancy and be prepared for the future, the HMC must access IBM through the Internet through RSF in the following manner: Transmission to the enhanced IBM Support System requires a domain name server (DNS). If a proxy for RSF is not used, the DNS must be configured on the HMC. If a proxy for RSF is used, the proxy must provide the DNS.
- ▶ The following hostnames and IP addresses are used, and your network infrastructure must allow the HMC to access RSF:
- ▶ esupport.ibm.com on port 443
- ▶ The use of IPv4 requires outbound connectivity to the following IP addresses:
  - 129.42.19.70
  - 129.42.18.70
  - 129.42.54.189
  - 192.148.6.11
- ▶ The use of IPv6 requires outbound connectivity to the following IP addresses:
  - 2620:1f7:c010:1:1:1:1:11
  - 2607:f0d0:2601:13:129:42:19:70
  - 2607:f0d0:1f01:9f:129:42:18:70
  - 2620:0:6c0:200:129:42:56:189

**Note:** All other previous IP addresses are no longer supported.

## 10.5 HMC and SE capabilities

The HMC and SE feature many capabilities. This section describes the key areas.

For more information about these capabilities, see the HMC and SE Driver 61/Version 2.17.0 console help system or see the [IBM Resource Link](#). At this web page, select **Library**, the applicable server and then select **Hardware Management Console Operations Guide** or **Support Element Operations Guide**.

With the introduction of the DPM mode, which is mainly for LinuxONE management, the user interface and user interaction with the HMC for hardware configuration changed significantly. The figures and descriptions in this section cover only the traditional Processor Resource/Systems Manager (PR/SM) mode.



## 10.5.1 Central processor complex management

The HMC is the primary place for CPC control. For example, the input/output configuration data set (IOCDs) includes definitions of LPARs, channel subsystems, control units, and devices and their accessibility from LPARs. IOCDs can be created and put into production from the HMC.

The HMC is used to start the system's power-on reset (POR). During the POR, processor units (PUs) are characterized and placed into their respective pools, memory is put into a single storage pool, and the IOCDs is loaded into and started in the hardware system area (HSA).

The hardware messages task displays hardware-related messages at the CPC, LPAR, or SE level. It also displays hardware messages that relate to the HMC.

## 10.5.2 LPAR management

Use the HMC to define LPAR properties, such as the number of processors of each type, how many are reserved, and how much memory is assigned. These parameters are defined in LPAR profiles and stored on the SE.

Because PR/SM must manage LPAR access to processors and the initial weights of each partition, weights are used to prioritize partition access to processors.

You can use the Load task on the HMC to perform an IPL of an operating system. This task causes a program to be read from a designated device and starts that program. You can perform the IPL of the operating system from storage, the USB flash memory drive (UFD), or an FTP server.

When an LPAR is active, and an operating system is running, you can use the HMC to dynamically change specific LPAR parameters. The HMC provides an interface to change partition weights, add logical processors to partitions, and add memory.

LPAR weights can also be changed through a scheduled operation. Use the **Customize Scheduled Operations** task to define the weights set to LPARs at the scheduled time.

Channel paths can be dynamically configured on and off (as needed for each partition) from an HMC.

Partition capping values can be scheduled and are specified on the Change LPAR Controls scheduled operation support. More information about a Change LPAR Controls scheduled operation is available on the SE.

- ▶ One example of managing the LPAR settings is the absolute physical hardware LPAR capacity setting. Driver 15 (zEC12/zBC12) introduced the capability to define (in the image profile for shared processors) the absolute processor capacity that the image is allowed to use (independent of the image weight or other cappings).
- ▶ To indicate that the LPAR can use the non-dedicated processor absolute capping, select **Absolute Capping** in the Image Profile Processor settings to specify an absolute number of processors to cap the LPAR's activity. The absolute capping value can be "None" or a value for the number of processors (0.01 - 255.0).
- ▶ Following on to LPAR absolute capping, LPAR group absolute capping uses a similar method to enforce the following components:
  - ▶ Customer licensing
  - ▶ Non-z/OS partitions where group soft capping is not an option
  - ▶ z/OS partitions where ISV does not support software capping



A group name, processor capping value, and partition membership are specified at the hardware console, along with the following properties:

- ▶ Set an absolute capacity cap by CPU type on a group of LPARs.
- ▶ Allow each partition to use capacity up to its individual limits if the group's aggregate consumption does not exceed the group's absolute capacity limit.
- ▶ Include updated SysEvent QVS support (used by vendors who implement software pricing).
- ▶ Only shared partitions are managed in these groups.
- ▶ Specify caps for one or more processor types in the group.
- ▶ Specify the absolute processor capacity (for example, 2.5 processors).
- ▶ Use Change LPAR Group Controls (as with windows that are used for software group-defined capacity), as shown in Figure 10-18.

Group Name	Member Partitions	Group Capacity Value	Absolute Capping for CPs	Absolute Capping for CBPs	Absolute Capping for ICFs	Absolute Capping for IFLs	Absolute Capping for zIIPs
CACPOOL	CETUS0C CETUS0E	0	None	None	None	20.00	None
CFBPOOL	CETUS0D CETUS0F	0	None	None	None	None	None
COCPOOL	CETUS06 CETUS07 CETUS08 CETUS09	0	None	None	None	33.00	None
CRUHPPOOL		0	None	None	None	69.00	None
DEFAULT	CETUS05	0	None	None	None	None	None
DEVGROU		50	None	None	None	None	None
PRVEPOOL		0	None	None	None	None	None
ROBSPool		0	None	None	None	None	None

Buttons: Save to Profiles, Change Running System, Save and Change, Reset, Cancel, Help

Figure 10-18 Change LPAR Group Controls: Group absolute capping

Absolute capping is specified as the absolute number of processors to which the group's activity is capped. The value is specified as a hundredth of a processor's capacity (for example, 4.56 processors).

The value is not tied to the Licensed Internal Code (LIC) configuration code (LICCC). Any value 0.01 - 255.00 can be specified. This configuration makes the profiles more portable, so you do not encounter problems when profiles are migrated to new machines.

Although the absolute cap can be specified to a hundredths of a processor, the exact amount might not be that precise. The same factors that influence the "machine capacity" also affect the precision with which the absolute capping works.

LPAR absolute capping can be changed through scheduled operations start.

### 10.5.3 HMC and SE remote operations

Because stand-alone, outside the IBM z17 HMCs (Tower or Rack Mount) are no longer supported with IBM z17, connections to the HMC (and SE) are made by a browser. However,



because the SE cannot be directly accessed by using a browser, the Single Object Operations (SOO) task on the HMC must be used.

**Note:** Remote web browser access is the default for the HMA HMCs.

The local HMC user provides log-on security for web browser log-on procedures. Certificates for secure communications are provided and can be changed by you.

Web browser access can be limited by specifying an IP address from the Customize Console Services task. To turn the Remote operation service on or off, click **Change** in the Customize Console Services window, as shown in Figure 10-19.

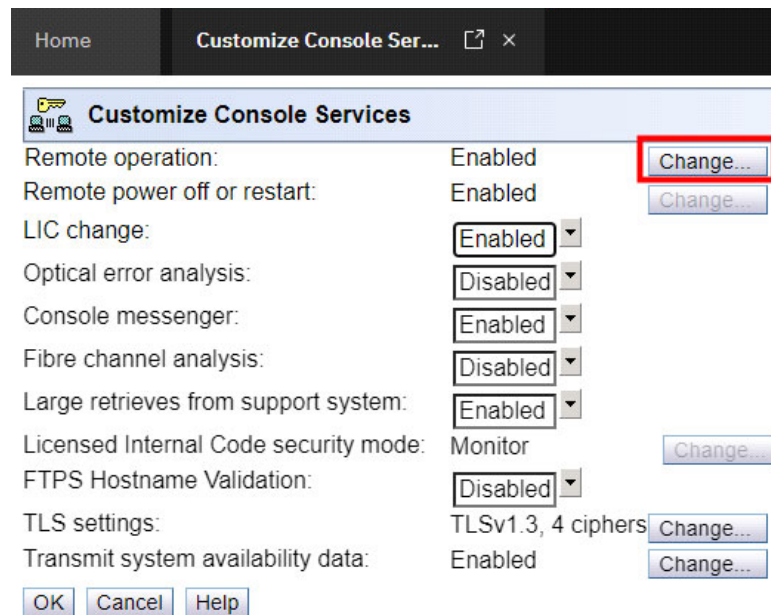


Figure 10-19 Customizing HMC remote operation

**Note:** If the **Change Remote Access Setting** → **IP Access Control** is set to Allow specific IP addresses, but none or incorrect IP addresses are in the list, a remote HMC connection is not available by using a web browser.

Microsoft Edge, Mozilla Firefox, Safari, and Goggle Chrome were tested as remote web browsers. For more information about web browser requirements, see the HMC and SE console help system or see the [IBM Resource Link](#). On this web page, select **Library**, the applicable server and then, select **Hardware Management Console help system content** or **Support Element help system content**.

## Single Object Operations

It is not necessary to be physically close to a SE to use it. The HMC can be used to access the SE remotely by using the SOO task. The interface is the same as the one used on the SE. For more information, see the [IBM Resource Link](#). At this web page, select **Library**, the applicable server, and then select **Hardware Management Console Operations Guide** or **Support Element Operations Guide**.



**Note:** Beginning with HMC Driver 41/Version 2.15.0, specific tasks previously required to access the SE in SOO mode were implemented as HMC tasks. With this enhancement, the HMC runs the tasks on the SE directly without accessing the SE in SOO mode.

## IBM HMC Mobile

IBM HMC Mobile is an iOS and Android application that allows you to monitor all of your IBM Z systems and partitions and receive alerts when messages or status changes occur.

You also can start, stop, or change the activation profile for a partition.

The HMC provides a full set of granular security controls, including MFA. This mobile interface is optional and disabled by default. More functions from the HMC are also planned to be available on IBM HMC Mobile.

For more information about HMC Mobile, see the following resources:

- ▶ This [IBM video](#)
- ▶ The HMC Mobile [website](#)

## 10.5.4 Operating system communication

The Operating System Messages task displays messages from an LPAR. You also can enter operating system commands and interact with the system. This feature is especially valuable for entering Coupling Facility Control Code (CFCC) commands.

The HMC also provides integrated 3270 and ASCII consoles. These consoles allow an operating system to be accessed without requiring other network or network devices, such as TCP/IP or control units.

### Updates to x3270 support

The Configure 3270 Emulators task on the HMC and TKE consoles was enhanced with Driver 15 to verify the authenticity of the certificate returned by the 3270 server when a secure and encrypted SSL connection is established to an IBM host. This 3270 Emulator with an encrypted connection is also known as Secure 3270.

Use the Certificate Management feature if the certificates returned by the 3270 server are not signed by a well-known trusted certificate authority (CA) certificate, such as VeriSign or Geotrust. An advanced action within the Certificate Management task, Manage Trusted Signing Certificates, adds trusted signing certificates.

For example, if the certificate that is associated with the 3270 server on the IBM host is signed and issued by a corporate certificate, it must be imported, as shown in Figure 10-20.

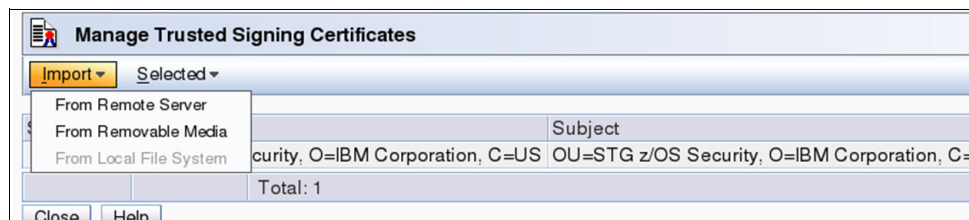


Figure 10-20 Manage Trusted Signing Certificates



The import from the remote server option can be used if the connection between the console and the IBM host can be trusted when the certificate is imported, as shown in Figure 10-21. Otherwise, import the certificate using removable media.

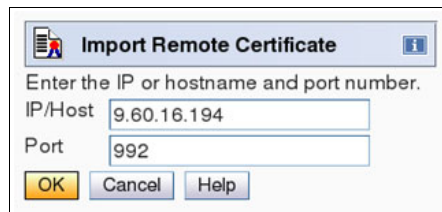


Figure 10-21 Import Remote Certificate example

A secure Telnet connection is established by adding the prefix L: to the IP address:port of the IBM host, as shown in Figure 10-22.

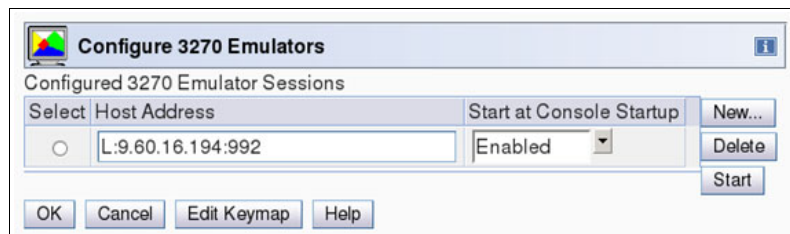


Figure 10-22 Configure 3270 Emulators

## 10.5.5 Monitoring

This section describes monitoring considerations.

### Monitor task group

The Monitor task group on the HMC and SE includes monitoring-related tasks for IBM Z CPCs, as shown in Figure 10-23.

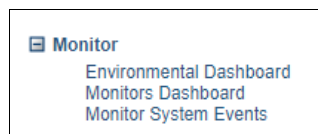


Figure 10-23 HMC Monitor Task Group

### Monitors Dashboard

The Monitors Dashboard in the Monitor group provides a tree-based view of resources.

Multiple graphical views are available for displaying data, including history charts. The Monitors Dashboard monitors processor and channel usage. It also produces data that includes power monitoring information, power consumption, and the air input temperature for the system.

You can display more information for the following components:

- ▶ Power consumption
- ▶ Environmental
- ▶ Aggregated processors
- ▶ Processors (with SMT information)



- ▶ System Assist Processors
- ▶ Logical Partitions
- ▶ Channels
- ▶ Adapters: Crypto use percentage is displayed according to the physical channel ID (PCHID number)

For more information please see “Monitors Dashboard task” on page 485.

## Environmental Dashboard

The Environmental Efficiency Statistics is part of the Monitor group. It provides the IBM Z CPC's historical power consumption and thermal information and is available on the HMC.

For more information please see “Environmental dashboard task” on page 486.

## Monitor System Events

The Monitor System Events task allows you to create and manage event monitors. An event monitor listens for events from managed objects. When an event is received, the monitor tests it with user-defined criteria. If the event passes the tests, the monitor sends an e-mail to interested users. For example, you can be notified via e-mail if a Hardware Message appears.

With IBM z17, you can generate system power consumption notifications if a certain power consumption threshold is reached on the LPAR or System level.

An example of setting an Event Monitor for the power consumption on the LPAR level can be seen in Figure 10-24.

The screenshot shows the 'Event Monitor Editor' window. The 'Name' field is 'Partition zOS Power Consumption'. The 'Description' is 'Notify if power consumption exceeds 10kW for 15min on ZOS partition (between 9-5)'. The 'Type' is 'Partition' and the 'Metric' is 'Power Consumption'. A table lists partitions: T14E11, T14E12, T14E13, ZOS (selected), ZVM1, and ZVM2, all on B32 system with LPAR Image description. The 'Consumption threshold (W)' is 10000 and 'Duration (minutes)' is 5. The 'Schedule' is set to 'Limit to times' from 9:00:00 AM to 5:00:00 PM, every day from 10/17/24 to 10/18/24. The 'Notification list' includes 'jane@sample.com ed@someeserver.com # Jane and Ed'.

Select	Partition Name	System	Description
<input type="checkbox"/>	T14E11	B32	LPAR Image
<input type="checkbox"/>	T14E12	B32	LPAR Image
<input type="checkbox"/>	T14E13	B32	LPAR Image
<input checked="" type="checkbox"/>	ZOS	B32	LPAR Image
<input type="checkbox"/>	ZVM1	B32	LPAR Image
<input type="checkbox"/>	ZVM2	B32	LPAR Image

Figure 10-24 Example of Event Monitor for power consumption on LPAR level



## 10.5.6 Capacity on-demand support

All capacity on demand (CoD) upgrades are performed by using the **Perform a Model Conversion** task on the HMC or SE. Use the task to retrieve and activate a permanent upgrade and to retrieve, install, activate, and deactivate a temporary upgrade. The task lists all installed or staged LICCC records to help you manage them. It also shows a history of recorded activities.

The HMC for IBM z17 features the following CoD capabilities:

- ▶ SNMP API support:
  - API interfaces for granular activation and deactivation
  - API interfaces for enhanced CoD query information
  - API event notification for any CoD change activity on the system
  - CoD API interfaces, such as On/Off CoD and Capacity Back Up (CBU)
- ▶ HMC / SE window features:
  - Window controls for granular activation and deactivation
  - History window for all CoD actions
  - Description editing of CoD records
- ▶ HMC/SE provides the following CoD information:
  - Millions of service units (MSU) and processor tokens
  - Last activation time
  - Pending resources that are shown by processor type instead of only a total count
  - Option to show more information about installed and staged permanent records
  - More information for the Attention state by providing seven more flags

HMC and SE are a part of the z/OS Capacity Provisioning environment. The Capacity Provisioning Manager (CPM) communicates with the HMC through IBM Z APIs, and enters CoD requests.

### Notes:

- ▶ The HWMCA\_ADD\_CAPACITY\_COMMAND and HWMCA\_REMOVE\_CAPACITY\_COMMAND APIs allow applications to add and remove temporary capacity for defined CPC objects. You can use the HWMCA\_ACTIVATE\_CBU\_COMMAND and HWMCA\_ACTIVATE\_OOCOD\_COMMAND APIs to allow applications to activate a CBU or On/Off CoD record for a defined CPC object.
- ▶ When activating a CBU record, the API activates all the resource in the default CBU record. If there is no default CBU record specified, the oldest CBU record is used. To set a CBU record as the default CBU record, select the Set as Default CBU button located at the bottom of the Record Details window.

For more information, see *Capacity on-Demand User's Guide*, SC28-7058.

For more information about the use of and setting up CPM, see the following publications:

- ▶ *z/OS MVS Capacity Provisioning User's Guide*, SC34-2661
- ▶ *Capacity on-Demand User's Guide*, SC28-7058

## 10.5.7 Server Time Protocol support

Server Time Protocol (STP) can be managed on the HMC via task **Manage System Time**.



**Important:** The Sysplex Time task on the SE was discontinued with IBM z15.

Starting with IBM z16 and also with IBM z17, the following significant enhancements are essential for the support of the Server Time Protocol function:

### **CPC direct Ethernet connectivity for the external time source (ETS)**

In previous IBM Z generations, the ETS for the STP was provided by connecting the Support element to the client network.

Starting with IBM z16, the ETS, PTP (IEEE 1588), or NTP network connectivity is provided by using the CPC oscillator (OSC) cards dedicated network ports (RJ45) to client LAN for accessing the time synchronization information. ETS direct connectivity to the IBM z17 CPC is provided for both supported ETS protocols: NTP and PTP.

Pulse-per-second connectivity is also provided for the highest accuracy in timing information.

Connection of the ETS directly to the IBM Z CPC provides less delay in accessing the time source than connection through the Support Element.

For more information, please see 2.2.2, “Oscillator (OSC) and Baseboard Management Controller (BMC) cards” on page 31.

### **n-mode power sensing for STP recovery / Set time server power failover**

On IBM Z, losing a Preferred Time Server (PTS) has significant consequences for the timing network and the overall workload execution environment of the IBM Z sysplex.

Starting with IBM z16, because an integrated battery facility (IBF) is no longer available, support was added to allow you to configure an option to monitor for n-mode power conditions (wall power or power cord loss). If detected, an automated failover occurs to the Backup Time Server (BTS) from the Preferred Time Server (PTS) / the Current Time Server (CTS) changes from the BTS to the PTS.

In the task **Manage System Time** -> Advanced Action -> **Set time server power failover**, you can turn automatic failover on or off when STP detects a loss of power on the PTS. STP can detect various power losses, including a loss as small as a single line cord failure, which could leave a portion of the system without power.

IBM supports the use of an external battery or Uninterruptible Power Supply (UPS) to ensure that the system remains running in the event of a power outage. The PTS and BTS should be running with an external power source and have automatic failover enabled to be completely safe from power interruptions. The failover to the BTS occurs within 1 minute after the power loss is detected and the external power supply takes over.

After resolving the power issue, you should inspect the system before returning it to operation. For this reason, an option allows you to specify the time to wait before the system reacts to the restored power. Then, after the specified waiting period has ended, the Current Time Server (CTS) is automatically switched back to the PTS.

Figure 10-25 on page 462 shows your options to configure **Set time server power failover**.



**Set time server power failover - PCC-TIME**

Setting a Preferred Time Server (PTS) power failover behavior allows the Current Time Server (CTS) to be automatically switched to the Backup Time Server (BTS) when a power issue is detected.

Choose the failover behavior for any power issues detected on the PTS.

☐ Automatically switch the CTS to the BTS if a power issue is detected on the PTS

Set the amount of time to delay before switching the CTS back to the PTS if full power is detected.

Switch CTS to PTS delay (d:hh:mm):  Range is 5 min - 1 week (0:00:05 - 7:00:00)

**CANCEL** **APPLY** **HELP**

Figure 10-25 Task Set time server power failover for STP

- More enhancements specific with IBM z17 can be found at “External Time Source (ETS) enhancements for STP and HMC NTP” on page 433.

For more information about planning and understanding STP and ETS, see the following publication: *IBM Z Server Time Protocol Guide*, [SG24-8480](#)

## 10.5.8 Security and user ID management

This section addresses security and user ID management considerations.

### HMC and SE HD encryption

Consider the following points (with continued emphasis on encryption):

- Passwords are never stored in clear (one-way hash)
- HMC/SE are a closed appliances;
- *All* network traffic is TLS encrypted
- HMC and SE feature embedded firewall
- Firmware is digitally signed and validated for delivery
- Firmware Integrity Monitoring is used for any attempted tempering post delivery

HDD encryption uses Trusted Platform Module (TPM) and Linux Unified Key Setup (LUKS) technology.

### HMC and SE security audit improvements

With the Audit and Log Management task, audit reports can be generated, viewed, saved, and off-loaded. The Customize Scheduled Operations task allows you to schedule audit report generation, saving, and off loading.

The Monitor System Events task allows Security Logs to send e-mail notifications by using the same type of filters and rules that are used for hardware and operating system messages.

With IBM z17, you can off load the following HMC and SE log files for customer audit:

- Console event log
- Console service history
- Tasks performed log
- Security logs
- System log



Full log off load and delta log off load (since the last off load request) are provided. Off loading to removable media and to remote locations by FTP is available. The off loading can be manually started by the new Audit and Log Management task or scheduled by the Customize Scheduled Operations task. The data can be off-loaded in the HTML and XML formats.

### **HMC user ID templates and LDAP user authentication**

Lightweight Directory Access Protocol (LDAP) user authentication and HMC user ID templates enable the addition and removal of HMC users according to your own corporate security environment. These processes use an LDAP server as the central authority.

Each HMC user ID template defines the specific authorization levels for the tasks and objects for the user who is mapped to that template. The HMC user is mapped to a specific user ID template. The system then obtains the name of the user ID template from content in the LDAP server schema data.

### **Default HMC user IDs**

For HMC Driver 61/Version 2.17.0 the default user IDs are limited to ACSADMIN and SERVICE.

ADVANCED, OPERATOR, STORAGEADMIN, and SYSPROG default users are no longer shipped. Default user roles for ADVANCED, OPERATOR, STORAGEADMIN, and SYSPROG are provided, and user IDs can be created from those roles.

Any default user IDs that are part of a previous HMC level can be carried forward to new HMC levels as part of a MES Upgrade or by way of selecting User Profile Data for the Save/Restore Customizable Console Data or Configure Data Replication tasks.

We recommend that you create your own individual user IDs using “New based on”, except for the local ACSADMIN or a concept to have a user ID with admin role to change/create users as an emergency.

### **HMC and SE secure FTP support**

You can use a secure FTP connection from an HMC/SE FTP client to a customer FTP server location. This configuration is implemented using the Secure Shell (SSH) File Transfer Protocol, which is an extension of SSH. You can use the Manage SSH Keys console feature that is available to the HMC and SE to import public keys that are associated with a host address.

The Secure FTP infrastructure allows HMC and SE applications to query whether a public key is associated with a host address and to use the Secure FTP interface with the suitable public key for a host. Tasks that use FTP now provide a selection for the secure host connection.

When selected, the task verifies that a public key is associated with the specified hostname. If a public key is not provided, a message window opens that points to the Manage SSH Keys task to enter a public key. The following tasks provide this support:

- ▶ Import/Export IOCDS
- ▶ Advanced Facilities FTP IBM Content Collector Load
- ▶ Audit and Log Management (Scheduled Operations only)
- ▶ FCP Configuration Import/Export
- ▶ OSA view Port Parameter Export
- ▶ OSA-Integrated Console Configuration Import/Export

With IBM z17 Driver 61/Version 2.17.0 you have the option for Import/Export from Remote Browsing File System (see more at “Import/Export from Remote Browsing File System” on page 431), so maybe secure FTP is no longer needed.



## 10.5.9 Automated operations via APIs

As an alternative to manual operations, an application can interact with the HMC and SE through an API. The interface allows a program to monitor and control the hardware components of the system in the same way a user on the UI performs these tasks.

On IBM z17, the APIs provide monitoring and control functions through SNMP, Web Services, and BCPii.

For more information about APIs, see:

- ▶ *Hardware Management Console Web Services API (Version 2.17.0)*, SC27-2646
- ▶ HMC Dashboard -> Helpful links -> APIs
- ▶ *SNMP Application Programming Interfaces*, SB10-7185
- ▶ “BCPii enhancements” on page 437

## 10.5.10 Cryptographic support

This section describes the cryptographic management and control functions that are available in the HMC and SE.

### Cryptographic hardware

IBM z17 systems include standard cryptographic hardware and optional cryptographic features for flexibility and growth capability.

The HMC/SE interface provides the following capabilities:

- ▶ Defining the cryptographic controls
- ▶ Dynamically adding a Crypto feature to a partition:
  - For the first time
  - That already uses Crypto
- ▶ Dynamically removing a Crypto feature from a partition

The Crypto Express8S, which is the new Peripheral Component Interconnect Express (PCIe) cryptographic coprocessor, is an optional IBM z17 feature. Crypto Express8S provides a secure programming and hardware environment on which crypto processes are run.

Each Crypto Express8S adapter can be configured by the installation as a Secure IBM CCA coprocessor, a Secure IBM Enterprise Public Key Cryptography Standards (PKCS) #11 (EP11) coprocessor, or an accelerator.

When EP11 mode is selected, a unique Enterprise PKCS #11 firmware is loaded into the cryptographic coprocessor. It is separate from the Common Cryptographic Architecture (CCA) firmware that is loaded when a CCA coprocessor is selected. CCA firmware and PKCS #11 firmware cannot coexist in a card.

An example of the Cryptographic Configuration window on the HMC is shown in Figure 10-26 on page 465.



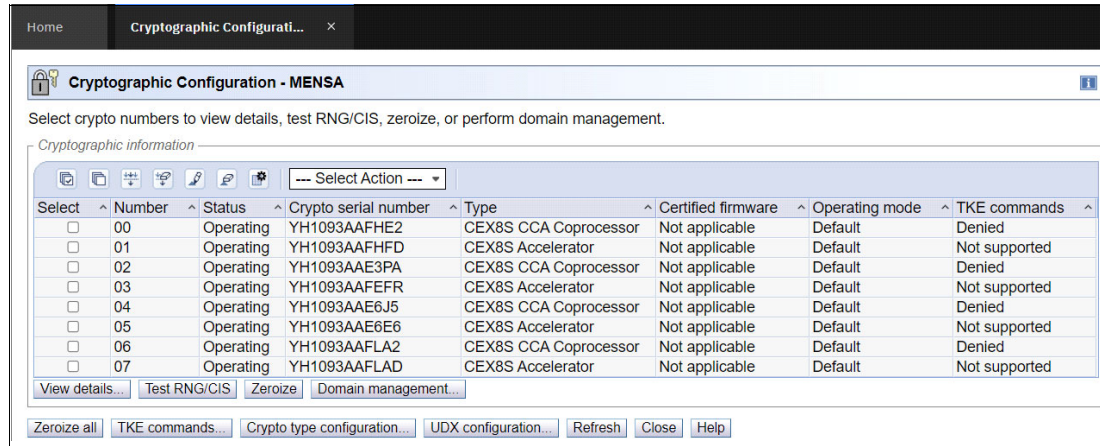


Figure 10-26 Cryptographic Configuration window

The Usage Domain Zeroize feature is provided to clear the suitable partition crypto keys for a usage domain when you remove a crypto card from a partition. Crypto Express8S or 7S in EP11 mode is configured to the standby state after the Zeroize process.

For more information, see *IBM z17 (9175) Configuration Setup*, [SG24-8960](#).

### Digitally signed firmware

Security and data integrity are critical issues with firmware upgrades. Procedures are in place to use a process to digitally sign the firmware update files that are sent to the HMC, SE, and TKE. By using a hash algorithm, a message digest is generated that is then encrypted with a private key to produce a digital signature. This operation ensures that any changes that are made to the data are detected during the upgrade process by verifying the digital signature. It helps ensure that no malware can be installed on IBM Z products during firmware updates. It also enables the IBM z17 Central Processor Assist for Cryptographic Function (CPACF) functions to comply with Federal Information Processing Standard (FIPS) 140-3 Level 1 planned for Cryptographic LIC changes. The enhancement follows the IBM Z focus of security for the HMC and the SE.

## 10.5.11 Installation support for z/VM that uses the HMC

Starting with z/VM V5R4 and z10, Linux on IBM Z can be installed in a z/VM virtual machine from HMC workstation media. This Linux on IBM Z installation can use the communication path between the HMC and the SE. No external network or extra network setup is necessary for the installation.

## 10.5.12 Dynamic Partition Manager

DPM is an administrative mode (front end to PR/SM) that was introduced for Linux-only systems for IBM z13 and following systems. With DPM, you can use your Linux and virtualization skills while taking advantage of the full value of IBM Z hardware, robustness, and security in a workload optimized environment.

A system can be configured in DPM mode or in PR/SM mode (POR is required to switch modes). In general, DPM supports the following functions:



- ▶ Create, provision, and manage partitions (processor, memory, and adapters) and storage
- ▶ Monitor and troubleshoot the environment

The following LPAR modes are available for DPM:

- ▶ z/VM
- ▶ Linux on IBM Z (also used for KVM deployments)
- ▶ Secure Service Container (SSC)

If DPM is enabled, the IBM z17 system cannot run z/OS, IBM z/VSE, and z/TPF LPARs.

The IBM z17 can be initialized in PR/SM mode or in DPM mode, but not both.

DPM provides a GUI for PR/SM (to manage resources). Tools, such as HCD are in DPM not necessary.

This IBM Redbooks publication does not cover scenarios that use DPM.

For more information about the use of DPM, see *IBM Dynamic Partition Manager (DPM) Guide*, SB10-7188.

**Important:** Consider the following points:

- ▶ The Enabling DPM function is run on the SE and is performed by your IBM system Service Representative at installation time.
- ▶ If DPM is enabled, the IBM z17 system cannot run z/OS, 21st CL VSE<sup>n</sup>, nor z/TPF LPARs.

## 10.6 HMC,SE, and CPC microcode

The microcode for the HMC, SE, and CPC is included in the driver/version. The HMC provides the management of the driver upgrade through Enhanced Driver Maintenance (EDM). EDM also provides the installation of the latest functions and the patches (MCLs) of the new driver.

When a driver is upgraded, always check the Driver (61) Customer Exception Letter option in the Fixes section at the IBM Resource Link.

### Microcode Change Level

Regular installation of Microcode Change Levels (MCLs) is key for RAS, optimal performance, and to enable new functions.

- ▶ Install MCLs on a quarterly basis at a minimum.
- ▶ Review hiper MCLs continuously to decide whether to wait for the next scheduled fix application session or to schedule one earlier if the risk assessment warrants.
- ▶ Sign on to the IBM Z Security Portal website and review for security alerts and related MCL fixes.

**Tip:** The Call Home Connect Cloud (CHCC)

<https://www.ibm.com/support/call-home-connect/cloud/> provides you access to important information about your IBM Z Systems. A lot of information you have found so far on IBM Resource Link has moved to CHCC. In the overview for a Machine profile you will also see the according current MCL information.



## Microcode terms

The microcode features the following characteristics:

- ▶ The driver contains engineering change (EC) streams.
- ▶ Each EC stream covers the code for a specific component of IBM z17. It includes a specific name and an ascending number.
- ▶ The EC stream name and a specific number are one MCL.
- ▶ MCLs from the same EC stream must be installed in sequence.
- ▶ MCLs can include installation dependencies on other MCLs.
- ▶ Combined MCLs from one or more EC streams are in one Bundle.
- ▶ An MCL contains one or more Microcode Fixes (MCFs).

How the driver, bundle, EC stream, MCL, and MCFs interact with each other, is shown in Figure 10-27.

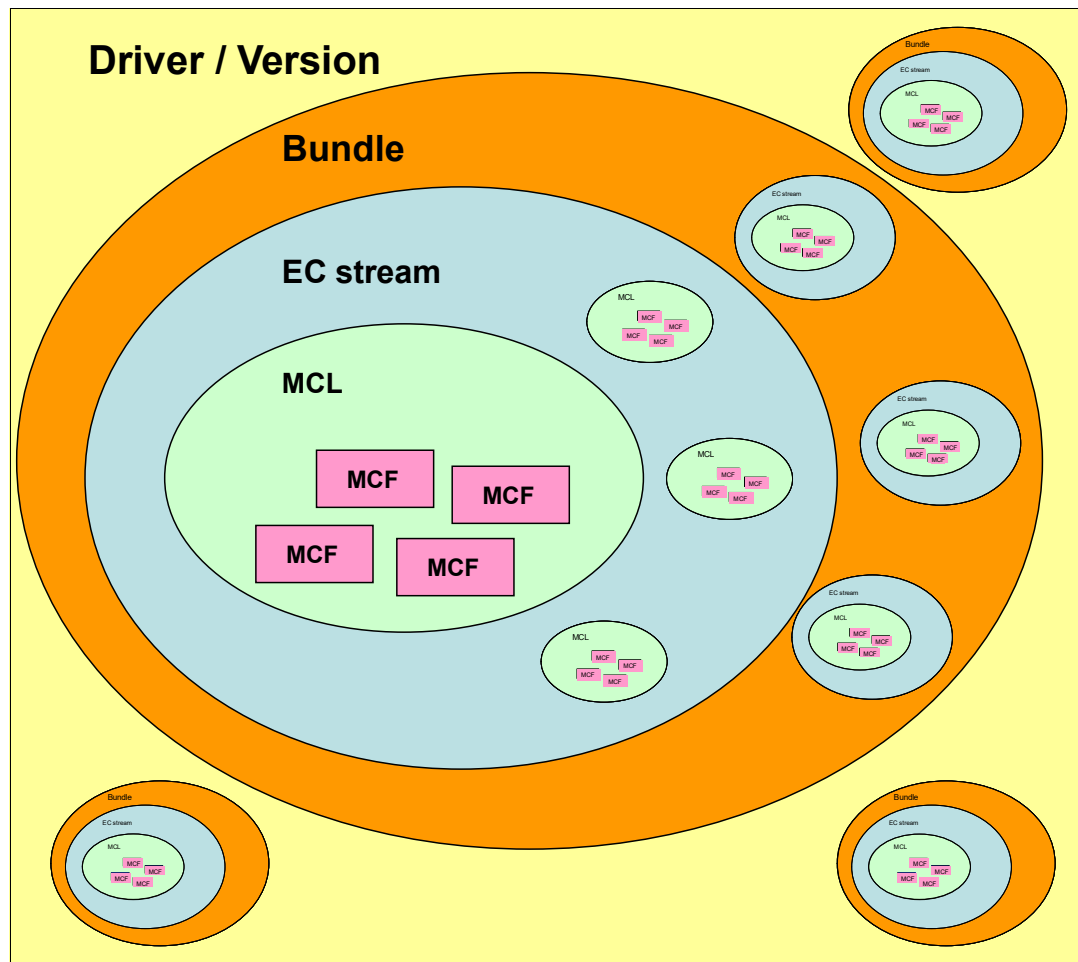


Figure 10-27 Microcode terms and interaction



## MCL Activation

By design and with planning, MCLs can be activated concurrently. Some MCLs need a disruptive configure off/on of resources (PCHIDs, cards, I/O elements etc.) to activate the new loaded microcode.

To check pending condition, you can go to **System Information** -> *Query Additional Actions....* or you can check on Call Home Cloud Connect (CHCC)  
<https://www.ibm.com/support/call-home-connect/cloud/> Information Reports (Reports -> Pending firmware updates).

## Microcode installation by MCL Bundle target

A *Bundle* is a set of MCLs that are grouped during testing and released as a group on the same date. You can install MCLs to a specific target Bundle level.

The System Information window is enhanced to show a summary Bundle level for the activated level, as shown in Figure 10-28.

**System Information - MENS**

**Machine Information**

EC number: P30875 LIC control level: 0001 Engineering Changes AROM  
 Type: 9175 Model number: ME1 Serial number: 0000200B9FB8  
 Version: 2.17.0 Driver level: 60 Bundle level: S01

**Internal Code Change Information**

Select	EC Number	Retrieved Level	Installable Concurrent	Activated Level	Accepted Level	Description
<input type="radio"/>	P30875	001	001	001	000	SE Framework
<input type="radio"/>	P30876	001	001	001	000	Firmware Management
<input type="radio"/>	P30877	000	000	000	000	PE messaging alert
<input type="radio"/>	P30878	001	001	001	000	I390
<input type="radio"/>	P30879	000	000	000	000	LPAR HV LIC
<input type="radio"/>	P30880	000	000	000	000	Coupling CFCC
<input type="radio"/>	P30881	001	001	001	000	PCX LIC
<input type="radio"/>	P30882	001	001	001	000	PSCN code deliverable for code on BMC and code on SE
<input type="radio"/>	P30883	001	001	001	000	Support Element Microcode
<input type="radio"/>	P30884	001	001	001	000	Millicode
<input type="radio"/>	P30885	001	001	001	000	PPC/IPC power FRUs
<input type="radio"/>	P30886	000	000	000	000	Feature enablement stream for CEC
<input type="radio"/>	P30887	000	000	000	000	Feature enablement stream for CEC
<input type="radio"/>	P30889	000	000	000	000	EnMA PUMA CUB (CPC Utility for Broadcasting)
<input type="radio"/>	P30892	000	000	000	000	Ficon Express 16SE LIC
<input type="radio"/>	P30893	000	000	000	000	FCP Express 16SE LIC
<input type="radio"/>	P30894	000	000	000	000	Ficon Express 32E LIC
<input type="radio"/>	P30895	000	000	000	000	FCP Express 32E LIC
<input type="radio"/>	P30896	001	001	001	000	Shadow FICON/HPF
<input type="radio"/>	P30897	001	001	001	000	Shadow FCP
<input type="radio"/>	P30898	001	001	001	000	Shadow OSA networking
<input type="radio"/>	P30900	000	000	000	000	OSA Express 7S - OSD
<input type="radio"/>	P30902	000	000	000	000	Shadow OSC 3215
<input type="radio"/>	P30903	001	001	001	000	Shadow OSC
<input type="radio"/>	P30904	001	001	001	000	Firmware Partition

**EC Details...**

**Pending Actions**

Some actions might be pending. Click **Query Additional Actions...** for more information.

[Query Additional Actions...](#)

OK Help

Figure 10-28 System Information: MCLs and Bundle level

### 10.6.1 Remote Code Load (RCL)

Remote Code Load (RCL) is a no-charge feature introduced with IBM z15 to remotely schedule and install firmware updates on IBM Z systems and HMC.

This feature allows an authorized client user to remotely schedule an automated firmware update operation. This secure firmware update operation will call home with live progress updates that are monitored by IBM support. RCL eliminates the need to schedule on-site



access for an IBM System Services Representative (SSR), for the duration of the firmware updates.

For a comprehensive guide on how to utilize RCL, please see the Remote Code Load for IBM Z Firmware publication (SC28-7044-02) on Resource Link found here: [Remote Code Load for IBM Z](#).

The client generates a token on the HMC/HMA and then use the IBM Resource Link site to create a scheduling request. The process flow is show in Figure 10-29.

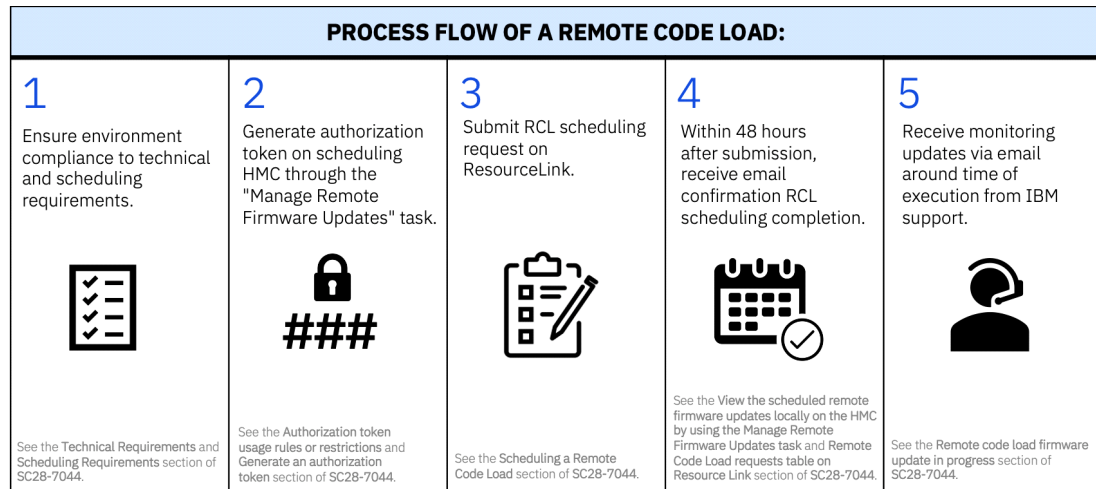


Figure 10-29 Process flow of a Remote Code Load (RCL)

Both RCL and IBM SSR onsite are options for the installation of Firmware update of MCL Bundles.

There is significant value from implementing RCL:

- ▶ Flexible scheduled to MCL Bundles via IBM Resource Link.
- ▶ No need for clients or IBM SSR to be onsite for IBM updates
- ▶ Site access requests not required for IBM SSR or client support staff.
- ▶ Clients can perform subsequent testing remotely.

The IBM z17 RCL health checks are categorized as blocking conditions and warning conditions.

A blocking condition will terminate the scheduled RCL, and a warning condition will allow the scheduled RCL to go ahead, but there may be potential issues if it is not addressed before the RCL.

Blocking Conditions are listed below:

- ▶ Communications are not currently active between the Primary/Alternate SE Pair.
- ▶ Automatic Service Reporting is not enabled on the targeted HMC.
- ▶ Automatic Service Reporting is not enabled on the peer HMA HMC.
- ▶ Change management is not fully enabled on this HMC.
- ▶ Change management is not fully enabled on the peer HMA HMC.
- ▶ The target HMA HMC is failing to establish communication with the peer HMA HMC.
- ▶ There are configuration issues for the SE defined on the targeted HMA HMC.
- ▶ There are configuration issues for the SE defined on the peer HMA HMC.
- ▶ One or more Alternate SE switch capability conditions would prevent automatic switchover for the HMA HMC pair update.
- ▶ There are one or more Service Required State conditions.



- One or more channels have not completed concurrent patch from a previous Bundle installation.

Some of the enhancements to RCL with the IBM z17 can be found “Remote Code Load (RCL) enhancements” on page 434.



## 11



# Environmental requirements

This chapter describes the environmental requirements for IBM z17 servers. It also lists the dimensions, weights, power, and cooling requirements that are needed to plan for the installation of an IBM z17.

**Naming:** Throughout this chapter, *IBM z17* refers to IBM z17 Model ME1 (Machine Type 9175), unless otherwise specified.

This chapter includes the following topics:

- ▶ 11.1, “Introduction” on page 472
- ▶ 11.3, “Physical specifications” on page 478
- ▶ 11.4, “Physical planning” on page 479
- ▶ 11.5, “Energy management” on page 484



## 11.1 Introduction

The following options are available for physically installing the server:

- ▶ Radiator cooling
- ▶ Power Distribution Unit (PDU)
- ▶ On a raised floor or nonraised floor
- ▶ I/O and power cables can exit under the raised floor or off the top of the server frames

**Bulk Power Assembly (BPA):** IBM z17 does not support the BPA.

For more information about physical planning, see *9175 Installation Manual for Physical Planning*, GC28-7049.

## 11.2 Power and cooling

The IBM z17 server can be a 1- 4 19-inch rack system, depending on the configuration. Frames are shipped separately in Arbo crates, along with separate front and rear cover sets for each frame boxed on a shipping pallet. The frames are bolted together during the installation procedure.

IBM z17 servers support installation on a raised floor or nonraised floor and are only available with the following power and cooling options:

- ▶ Intelligent Power Distribution Unit-based power (iPDU) or PDU
- ▶

All IBM z17 models include radiator-based cooling (air cooled system).

### 11.2.1 Intelligent Power Distribution Unit

The iPDUs can be ordered as the following feature codes (FCs), per customer data center power infrastructure requirements:

- ▶ 200-208V 60A/3 Phase “Delta” PDU (FC 0563)
- ▶ 380-415V 32A/3 Phase “Wye” PDU (FC 0564)

A PDU-based system can have 2 - 8 power cords, depending on the configuration. The use of iPDU on IBM z17 might enable fewer frames, which allows for more I/O slots to be available and improves power efficiency to lower overall energy costs. It also offers some standardization and ease of data center installation planning, which allows the IBM z17 to easily coexist with other platforms within the data center.

#### Power requirements

The IBM z17 operates from 2 or 4 fully redundant power supplies. These redundant power supplies each have their own power cords, or pair of power cords, which allows the system to survive the loss of customer power to either power cord or power cord pair.

The IBM z17 is designed with a fully redundant power system. To make full use of the redundancy that is built into the server, the PDUs within one pair must be powered from different power distribution panels. In that case, if one PDU in a pair fails, the second PDU ensures continued operation of the server without interruption.



The second, third, and fourth PDU pairs are installed dependent on other CPC or PCIe+ I/O drawers installed. The locations of the PDU pairs and frames are listed in Table 11-1.

*Table 11-1 PDU pairs and frames location*

Frame C	Frame B	Frame A
N/A	N/A	PDU A3 / A4
PDU C1 / C2	PDU B1 / B2	PDU A1 / A2

Power cords for the PDUs are attached to the options that are listed in Table 11-2.

*Table 11-2 Power cords for PDUs*

Supply type	Input voltage	Input frequency	Input current rating
2, 4, 6, or 8, 3-phase power cords	200 - 240 V AC	50/60 Hz (47 - 63 Hz with tolerance)	48 A
2, 4, 6, or 8, 3-phase power cords	380 - 415 V AC	50/60 Hz (47 - 63 Hz with tolerance)	24 A

A rear view of a maximum configured PDU-powered system with four CPC drawers and 12 PCIe+ I/O drawers is shown in Figure 11-1 on page 474.

:



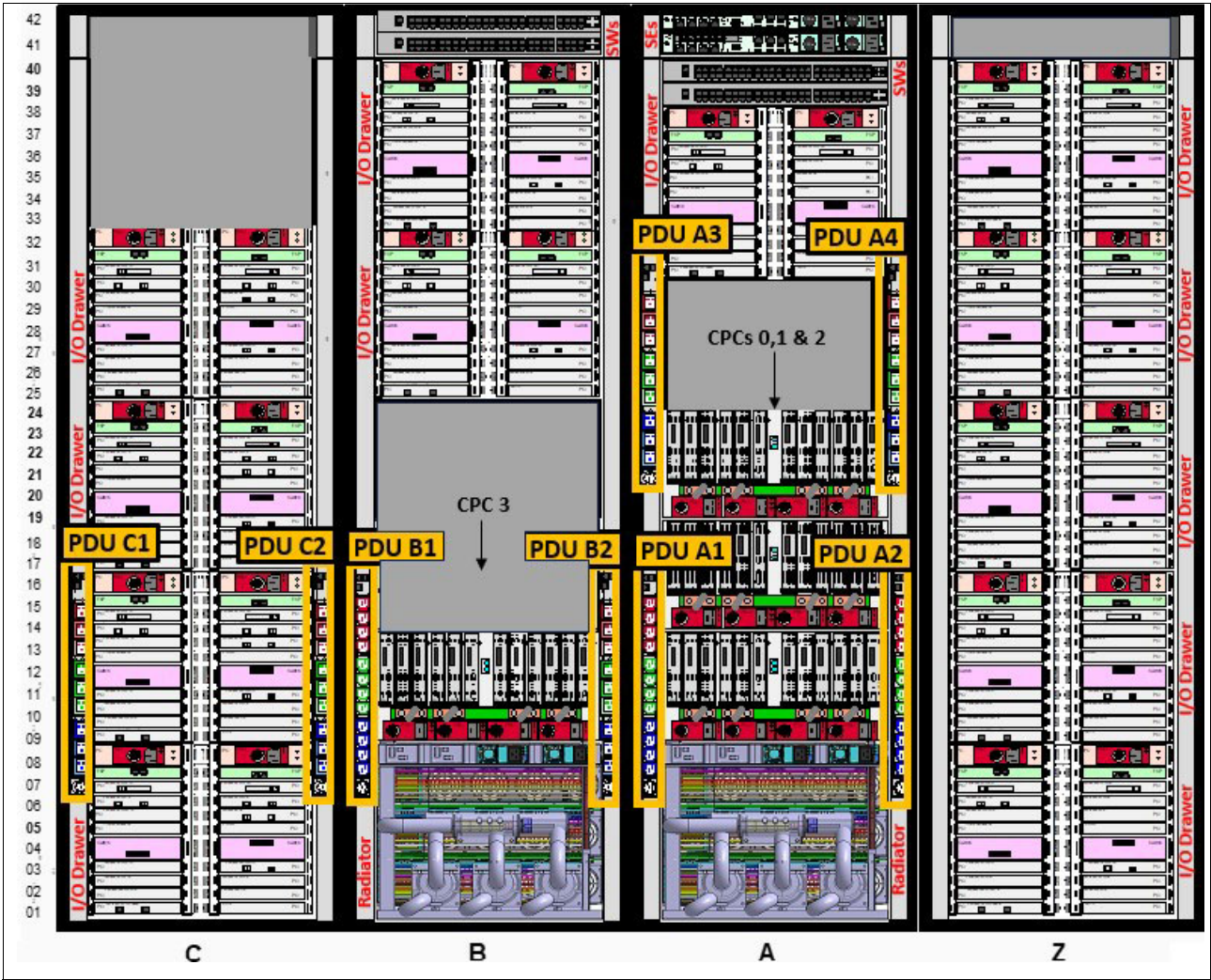


Figure 11-1 Rear View of maximum configured PDU system

The number of PDUs and power cords that are required based on the number of CPC drawers and PCIe+ I/O drawers are in Table 11-3.

Table 11-3 Number of PDUs installed

Number of CPCs	Number of PCIe+ I/O drawers												
	0	1	2	3	4	5	6	7	8	9	10	11	12
1	2	2	2	2						6	6	6	6
2	4	4	4	4	4	4	4	4	6	6	6	6	6
3	4	4	4	4	4	4	4	6	6	6	6	6	N/A
4	6	6	6	6	6	6	6	6	6	6	8	8	8

Power consumption

The utility power consumption for the IBM z17 for PDU option is listed in Table 11-4 on page 475.



Table 11-4 IBM z17 utility power consumption for PDU

FC CPC number	Number of PCIe+ I/O drawers												
	0	1	2	3	4	5	6	7	8	9	10	11	12
max43 FC0571	5.1 kw	6.0 kw	7.0 kw	7.9 kw	8.8 kw	9.7 kw	10.6 kw	N/A	N/A	N/A	N/A	N/A	N/A
max90 FC0572	8.8 kw	9.8 kw	10.8 kw	11.7 kw	12.6 kw	13.5 kw	14.5 kw	15.4 kw	16.4 kw	17.3 kw	18.2 kw	19.1 kw	20.0 kw
max136 FC0573	12.7 kw	13.7 kw	14.6 kw	15.6 kw	16.5 kw	17.4 kw	18.3 kw	19.3 kw	20.2 kw	21.1 kw	22.0 kw	22.9 kw	N/A
max183 FC0574	13.5 kw	14.5 kw	15.6 kw	16.6 kw	17.7 kw	18.7 kw	19.8 kw	20.8 kw	21.9 kw	22.9 kw	24.0 kw	25.0 kw	25.8 kw
max208 FC0575	17.1 kw	18.1 kw	19.0 kw	20.0 kw	20.9 kw	21.8 kw	22.7 kw	23.6 kw	24.5 kw	25.4 kw	26.3 kw	27.2 kw	28.1 kw
<b>Note:</b> Consider the following points: <ul style="list-style-type: none"> <li>▶ The power values that are listed in this table assume the CPC process drawer and PCIe+ I/O drawers are plugged to the maximum with highest power features (that is, memory and I/O adapters and fan-outs). Also assumed is that maximum ambient temperature is used.</li> <li>▶ Typical configurations and data center conditions result in lower power. A calculator available on Resource Link calculates power and weight for specific configurations and environmental conditions.</li> </ul>													

**Considerations:** Power consumption is lower in a normal ambient temperature room, and for configurations that feature a lesser number of I/O slots, smaller amount of memory, and fewer PUs.

Power estimation for any configuration, power source, and room condition can be obtained by using the power estimation tool that is available at the [IBM Resource Link website](#) (login required).

On the Resource Link page, click **Tools** → **Power and weight estimation**.

## Power requirements

The 9175 operates from 2 or 4 fully redundant power supplies. These redundant power supplies each have their own power cords, or pair of power cords, which allows the system to survive the loss of customer power to either power cord or power cord pair.

If power is interrupted to one of the power supplies, the other power supply assumes the entire load and the system continues to operate without interruption. Therefore, the power cords for each power supply must be wired to support the entire power load of the system.

For the most reliable availability, the power cords in the rear of the frame should be powered from different PDUs. All power cords exit through the rear of the frame. The utility current distribution across the phase conductors (phase current balance) depends on the system configuration. Each front-end power supply is provided with phase switching redundancy.

The loss of an input phase is detected and the total input current is switched to the remaining phase pair without any power interruption. Depending on the configuration input power draw, the system can run from several minutes to indefinitely in this condition. Because most single phase losses are transients that recover in seconds, this redundancy provides protection against virtually all single phase outages.



## 11.2.2 Cooling requirements

The IBM z17 cooling system is manufactured as a radiator (air) cooled system. The dual chip modules (DCMs) are cooled with an internal water loop. The liquid in the internal water circuit is cooled by using a radiator. I/O drawers, PCIe I/O drawers, power enclosures, and CPC drawers are cooled by chilled air with blowers.

The IBM z17 servers include a recommended (long-term) ambient temperature range of 18°C (64.4°F) - 27°C (80.6°F). The minimum allowed ambient temperature is 5°C (41°F); the maximum allowed temperature is 40°C (104°F).

For more information about cooling requirements, see *9175 Installation Manual for Physical Planning*, GC28-7049.

### Radiator (air) cooling

The following radiator (air) cooling options are available:

- ▶ A-Frame radiator air-cooled feature code (FC 4045)
- ▶ B-Frame radiator air-cooled feature code (FC 4046)

The radiator cooling system requires chilled air to fulfill the air-cooling requirements. IBM z17 system airflow is from the front (intake, chilled air) to the rear (exhausts, warm air) of the frames. The chilled air is provided through perforated floor panels in front of the system

The hot and cold airflow and the arrangement of server aisles are shown in Figure 11-3 on page 477.

### Cooling design

The IBM z17 continues to offer a client air cooled (RCU, internal radiator) system for cooling the high performance processor modules.

Processor heat is picked up by an internal liquid cooling loop and transferred to data center air. The cooling loop contains a 40% Propylene Glycol - 60% water solution. This solution allows IBM to ship IBM z17 systems filled with liquid. IBM z17 will not have an IBM Fill and Drain tool (FDT), and will not require the handling or storage of an liquid in the field.

The IBM z17 is a closed loop, preassembled and pretested enterprise-class cooling solution that once filled requires no scheduled maintenance.

All liquid carrying components are very robust and designed not to leak, qualified not to leak and tested not to leak. IBM z17 is designed to detect, report and contain a leak if in the extremely unlikely case one were to occur. With this design, no additional data center infrastructure is needed to support the IBM z17 radiator based cooling system.



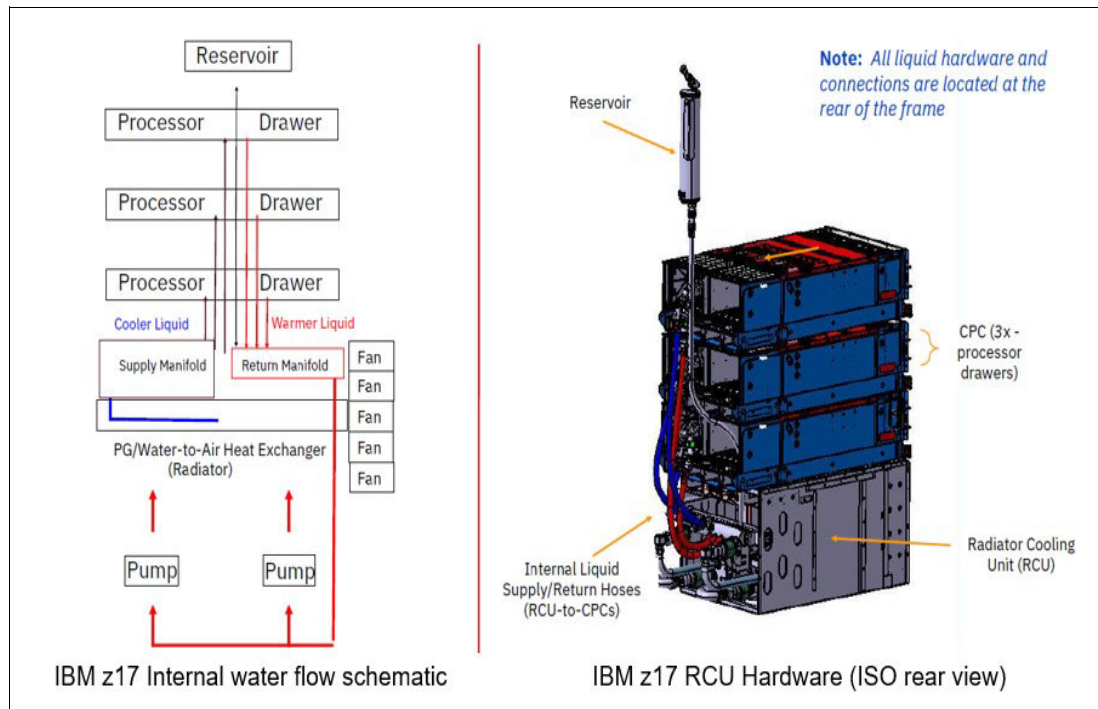


Figure 11-2 IBM z17 internal water flow schematics

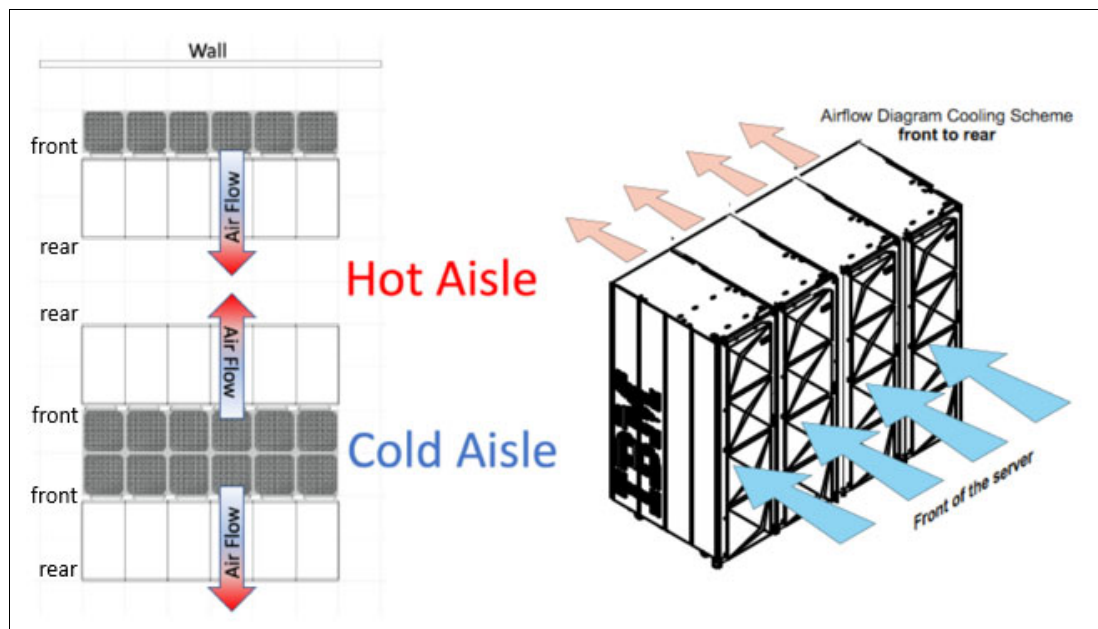


Figure 11-3 Hot and cold aisles

As shown in Figure 11-3, rows of servers must be placed front-to-front. Chilled air is provided through perforated floor panels that are placed in rows between the fronts of servers (the cold aisles). Perforated tiles generally are not placed in the hot aisles.



If your computer room causes the temperature in the hot aisles to exceed a comfortable temperature, add as many perforated tiles as necessary to create a satisfactory comfort level. Heated exhaust air exits the computer room above the computing equipment.

For more information about the requirements for air-cooling options, see *9175 Installation Manual for Physical Planning*, GC28-7049.

## 11.3 Physical specifications

This section describes the weights and dimensions of IBM z17 server.

The IBM z17 can be installed on a raised or nonraised floor. (For more information about weight distribution and floor loading tables, see *IBM 9175 Installation Manual for Physical Planning*, GC28-7049. This data is used with the maximum frame weight, frame width, and frame depth to calculate the floor loading.

Weight estimates for the maximum system configurations on the 9175 PDU-based system are listed in Figure 11-10 on page 485.

Table 11-5 Weights for maximum machine configurations (PDU)

Frame configurations	Individual frame weights kg (lbs)				Total weight kg (lbs)
	Z Frame	A Frame	B Frame	C Frame	
<b>A</b>	-	795 kg (1753 lbs)	-	-	795 kg (1753 lbs)
<b>ZA</b>	712 kg (1569 lbs)	739 kg (1629 lbs)	-	-	1451 kg (3198 lbs)
<b>ZAC</b>	712 kg (1569 lbs)	720 kg (1587 lbs)	-	740 kg (1632 lbs)	2172 kg (4788 lbs)
<b>AB</b>	-	739 kg (1629 lbs)	704 kg (1552 lbs)	-	1443 kg (3181 lbs)
<b>ZAB</b>	712 kg (1569 lbs)	720 kg (1587 lbs)	704 kg (1552 lbs)	-	2136 kg (4709 lbs)
<b>ZABC</b>	712 kg (1569 lbs)	720 kg (1587 lbs)	685 kg (1510 lbs)	550 kg (1213 lbs)	2667 kg (5880 lbs)
<b>Notes:</b> Consider the following points: <ul style="list-style-type: none"> <li>▶ Weight is based on the maximum system configuration.</li> <li>▶ All weights are approximate and do not include Earthquake Kit hardware.</li> <li>▶ Ensure that the raised floor on which you are installing the server can support the weight.</li> </ul>					

The power and weight estimation tool for IBM Z servers on [IBM Resource Link](#) (log in required) covers the estimated weight for your designated configuration.

On the Resource Link web page, click **Tools** → **Power and weight estimation**.



## 11.4 Physical planning

This section describes the floor mounting, power, and I/O cabling options. For more detailed information, see *IBM 9175 Installation Manual for Physical Planning*, GC28-7049.

**Note:** On the IBM z17, all I/O cabling and power cords enter the rear of the machine; therefore, all related features for Bottom and Top Exit cabling are in the rear of the frame.

IBM z17 servers can be installed on either a raised or a nonraised floor. Figure 11-4 shows all Feature Codes combinations for Raised and Non-Raised Floor installations.

Customer Environment	Bottom Exit Cabling	Top Exit Cabling	FCs to be Ordered	Comments
Raised Floor	Yes	No	7804 only	Ships with Bottom Exit Tailgate & supports Bottom FQC FC 5827
Raised Floor	Yes	Yes, no Top Exit Enclosure	7803 & 7804	Ships with Bottom Exit Tailgate & supports Bottom FQC FC 5827
Raised Floor	Yes	Yes, with Top Exit Enclosure	7804 & 5823	Ships with Bottom Exit Tailgate & supports FQC FCs 5824, 5826 & 5827
Raised Floor	No	Yes, no Top Exit Enclosure	7803	Ships with Bottom Seal Plate and does not support FQC FCs
Raised Floor	No	Yes, with Top Exit Enclosure	5823 & 7803	Ships with Bottom Seal Plate and only supports FQC FC 5824 & 5826
Non-Raised Floor	No (not supported)	Yes, no Top Exit Enclosure	7998* & 7803	Ships with Bottom Seal Plate and does not support FQC FCs
Non-Raised Floor	No (not supported)	Yes, with Top Exit Enclosure	7998* & 5823	Ships with Bottom Seal Plate and supports FQC FCs 5824 & 5826

*\*FC 7998: Non-Raised Floor Support (flag)*

Figure 11-4 External Cabling Feature Codes Combinations

- **Important:** The Bottom Exit Cabling Feature Code 7804 is REQUIRED for routing any external cables through the bottom of the system frame(s). By itself, Feature Code 7804 supports routing of power line cords and point-to-point (pass-through) customer cables. To support structured cabling (fiber quick-connect) with Key Up / Key Down interconnect polarity, Feature Code 5827 must also be ordered. Key Up / Key Up interconnect polarity is not supported in bottom exit cabling.

To help manage the cabling when using the Top Exit Enclosure or the Bottom Exit Cabling features, the following optional Fiber Quick Connect (FQC) features are available:

- ▶ FC 5824: key-up/key-down cabling polarity
- ▶ FC 5826: key-up/key-up cabling polarity

**Important:** The Fiber Quick-Connect Bracket Feature Code 5824 is an optional addition to Top Exit Enclosure Feature Code 5823, supporting structured cabling with only Key Up / Key Down interconnect polarity. These Brackets must be ordered in quantities of 2, 4, or 6 per server frame, and the customer must specify the total order quantity for their entire IBM Z system. Each Bracket provides 16 fiber quick-connect ports, for a maximum of 96 ports per system frame.



### 11.4.1 Top Exit Cabling without Tophat (FC 7803)

This feature allows power and cables still can be run out the top of the rack through two adjustable openings at the top rear of the rack, as shown in Figure 11-5.

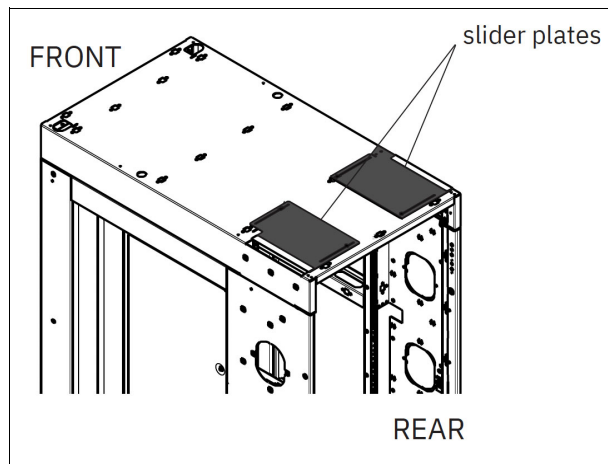


Figure 11-5 Slider plates

If this Top Exit Cabling feature is not ordered in conjunction with the Bottom Exit Cabling feature, a bottom seal plate will be shipped in order to seal the bottom side of the frame.

**Note:** The optional Top Exit Cabling feature is mandatory when ordering one of the Top Exit Fiber Quick Connect (FQC) features FC 5824 and FC 5826.



### 11.4.2 Bottom Exit Cabling feature (FC 7804)

The Bottom Exit cabling feature is required for raised floor environments, where I/O cabling or power cords must exit from the bottom of the frame. This feature includes the hardware to allow bottom exit, and other components for cable management and filler plates to preserve the recommended air circulation, as shown in Figure 11-6.

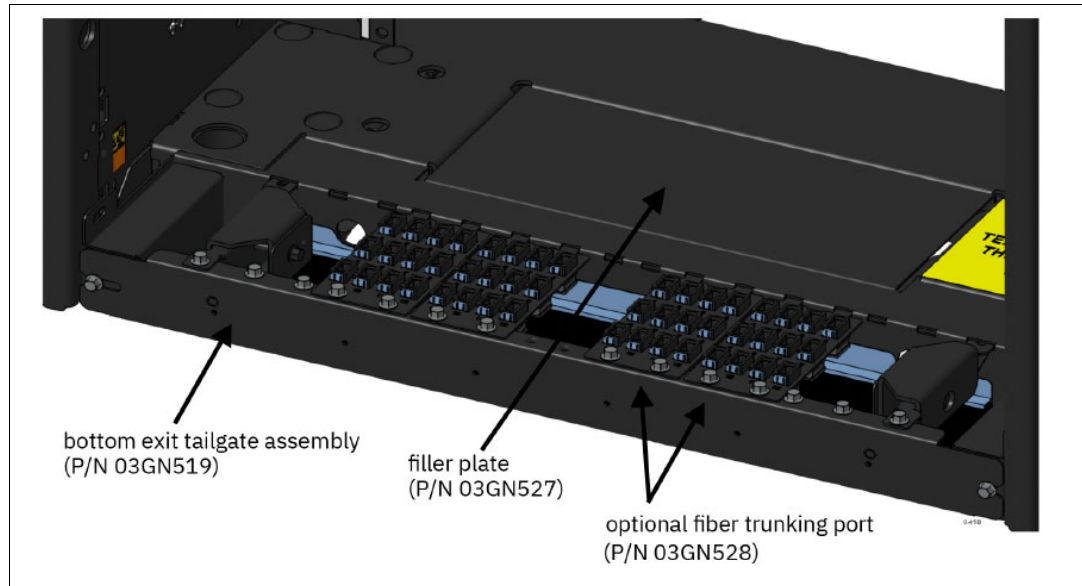


Figure 11-6 Bottom exit cabling feature (FC7804)

**Note:** This feature is mandatory when ordering the Bottom Exit Fiber Quick Connect (FQC) feature FC 5827.

### 11.4.3 Top Exit Enclosure feature (FC 5823)

This optional feature, sometimes called “Tophat”, includes additional top exit enclosure hardware for cable egress & strain relief as well as to provide a cold/hot air containment barrier’s contact location to the top system, as shown in Figure 11-7 on page 482.



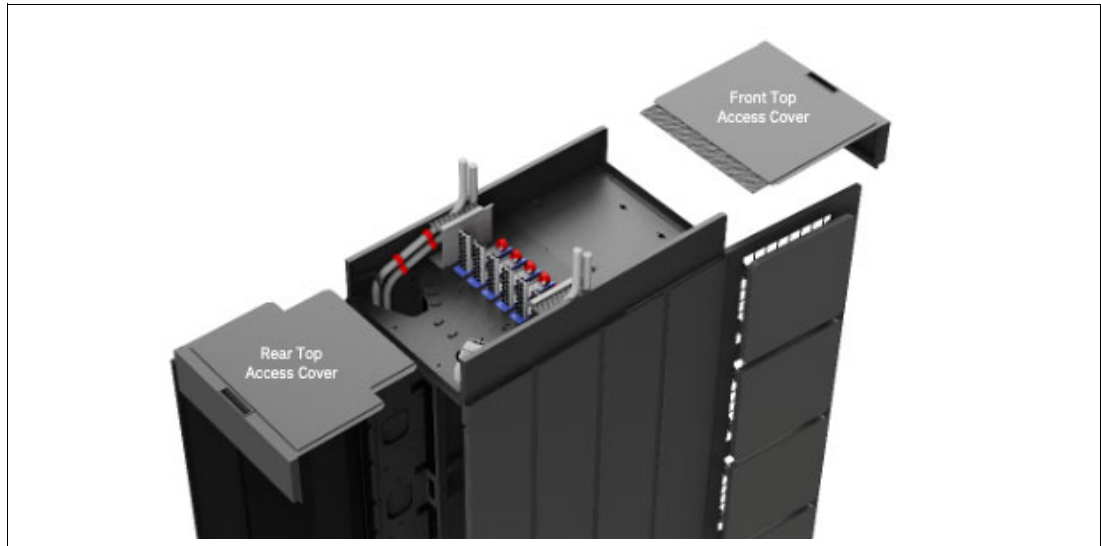


Figure 11-7 Top Exit Enclosure feature

**Notes:**

- ▶ The same extra hardware is provided for every frame in the configuration.
- ▶ This feature is mandatory when ordering one of the Top Exit Fiber Quick Connect (FQC) features FC 5824 and FC 5826.

The Top Exit Enclosure feature adds 177.5 mm (6.98 in.) to the height of the frame and approximately 12.2 kg (27 lbs) to the weight.

If the Top Exit Enclosure feature is not ordered, two sliding plates are available on the top of the frame (one on each side of the rear of the frame) that can be partially opened. By opening these plates, I/O cabling and power cords can exit the frame. The plates should be removed to install the Top Exit Enclosure feature as shown in Figure 11-8 on page 483.



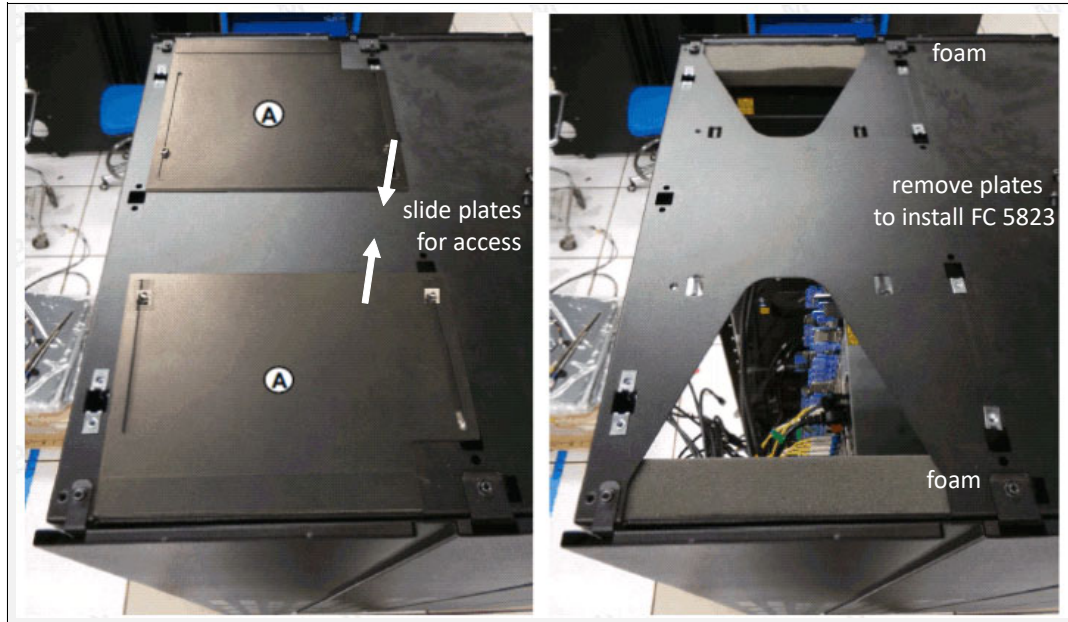


Figure 11-8 Top Exit access plate

#### 11.4.4 Frame Bolt-down kit

An Earthquake Kit, RF (FC 8014 and FC 8015) is available for the IBM z17. The kit provides hardware to enhance the ruggedness of the frame, the frame stiffener, and to tie down the frame to a concrete floor.

The frame tie-down kit can be used on a nonraised floor (FC 8015) where the frame is secured directly to a concrete floor, or on a raised floor (FC 8014) where the frame is secured to the concrete floor underneath the raised floor.

Raised floors 241.3 mm (9.5 inches) - 1270 mm (50 inches) are supported.

The kits help secure the frames and their contents from damage when they are exposed to shocks and vibrations, such as in a seismic event. The frame tie-downs are intended for securing a frame that weighs up to 1308 kg (2885 lbs).

For more information, see *IBM 9175 Installation Manual for Physical Planning*, GC28-7049.

#### 11.4.5 Service clearance areas

IBM z17 servers require specific service clearance (see Figure 11-9 on page 484) to ensure the fastest possible repair in the unlikely event that a part must be replaced. Failure to provide enough clearance to open the front and rear covers results in extended service times or outages.



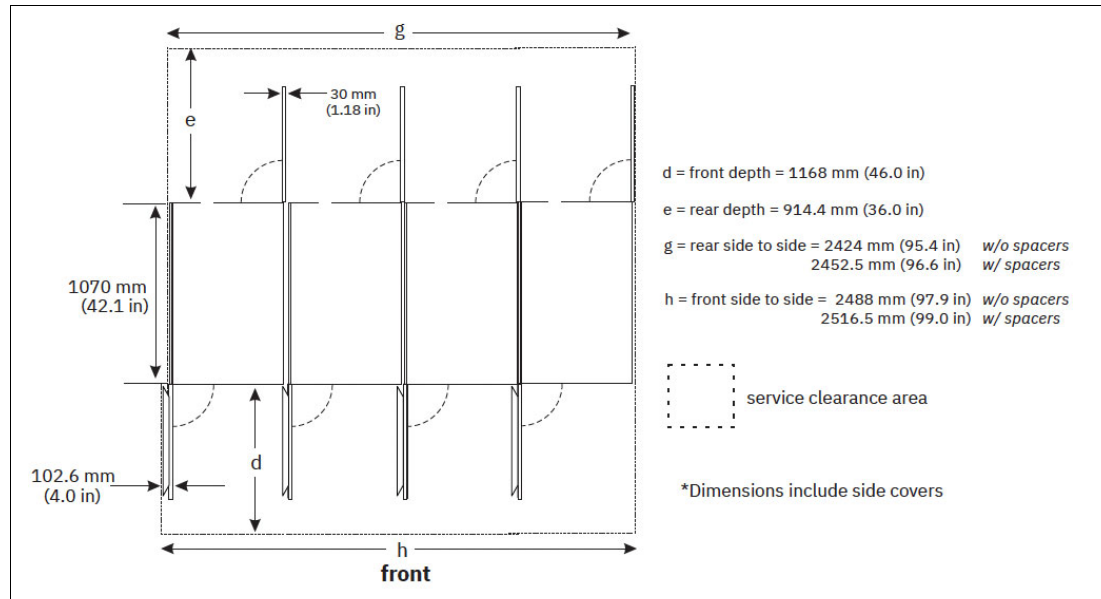


Figure 11-9 Service clearance area (four frames)

For more information, see *IBM 9175 Installation Manual for Physical Planning*, GC28-7049.

## 11.5 Energy management

This section describes the elements of energy management to help you understand the requirements for power and cooling, monitoring and trending, and reducing power consumption.

The hardware components in the IBM z17 are monitored and managed by the energy management component in the Support Element (SE) and Hardware Management Console (HMC). The user interfaces (UIs) of the SE and HMC provide views, such as the Monitors Dashboard, Environmental Dashboard and Energy Optimization Advisor.

Also via API these information are available. For more information see *10.5.9, “Automated operations via APIs” on page 464*.

The following tools are available to plan and monitor the energy consumption of IBM z17 servers:

- Power and Weight estimation tool on [Resource Link](#).
- Management -> **Energy Optimization Advisor** task for maximum potential power on HMC
- Monitor -> **Monitors Dashboard** and **Environmental Dashboard** tasks on HMC



## 11.5.1 Environmental monitoring

This section describes energy monitoring for IBM Z systems.

### Energy Optimization Advisor

This window is started from the HMC targeting the system and task under Energy Management. This window displays recommendations that reduces power consumption based on the present system operation. Select the advice hyperlink to provide specific recommendations for your system.

The window displays the following recommendations:

- ▶ Thermal Advice
- ▶ Processor Utilization Advice

Select the advice hyperlinks to provide specific recommendations for your system, as shown in Figure 11-10.

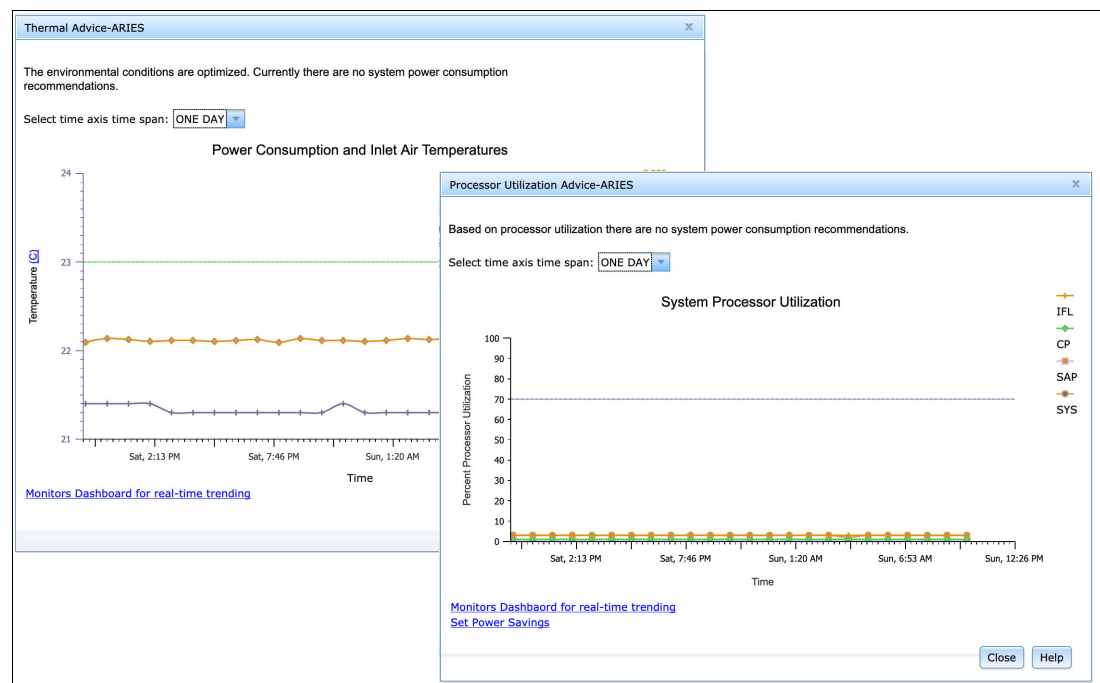


Figure 11-10 Energy Optimization Advisor

### Monitors Dashboard task

In IBM z17 servers, the **Monitors Dashboard** task in the Monitor task group provides a tree-based view of resources. Multiple graphical views display data, including history charts. This task monitors processor and channel usage. It produces data that includes power monitoring information, power consumption, and the air input temperature for the server. Since IBM z16 also the Power Consumption (kW) for LPARs are shown.

An example of the Monitors Dashboard task is shown in Figure 11-11 on page 486.



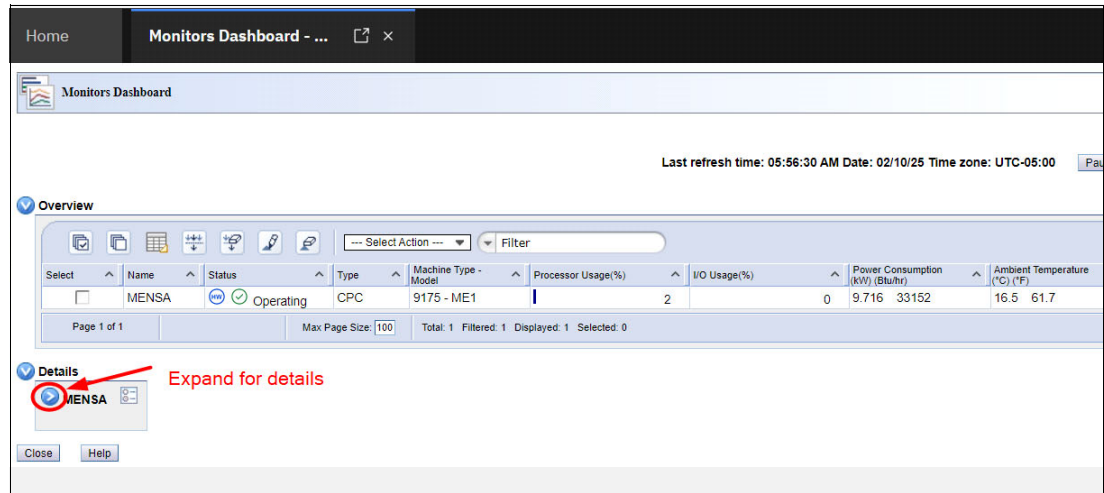


Figure 11-11 Example of Monitors Dashboard

## Environmental dashboard task

The **Environmental dashboard** task (see Figure 11-12 on page 487) is part of the Monitor task group. It provides historical power consumption, Processor utilization and thermal information for the CPC.

The data is presented in table format and graphical “histogram” format. The data also can be exported to a .xlsx-formatted file so that the data can be imported into a spreadsheet. For this task, you must use a web browser to connect to an HMC.



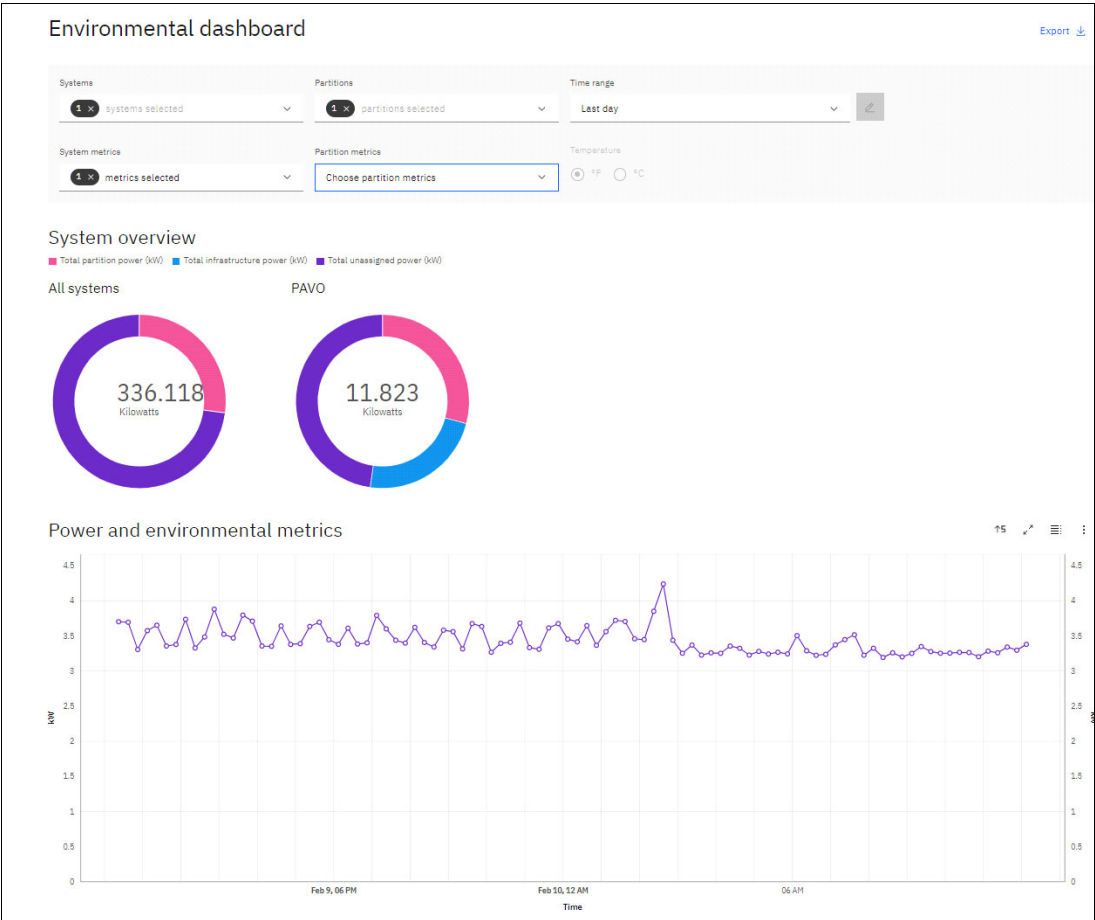


Figure 11-12 Example of Environmental dashboard

Figure 11-13







## 12



# Performance and capacity planning

This chapter describes the performance and capacity planning of IBM z17.

**Note:** Throughout this chapter, *IBM z17* refers to IBM z17 Model ME1 (Machine Type 9175) unless otherwise specified.

This chapter includes the following topics:

- ▶ 12.1, “IBM z17 performance characteristics” on page 490
- ▶ 12.2, “IBM z17 Large System Performance Reference ratio” on page 492
- ▶ 12.3, “Fundamental components of workload performance” on page 493
- ▶ 12.4, “Relative Nest Intensity” on page 495
- ▶ 12.5, “LSPR workload categories based on L1MP and RNI” on page 497
- ▶ 12.6, “Relating production workloads to LSPR workloads” on page 497
- ▶ 12.7, “CPU MF counter data and LSPR workload type” on page 498
- ▶ 12.8, “Workload performance variation” on page 499
- ▶ 12.9 “Capacity planning considerations for IBM z17”, on page 487



## 12.1 IBM z17 performance characteristics

The largest IBM z17 Model ME1 Feature Max208 (9175-7K8) is expected to provide approximately 15% more capacity than the largest IBM z16 Model A01 Feature Max200 (3931-7K0) with some variation based on workload and configuration.

Single processor capacity of IBM z17 for equal n-way at common client configurations is approximately 11% greater than on IBM z16 with some variation based on workload and configuration.

Figure 12-1 shows a system performance comparison of successive IBM Z servers.

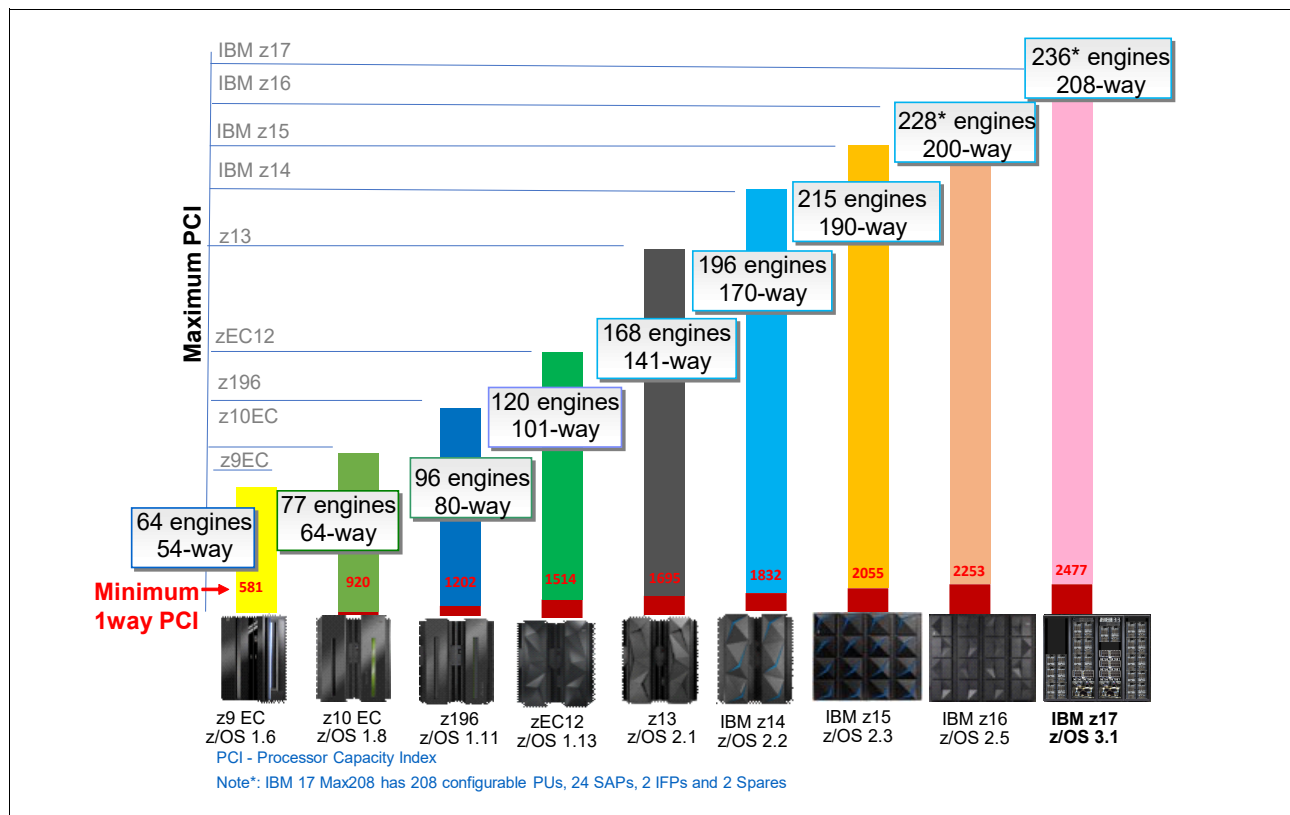


Figure 12-1 System performance comparison of successive IBM Z servers

### 12.1.1 IBM z17 single-thread capacity

**Note:** The acronym PCI stands for *Processor Capacity Index*.

Operating system support varies for the number of “engines” that are supported.

The IBM z17 processor chip runs at 5.5 GHz clock speed, which is a 5.8% improvement over the 5.2 GHz IBM z16 processor chip, and the performance is increased. For N-way



processors model, it increases 1.11x plus +/- 2.0% on average at equal N-way configuration. These numbers differ depending on the workload type and LPAR configuration.

### 12.1.2 IBM z17 SMT capacity

From IBM z13 to IBM z17, customers can choose to run two threads on IFL and zIIP cores by using SMT mode. SMT increases throughput by 10 - 40% (average 25%), depending on workload.

For zIIPs and IFLs, the performance improvement per SMT enabled core for z17 (9175) versus z16 (3931) is approximately 11%, which is the same improvement for zIIPs and IFLs in a single thread.

### 12.1.3 IBM Integrated Accelerator for zEnterprise Data Compression (zEDC)

Starting with z13, IBM introduced the zEnterprise Data Compression (zEDC) Express PCIe feature, which increases efficiency for data storing and data transfers.

The zEDC Express feature was adopted by enterprises because it helps with software costs for compression and decompression operations (by offloading these operations), and increases data encryption (compression before encryption) efficiency.

With IBM z15, the zEDC Express functions were moved off from the PCIe infrastructure into the processor chip. By moving the compression and decompression into the processor on-chip, IBM z15, IBM z16 and IBM z17 processors provides a new level of performance for these tasks and eliminates the need for the zEDC Express feature virtualization. It also brings new use cases to the platform.

The IBM z17 continues to support IBM Integrated Accelerator for zEDC. For more information, see Chapter B, “IBM Integrated Accelerator for zEnterprise Data Compression” on page 525.

### 12.1.4 Primary performance improvement drivers with IBM z17

The attributes and design points of IBM z17 contribute to overall performance and throughput improvements as compared to the IBM z16. The following major items contribute to IBM z17 performance improvements:

- ▶ IBM z17 microprocessor architecture:
  - 5.5GHz frequency
  - Private 128K L1-I, 128K L1-D
  - Branch prediction enhancements, I-Cache prefetching
  - Increased physical renaming registers, vector registers, Store shadow-cache enhancements
- ▶ Cache:
  - L2 Private cache (unified) increased to 36 MB
  - 10 L3's per CP chip -> 360MB vL3 cache, 2.88GB vL4 cache
  - X-bus bandwidth improvement (33% faster)
  - A-Bus link encryption, reduced cross-drawer latency
- ▶ Memory:
  - DDR5 DIMMs (DDR4 Carry-forward)



- Improved memory encryption (AES-256)
- ▶ I/O:
  - New on-chip Data Processing Unit
    - Onboarding of legacy ASIC Functionality
    - Reduced system power
  - PCIe Gen5 support
  - OSA-RoCE SMCR support (single networking port, reduced cost)
- ▶ IBM 2nd Gen Integrated Accelerator for Z AI:
  - INT8 Quantization support
  - Distribution of AI operations requests to all 8 integrated Accelerators for Z AI within a drawer for increased inference capacity
- ▶ Software and hardware:
  - z/OS HiperDispatch Optimizations
  - PR/SM Algorithm Improvements (including LPAR Resource Placement)

## 12.2 IBM z17 Large System Performance Reference ratio

The Large System Performance Reference (LSPR) provides capacity ratios among various processor families that are based on various measured workloads. It is a common practice to assign a capacity scaling value to processors as a high-level approximation of their capacities.

For z/OS studies, the capacity scaling factor that is commonly associated with the reference processor is set to a 2094-701 with a Processor Capacity Index (PCI) value of 593. This value is unchanged since z/OS V1R11 LSPR. The use of the same scaling factor across LSPR releases minimizes the changes in capacity results for an older study and provides more accurate capacity view for a new study.

Performance data for IBM z17 servers were obtained with z/OS 3.1 (running DB2 for z/OS V13, CICS TS6.1, Enterprise COBOL V6.4, and Websphere Application Server for z/OS V9.0.5.14). All IBM Z server generations are measured in the same environment with the same workloads at high usage.

**Note:** If your software configuration is different from what is described here, the performance results might vary.

The largest IBM z17 208 way configuration (9175-7K8) is expected to provide approximately 15% more capacity than the largest IBM z16 200 way (3931-7K0). However, the observed performance increase varies depending on the workload type.

Consult the LSPR when you consider performance on the IBM z17. The range of performance ratings across the individual LSPR workloads is likely to include a large spread. Performance of the individual logical partitions (LPARs) varies depending on the fluctuating resource requirements of other partitions and the availability of processor units (PUs). Therefore, it is important to know which LSPR workload type suite your production environment. For more information, see 12.8, “Workload performance variation” on page 499.

For more information about performance, see [this web page](#) of the IBM Support Docs website.



For more information about millions of service units (MSU) ratings, see this [IBM Z resources](#) web page.

## 12.2.1 LSPR workload suite

Historically, LSPR capacity tables, including pure workloads and mixes, were identified with application names or a *software* characteristic; for example, CICS, IMS, OLTP-T,<sup>1</sup> CB-L,<sup>2</sup> LoIO-mix,<sup>3</sup> and TI-mix.<sup>4</sup> However, capacity performance is more closely associated with how a workload uses and interacts with a specific processor *hardware* design.

The CPU Measurement Facility (CPU MF) data that was introduced on the z10 provides insight into the interaction of workload and *hardware design* in production workloads. CPU MF data helps LSPR to adjust workload capacity curves that are based on the underlying hardware sensitivities; in particular, the processor access to caches and memory.

This processor access to caches and memory is called *Relative Nest Intensity* (for more information, see 12.4, “Relative Nest Intensity” on page 495). By using this data, LSPR introduces three workload capacity categories that replace all older primitives and mixes.

LSPR contains the internal throughput rate ratios (ITRRs) for the IBM z17 and the previous generation processor families. These ratios are based on measurements and projections that use standard IBM benchmarks in a controlled environment.

**Note:** The throughput that any user experiences can vary depending on the amount of multiprogramming in the user’s job stream, the I/O configuration, and the workload that is processed. Therefore, no assurance can be given that an individual user can achieve throughput improvements that are equivalent to the performance ratios that are stated.

## 12.3 Fundamental components of workload performance

Workload performance is sensitive to the following major factors:

- ▶ Instruction path length
- ▶ Instruction complexity
- ▶ Memory hierarchy and memory nest

These factors are described next.

### 12.3.1 Instruction path length

A *transaction* or *job* runs a set of instructions to complete its task. These instructions are composed of various paths through the operating system, subsystems, and application. The total count of instructions that are run across these software components is referred to as the *transaction* or *job path length*.

The path length varies for each transaction or job, and depends on the complexity of the tasks that must be run. For a specific transaction or job, the application path length tends to stay the same, assuming that the transaction or job is asked to run the same task each time.

<sup>1</sup> Traditional online transaction processing workload (formerly known as IMS).

<sup>2</sup> Commercial batch with long-running jobs.

<sup>3</sup> Low I/O Content Mix Workload.

<sup>4</sup> Transaction Intensive Mix Workload.



However, the path length that is associated with the operating system or subsystem can vary based on the following factors:

- ▶ Competition with other tasks in the system for shared resources. As the total number of tasks grows, more instructions are needed to manage the resources.
- ▶ The number of logical processors (*n-way*) of the image or LPAR. As the number of logical processors grows, more instructions are needed to manage resources that are serialized by latches and locks.

### 12.3.2 Instruction complexity

The type of instructions and the sequence in which they are run interacts with the design of a microprocessor to affect a performance component. This factor is defined as *instruction complexity*. The following design alternatives affect this component:

- ▶ Cycle time (GHz)
- ▶ Instruction architecture
- ▶ Pipeline
- ▶ Superscalar
- ▶ Out-of-order execution
- ▶ Branch prediction
- ▶ Translation Look-aside Buffer (TLB)
- ▶ Transactional Execution (TX)
- ▶ Single instruction multiple data instruction set (SIMD)
- ▶ Simultaneous multi-threading (SMT)<sup>5</sup>

As workloads are moved between microprocessors with various designs, performance varies. However, when on a processor, this component tends to be similar across all models of that processor.

### 12.3.3 Memory hierarchy and memory nest

The *memory hierarchy* of a processor generally refers to the caches, data buses, and memory arrays that stage the instructions and data that must be run on the microprocessor to complete a transaction or job.

The following design choices affect this component:

- ▶ Cache size
- ▶ Latencies (sensitive to distance from the microprocessor)
- ▶ Number of levels, the Modified, Exclusive, Shared, Invalid (MESI) protocol, controllers, switches, the number and bandwidth of data buses, and so on.

With IBM z17, physical L3/L4 caches no longer exist. L2 caches that are on each processor core are virtual L3/L4 caches on IBM z17. For more information, see Chapter 3, “Central processor complex design” on page 71.

---

<sup>5</sup> Available for IFL, zIIP, and SAP processors only.



A memory nest in an IBM z17 CPC drawer is shown in Figure 12-2.

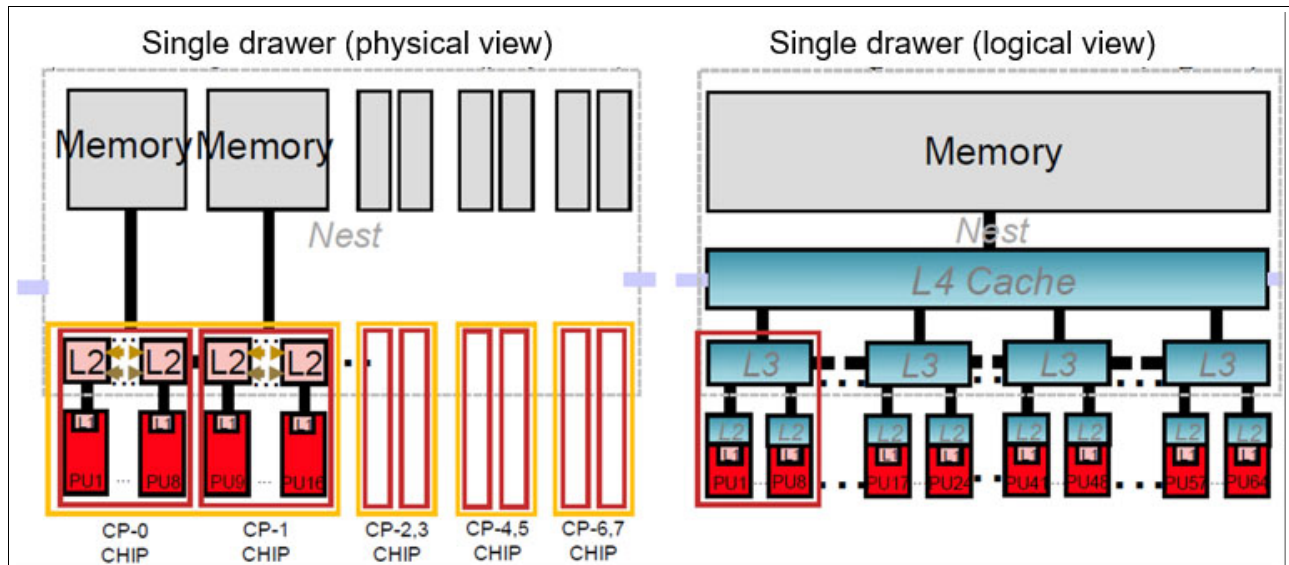


Figure 12-2 IBM z17 physical and virtual single drawer memory hierarchy

Workload performance is sensitive to how deep into the memory hierarchy the processor must go to retrieve the workload instructions and data for running. The best performance occurs when the instructions and data are in the caches that are nearest the processor because little time is spent waiting before running. If the instructions and data must be retrieved from farther out in the hierarchy, the processor spends more time waiting for their arrival.

As workloads are moved between processors with various memory hierarchy designs, performance varies because the average time to retrieve instructions and data from within the memory hierarchy varies. Also, when on a processor, this component continues to vary because the location of a workload's instructions and data within the memory hierarchy is affected by several factors that include, but are not limited to, the following factors:

- ▶ Locality of reference
- ▶ I/O rate
- ▶ Competition from other applications and LPARs

## 12.4 Relative Nest Intensity

The most performance-sensitive area of the memory hierarchy is the activity to the memory nest. This area is the distribution of activity to the shared caches and memory.

The term *Relative Nest Intensity* (RNI) indicates the level of activity to this part of the memory hierarchy. By using data from CPU MF, the RNI of the workload that is running in an LPAR can be calculated. The higher the RNI, the deeper into the memory hierarchy the processor must go to retrieve the instructions and data for that workload.



RNI reflects the distribution and latency of sourcing data from shared caches and memory, as shown in Figure 12-3.

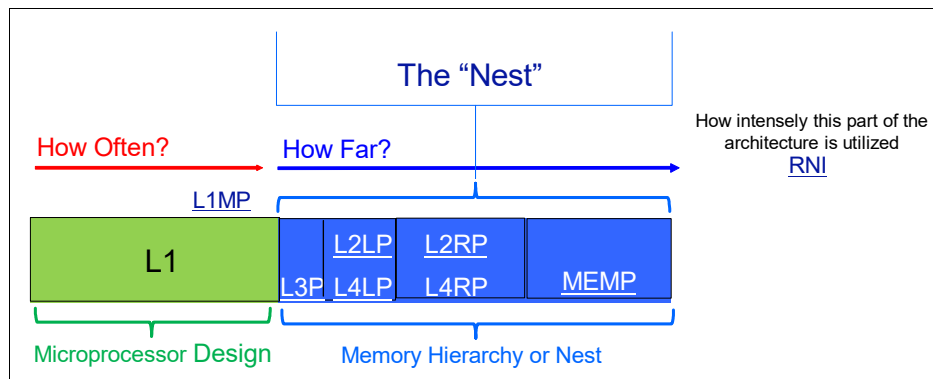


Figure 12-3 Relative Nest Intensity

Many factors influence the performance of a workload. However, these factors often are influencing the RNI of the workload. The interaction of all these factors results in a net RNI for the workload, which in turn directly relates to the performance of the workload.

These factors are tendencies, not absolutes. For example, a workload might have a low I/O rate, intensive processor use, and a high locality of reference, which all suggest a low RNI. However, it might be competing with many other applications within the same LPAR and many other LPARs on the processor, which tends to create a higher RNI. It is the net effect of the interaction of all these factors that determines the RNI.

The traditional factors that were used to categorize workloads in the past are shown with their RNI tendency in Figure 12-4.

Relative Nest Intensity		
Low		High
Batch	Application Type	Transactional
Low	I/O Rate	High
Single	Application Mix	Many
Intensive	CPU Usage	Light
Low	Dispatch Rate	High
High locality	Data Reference Pattern	Diverse
Simple	LPAR Configuration	Complex
Extensive	Software Configuration Tuning	Limited

Figure 12-4 Traditional factors that were used to categorize workloads

Little can be done to affect most of these factors. An application type is whatever is necessary to do the job. The data reference pattern and processor usage tend to be inherent to the nature of the application. The LPAR configuration and application mix are mostly a function of what must be supported on a system. The I/O rate can be influenced somewhat through buffer pool tuning.

However, one factor, *software configuration tuning*, is often overlooked but can have a direct effect on RNI. This term refers to the number of address spaces (such as CICS



application-owning regions [AORs] or batch initiators) that are needed to support a workload. This factor always existed, but its sensitivity is higher with the current high frequency microprocessors. Spreading the same workload over more address spaces than necessary can raise a workload's RNI. This increase occurs because the working set of instructions and data from each address space increases the competition for the processor caches.

Tuning to reduce the number of simultaneously active address spaces to the optimum number that is needed to support a workload can reduce RNI and improve performance. In the LSPR, the number of address spaces for each processor type and *n-way* configuration is tuned to be consistent with what is needed to support the workload. Therefore, the LSPR workload capacity ratios reflect a presumed level of software configuration tuning. Retuning the software configuration of a production workload as it moves to a larger or faster processor might be needed to achieve the published LSPR ratios.

## 12.5 LSPR workload categories based on L1MP and RNI

A workload's L1MP (Level 1 Miss Per 100 instructions or percentage of data and instruction references that miss the L1 cache) and RNI are the most influential factors in determining workload performance. Other more traditional factors, such as application type or I/O rate, include RNI tendencies. However, it is the L1MP and the net RNI of the workload that is the underlying factor in determining the workload's performance.

The LSPR now runs various combinations of former workload primitives, such as CICS, Db2, IMS, OSAM, VSAM, WebSphere, COBOL, and utilities, to produce capacity curves that span the typical range of RNI.

The following workload categories are represented in the LSPR tables:

- ▶ **LOW** (relative nest intensity)  
A workload category that represents light use of the memory hierarchy.
- ▶ **AVERAGE** (relative nest intensity)  
A workload category that represents average use of the memory hierarchy. This category is expected to represent most production workloads.
- ▶ **HIGH** (relative nest intensity)  
A workload category that represents a heavy use of the memory hierarchy.

These categories are based on the L1MP and the RNI. The RNI is influenced by many variables, such as application type, I/O rate, application mix, processor usage, data reference patterns, LPAR configuration, and the software configuration that is running. CPU MF data can be collected by z/OS System Measurement Facility on SMF 113 records or by z/VM Monitor starting with z/VM V5R4.

For more information about how z/VM Monitor captures CPU MF records visit the following link: <https://www.ibm.com/perf/tips/cpumf.html>

## 12.6 Relating production workloads to LSPR workloads

Historically, the following techniques were used to match production workloads to LSPR workloads:

- ▶ Application name (a client that is running CICS can use the CICS LSPR workload)
- ▶ Application type (create a mix of the LSPR online and batch workloads)



- I/O rate (the low I/O rates that are used a mix of low I/O rate LSPR workloads)

The IBM z Processor Capacity Reference (IBM zPCR) tool supports the following workload categories:

- Low
- Low-Average
- Average
- Average-high
- High

For more information about the no-charge IBM zPCR tool (which reflects the latest IBM LSPR measurements), see the [Getting Started with IBM z Processor Capacity Reference](#).

As described in 12.5, “LSPR workload categories based on L1MP and RNI” on page 497, the underlying performance sensitive factor is how a workload interacts with the processor hardware.

## 12.7 CPU MF counter data and LSPR workload type

Beginning with the z10 processor, the hardware characteristics can be measured by using CPU MF (SMF 113) counters data. A production workload can be matched to an LSPR workload category through these hardware characteristics.

For more information about RNI, see 12.5, “LSPR workload categories based on L1MP and RNI” on page 497.

The AVERAGE RNI LSPR workload is intended to match most client workloads. When no other data is available, use the AVERAGE RNI LSPR workload for capacity analysis.

Low-Average and Average-High categories allow better granularity for workload characterization but these categories can apply on IBM zPCR only.

The CPU MF data can be used determine workload type. When available, this data allows the RNI for a production workload to be calculated.

By using the RNI and another factor from CPU MF, the L1MP (Level 1 Miss Per 100 instructions or percentage of data and instruction references that miss the L1 cache), a workload can be classified as LOW, AVERAGE, or HIGH RNI. This classification and resulting hit are automated in the IBM zPCR tool. It is preferable to use IBM zPCR for capacity sizing.

Refer to Table 12-1 for the LSPR Workload Decision Table, based in L1MP and RNI.

L1MP	RNI	LSPR Workload Match
< 3%	>= 0.75 < 0.75	AVERAGE LOW
3% to 6%	> 1.0 0.6 to 1.0 < 0.6	HIGH AVERAGE LOW
> 6%	>= 0.75 < 0.75	HIGH AVERAGE

Table 12-1 L1MP and RNI-based LSPR Workload Decision Table



**Reminder:**

- ▶ RNI is **not** a performance metric.
- ▶ RNI and L1MP allows one to match their workload to an LSPR workload
  - Any other use of RNI is not valid

Starting with z/OS V2R1 with APAR OA43366, zFS file is no longer required for CPU MF and Hardware Instrumentation Services (HIS). HIS is a z/OS function that collects hardware event data for processors in SMF records type 113, and a z/OS UNIX System Services output files.

Only SMF 113 record is required to know proper workload type by using CPU MF counter data. CPU overhead of CPUMF is minimal. Also, the amount of SMF 113 record is 1% of typical SMF 70 and 72 which RMF writes.

CPU MF and HIS can be used for deciding workload type and other purposes. For example, starting with z/OS V2R1, you can record Instruction Counts in SMF type 30 record when you activate CPU MF. Therefore, we strongly recommend that you *always* activate CPU MF.

For more information about getting CPUMF counter data, see [CPU MF - 2022 Update and WSC Experiences](#) at the IBM Techdoc Library website.

## 12.8 Workload performance variation

As the size of transistors approaches the size of atoms that stand as a fundamental physical barrier, a processor chip's performance can no longer double every two years (Moore's Law<sup>6</sup> does not apply).

A holistic performance approach is required when the performance gains are reduced because of frequency. Therefore, hardware and software synergy becomes an absolute requirement.

Starting with z13, Instructions Per Cycle (IPC) improvements in core and cache became the driving factor for performance gains. As these microarchitectural features increase (which contributes to instruction parallelism), overall workload performance variability also increases because not all workloads react the same way to these enhancements.

Because of the nature of the IBM z17 multi-CPC drawer system and resource management across those drawers, performance variability from application to application is expected.

Also, the memory and cache designs affect various workloads in many ways. All workloads are improved, with cache-intensive loads benefiting the most. For example, having more PUs per CPC drawer, each with higher capacity than IBM z16, more workload can fit on an IBM z17 CPC drawer. This configuration can result in better performance. For example, IBM z16 two drawer system model A01 can populate maximum 82 PUs (Max82).

In contrast, IBM z17 two drawer system Max90 can populate maximum 90 PUs. Therefore, eight more PUs can share caches and memories within the first and second drawers respectively, so the performance improvements is expected.

The workload variability for moving from IBM z16 to IBM z17 is expected to be stable. Workloads that are migrating from z10 EC, z196, and zEC12 to IBM z17 can expect to see similar results with slightly less variability than the migration from IBM z16 and IBM z17.

<sup>6</sup> For more information, see the [Moore's Law website](#).



Experience demonstrates that IBM Z servers can be run at up to sustained 100% utilization levels. However, most customers prefer to leave some room and run at 90% or slightly under.

## 12.9 Capacity planning considerations for IBM z17

In this section, we describe recommended ways conduct capacity planning for IBM z17.

**Do not use MIPS or MSUs for capacity planning:** Do *not* use “one number” capacity comparisons, such as MIPS or MSUs. IBM does not officially announce the processor performance as “MIPS”. MSU is only a number for software license charge and it does *not* represent the processor’s performance.

### 12.9.1 Collect CPU MF counter data

It is important to recognize the LSPR workload type of your production system. As described in 12.7, “CPU MF counter data and LSPR workload type” on page 498, the capacity of the processor is different from the LSPR workload type. By collecting the CPU MF z/OS SMF 113 record, you can recognize the workload type in a specific IBM-provided capacity planning tool. Therefore, collecting CPU MF counter data is a first step to begin the capacity planning.

### 12.9.2 Creating EDF files with CP3KEXTR

An EDF file is an input file of the IBM Z capacity planning tools. You create this file for individual z/OS SYSIDs / LPARs by using the CP3KEXTR program. The CP3KEXTR program reads SMF records and extracts needed data as input to the IBM z Processor Capacity Reference (IBM zPCR) and IBM Z Batch Network Analyzer (IBM zBNA) tools.

**Note:** You should create an EDF file for each z/OS SYSID and load all the EDFs for the same CPC into IBM zPCR at the same time so to ensure that the correct LSPR Workload is assigned to each LPAR. IBM zPCR supports using drag-n-drop for multiple EDF files.

CP3KEXTR is offered as a no-charge application. It can also create the EDF files for IBMzCP3000. IBM zCP3000 is an IBM internal tool, but you can create the EDF files for it on your system.

For more information about CP3KEXTR, see the IBM Techdoc [z/OS Data Extraction Program \(CP3KEXTR\) for IBM zPCR and IBM zBNA](#).

#### **Creating EDF file with CP3KVMXT (z/VM)**

CP3KVMXT is the VM Extract Utility for IBM zCP3000 and IBM zPCR Capacity Planning Support Tools. CP3KVMXT reads CP Monitor data from a z/VM system, and generates an Enterprise Data File (EDF) of PR/SM, system image, and workload-related measurements for input into the IBM zCP3000 and IBM zPCR capacity planning tools. A CP3KVMXT-created EDF can be used to model interactive VM workloads or workloads under guest operating systems such as Linux and can be concurrently loaded with a CP3KEXTR-created z/OS EDF for the same data intervals.



**Note:** You should create an EDF file for each z/VM system and load all the EDFs for the same CPC into IBM zPCR at the same time so to ensure that the correct LSPR Workload is assigned to each LPAR. IBM zPCR supports using drag-n-drop for multiple EDF files.

For additional information, see [CP3KVMXT - VM Extract Utility for zCP3000 and zPCR Capacity Planning Support Tools](#).

### 12.9.3 Loading EDF files to the capacity planning tool

By loading EDF files to IBM capacity planning tool, you can see the LSPR workload type that is based on CPU MF counter data.

Figure 12-5 on page 502 shows a sample IBM zPCR window of a workload type. In this example, the workload type displays in the “Assigned Workload” column. The example shows only one partition, PX11, selected. Note that all active partitions can be selected on this panel. When you load the EDF file to IBM zPCR, it automatically sets your LPAR configuration. It also makes easy to define the LPAR configuration to the IBM zPCR.



**Create LPAR Configuration from EDF**

**LPAR Configuration from EDF**

z/VM Monitor Data Set Name: AIUZ17.T150043.H AIUZ17.T160044.H  
 Extract Version: CP3KVMXT.V2R9J.02/05/25  
 EDF File Name: /Users/priyalshah/Downloads/EDF Files-2/AIUZ17.v2r9i.edf  
 Interval #3: Date=2025-01-25 Time=15:10:00 Length=00:05:00  
**CPC ID: CPCB9FB8; GP Processor Model = 9175-724**  
**z17 Host = 9175-ME1(Max136)/700 with 136 CPs: GP=24 zIIP=48 IFL=60 ICF=4**

**Create LPAR Configuration**

#1 Configuration #1

LPAR Host as specified above  
 Partition Configuration as specified below

Copy LP	LP is Active	LP from EDF	Partition Identification				Assigned Workload	Partition Configuration				HyperDispatch			CPU MF	Method Used	
			No.	Type	Name	SCP		Mode	Total LCPs	Weight	Weight %	Capping	SMT	Is Active			Parked LCPs
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		1	GP	MENSA02	z/OS-3.1	Average	SHR	4.0	10	20.0%						Default
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			zIIP	MENSA02	z/OS-3.1	Average	SHR	2.0	10	20.0%						
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		2	GP	MENSA08	z/OS-3.1	Average	SHR	4.0	10	20.0%						Default
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			zIIP	MENSA08	z/OS-3.1	Average	SHR	2.0	10	20.0%						
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		3	GP	MENSA3A	z/OS-3.1	Average	SHR	4.0	10	20.0%						Default
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			zIIP	MENSA3A	z/OS-3.1	Average	SHR	8.0	10	20.0%						
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		4	GP	MENSA32	z/OS-3.1	Average	SHR	4.0	10	20.0%						Default
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			zIIP	MENSA32	z/OS-3.1	Average	SHR	4.0	10	20.0%						
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		5	GP	MENSA38	z/OS-3.1	Average	SHR	4.0	10	20.0%						Default
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			zIIP	MENSA38	z/OS-3.1	Average	SHR	8.0	10	20.0%						
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	6	IFL	MENSA3B	z/VM-7.3	Low/LV	SHR	8.0	10	50.0%		10.4%	<input checked="" type="checkbox"/>	0.0	Low/LV	CPU MF
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		7	IFL	MENSA0A	z/VM-7.3	Average/LV	SHR	12.0	10	50.0%						Default
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		8	ICF	MENSA3E	CFCC	CFCC	SHR	2.0	10	25.0%						Default
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		9	ICF	MENSA3F	CFCC	CFCC	SHR	2.0	10	25.0%						Default
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		10	ICF	MENSA0C	CFCC	CFCC	SHR	2.0	10	25.0%						Default
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		11	ICF	MENSA0D	CFCC	CFCC	SHR	2.0	10	25.0%						Default

Default SCP for GP Partitions: ☐ z/OS ☒ z/VM

IFL Partitions: ☐ z/OS ☒ z/VM

Estimate parked LCPs where unknown for: ☐ GP partitions ☐ IFL partitions

Select All Select Active Remove All Choose Another EDF Interval

Create LPAR Configuration ☐ Remove Parked LCPs from the LCP Count when copying partitions into zPCR

Click on "Copy LP" checkbox to select partitions to be copied to the LPAR configuration

Figure 12-5 IBM zPCR LPAR Configuration from EDF window

## 12.9.4 IBM z17 Performance Best Practices

Understand and follow the IBM z17 best practices and use the following tips to get the most efficiency, best performance and best throughput of the IBM z17:

- ▶ Collect the CPU MF counters (z/OS SMF 113s) on all LPARs. (CPU MF runs on every IBM z17 supported Operating Systems OS), see: <http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TC000066>
  - For z/OS, capture correlator SMF 98 subtype 1.
    - Provides supervisor information in five seconds intervals and helps perform performance problem analysis. Check the following White paper: <https://www.ibm.com/support/pages/node/6437547>
- ▶ z/VM CPU MF is collected in Monitor records. (there is no SMF). For more information, refer to: <https://www.vm.ibm.com/perf/tips/cpumf.html>



- ▶ Set your weights and Logicals for all your partitions to match your business needs.
- ▶ Set the efficient number of logicals to support the engines by weight
  - Assign one or two more logicals than engines by weight.
    - For instance, a 20 GCP LPAR processors with a 50% weight, set your logicals to one or two 2 more than 10. Assign the weights and logicals also applies to specialty engines.
- ▶ Assign a suitable number of logical CPs. If you assign too many logical CPs to the LPAR, PR/SM may place them at further distances, reducing efficiency and increasing unnecessary LPAR management. This issue reduces the efficiency of the cache. For more information about the number of Logical CPs defined for an LPAR, please refer to this document: [Number of Logical CPs Defined for an LPAR](#).
- ▶ The server capacity declines relative to the LCP:RCP ratio (sum of logical CPs that is defined in all LPARs: the number of physical CPs on your configuration). Therefore, assigning the correct number of logical CPs to an LPAR is important.
- ▶ Utilize IBM zPCR to size IBM Processors. (Don't use MIPS tables to do capacity sizing).
- ▶ Design your LPARs to "fit" in a single drawer with room to grow.
  - When the number of logicals exceeds the drawer boundary all physics come into play, slowing down, and that CPU time is clocked to your applications and to your bill.
- ▶ For the larger partitions, start the strategy to split them into smaller partitions.
  - Use IBM zPCR to show potential capacity savings for "more smaller" LPARs.
- ▶ Utilize HiperDispatch in every z/OS and z/VM LPAR. HiperDispatch optimizes processor cache usage by creating an affinity between a PU and the workload.

### 12.9.5 IBM zPCR HiperDispatch Report

IBM zPCR HiperDispatch report supports the ability to show how many Vertical High, Medium, and Low CPs will be assigned to each partition, based on current weights and LCPs. Figure 12-6 on page 504 shows how to select the HiperDispatch Assignment Report from the Partition Detail Report.

For additional information refer to: "HiperDispatch" on page 80.



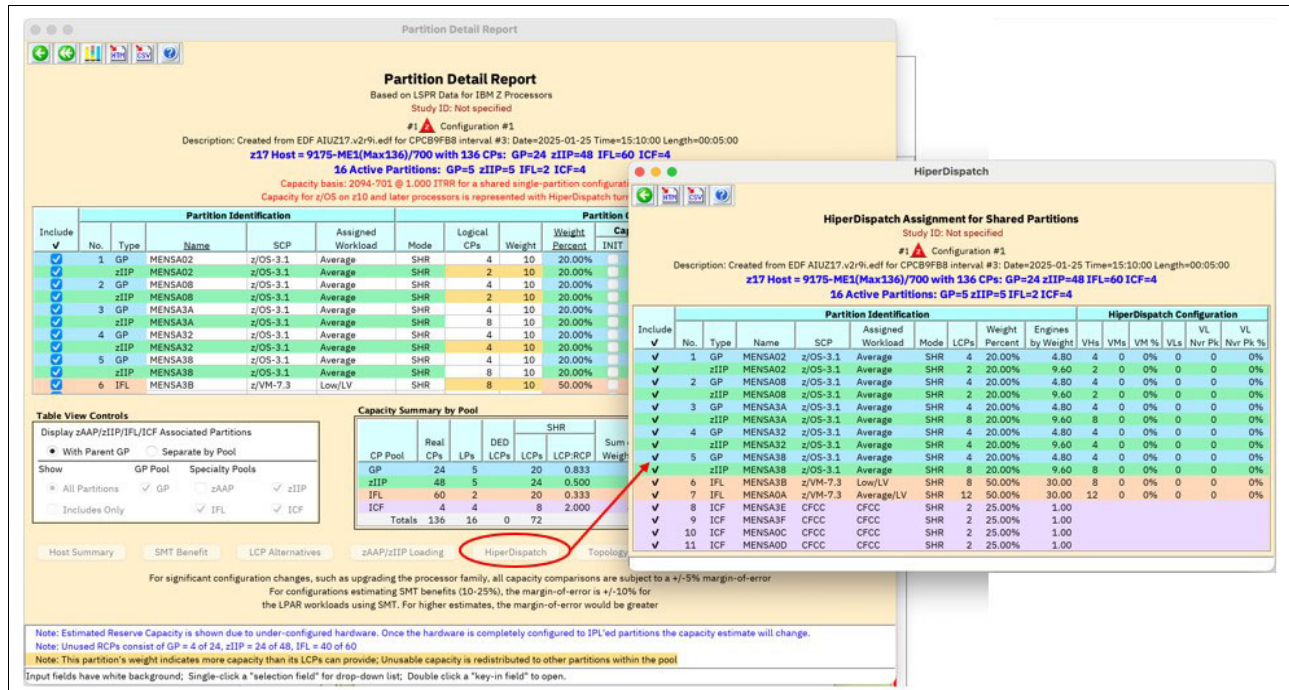


Figure 12-6 Selecting HiperDispatch Assignment for Shared Partitions from the Partition Detail Report

Both windows shown in Figure 12-6 will remain visible. Changes to the Partition Detail Report will reflect in the HiperDispatch window.

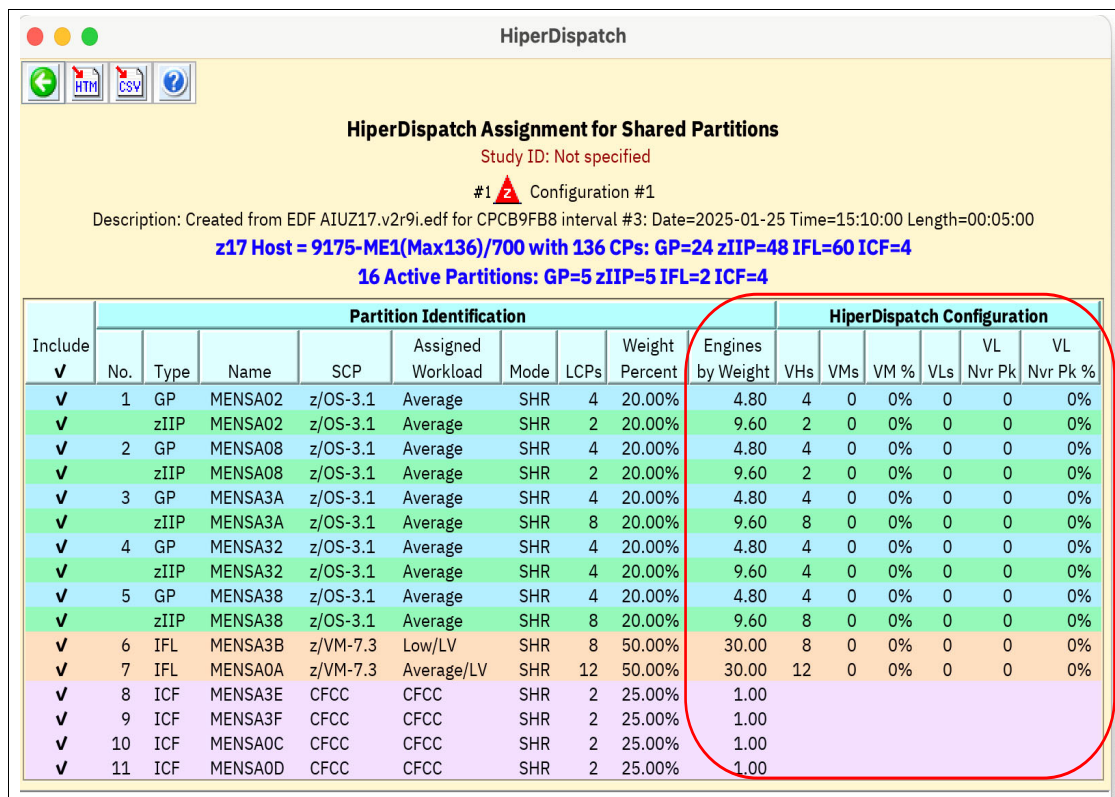


Figure 12-7 HiperDispatch Assignment for Shared Partitions Report Example



Figure 12-7 on page 504 shows all defined partitions and how HiperDispatch is expected to manage their logical CPs.

HiperDispatch supports logical CPs running z/OS v1.7 and later and z/VM v6.3 and later. For z/OS partitions, zIIPs and shared CPs are affected similarly. For z/VM partitions, IFLs and associated logical CPs are also affected similarly.

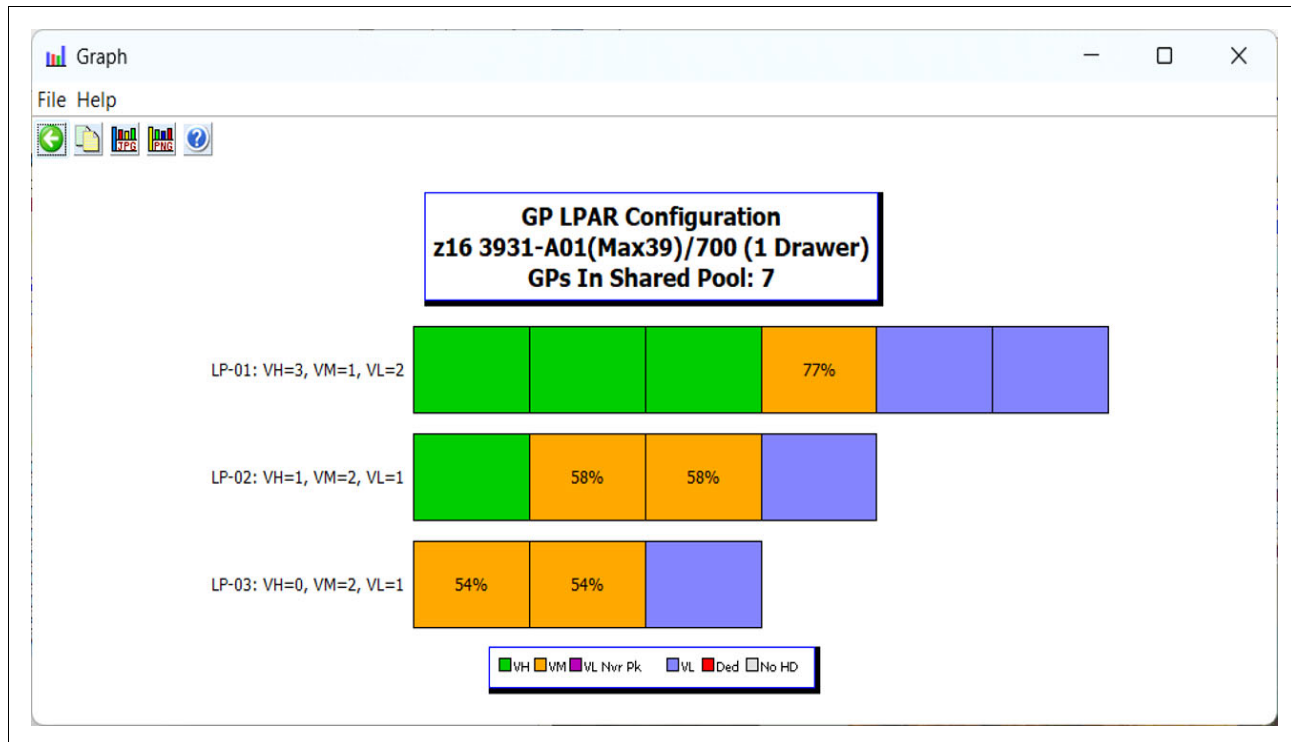


Figure 12-8 HiperDispatch Graph for the GP LCP HD Assignments Processor Topology Support

The HiperDispatch window shown in Figure 12-7 on page 504, contains most of the information from the Partition Detail Report, plus:

- ▶ Engines by Weight: Partition Weight% times the number of real CPs in the pool
- ▶ VHs: Number of LCPs categorized as Vertical High
- ▶ VMs: Number of LCPs categorized as Vertical Medium.
- ▶ VM%: Percent of time the partition's Vertical Medium LCPs are committed
- ▶ VLs: Number of LCPs categorized as Vertical Low
- ▶ VL Nvr Pk: Number of LCPs categorized as Vertical Low Never Parked
- ▶ VL Nvr Pk%: Percent of time the partition's Vertical Low Never Parked LCPs are committed.

When input fields are modified on the Partition Detail Report window, results on the HiperDispatch window will also be updated. Note that when exiting the HiperDispatch window, any changes made to the Partition Detail Report window are not automatically reset.

**Note:** For GP or IFL partitions where HiperDispatch is not supported, only the VMs and VM% columns apply. For ICF partitions, none of the HiperDispatch columns apply.



## 12.9.6 IBM zPCR Topology Report

Starting with the IBM z16, a new Topology Report window is available. In order to view it, the IBM z17 configuration must have been generated via EDF for a configuration where CPU MF was enabled. The Topology window portrays how the partition's logical CPs and their classification are distributed:

- ▶ Across the installed drawers (maximum of 4)
- ▶ Across the 4 Dual Chip Modules (DCMs) on each drawer
- ▶ Across the 2 chips on each DCM

See Figure 12-10 on page 507.

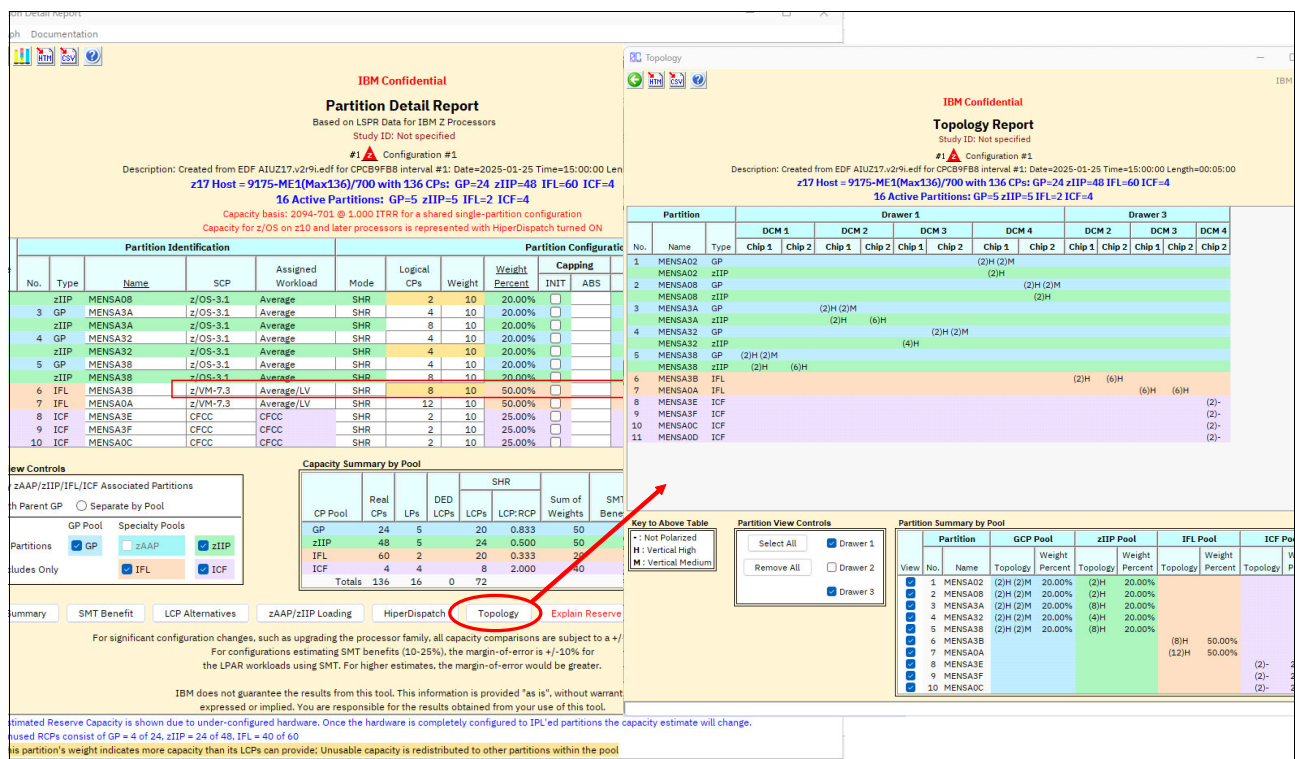


Figure 12-9 Selecting Topology Report from the Partition Detail

The IBM zPCR topology report is shown in Figure 12-10 on page 507, where:

- ▶ LPARs are identified by row
- ▶ IBM z17 Drawer/DCM/CHIP appears at the top lines
- ▶ Topology report displays warning messages
- ▶ LPARs Totals by Pool table is displayed at the bottom **with support to filter by partition**
- ▶ Report is accessed from the Partition Detail Window
- ▶ Latest versions of extract are required:
  - available here: [IBM Support page](#).

**Note:** The Topology report in Figure 12-10 on page 507 is showing all active partitions. Information about a specific partition can be obtained by clicking on the “Remove all” button to the left of the Partition Totals by Pool table at the bottom right and then clicking on the “View” check-box for a specific partition.



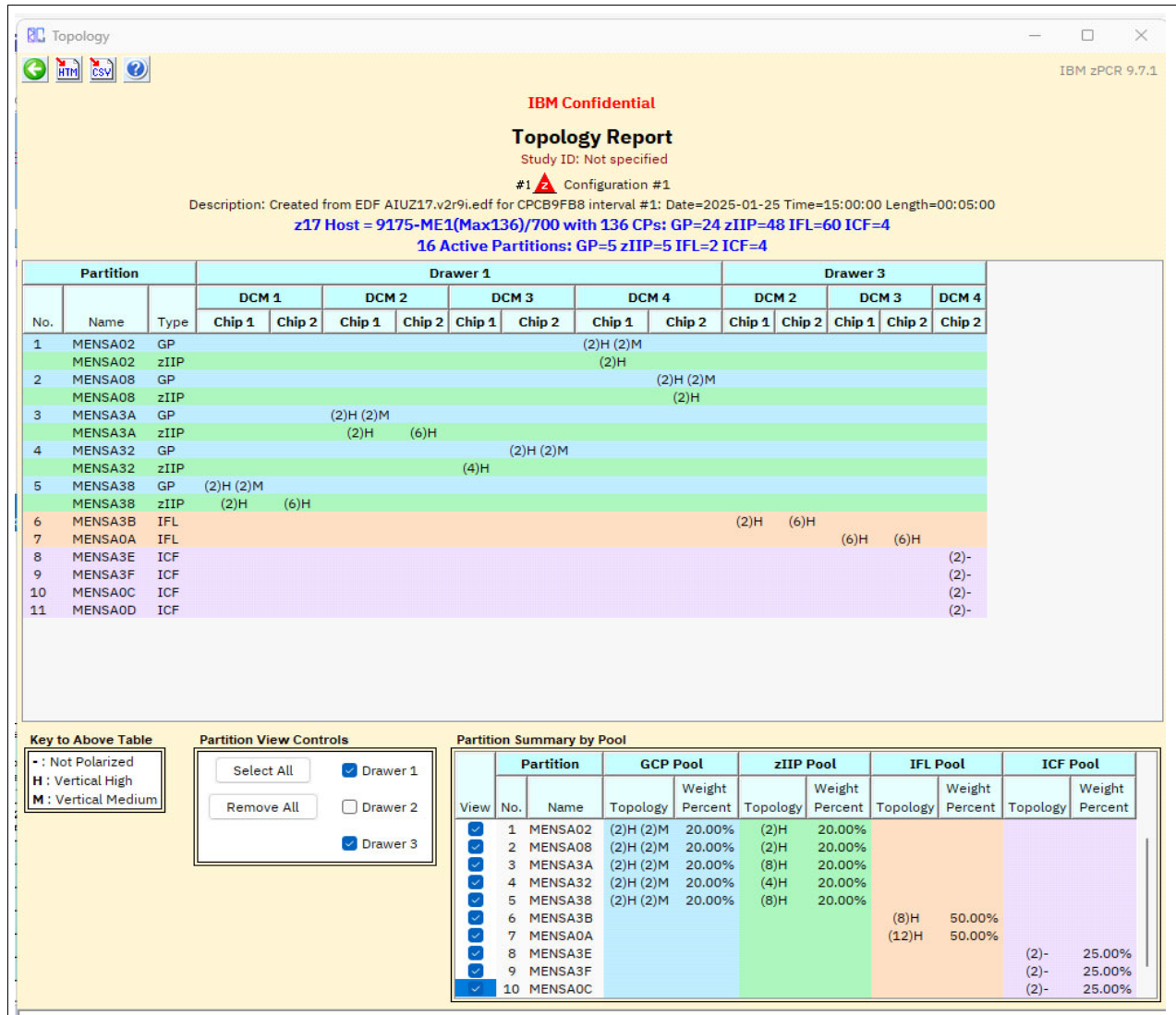


Figure 12-10 IBM zPCR Topology report example

IBM zPCR Topology Report is based in the z/OS new Data Gatherer functionality delivered with APAR A62064, PTF available for z/OS 2.5 and z/OS 2.4. z/OS data is in the SMF 70.1 record.

z/VM support is provided in the 7.3 version and later, and APARs are available for z/VM 7.1 and 7.2.

Additionally, consider collecting the z/OS SMF 99 Subtype 14 records for all LPARs in the IBM z17. Record has a single LPAR scope, so need all LPARs to get the total picture.

z/OS SMF 99 Subtype 14 contains HiperDispatch Topology data including:

- ▶ Logical Processor characteristics: Polarity (VH, VM, VL) and Affinity Node
- ▶ Physical topology information: Drawer/ DCM/ Chip location data for each logical CP.
- ▶ z/OS SMF 99 Subtype 14 is written every 5 minutes or when a Topology change occurs
  - Topology changes:
    - Configuration Change or weight change
    - Driven by IRD weight management
    - Record provides a Topology Change indicator to show when the topology changed



- LPAR topology may have a very significant impact on processor CPU efficiency. Remote cache accesses may take hundreds of machine cycles. SMF 99.14 records are produced every 5 minutes and capture drawer/DCM/chip location data for each logical CP.

In the Topology Reports, IBM zPCR reports “*measured*” data and it shows the “*what-is*” and not the “*what-if*” topology scenarios.

## 12.9.7 IBM z17 HMC - View Partition Resource Assignments

**Note:** To access HMC *Partition Resource Assignments*, you must use the “Service” logon id or any other user id with the same authority.

At the HMC “Home” screen click Systems Management, under the task bar and then select your listed target system.

Under “Tasks” at the bottom right click the Configuration (+) sign. Next, click “View Partition Resource Assignments”. The panel shown in Figure 12-11 will display:

	HD MENSA02	MENSA03	MENSA04	MENSA08	MENSA0A	MENSA0C	MENSA0D	HD MENSA32	HD MENSA38	HD MENSA3A	HD MENSA3B	MENSA3C	MENSA3D	MENSA3E
DCM 00														
DCM 01														
DCM 02														
DCM 07														
DCM 08														
DCM 09														
DCM 10														
DCM 11														
DCM 00: Chip 0		6 G D												
DCM 00: Chip 1		6 G D												
DCM 01: Chip 0	2 G M	4 G D						2 G M	2 G M	2 G M				
DCM 01: Chip 1	2 G L			4 G D		2 G SH		2 G L	2 G L	2 G L				
DCM 02: Chip 0						2 G SH						2 G SH	2 G SH	2 G SH
DCM 07: Chip 0											2 I H			
DCM 07: Chip 1											6 I H			
DCM 08: Chip 1		6 I D												
DCM 09: Chip 0		4 I D												
DCM 09: Chip 1		6 I D												
DCM 10: Chip 0					4 I D									
DCM 10: Chip 1					6 I D									

Figure 12-11 HMC - View Partition Resource Assignments for the IBM z17<sup>7</sup>

Use the Partition Resource Assignments to view processor allocations to partitions in your system. The active logical partitions are identified at the top of the table. The Node and Chip numbers associated with each active logical partition are identified on the left. You can view the Node and Chips assignments using the **Expand A11** and **Collapse A11** options under the “Actions” pull down view or hide sections.

To view the resource assignments for partitions:

- Logical partition name:

<sup>7</sup> The image has been reduced from its original size horizontally, due to the number of active partitions, and vertically due to the number of installed drawers.

<sup>9</sup> NUMA, Non-Uniform Memory Access is a multiprocessor design where memory access time depends on the memory location relative to the processor.



- Displays the active logical partition and if HiperDispatch (image icon of the logical partition name entry) is enabled.
- ▶ Node:
  - Displays the processor Node number in your system.
- ▶ Chip:
  - Displays the processor Chip number associated with the Node and lists the processor types associated with each active logical partitions. The Chip Collapse All icon displays a summary view. The following physical processor types are:
    - General processors: (G)
    - Coupling facility processors: (C)
    - Integrated Facilities for Linux IFLs: (I)
    - z Integrated Information Processors (zIIPs): (Z)
    - Integrated Firmware Processor (IFPs): (F)

The physical processor types may have some of the following conditions:

- ▶ Indicates the physical processor types are shared: (SH)
- ▶ Indicates the physical processor is dedicated: (D)
- ▶ Indicates the vertical polarity for the physical processor types (H / M / L)

## 12.9.8 IBM zPCR Large Partition Support

IBM zPCR version 9.7.1<sup>8</sup> implemented an important change to the LPAR Configuration Capacity Planning function. For the IBM z17, 9175 (Max208/700) a change was made to match the way RCPs per drawer are assigned when the total number RCPs in the configuration is up to 182. This change compensates for the way the RCP count per drawer actually occurs.

This change was needed since it turns out that not all IBM z17 Max208's have 53 configurable PU drawers. Some Max208's have 47 PU drawers when there are 182 or less PU's in the configuration.

IBM zPCR shows the distribution of the IBM z17 PUs per drawer according to the Figure 12-16 on page 513, and there are very useful graphs called ***Estimated Distribution of RCPs Across Drawers*** that are available as an option from the Partition Detail Report, which identifies the number of PUs on each drawer.

See Figure 12-14 on page 512 and Figure 12-15 on page 512 for details about the new available graphs.

Figure Figure 12-12 on page 510 shows the PUs distribution per drawer for an IBM z16 Max200 according to the number of available configurable PUs.

As shown in Figure 12-13 on page 511, the subject IBM z16 has 165 CPs, the PU resources are distributed according to the Figure 12-12 on page 510, and are also shown in the Estimated Distribution of RCPs Across Drawers graph in Figure 12-14 on page 512.

<sup>8</sup> Your currently installed version of IBM zPCR must be uninstalled before installing IBM zPCR 9.7.1. This step is necessary to facilitate conversion to the latest IBM Java 17 Semeru 64-bit runtime environment that is included with IBM zPCR.



IBM z16		1 <sup>st</sup> Drawer				2 <sup>nd</sup> Drawer				3 <sup>rd</sup> Drawer				4 <sup>th</sup> Drawer			
Feature	Cust PUs	Cust PUs	SAP	IFP	Spare	Cust PUs	SAP	IFP	Spare	Cust PUs	SAP	IFP	Spare	Cust PUs	SAP	IFP	Spare
Max200	200	47	6	2	2	51	6	0	0	51	6	0	0	51	6	0	0
Max168	168	39	5	2	2	43	5	0	0	43	5	0	0	43	5	0	0
Max125	125	39	5	2	2	43	5	0	0	43	5	0	0				
Max82	82	39	5	2	2	43	5	0	0								
Max39	39	39	5	2	2												

Figure 12-12 IBM z16 configurable PU distribution per Model (feature) and CPC drawer

**Note:** Screen Captures from Figures 12-12 to Figure 12-15 were taken from an IBM z16 server zPCR study. All the examples shown are valid and compatible with IBM z17.

### Partition Detail Report warnings

IBM zPCR implemented a warning indicating that on IBM z13 and later, the best performance is achieved when the partition's logical CP count does not exceed the number of RCPs in a single drawer.

Additionally, IBM zPCR version 9.6.4 and newer implements a new notice for partitions approaching, within 10%, the maximum drawer size. This critical notice indicates that one or more partitions partition are getting close to a drawer boundary. When that happens, capacity growth by adding LCPs is very limited.

The new notice appears as a "Note" message in the Partition Detail Report. The "Note" and the partition LCPs are shaded with the same **violet** color, as shown for partition IFL-01 in Figure 12-12 above and for partition GP-02 in Figure 12-13 on page 511.

Figure 12-13 on page 511 shows an example of a Partition Detail Report this time for an **IBM z16, 3931-A01**, (Max200)/700 with 165 CPS (45 GPs, 44 zIIPs, 60 IFLs and 16 ICFs), and six active partitions (2 GP, 2 zIIP, 1 IFL and 1 ICF). Resources are allocated as shown in the Partition Identification and Partition Configuration fields.

As the IBM z16 has 165 CPs, the PU resources are distributed according to the Figure 12-12, and are shown in the Estimated Distribution of RCPs Across Drawers graph in Figure 12-14 on page 512.

**Important:** Please pay special attention to the colors assigned to the Logical CPs column and relate them to the "warnings" and "notes" at the bottom of the report. See Figure 12-13 on page 511.



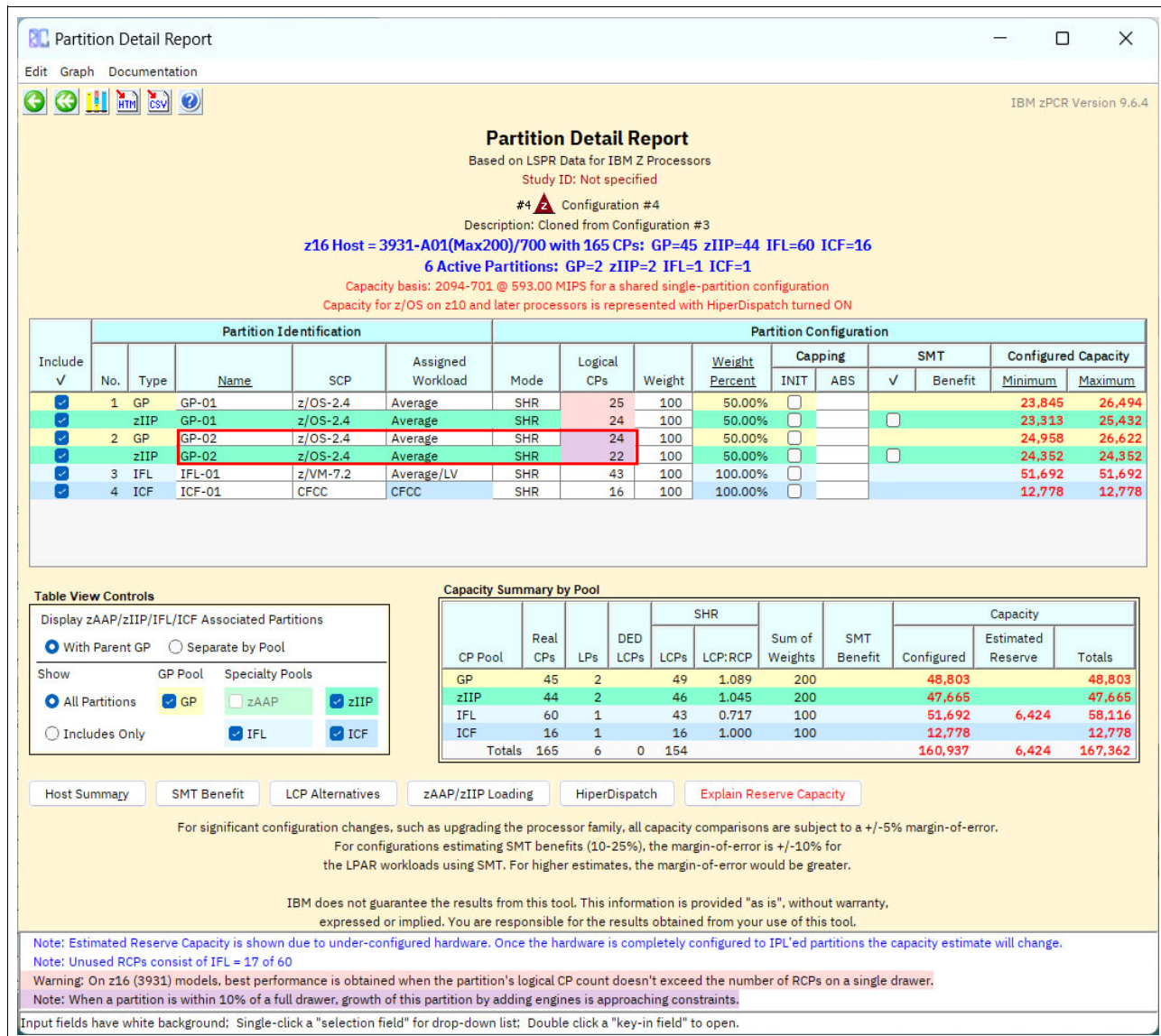


Figure 12-13 Partition Detail Report example (Max200)/700 with 165 CPs

The Partition Detail Report in Figure 12-13 above highlights the partition GP-02 to indicate it is within 10% of the maximum drawer size in the number of CPs. The GP-02 partition and the "Note" at the bottom are shaded with the same violet color.

Figure 12- Figure 12-15 on page 512 expands the information about the highlighted messages shown at the bottom of the Figure 12-13.

On IBM z16 and IBM z17 models, the best practice is a partitions's logical CP count should not exceed the number of RCP (Real CPs) in the largest drawer. Partitions which exceed a drawer boundary have special capacity considerations. For IBM z16 see Figure 12-12 on page 510, and for IBM z17 see Figure 12-16 on page 513.



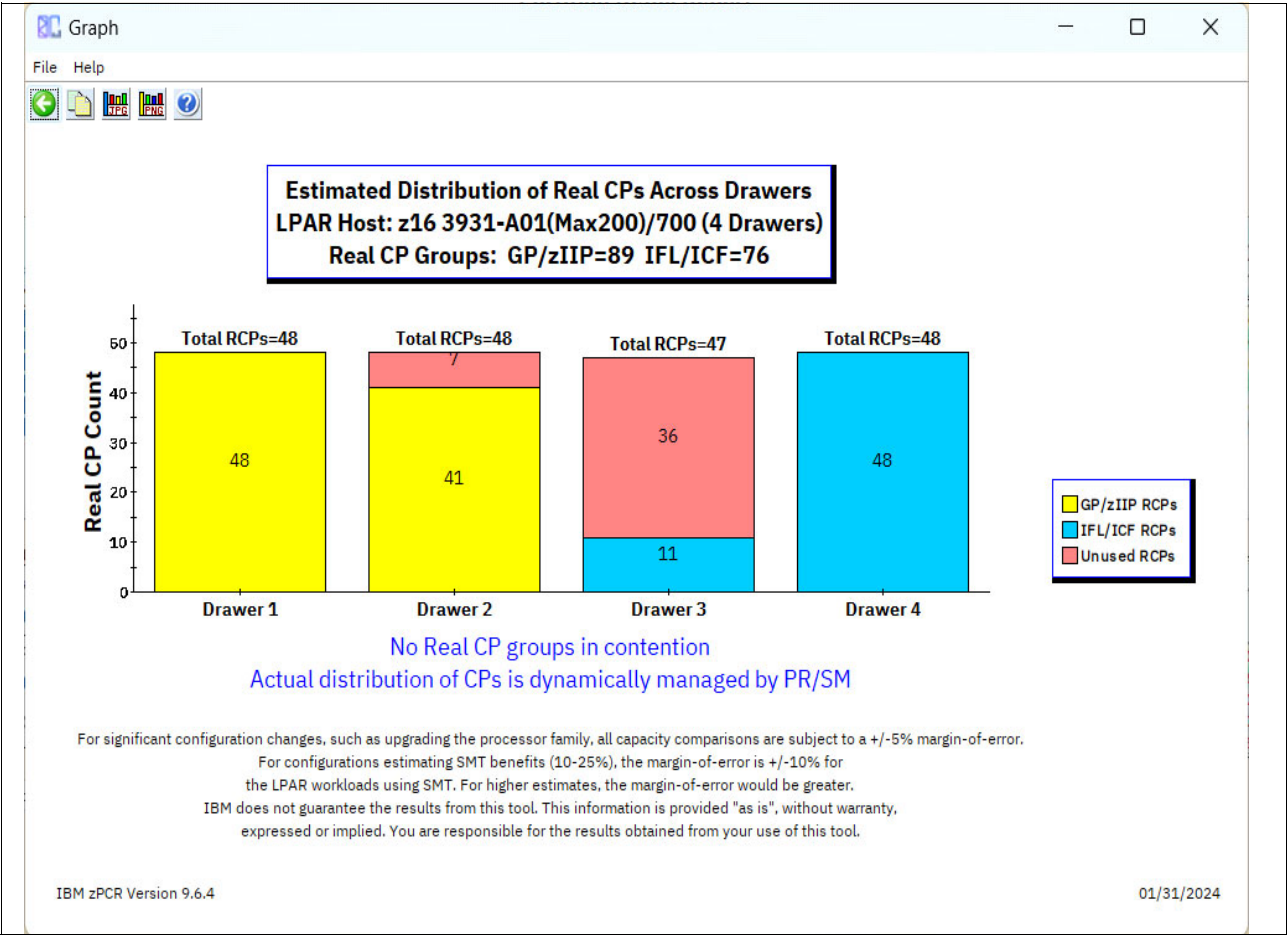


Figure 12-14 Estimated distribution of real CPs across drawers report

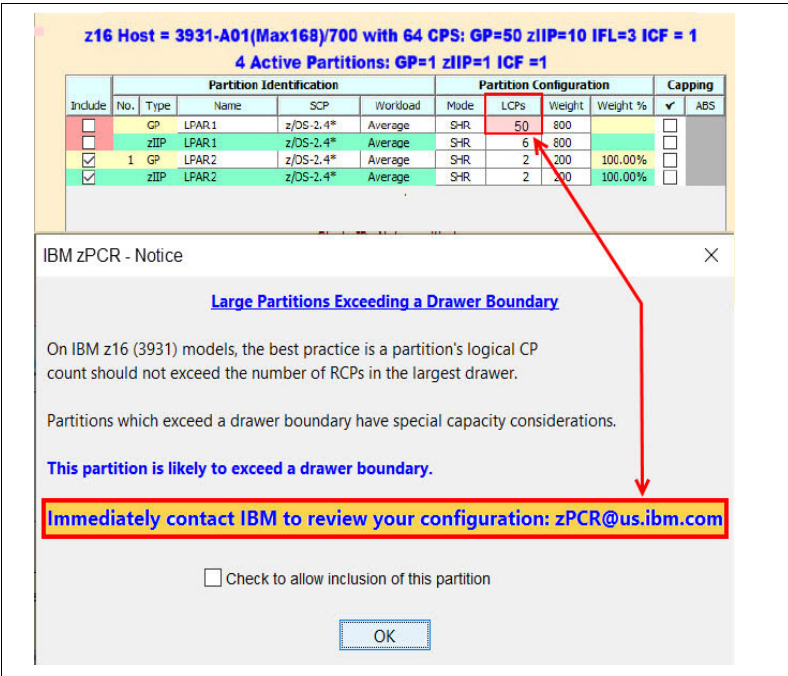


Figure 12-15 Large Partition Exceeding a Drawer Boundary warning Message



IBM z17 continues the IBM z15 and IBM z16 NUMA<sup>9</sup> design. IBM z17 has two clusters and four DCMs per drawer (refer to Figure 12-16) for the number of Configurable PUs per drawer).

IBM z17		1 <sup>st</sup> Drawer				2 <sup>nd</sup> Drawer				3 <sup>rd</sup> Drawer				4 <sup>th</sup> Drawer			
Feature	Cust PUs	Cust PUs	SAP	IFP	Spare	Cust PUs	SAP	IFP	Spare	Cust PUs	SAP	IFP	Spare	Cust PUs	SAP	IFP	Spare
Max208	208	49	6	2	2	53	6	0	0	53	6	0	0	53	6	0	0
Max183	183	42	6	2	2	47	5	0	0	47	5	0	0	47	5	0	0
Max136	136	46	6	0	0	43	5	2	2	47	5	0	0				
Max90	90	43	5	2	2	47	5	0	0								
Max43	43	43	5	2	2												

Figure 12-16 IBM z17 configurable PU distribution per Model (feature) and CPC drawer

In the case where a single partition spans from one drawer into a second, the cross-drawer penalty has increased on IBM z17. However, this is offset by more cores per drawer and higher capacity than IBM z15, which allows more work to “fit” on a single drawer.

As discussed in 3.5.9, “Processor unit assignment” on page 115, and under “Memory allocation” on page 117, PR/SM memory and logical processor allocation goal is to place all logical partition resources on a single CPC drawer, if possible. There can be negative impacts on a logical partition’s performance when CPs are allocated in different drawers.

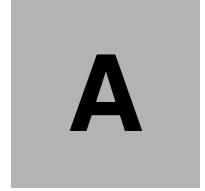
IBM zPCR implements a warning notice when logical partition’s number of Logical CPs defined is larger than the number of Real CPs in a single drawer. When that scenario occurs, it is advisable to contact IBM to review your configuration. See Figure 12-15 on page 512.

<sup>9</sup> NUMA, Non Uniform Memory Access, is a computer memory design used in multiprocessing where the access time varies according to the memory location relative to the processor.









# **IBM Z Integrated Accelerator for AI and IBM Spyre AI Accelerator**

This appendix provides an overview of the next generation of on-chip AI acceleration implemented in the IBM z17 processor (Telum II), and the new PCIe attached IBM Spyre AI Accelerator.



## A.1 Overview

Each generation of the IBM Z processing chip is enhanced with new on-chip functions, such as compression, sort, cryptography, and vector processing. The purpose-built accelerators that provide these functions mean lower latency and higher throughput for specialized operations. These accelerators work together with advanced chip design features such as data prefetch, high capacity L1 and L2 caches, branch prediction, and other innovations.

The on-chip accelerators provide support for and enable compliance with security policies because the data is not leaving the platform to be processed. The hardware, firmware, and software are vertically integrated to seamlessly deliver this function to the applications.

In August 2021, [IBM announced a new generation](#) of IBM Z processor, Telum, with a new Artificial Intelligence (AI) accelerator (see Figure A-1). This innovation brings incredible value to the applications and workloads that are running on IBM Z.

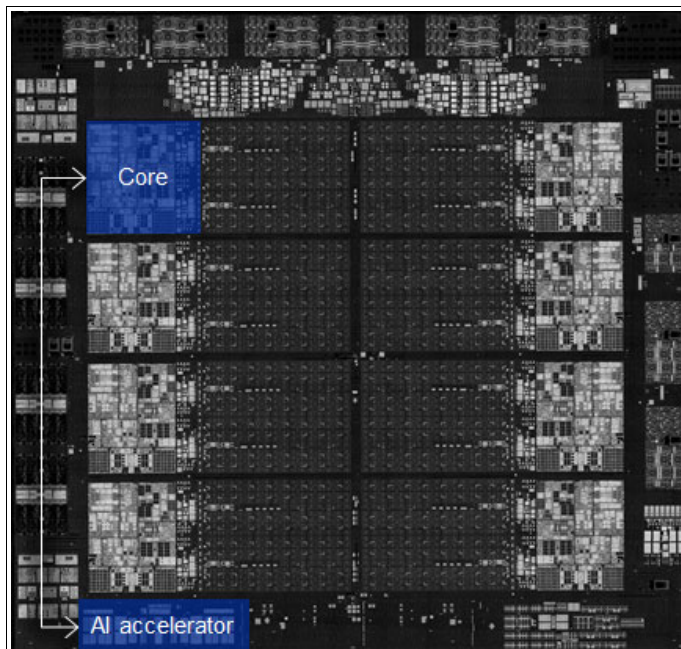


Figure A-1 IBM z16 Processor chip: AIU location

With the new IBM Z Integrated Accelerator for AI, customers can benefit from the acceleration of AI operations, such as fraud detection, customer behavior predictions, and streamlining of supply chain operations, all in real time. Customers can derive the valuable insights from their transactional data instantly.

The Integrated Accelerator for AI delivers AI inference in real time, at large scale and rate, with no transaction left behind, all without needing to offload data from IBM Z to perform AI inference.

The AI capability is applied directly into the running transaction, shifting the traditional paradigm of applying AI to the transactions that were completed. This innovative technology can be used for intelligent IT workloads placement algorithms, which contributes to the better overall system performance. The accelerator is driven by new Neural Networks Processing Assist (NNPA) instructions.



## A.2 NNPA and IBM z16 Hardware

NNPA instructions are a new set of nonprivileged Complex Instruction Set Computer (CISC) memory-to-memory instructions that operates on tensor objects that are in user programs' memory. In contrast, other accelerators are implemented through some form of memory mapped IO, which requires memory management across a hypervisor and operating system stack, often adding significant latency. AI functions and macros are abstracted by NNPA instructions which perform common mathematical operations associated with processing an AI prediction such as matrix multiplication, activation, and convolution, among many others.

Figure A-2 shows the IBM z16 AI accelerator and its components: the data movers that surround the compute arrays are composed of the Processor Tiles (PT), Processing Elements (PE), and Special Function Processors (SFP).

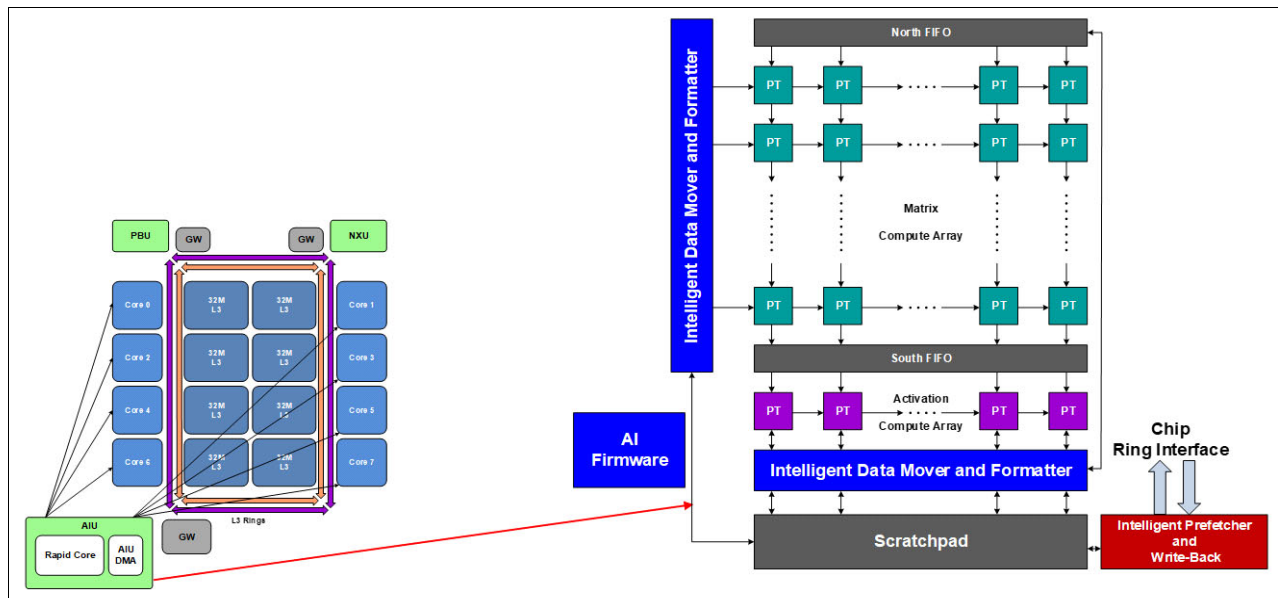


Figure A-2 IBM z16 AIU logical diagram

Intelligent data movers and prefetchers are connected to the chip by way of a ring interface for high-speed, low-latency, read/write cache operations (200+ GBps read/store bandwidth, and 600+GBps bandwidth between engines).

Compute Arrays consist of 128 processor tiles with 8-way FP-16 FMA SIMD, which are optimized for matrix multiplication and convolution, and 32 processor tiles with 8-way FP-16/FP-32 SIMD, which is optimized for activation functions and complex functions.

On the IBM z16, the integrated AI accelerator delivered more than 6 TOPS (trillion operations per second) per chip and over 192 TOPS in the 32 chip system (a fully configured IBM z16 with four CPC drawers).

The AI accelerator on the IBM z16 is shared by all cores on the chip. The firmware, which is running on the cores and accelerator, orchestrates and synchronizes the execution on the accelerator.



## A.3 How to use IBM Z Integrated AI Accelerator in your enterprise

This chart that is shown in Figure A-3 shows the high level of seamless integration of AI accelerator into enterprise AI/ML solution stack. Great flexibility and interoperability are realized for training and building models.

Acknowledging the diverse AI training frameworks, customers can train their models on platforms of their choice, including IBM Z (on-premises and in hybrid cloud) and then, deploy it efficiently on IBM Z in colocation with the transactional workloads. No other development effort is needed to enable this strategy.

To allow this flexible “Train anywhere, Deploy on IBM Z” approach, IBM invests in the [Open Neural Network Exchange](#) (ONNX) technology (see Figure A-3).

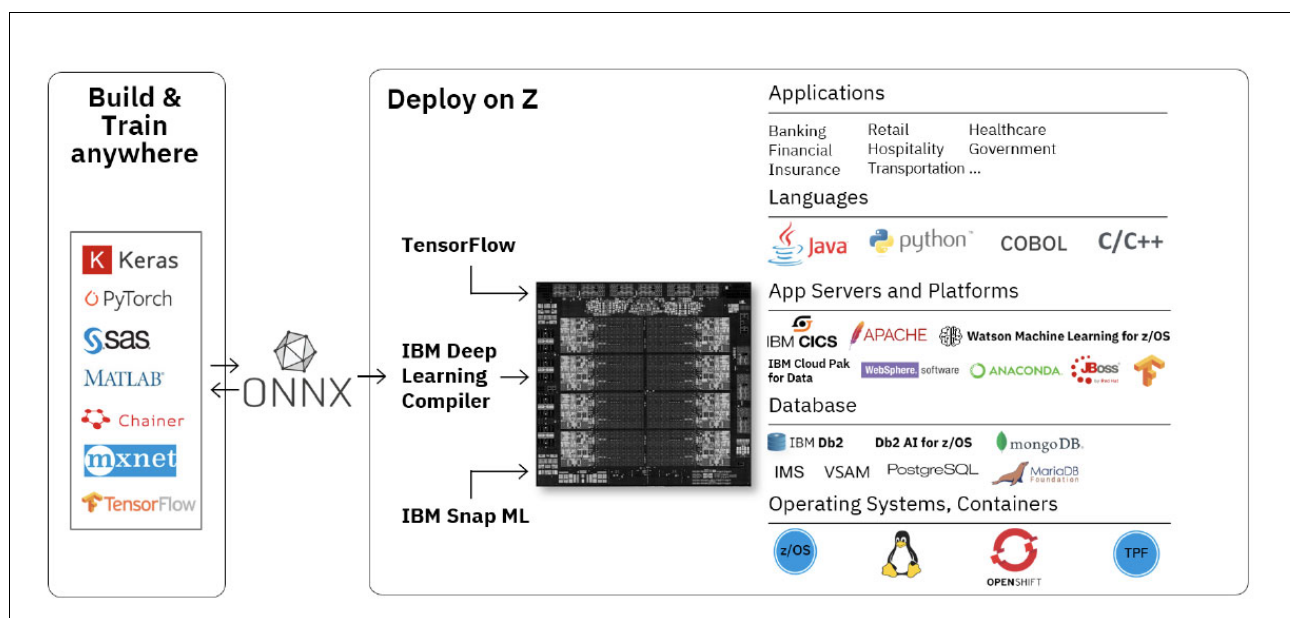


Figure A-3 ONNX ecosystem

This standard format represents AI models, with which a data scientist can build and train a model in the framework of choice without worrying about the downstream inference implications. To enable deployment of ONNX models, IBM provides an ONNX model compiler that is optimized for IBM Z. In addition, IBM is optimizing key open source AI and model serving frameworks, such as PyTorch, TensorFlow, TensorFlow Serving, and Nvidia Triton Inference Server for use on IBM Z.

IBM's open source [zDNN library](#) provides common APIs for the functions that enable conversions from the Tensor format to the accelerator required format. Customers can run zDNN under z/OS<sup>1</sup> and Linux on IBM Z. A Deep Learning Compiler (DLC) for z/OS and Linux also is available to compile ONNX deep learning models into shared libraries which can be integrated into C, C++, Java, or Python applications. In addition to leveraging the Integrated Accelerator for AI, the compiled models take advantage of other hardware acceleration capabilities such as SIMD (Single Instruction Multiple Data).

<sup>1</sup> zDNN is zCX eligible, which it runs on zIIPs under z/OS



## A.4 Next Generation Artificial Intelligence Unit - AIU

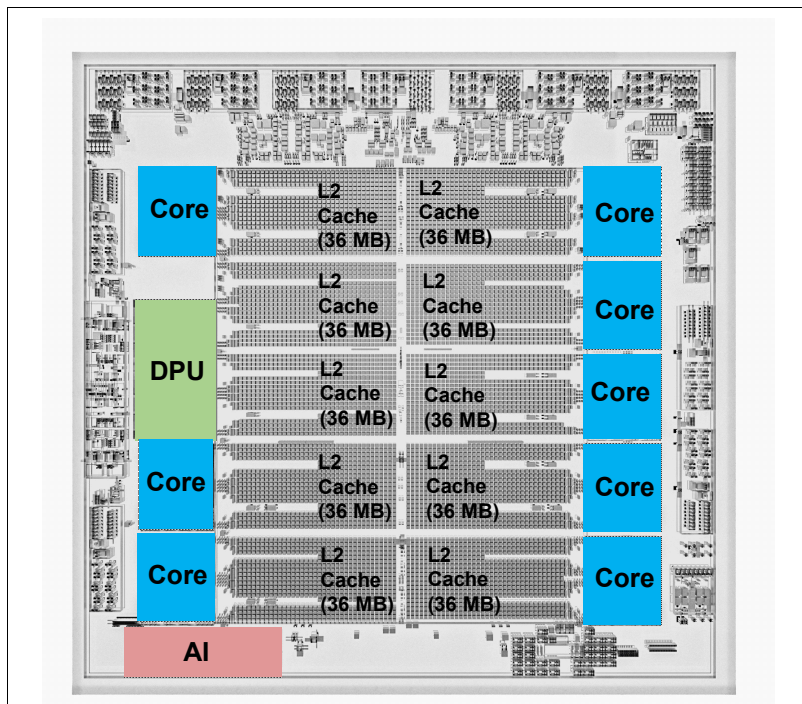
The IBM z17 introduced the new Telum II processor which includes the next generation on-chip AI Accelerator.

Developed using Samsung 5nm technology, the new IBM Telum II processor features eight high-performance cores running at 5.5 GHz. Telum II will include a 40% increase in on-chip cache capacity, with the virtual L3 and virtual L4 growing to 360MB and 2.88GB respectively. The processor integrates a new data processing unit (DPU) specialized for IO acceleration and the next generation of on-chip AI acceleration. These hardware enhancements are designed to provide significant performance improvements for clients over previous generations.

Infusing AI into enterprise transactions has become essential for many clients' workloads. For instance, AI-driven fraud detection solutions save clients millions of dollars annually. With the introduction of the AI accelerator on the Telum processor, there has been active AI adoption on the z16 platform with the implementation of unique use cases by clients belonging to a wide variety of industries. Building on this success and to meet growing performance demands, IBM has significantly enhanced the Integrated AI accelerator on the Telum II processor.

The compute power of each accelerator is improved by 4x, reaching 24 trillion operations per second (TOPS). But TOPS alone don't tell the whole story, additional enhancements have been made to the accelerator's architectural design plus optimizations to the AI ecosystem that sits on top of the accelerator. Additionally, support for INT8 as a data type has been added to enhance compute capacity and efficiency for applications where INT8 is preferred, thereby enabling the use of newer models.

New NNPA instructions have also been incorporated to better support large language models (LLMs) within the accelerator. They are designed to support an increasingly broader range of AI models for a comprehensive analysis of both structured and textual data.





On Telum, the cores on each processor chip only had access to their local AI accelerator, and in the event where two cores on the same chip issued an NNPA instruction at the same time, access to the AI accelerator was time-sliced, resulting in a wait-time to get access to the AI accelerator.

Telum II was designed so that a processor core can offload AI operations to any of the other integrated AI accelerators in the 7 adjacent processor chips in the drawer. This architectural design provides each core access to a much larger pool of AI compute resources, and reduces the contention for an Integrated Accelerator for AI. This represents a significant enhancement to the on-processor AI capabilities.

When it comes to AI acceleration in production enterprise workloads, a fit-for-purpose architecture matters. Telum II is engineered to enable model runtimes to sit side by side with the most demanding enterprise workloads, while delivering high throughput, low-latency inferencing.

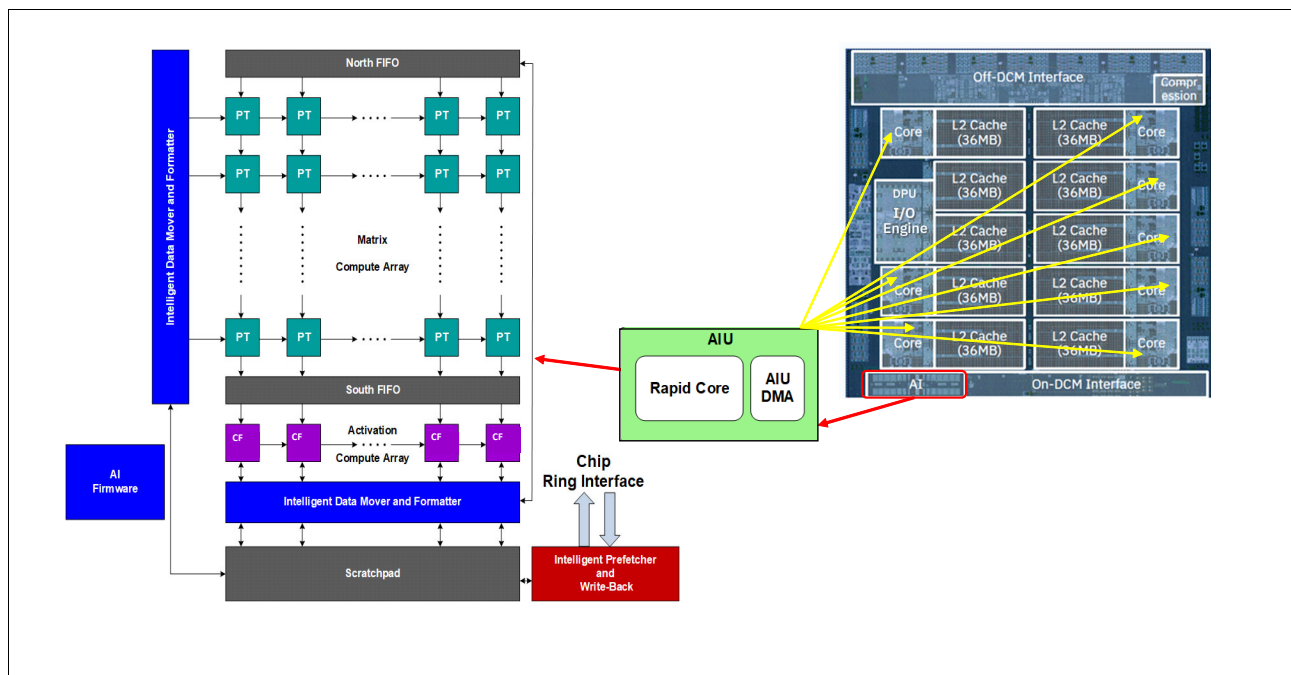


Figure A-4 IBM z17 AIU logical diagram

On the IBM z17, the integrated AI accelerator delivered more than 24 TOPS (trillion operations per second) per chip and over 192 TOPS per drawer and 768 TOPS in the 32 chip system (a fully configured IBM z17 with four CPC drawers).

#### A.4.1 Spyre Accelerator

The IBM Spyre Accelerator is delivered via a 75W gen 5x PCIe-attached AI accelerator with 128 GB of LPDDR5 memory. The IBM Spyre Accelerator is geared to handle larger, more complex AI use cases which leverage large foundation models for code generation, AI assistants, document processing, and classification. It is worth noting that Spyre is purpose-built to scale foundation model inferencing, not training. Additionally, the card enables scalable, sustainable, and secure AI through the accelerator compute capacity and



architecture, low power consumption, and execution in a Confidential Computing environment.

The Spyre card has 32 accelerator cores, which share a similar architecture to the Telum II AI accelerator. Multiple Spyre Accelerator cards can be clustered together, for example a cluster of eight cards enables workloads to transparently leverage 256 accelerator cores. Scalable by card and drawer, clients are able to take advantage of this efficient solution for next-generation AI workloads.

The IBM Spyre accelerator assembly is a PCI adapter plugged in the I/O drawer. Up to 48 features are offered with a maximum of eight per I/O drawer. The Spyre accelerator is an Enterprise Grade AI chip which enables generative AI capabilities on the IBM z17.

The adapter has 1 TB of Memory with 1.6 TB per second aggregate memory bandwidth. Figure 12-17 shows the Spyre adapter and the IBM Spyre chip.

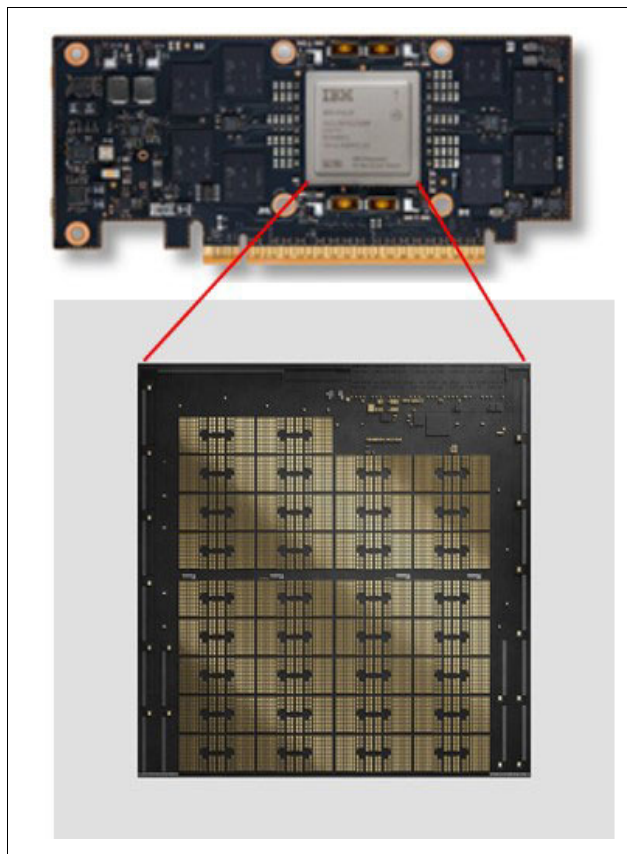


Figure 12-17 IBM Spyre Accelerator Adapter and Chip

Telum II and the Spyre Accelerator card together support a broader set of models, enabling composite AI use cases. Composite AI refers to leveraging the strengths of multiple AI models, traditional and foundational, to improve the overall accuracy of a prediction, as compared to just using one. Composite AI can be used in insurance claim fraud detection, with a traditional neural network model being used to analyze and make a prediction based on structured data, while a Large Language Model (LLM) could then be used to extract information from the unstructured textual data associated with a claim. This composite AI method can be applied for advanced detection of suspicious financial activities, support compliance with regulatory requirements, and mitigate the risk of financial crimes.



## A.4.2 Planning for the IBM Spyre Accelerator Cards

The firmware, management, operation, monitoring, and diagnostics for the card will be provided by a software appliance, which we will refer to as the Spyre Support Appliance (SSA). Firmware for the card will be decoupled from the regular IBM Z firmware cycle, meaning that firmware updates to the card will not be delivered as Microcode Change Level (MCL) bundles, but instead will be delivered through the SSA.

Two Spyre Support Appliances will be necessary to ensure that IBM Spyre Accelerator cards provide the expected IBM Z level of enterprise grade reliability, availability, and security. A pair of Spyre Support Appliances can manage multiple Spyre Accelerator cards on the CEC. An organization may consider provisioning more Spyre Support Appliances if there are isolation requirements for development, test, and production environments.

Spyre Support Appliances will need to run in dedicated Secure Service Container type LPARs. The IBM Secure Service Container is a container technology through which you can quickly and securely deploy firmware and software appliances on the server. Unlike most other types of partitions, a Secure Service Container partition contains its own embedded operating system, security mechanisms, and other features that are specifically designed for simplifying the installation of appliances, and for securely hosting them.

A Secure Service Container partition is a specialized container for installed and running specific software appliances. An appliance is an integration of operating system, middleware, and software components that work autonomously and provide core services and infrastructures that focus on consumability and security.

In addition to the Spyre Support Appliances, an additional appliance will be needed, the Appliance Control Center, a new component responsible for managing all appliances under an HMC, even those that span CECs. The Appliance Control Center will be responsible for appliance management tasks such as performing updates to the Spyre Support Appliance and pulling code dumps. The Appliance Control Center will need to run in a dedicated Secure Service Container type LPAR. Only one Appliance Control Center is needed as long as there is network connectivity to where the Spyre Support Appliances will be deployed.

The table below summarizes the LPAR hardware requirements for the Spyre Support Appliances and the Appliance Control Center. Note that these values could change, and are provided to assist in the planning process for Spyre Accelerator Cards.

LPAR	IFL Count	Memory	Disk Storage
SSA #1	2 Shared IFLs	50 GB	50 GB
SSA #2	2 Shared IFLs	50 GB	50 GB
ACC	2 Shared IFLs	16 GB	50 GB

In addition to assigning the necessary resources to the LPARs, the LPARs will also need to have network connections defined.

### I/O Configuration

One physical function and one virtual function will need to be assigned to each Spyre Support Appliance LPAR. In addition, a virtual function will need to be assigned to the z/OS or Linux LPAR which will host the workload which uses Spyre Accelerator Cards.

Spyre Accelerator Cards can be added to the system in sets of eight, from one to six sets. This means that a minimum of 8 Spyre Accelerator Cards can be installed on a z17 system. A maximum of 8 Spyre Accelerator Cards can be placed in a single I/O drawer. A z17 system



will support a maximum of 48 total Spyre Accelerator cards, which could be installed in as little as 6 I/O drawers.

IBM Spyre Accelerator Cards will need to be placed in certain slots due to power and cooling impacts, with the goal of spreading across the fewest I/O drawers. IBM will provide guidance regarding which slots to use for Spyre Accelerator Cards.

When ordering a z17, there is an option to select a feature code which will allow the reservation of I/O capacity for a variable number of Spyre Accelerator Card sets, from one to six sets. Note that this process is necessary for ensuring an adequate number of I/O drawers and enough capacity in an I/O drawer(s) to accommodate the cards.

To assist with planning, IBM will provide guidance regarding the number of Spyre Accelerator Cards needed to support certain use cases.

Clients will need to refrain from applying the hardware Miscellaneous Equipment Specifications (MES) that add any adapters to the I/O drawer(s), until the Spyre Accelerator Cards become Generally Available. This is to prevent interfering with reserved I/O slots for Spyre Accelerator cards, as any unrelated I/O MES will default placement in the first available I/O slot(s).







**B**

# IBM Integrated Accelerator for zEnterprise Data Compression

This appendix describes the IBM Integrated Accelerator for zEnterprise Data Compression (zEDC) that is implemented in IBM Z hardware.

The appendix includes the following topics:

- ▶ “Client value of IBM Z compression” on page 525
- ▶ “IBM z17 IBM Integrated Accelerator for zEDC” on page 526
- ▶ “IBM z17 migration considerations” on page 527
- ▶ “Software support” on page 528
- ▶ “Compression acceleration and Linux on IBM Z” on page 529
- ▶

## B.1 Client value of IBM Z compression

The amount of data that is captured, transferred, and stored continues to grow. Software-based compression algorithms can be costly in terms of processor resources, storage costs, and network bandwidth.

An optional PCIe feature that was available for IBM z14 servers, zEDC Express, addressed customer requirements by providing hardware-based acceleration for data compression and decompression. zEDC provided data compression with lower CPU consumption than compression technology that was available on the IBM Z server.

Customers deployed zEDC compression to deliver the following types of compression:

- ▶ Storage
- ▶ Data transfer
- ▶ Database
- ▶ In-application



Data compression delivers the following benefits:

- ▶ Disk space savings
- ▶ Improved elapse times
- ▶ Reduced CPU consumption
- ▶ Reduced network bandwidth requirements and transfer times

Although the zEDC PCIe feature provided CPU savings by offloading the select compression and decompression operations, it included the drawback of limited virtualization capabilities (one zEDC PCIe feature can be shared across a maximum of 15 LPARs) and limited bandwidth.

IBM z15 introduced an on-chip accelerator (implemented in the PU chip) for compression and decompression operations, which was tied directly into processor's L3 cache. As a result, it provided much higher bandwidth and removed the virtualization limitations of a PCIe feature.

The IBM z17 further addresses the growth of data compression requirements with the integrated on-chip compression unit (implemented in processor Nest, one per PU chip) that significantly increases compression throughput and speed compared to previous zEDC deployments.

## B.2 IBM z17 IBM Integrated Accelerator for zEDC

The IBM z17 Integrated Accelerator for zEDC delivers industry-leading throughput and provides value for existing and new compression users by bringing the compression facility onto the PU chip, which is tied into L3<sup>1</sup> cache.

IBM z17 compression and decompression are both implemented in the Nest Accelerator Unit (NXU, see Figure 3-12 on page 91) on each processor chip.

One NXU is available per processor chip, which is shared by all cores on the chip and features the following benefits:

- ▶ New concept of sharing and operating an accelerator function in the nest
- ▶ Supports DEFLATE-compliant compression and decompression and GZIP CRC/ZLIB Adler
- ▶ Low latency
- ▶ High bandwidth
- ▶ Problem state execution
- ▶ Hardware and firmware interlocks to ensure system responsiveness
- ▶ Designed instruction
- ▶ Run in millicode

The IBM z17 Integrated Accelerator for zEDC has an Improved Gasp hashing algorithm which increases the compression ratio.

---

<sup>1</sup> Virtual L3 (shared victim) cache for IBM z17. For more information, see Chapter 2, "Central processor complex hardware components" on page 19.



## B.2.1 Compression modes

Compression is run in one of the following modes:

- ▶ Synchronous

Execution occurs in problem state where the user application starts the instruction in its virtual address space.

- ▶ Asynchronous

Execution is optimized for Large Operations under z/OS for authorized applications (for example, BSAM) and issues I/O by using EADMF for asynchronous execution.

## B.3 IBM z17 migration considerations

The IBM Integrated Accelerator for zEDC is fully compatible with previous versions of zEDC. Data that is compressed by zEDC on previous IBM Z generations can be read by IBM z17 (the on-chip) NXU and vice versa.

### B.3.1 All z/OS configurations stay the same

No changes are required when moving from earlier systems using zEDC to IBM z17.

The IFAPRDxx chargeable feature is still required for authorized services. For problem state services, such as zlib use of Java, it is not required.

### B.3.2 Consider fail-over and disaster recovery sizing

The throughput increase on IBM z17 means that the throughput requirements need to be considered whether failing over to earlier IBM Z systems with zEDC in a Disaster recovery situation.

### B.3.3 Performance metrics

On-chip compression introduces the following system reporting changes:

- ▶ Synchronous executions are not recorded (just an instruction invocation)
- ▶ Asynchronous executions are recorded:
  - SMF30 information is captured for asynchronous usage
  - RMF EADM reporting is enhanced (RMF 74.10)
  - SAP use is updated to include the time that is spent compressing and decompressing

### B.3.4 zEDC to IBM z17 zlib Program Flow for z/OS

The z/OS-provided zlib library is statically linked into many IBM and ISV products and remains functional. However, to realize the best optimization for IBM Integrated Accelerator for zEDC, some minor changes are made to zlib.



The current zlib and the new zlib function are available for the IBM z15, z16 and z17 hardware. It functions with or without the IBM z15 / IBM z16 / IBM z17 z/OS PTFs on IBM z14 and below.

## B.4 Software support

On-Chip Compression function is compatible with zEDC support and is available in z/OS V2R5 and later for data compression and decompression. Although support for data recovery (decompression) if zEDC or On-Chip Compression not available, it is provided through software in z/OS V2R5 and later, with the appropriate program temporary fixes (PTFs).

Software decompression is slow and can involve considerable processor resources. Therefore, it is not recommended for production environments.

A specific fix category that is named `IBM.Function.zEDC` identifies the fixes that enable or use the zEDC and On-Chip Compression function.

z/OS guests that run under z/VM V7R3 with PTFs and later can use the zEDC Express feature and IBM z17 On-Chip Compression.

For more information about how to implement and use the IBM Z compression features, see *Reduce Storage Occupancy and Increase Operations Efficiency with IBM zEnterprise Data Compression*, [SG24-8259](#).

### B.4.1 IBM Z Batch Network Analyzer

IBM Z Batch Network Analyzer (zBNA) is a no-charge, “as-is” tool. It is available to customers, IBM Business Partners, and IBM employees.

zBNA can run on Microsoft Windows or Apple MacOS. It provides graphical and text reports, including Gantt charts, and support for alternative processors.

zBNA can be used to analyze client-provided System Management Facilities (SMF) records to identify jobs and data sets that are candidates for zEDC and IBM z17 On-Chip Compression across a specified time window (often a batch window).

zBNA can generate lists of data sets by the following jobs:

- ▶ Jobs that perform hardware compression and might be candidates for On-Chip Compression.
- ▶ Jobs that might be On-Chip Compression candidates, but are not in extended format.

Therefore, zBNA can help you estimate the use of On-Chip Compression features and help identify savings.

The following resources are available:

- ▶ IBM Employees can obtain zBNA and other CPS tools at the IBM Z Batch Network Analyzer (zBNA) Tool page of the IBM Techdoc website. For more information, contact your IBM Representative.
- ▶ IBM Business Partners can obtain zBNA and other CPS tools at the [IBM PartnerWorld website](#) (log in required).
- ▶ IBM customers can obtain zBNA and other CPS tools at the [IBM Z Batch Network Analyzer \(zBNA\) Tool web page](#) of the IBM Techdoc Library website.



## B.5 Compression acceleration and Linux on IBM Z

The IBM z17 On-Chip Compression accelerator solves the virtualization limitations in IBM z14 because the function is no longer an I/O device and is available as a problem state instruction to all Linux on IBM Z guests without constraints.

This feature enables pervasive usage in highly virtualized environments.

IBM z17 On-Chip Compression is available to open source applications by way of zlib.









# Tailored Fit Pricing and IBM Z Flexible Capacity for Cyber Resiliency

This appendix describes IBM Z Tailored Fit Pricing and IBM Z Flexible Capacity for Cyber Resiliency.

Although these are two different offerings, Tailored Fit Pricing for Software is a prerequisite for Flexible Capacity for Cyber Resiliency (or 'Flexible Capacity' in short).

This appendix includes the following topics:

- ▶ "Tailored Fit Pricing" on page 532
- ▶ "Software Consumption Model" on page 533
- ▶ "Hardware Consumption Model" on page 534
- ▶ "Conclusion" on page 537
- ▶ "IBM Z Flexible Capacity for Cyber Resiliency" on page 537
- ▶ "Use cases of IBM Flexible Capacity for Cyber Resiliency" on page 538
- ▶ "How does IBM Flexible Capacity for Cyber Resiliency work?" on page 539
- ▶ "Tailored fit pricing for hardware and IBM Z Flexible Capacity for Cyber Resiliency" on page 542
- ▶ "Ordering and installing IBM Z Flexible Capacity for Cyber Resiliency" on page 543
- ▶ "Terms and conditions of IBM Z Flexible Capacity for Cyber Resiliency" on page 544
- ▶ "IBM Z Flexible Capacity for Cyber Resiliency versus Capacity Back Up" on page 545



## C.1 Tailored Fit Pricing

**Note:** For more information about Tailored Fit Pricing, see this IBM Z Resources [web page](#).

The information that is included in this section about TFP is taken in part from the IBM publication *Tailored Fit Pricing - A sensible pricing model*, 75034875USEN-01.

IBM Z platform users were used to paying for IBM Z software and hardware at peak capacity required, and to manage their software costs by capping the machine usage. Traditionally, they capped the machine usage by using one or more of the following methods:

- ▶ Running batch workloads during off-shift hours
- ▶ Reducing machine resources accessed by development and test workloads
- ▶ Not introducing new workloads or applications onto the platform, even when it was the most logical technology for such workloads
- ▶ Investing in tools and resources to manage subcapacity capping

These approaches, while effective in predictable workload management, created a mind-set that stifles innovation and limits the ability for businesses to use the full value of Z technology, especially now as they adapt for digital transformation and the journey to hybrid cloud.

IBM introduced Tailored Fit Pricing (originally for IBM Z software) as a simpler pricing model to allow Z customers to better use their platform investments as their business demands, and in a more cost competitive way. Building on the success of this program, a variable IBM Z hardware model was introduced to extend the value of Tailored Fit Pricing for IBM Z.

With Tailored Fit Pricing models now available across hardware and software, customers can gain more flexibility and control with pricing solutions that can be tailored for business demands, which helps to balance costs while deriving even more value from hybrid cloud.

IBM's Tailored Fit Pricing model can address the following key concerns:

- ▶ Complexity of subcapacity pricing model that lead to IBM Z being managed as a cost center
- ▶ Difficulty in establishing the cost of new workload deployment and the effect on cost of existing workloads
- ▶ Investment in tools and resources to manage subcapacity that can inflate costs
- ▶ Lack of development and test resources
- ▶ Purchasing hardware for peak capacity to handle short term spikes

The software and hardware pricing models provide customers an opportunity to grow and more fully use their IBM Z investment for new opportunities.

IBM originally introduced DevTest Solution and New Application Solution in 2017. These solutions further evolved when in May 2019, IBM announced two significant other solutions: Enterprise Consumption (since renamed to the Software Consumption Solution), and Enterprise Capacity Solution.



In May 2021, IBM announced a new hardware solution, called Hardware Consumption Solution. All of these options were gathered into a new family of IBM Z pricing called Tailored Fit Pricing (TFP).

The Software Consumption Solution and the Hardware Consumption Solution are discussed next.

## C.2 Software Consumption Model

To meet the demands of modern workloads and provide a commercial confidence to match the technology confidence, the Consumption Solution was well received and adopted.

Customers typically transition onto TFP Consumption for the following reasons:

- ▶ It is a software pricing model that is better suited to today's workloads profiles (typically, where they are increasingly spiky). Also, it is a pricing model that is better suited to future uses; for example, inclusion in Hybrid Cloud architectures.
- ▶ A customer on TFP Consumption can confidently remove all forms of capping and expose all their workloads to *all* of the hardware infrastructure they own.
- ▶ Any form of growth (from a new workload to a 30-year-old COBOL application that is used more often) qualifies for a much-improved Price per MSU.

One key concept in a Software Consumption Model is the customer baseline. The IBM team works with the customer to review their previous 12 months production MSU consumption and billing to determine an effective price per MSU and establish a predictable price for all growth at discounted rate (see Figure C-1).

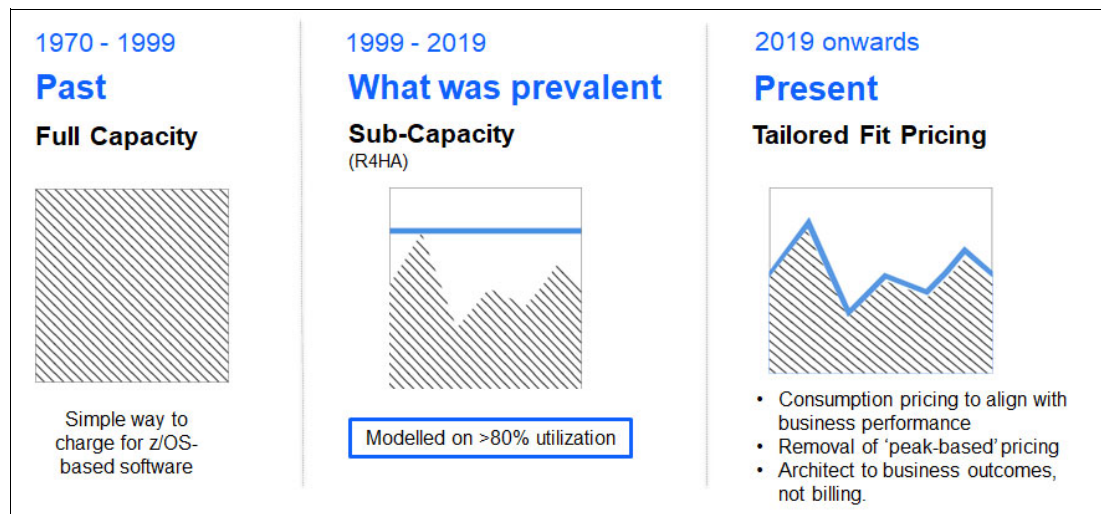


Figure C-1 Evolution of IBM Z Software pricing

With the Software Consumption model, no concept of peaks or white space is used that previously were integral to the sub-capacity-based model. Therefore, customers are free to remove capping and can use all of the owned capacity without worrying about penalties for peaking or spiking.

Although a customer does commit to an MSU baseline, if the MSUs for a specific year are not fully used, the customer can carry over any unused MSUs for use in the following year for the life of the contract. TFP consumption encourages and rewards growth; that is, MSUs that are processed above the baseline are charged at an aggressive Growth Price per MSU.



All workload processing can benefit by running with no capping and having all owned infrastructure available. Batch processing can also take advantage and reduce batch windows dramatically.

Without capping in place, customers can expect jobs to finish faster, yet at the same cost. Online processing can process more transactions simultaneously, which improves response times. This result is a function of the new billing approach, which is based on the actual amount of work that is performed, rather than the peak capacity consumption that is reached.

To provide improved economics for growth, TFP Consumption customers pay preferential pricing on the MSUs that are used above their baseline, regardless of whether that growth came from existing or new workloads. No other approval, qualification, or processing is required to use the growth pricing rate.

### **C.2.1 International Program License Agreement in the Software Consumption Model**

IBM Z enterprise customers often have One Time Charge (OTC) IBM products that are running in the same environment as MLC products. With the Software Consumption Model, IBM offers two choices for the handling of OTC software: A Full Capacity licensing model, and the option to use their existing OTC software licensing in a consumption model with the flexibility to use the entitlement on a consumption basis across a year as business demands.

With coverage of both MLC and capacity-based IPLA products, the Software Consumption Model offers a single and comprehensive software solution to IBM Z customers.

## **C.3 Hardware Consumption Model**

IBM introduced a hardware solution that provides added flexibility to the IBM Z IT infrastructure pricing and is a natural and valuable extension to the Tailored Fit Pricing models for IBM Z software.

To meet the demands of modern workloads, IBM Z hardware can now include, in addition to the base capacity, a subscription-based corridor of pay-for-use capacity. This always-on corridor of consumption-priced capacity helps alleviate the effect of short, unpredictable spikes in workload that are becoming more common in today's digital world.

The usage charges feature a granularity of 1 hour and are based on MSUs that are used, as measured by the subcapacity reporting Tool (SCRT), not full engine capacity.

Tailored Fit Pricing for IBM Z hardware enables customers to be ready for the unknown and unexpected. The presence of the always-on capacity contributes to better efficiency, reduced overhead, and shorter response times. This offering is available to IBM Z customers, who are using Tailored Fit Pricing for IBM Z software, for machines starting with the IBM z15, for z/OS general-purpose CP capacity. Starting from the z17 announced in 2025, Tailored Fit Pricing capacity will also become available for Integrated Facility for Linux (IFL) and IBM Z Integrated Information Processor (zIIP) specialty engines.



### C.3.1 Tailored Fit Pricing for IBM Z hardware in more detail

Tailored Fit Pricing for IBM Z Hardware (TFP HW pricing) includes two aspects: a Subscription charge, and a Usage charge (see Figure C-2).

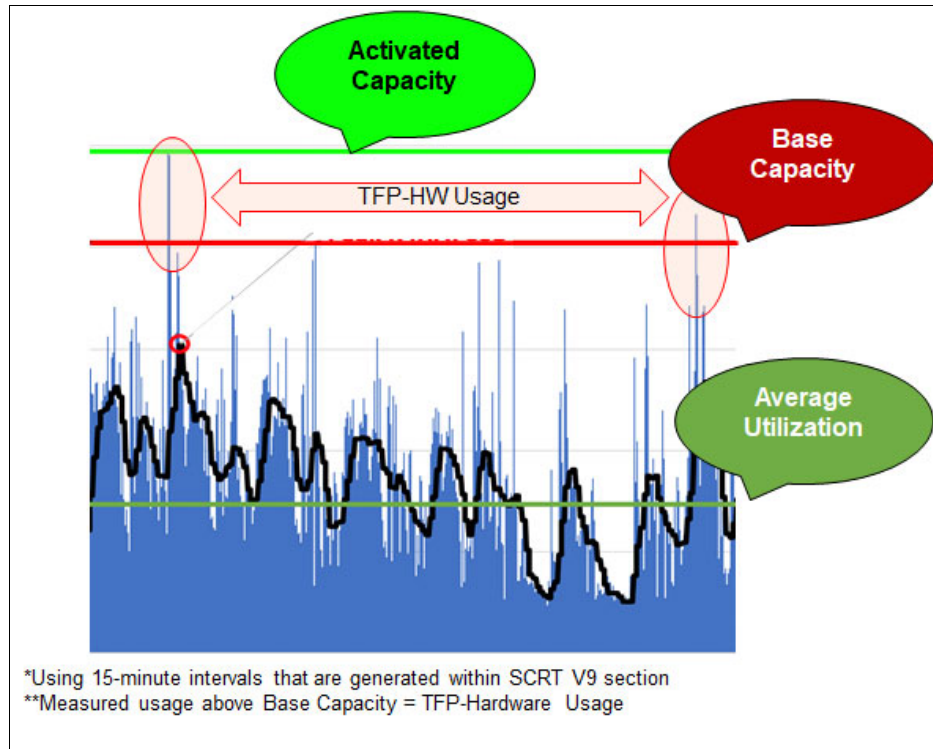


Figure C-2 TFP-Hardware use

One of the most important facts about TFP Hardware is that a single system or central processing complex (CPC) is always considered. It is on this level that usage is measured. The blue bars that are shown in Figure C-2 represent individual 15-minute intervals, and the usage is measured in each one.

In Figure C-2, the dark green line shows the average use of the machine over the entire period that is measure (normally a month). The black line shows the Rolling 4 Hour Average (R4HA), and it is displayed for information purposes only because the R4HA is not used for TFP-Hardware calculations.

The red line is the customer owned capacity (or Base capacity). The light green line shows the activated capacity of the machine, as reported by Sub-Capacity Reporting Tool (SCRT).

As shown in Figure C-2, some of the blue bars reach above the red line, into the corridor between the red line and light green lines, which represents the TFP-Hardware capacity. Therefore, use over the customer's owned (Purchased) capacity is measured, so a Usage charge is incurred. If no use is measured over the customer's owned (Purchased) capacity, no Usage charge is incurred.

In addition to the Usage charge, which might not be relevant, a Subscription charge also can be assessed. The Subscription charge is a flat, per system, per month payment that is based on the amount of TFP-Hardware capacity that is provided on the system. The Subscription charge is invoiced prepaid for the contract term, or postpaid monthly.



Because only entire engines can be activated, it is always full engine sizes that are based on IBM Large System Performance Reference (LSPR) capacity levels. The Subscription charge covers the value that the extra activated capacity brings, even if no measured usage exists (see Figure C-3).

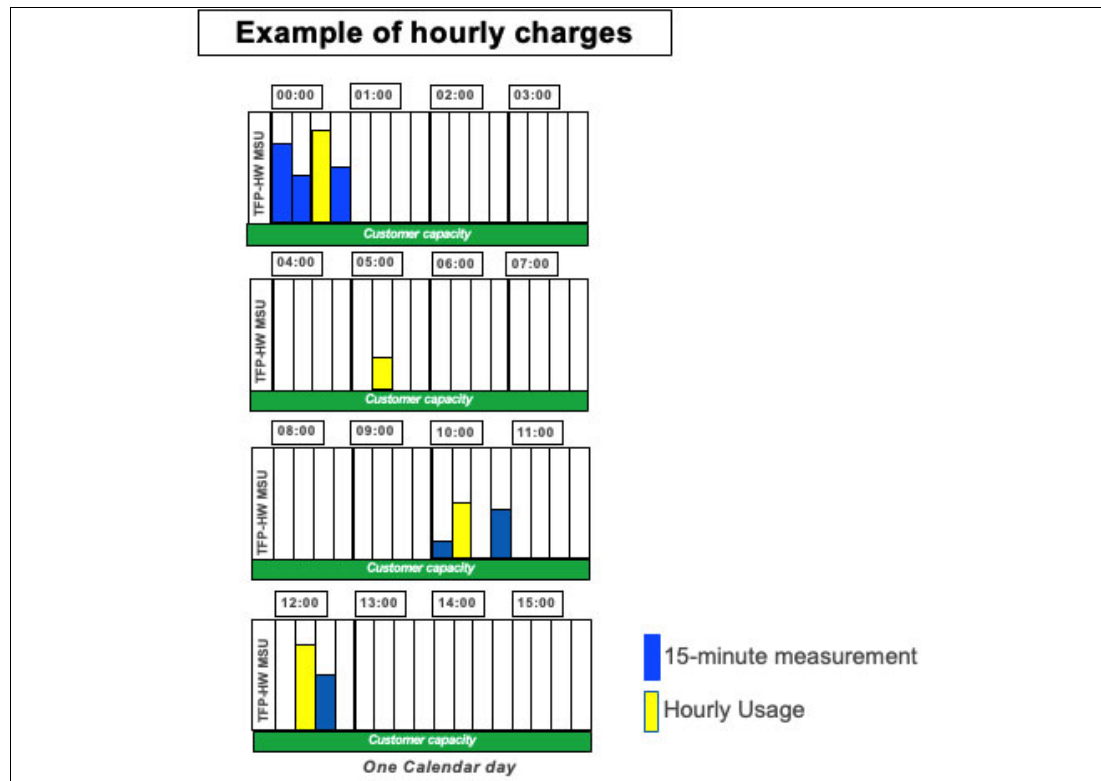


Figure C-3 Charge examples

The green bar in Figure C-3 represents the capacity that the customer owns (the so-called *Base capacity*). The transparent bars represent activated TFP-Hardware corridor.

The blue bars show the measured usage within the 15-minute intervals that are greater than the customer-owned capacity. The yellow bar is the highest measured TFP-Hardware use within the defined hour.

The spikes that are counted for invoicing are the yellow bars (the highest within each hour).

The yellow bars hold a measured million of service units (MSU) value. If the total MSU use of the yellow bars is 250 MSU, the customer receives an invoice of 250 times the hourly usage charge per MSU.

### C.3.2 Efficiency benefits of TFP-Hardware

Even if no usage measured, the following benefits are realized resulting from having the extra capacity activated:

- Improved and more predictable response times with lower latency (especially when compared to a public cloud solution)
- Faster transaction processing, with shorter spikes of high use



- ▶ Higher number of active processor engines have a positive n-way effect (higher parallelization) and delivers more cache, less contention, and overhead
- ▶ Optimized workload handling under customer-defined use thresholds
- ▶ Improved insight for future capacity planning
- ▶ Improved balance between physical and logical Central Processor (CPs)
- ▶ Reduced Processor Resource/System Manager (PR/SM) logical partitions (LPAR) management, less overhead

## C.4 Conclusion

The combination of TFP for the hardware and software is a powerful one for our customers to maximize their investment in IBM Z. For eligible customers, it allows them to use their hardware for their general capacity requirements with a consumption-based corridor on top.

When this TFP for hardware is combined with the TFP-SW solutions, it allows customers to unleash the full power of their IBM Z machine.

## C.5 IBM Z Flexible Capacity for Cyber Resiliency

Resiliency continues to be a hot topic for clients and a key value driver for IBM Z, especially for the largest IBM Z clients in regulated industries.

Flexible Capacity (Flex Cap) is designed to provide increased flexibility and control to shift workloads between participating IBM z16 and IBM z17 machines at different locations and stay for up to 12 months. This helps organizations to improve their resiliency and continue to comply to evolving regularly mandates by running a real failover scenario for workloads that demonstrates business continuity.

By using IBM GDPS, scripts can be run to fully automate Flexible Capacity for Cyber Resiliency workload shifts between participating IBM z16 and z17 machines at different locations.

The following Flexible Capacity features were introduced with IBM z16:

- ▶ **Flexible Capacity for Cyber Resiliency Enterprise Edition**

Allows active capacity flexibility for all engine types to allow reallocating workloads (including production) for a maximum period of 12 months and a maximum of 12 events (activations or deactivations) per contract year to facilitate intra- and inter-site workload/role swaps. A migration period of 24 hours is permitted when the flexible capacity record is activated on both machines.

- ▶ **Flexible Capacity for Cyber Resiliency Limited Term**

Allows active capacity flexibility for all engine types to allow reallocating workloads (including production) for maximum period of 30 to comply with regulator requirements, pro-active DR, facility maintenance, and 90 days in case of a real DR. A maximum of 4 events (activations or deactivations) per contract year, but only across different data centers (inter-site). A migration period of 24 hours is permitted when flexible capacity record is activated on both machines.

With IBM z17 the Flexible Capacity capabilities were extended with the following features:

- ▶ **Flexible Capacity Infrastructure Testing**



Enabling clients to perform infrastructure testing only with a copy of your workload within an isolated (network) environment, for example testing whether processes, connections, and automation works (SAN, Network, Distributed Systems, 3rd party license keys, lights-on, GDPS scenarios, etc.). This can be done for a maximum of 10 days and a minimum of 72 hours between tests. This feature can be ordered at increments of 1 per machine, with a maximum of 10.

► **Additional Flexible Capacity Events**

Enabling clients to optionally acquire additional events (activations or deactivations) beyond the standard 4 or 12 per contract year that Flexible Capacity for Cyber Resiliency offers by default. This can be achieved by purchasing prepaid features. The prepaid feature is available for the lifespan of the Flex Cap record.

► **Overrun features (for Flex Cap, Tailored Fit Pricing HW, and Test and Stress Testing)**

These are billing features that will be used in case a client:

- exceeds the migration time of 24 hours (Flex Cap)
- has more capacity active after the migration period than their owned or contracted capacity (Flex Cap / TFP HW / TST)
- runs an infrastructure test for longer than 10 days (Flex Cap)
- consumes more Stress Test days than contracted (TST)
- is running TST during the migration period of Flex Cap

**Note:** TST (Test and Stress Testing) capacity must be deactivated in case the TST Logical Partitions (LPARs) are being moved to a different machine as part of the migrations.

## C.6 Use cases of IBM Flexible Capacity for Cyber Resiliency

In this section, we present several IBM Flexibly Capacity use cases for Cyber Resiliency.

### C.6.1 Disaster recovery and DR testing

Activate required capacity at your DR site to continue running your business. Automate and test recovery procedures for unplanned outages, including cyber attacks, to help provide near continuous availability and DR

### C.6.2 Frictionless compliance

Meet the ever-evolving requirements of global regulators, allowing a highly automated and fast process to demonstrate a production site swap

### C.6.3 Facility maintenance

Run your production workload from your alternative site while you maintain your primary site with the capacity that you need.

### C.6.4 Pro-active avoidance

Protect your critical business services from natural disasters. Avoid rolling power outages. Migrate your critical workloads to an alternative site before your business is affected and remain in that state for up to one year.



## C.7 How does IBM Flexible Capacity for Cyber Resiliency work?

This section describes the high-level process for implementing IBM Flexible Capacity for Cyber Resiliency.

### C.7.1 Set up process

Flexible Capacity for Cyber Resiliency is facilitated through a Temporary Capacity record.

In the first step of the setup process (see Figure C-4), the active capacity of the participating IBM z16 and IBM z17 machines is changed to a base machine plus the temporary entitlement record (TER) up to the High Water Mark (HWM) of the machine. The base capacity is defined by the customer. The machines' HWM remains unchanged.

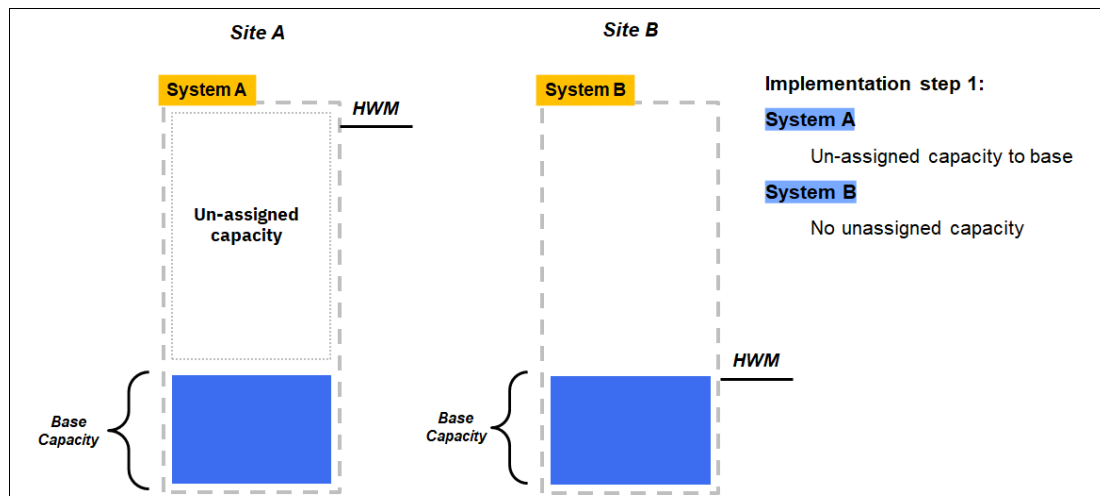


Figure C-4 Initial configuration

In the next step (see Figure C-5), the now unassigned capacity is restored with a new Flexible Capacity record. Another Flexible Capacity record is installed on System B to increase the capacity to the amount of MIPS that the customer licensed.

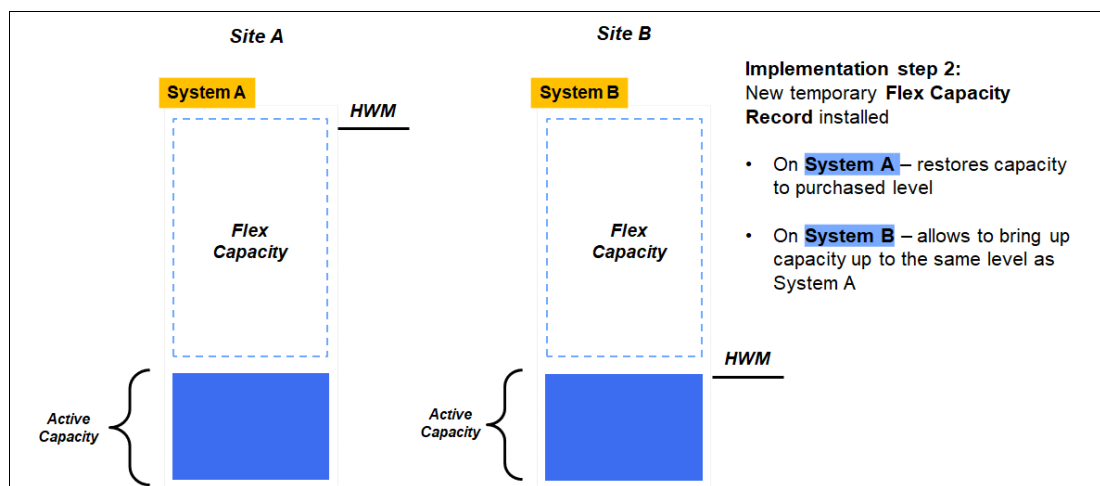


Figure C-5 Flexible Capacity Installed record



The IBM Flex Capacity record on System A then is activated until the machine's HWM. On System B, the Flexible Capacity record is installed, but not activated. The setup process is now complete (see Figure C-6).

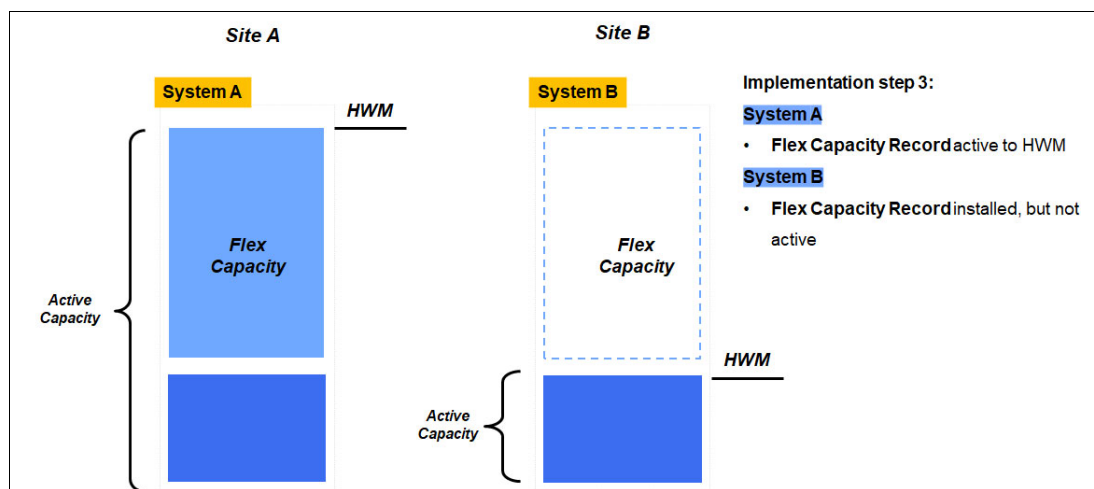


Figure C-6 Flexible Capacity activated record

## C.7.2 Transferring workloads

After the set up is complete, it is possible to transfer workload from Site A to Site B. This process can be automated by using GDPS technology.

GCP (general-purpose processor) capacity is measured by using licensed MIPS. However, you cannot activate half a CP. Therefore, it is the customer's responsibility to stay under the licensed amount of MIPS; otherwise, extra charges are incurred for the use that is greater than the licensed MIPS.

After the Flexible Capacity record is active on both sites, a time limit of 24 hours begins in which workloads can be transferred from Site A to Site B without leading to more charges.

During this time window, Flexible Capacity can remain active on both sites, as shown in Figure C-7

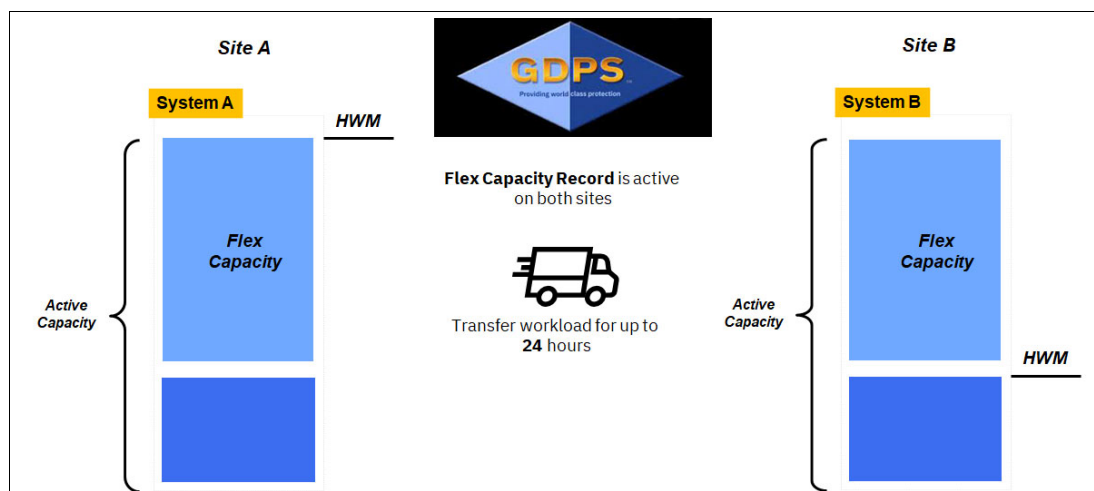


Figure C-7 Transfer window



After 24 hours, the Flexible Capacity in Site A must be deactivated and System A reduced to base capacity (see Figure C-8).

Flexible Capacity on System B can stay running for up to 12 months.



Figure C-8 Stay in DR for up to one year

### C.7.3 Multi-system environment

A two-site, two-system environment is only the base infrastructure. Flexible Capacity for Cyber Resiliency also is suited for complex multi-system environments.

Flexible Capacity from several systems can be consolidated to single systems in Site B or vice-versa (see Figure C-9).

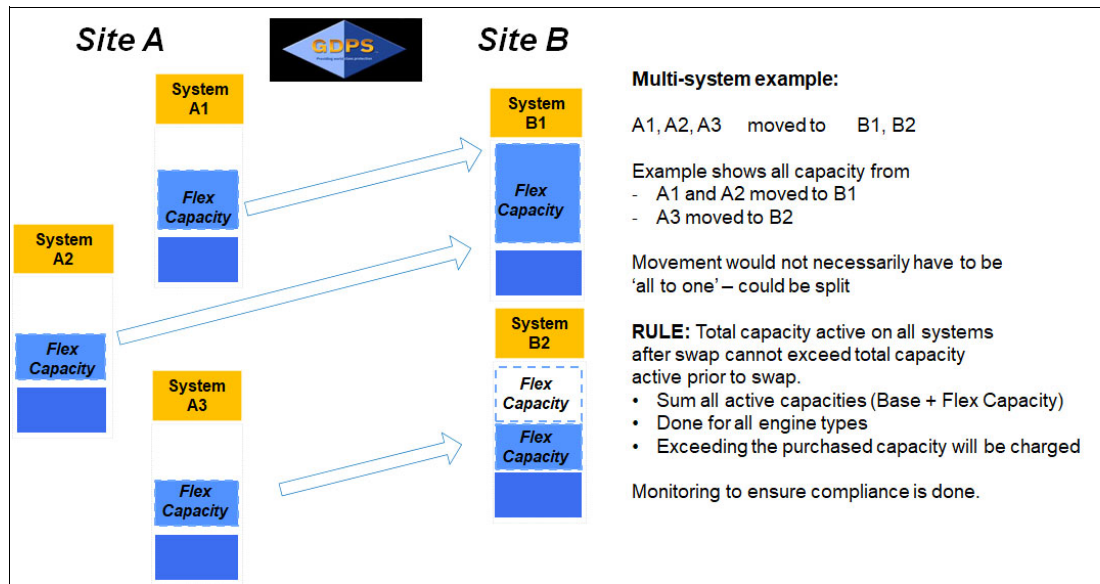


Figure C-9 Multi-system Flex Capacity example

Partial activations and deactivations of the Flexible Capacity record can be done.

However, the total capacity that is active on all participating systems after the migration period cannot exceed the total capacity owned by the client on those participating systems.



Flexibly Capacity is available for all engine types; exceeding the purchased capacity incurs charged. Monitoring to ensure compliance is done by way of Call Home data that IBM receives periodically. The new Flexible Capacity Overrun features will be used to invoice additional costs associated with the customer overrunning agreed terms and conditions:

- ▶ Migration period longer than 24 hours
- ▶ More capacity active than the customer owned capacity after the migration period
- ▶ FlexCap for Infrastructure testing is active longer than 10 days

## C.8 Tailored fit pricing for hardware and IBM Z Flexible Capacity for Cyber Resiliency

IBM Z Flexible Capacity for Cyber Resiliency can be used with TFP Hardware. The Flexible Capacity limit is always based on the Base capacity. The presence or activation of TFP-Hardware does not affect the amount of capacity that can be activated by a Flexible Capacity Transfer record.

The Flexible Capacity Transfer record always is considered to be the first record that was activated, regardless of the order in which temporary records or TFP-Hardware were activated.

The example that is shown in Figure C-10 shows an active Sysplex across Site A and Site B.

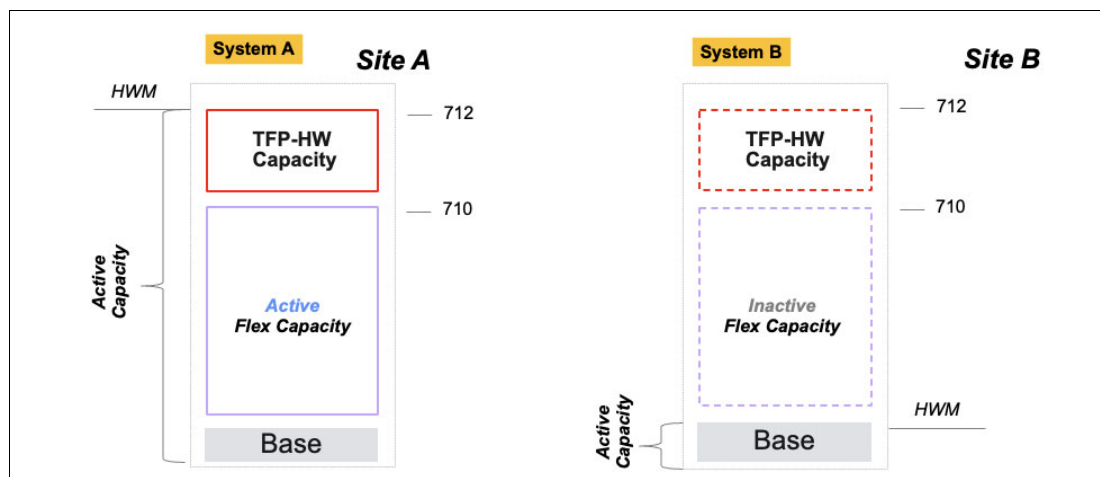


Figure C-10 Flexible Capacity: Two site sysplex

System A has a Base capacity of 401 and the Flexible Capacity Record is activated up to a 710. On top of the active Flexible Capacity sits a TFP Hardware capacity of two extra CPs to a maximum of 712.

System A is in a Sysplex with System B that has the same configuration.

Now, the data-center operator decides to perform maintenance on System A and activates the Flexible Capacity and TFP-HW capacity on System B. After migration of the workloads, the active Flexible Capacity and TFP-HW capacity on System A needs to be deactivated. This operation needs to be completed within the 24 hours of the migration period.



Figure C-11 shows the configuration of both machines after the transfer.

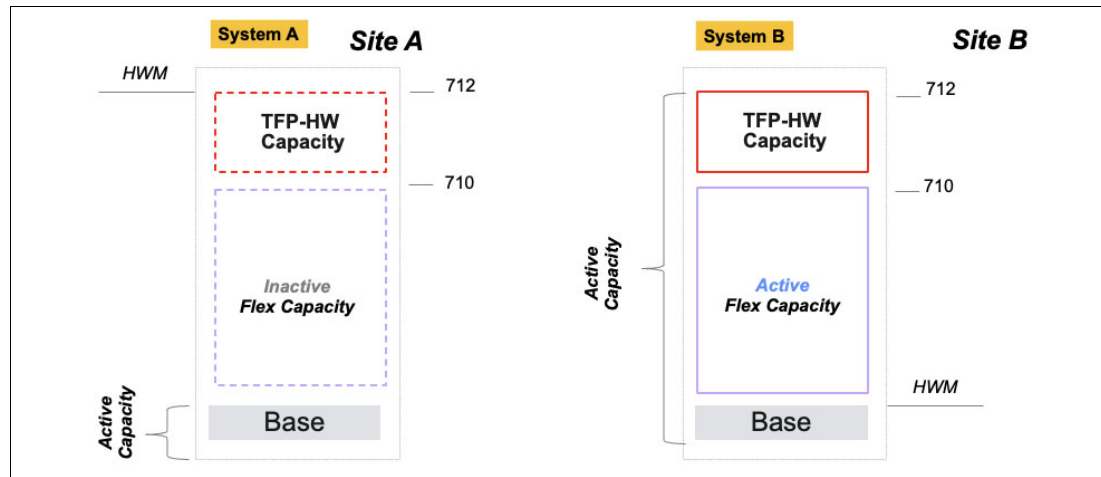


Figure C-11 Flexible Capacity with two site sysplex - transfer

After the migration period System A has deactivated all of its Flexible Capacity and TFP for HW capacity and is left with only the base capacity of 401. On System B all Flexible Capacity is active and on top of the Flexible Capacity record sits the TFP Hardware capacity.

Now, System A can safely perform maintenance.

System B has now the entire Flexible Capacity record activated and shows an active capacity of 710. The TFP-Hardware capacity sits again on top of the active capacity and adds now another 2 CPs to the 712.

This shows that the presence or activation of TFP-Hardware does not impact the amount of capacity that can be activated by a Flexible Capacity Transfer record.

The TFP-Hardware capacity always “floats on top” of any other activated capacity for TFP-Hardware usage charging. Therefore, no double charging can occur.

## C.9 Ordering and installing IBM Z Flexible Capacity for Cyber Resilience

To facilitate ordering Flexible Capacity, the following Capacity on-Demand feature codes (FC) were introduced:

- ▶ Flexible Capacity Authorization (FC 9933)
- ▶ Flexible Capacity Transfer Record (FC 0376)
- ▶ Billing feature codes (Feature Codes 0317 - 0322, and 0376 - 0386)

Optional features to extent Flexible Capacity functionality:

- ▶ Flexible Cap 10-day test (FC 0824)
- ▶ Prepaid Flexible Capacity Event action (FC 0837)
- ▶ Violation Flexible Capacity Event (FC 0838)

FC 9933 and 0376 must be ordered on each machine that is participating in Flexible Capacity for Cyber Resiliency.



FC 0824 (10 days test) must be purchased on each machine where Infrastructure Testing is required. This feature can be ordered at increments of 1 per machine, with a maximum of 10, and requires FC 9933 and FC 0376.

A new Capacity on-Demand contract attachment also must be signed by the customer.

## C.10 Terms and conditions of IBM Z Flexible Capacity for Cyber Resiliency

The terms and conditions of IBM Z Flexible Capacity for Cyber Resiliency are listed in Table C-1.

*Table C-1 Terms and conditions of IBM Flexible Capacity*

Term or condition	Description
<b>Cross machine (de)activation</b>	Enterprise Edition: Inter- and intra-site workload moves can be done, regardless of distance, mirroring, or coupling technology.  Limited Term: Intra-site moves are <i>not</i> allowed (two machines in the same data center cannot move capacity back and forth).
<b>Entitlement</b>	The owner of the machine holds a title to the physical hardware. The capacity of that machine is enabled and controlled by way of the LIC of the machine, which is licensed, not sold.
<b>Overlap period</b>	A 24-hour period in which the temporary record can be active on both systems.
<b>Activation limit</b>	4 (LiTe) or 12 (EE) activations or deactivations per serial number within 12 months.
<b>Activation period</b>	Keep the flexible capacity record active on your alternative site for up to 12 months.
<b>License transfer</b>	LIC is licensed only to one serial numbered machine, and its transfer to another machine is not permitted (but can be carried forward in case of generation upgrade).
<b>License expiration</b>	The record associated with the Flexible Capacity license can be ordered for a maximum period of 5 years, and up until Withdrawn from Marketing it can be extended, but always to a maximum of 5 years total.
<b>TFP for software</b>	Offering requires TFP for software; CMP is grandfathered in.
<b>Maintenance</b>	Maintenance is post-paid for Flexible Capacity activated (per day).
<b>Microcode only</b>	IBM Z Flexible Capacity for Cyber Resiliency is Microcode only. Additional Memory, I/O Cards, drawers and other infrastructure related components are not part of this solution.
<b>Call home</b>	Customer agrees to use Call Home data to monitor capacity usage.
<b>Charges for capacity exceeding the temp record</b>	Capacity that is used beyond the purchased capacity is charged at previously defined overrun prices.



## C.11 IBM Z Flexible Capacity for Cyber Resilience versus Capacity Back Up

Flexible Capacity and Capacity Back Up (CBU) feature different purposes and limitations, as listed in Table C-2.

Table C-2 Differences between Flexible Capacity and CBU

	IBM Z Flexible Capacity for Cyber Resiliency	IBM Z Flexible Capacity for Infrastructure Testing	Capacity Back Up
<b>Primary purpose and benefits</b>	<p>All CBU benefits at an alternative site, plus:</p> <ul style="list-style-type: none"> <li>▶ Run mainframe workload in alternative site for up to 1 year</li> <li>▶ Demonstrate ability to migrate entire mainframe workload to alternative site (for example, regulatory requirements)</li> <li>▶ Run entire workload in alternative machine or site for extended period (for example, facilities requirements)</li> <li>▶ Proactively swap production to alternative out-of-region site before a (natural) disaster occurs</li> </ul>	<p>Dynamically activate and deactivate capacity to facility testing only (see limitations below)</p>	<p><b>Real CBU activation:</b></p> <ul style="list-style-type: none"> <li>▶ Run full mainframe workload if a real DR event is declared.</li> </ul> <p><b>CBU Test activation:</b></p> <ul style="list-style-type: none"> <li>▶ Ability to test DR readiness.</li> <li>▶ Validate with non-production workload.</li> <li>▶ Validate with production workload. Requires primary site capacity to be disabled.</li> </ul>
<b>Limitations</b>	<ul style="list-style-type: none"> <li>▶ Alternative site can be used for maximum of 1 year (for LiTe this is 30 days and 90 days in case of a real DR)</li> <li>▶ Allows up to 24-hour overlap period, where full capacity is active on both sites for transferring workload</li> <li>▶ FlexCap EE: Allows 12 events (deactivations and activations) per serial number per contract year.</li> <li>▶ FlexCap LiTe: Allows 4 events per serial number per contract year,.</li> </ul>	<ul style="list-style-type: none"> <li>▶ Can only be performed with copies of production or test environments</li> <li>▶ Each activation is limited to a maximum of 10 days, with a minimum of 72 hours between activations</li> </ul>	<p><b>Real CBU activation:</b></p> <ul style="list-style-type: none"> <li>▶ A real CBU activation is limited to a maximum of 90 days</li> </ul> <p><b>CBU Test activation:</b></p> <ul style="list-style-type: none"> <li>▶ Each CBU test is limited to a maximum of 10 days, with a minimum of 72 hours between activations</li> <li>▶ When validating with production workload, the primary site capacity needs to be disabled.</li> </ul>



	IBM Z Flexible Capacity for Cyber Resiliency	IBM Z Flexible Capacity for Infrastructure Testing	Capacity Back Up
<b>Contracts and features</b>	<ul style="list-style-type: none"><li>▶ LIC is licensed to only one specific serial-numbered machine, and its transfer to another machine is not permitted</li><li>▶ Offering requires TFP for software</li><li>▶ The LIC license expires up to 5 years past WFM, depending on renewal date. An LIC license can be carried forward under the condition there is an announced upgrade path to the new machine and the contract is not expired.</li></ul>	<ul style="list-style-type: none"><li>▶ A supplement to the Flexible Capacity attachment is needed</li></ul>	<ul style="list-style-type: none"><li>▶ Capacity Back Up Agreement</li><li>▶ CBU capacity provided in annual increments</li><li>▶ CBU tests provided in individual increments</li></ul>





# Channel options

This appendix describes all channel attributes, required cable types, maximum unrepeated distance, and bit rate for IBM z17. The features that are hosted in the PCIe drawer for Cryptography also are listed.



## D.1 Available Channel options

For all optical links the connector type is LC Duplex, except for the zHyperLink and the ICA SR connections, which are established with multi-fiber push-on (MTP) connectors.

The MTP connector of the zHyperLink and the ICA SR connection feature two rows of 12 fibers and are interchangeable.

The electrical Ethernet cable for the Open Systems Adapter (OSA) connectivity is connected through an RJ45 jack.

The attributes of the channel options that are supported on IBM z17 are listed in Table D-1.

Table D-1 IBM z17 channel feature support

Channel feature	Feature codes	Bit rate <sup>a</sup> in Gbps (or stated)	Cable type	Maximum unrepeated distance <sup>b</sup>	Ordering information
<b>zHyperLink and Fiber Connection (FICON)</b>					
zHyperlink Express2.0	0351	8 Gbps	OM3, OM4	See Table D-2 on page 550	New build
FICON Express32-4P LX	0387	8, 16, or 32	SM 9 µm	5 km @ 32 Gbps <sup>c</sup> (3.1 miles) 10 km @ 16 or 8 Gbps (6.2 miles)	New build
FICON Express32-4P SX	0388	8, 16, or 32	OM1, OM2, OM3, OM4	See Table D-3 on page 550.	New build
FICON Express32S LX	0461	8, 16, or 32	SM 9 µm	5 km @ 32 Gbps <sup>c</sup> (3.1 miles) 10 km @ 16 or 8 Gbps (6.2 miles)	Carry forward
FICON Express32S SX	0462	8, 16, or 32	OM1, OM2, OM3, or OM4	See Table D-3 on page 550.	Carry forward
FICON Express16SA LX	0436	8, or 16	SM 9 µm	10 km (6.2 miles)	Carry forward
FICON Express16SA SX	0437	8, or 16	OM1, OM2, OM3, OM4	See Table D-3 on page 550.	Carry forward
<b>Network Express, Open Systems Adapter (OSA) and Remote Direct Memory over Converged Ethernet (RoCE)</b>					
Network Express LR 25G	0527	25	SM 9 µm	10 km (6.2 miles)	New build
Network Express SR 25G	0526	25	MM 50 µm	OM3: 70 m (229 feet) OM4: 100 m (328 feet)	New build
Network Express LR 10G	0525	10	SM 9 µm	10 km (6.2 miles)	New build
Network Express SR 10G	0524	10	MM 50 µm	OM1: 33 m (108 feet) OM2: 82 m (269 feet) OM3: 300 m (984 feet) OM4: 400 m (1312 feet)	New build
OSA-Express7S 1.2 25GbE LR	0460	25	SM 9 µm	10 km (6.2 miles)	New build, Carry Forward



Channel feature	Feature codes	Bit rate <sup>a</sup> in Gbps (or stated)	Cable type	Maximum unrepeat distance <sup>b</sup>	Ordering information
OSA-Express7S 1.2 25GbE SR	0459	25	MM 50 $\mu$ m	OM3: 70 m (229 feet) OM4: 100 m (328 feet)	New build, Carry Forward
OSA-Express7S 1.2 10GbE LR	0456	10	SM 9 $\mu$ m	10 km (6.2 miles)	New build, Carry Forward
OSA-Express7S 1.2 10GbE SR	0457	10	MM 62.5 $\mu$ m MM 50 $\mu$ m	OM1: 33 m (108 feet) OM2: 82 m (269 feet) OM3: 300 m (984 feet) OM4: 400 m (1312 feet)	New build, Carry Forward
OSA-Express7S 1.2 GbE LX	0454	1.25	SM 9 $\mu$ m	5 km (3.1 miles)	New build, Carry Forward
OSA-Express7S 1.2 GbE SX	0455	1.25	MM 62.5 $\mu$ m MM 50 $\mu$ m	OM1: 275 m (902 feet) OM2: 550 m (1804 feet)	New build, Carry Forward
OSA-Express7S 10GbE LR	0444	10	SM 9 $\mu$ m	10 km (6.2 miles)	Carry forward from z15 only
OSA-Express7S 10GbE SR	0445	10	MM 62.5 $\mu$ m MM 50 $\mu$ m	OM1: 33 m (108 feet) OM2: 82 m (269 feet) OM3: 300 m (984 feet) OM4: 400 m (1312 feet)	Carry forward from z15 only
OSA-Express7S GbE LX	0442	1.25	SM 9 $\mu$ m	5 km (3.1 miles)	Carry forward from z15 only
OSA-Express7S GbE SX	0443	1.25	MM 62.5 $\mu$ m MM 50 $\mu$ m	OM1: 275 m (902 feet) OM2: 550 m (1804 feet)	Carry forward from z15 only
OSA-Express7S 1000BASE-T	0446	1000 Mbps	Cat 5, Cat 6 unshielded twisted pair (UTP)	100 m (328 feet)	Carry forward from z15 only
<b>Parallel Sysplex</b>					
Coupling Express3 LR 25Gb	0499	25	SM 9 $\mu$ m	10 km (6.2 miles)	New Build
Coupling Express3 LR 10Gb	0498	10	SM 9 $\mu$ m	10 km (6.2 miles)	New Build
ICA SR2.0	0216	8 GBps	OM3, OM4	See Table D-2 on page 550	New Build

- a. The link data rate does not represent the performance of the link. The performance depends on many factors, including latency through the adapters, cable lengths, and the type of workload.
- b. Where applicable, the minimum fiber bandwidth distance in MHz-km for multi-mode fiber optic links is included in parentheses.
- c. For 32 Gbps links, point to point (to another switch, director, DWDM equipment or another FICON Express32S or FICON Express32-4P) distance is limited to 5 km (3.1 miles).

The unrepeat distances for different multimode (MM) fiber optic types for zHyperLink Express and ICA SR are listed in Table D-2 on page 550.



*Table D-2 Unrepeated distances*

Cable type (modal bandwidth)	ICA SR / zHyperLink transfer rate (8 GBps)
OM3 (50 µm at 2000 MHz·km)	100 meters
	328 feet
OM4 (50 µm at 4700 MHz·km)	150 meters
	492 feet

The maximum unrepeated distances for FICON SX features are listed in Table D-3.

*Table D-3 Maximum unrepeated distance for FICON SX features*

Cable type/bit rate	8 Gbps	16 Gbps	32 Gbps
OM1 (62.5 µm at 200 MHz·km)	21 meters	N/A	N/A
	69 feet	N/A	N/A
OM2 (50 µm at 500 MHz·km)	50 meters	35 meters	20 meters
	164 feet	115 feet	65 feet
OM3 (50 µm at 2000 MHz·km)	150 meters	100 meters	70 meters
	492 feet	328 feet	229 meters
OM4 (50 µm at 4700 MHz·km)	190 meters	125 meters	100 meters
	693 feet	410 feet	328 feet



**E**

# Frame configurations with Power Distribution Units

This appendix describes the various frame configurations for Power Distribution Units (PDUs) based systems. All of the figures that are included here are views from the rear of the system.

The common building blocks are displayed and range from 1 - 4 frames, with various numbers of CPC drawers and PCIe+ I/O drawers.

This chapter includes the following topic:

- ▶ “Power Distribution Unit configurations” on page 551
- ▶ New PCIe+ I/O drawers / PCHIDs location Options on page 555

## E.1 Power Distribution Unit configurations

The various PDU-based system configurations are shown in Figure E-1 on page 552 - Figure E-6 on page 554.



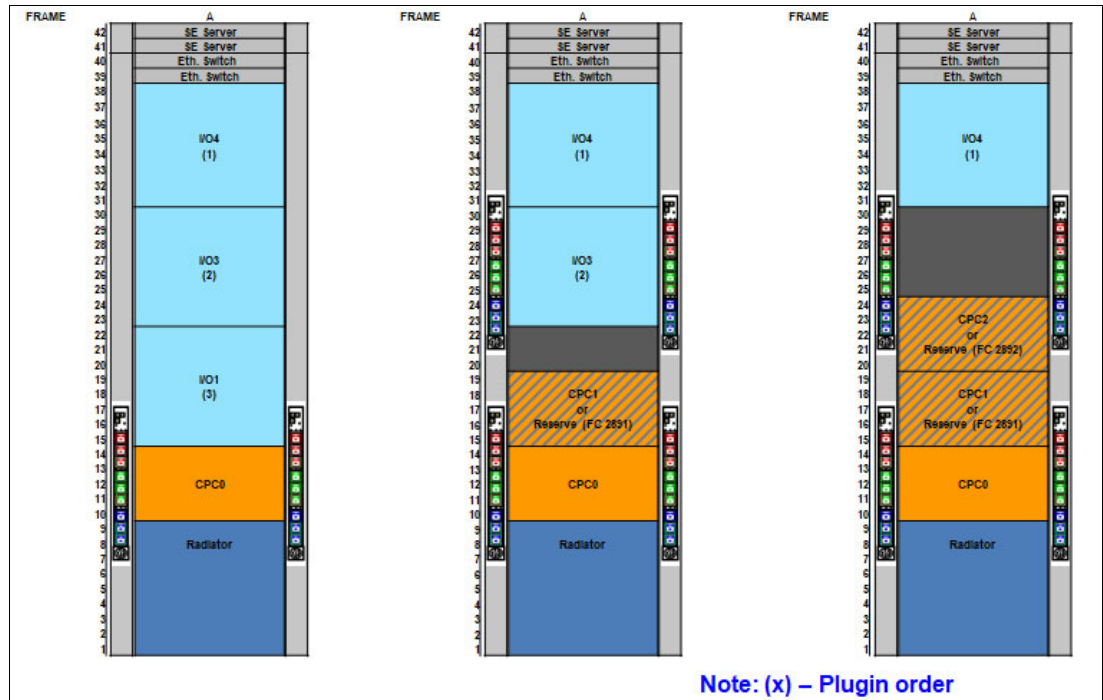


Figure E-1 Single frame

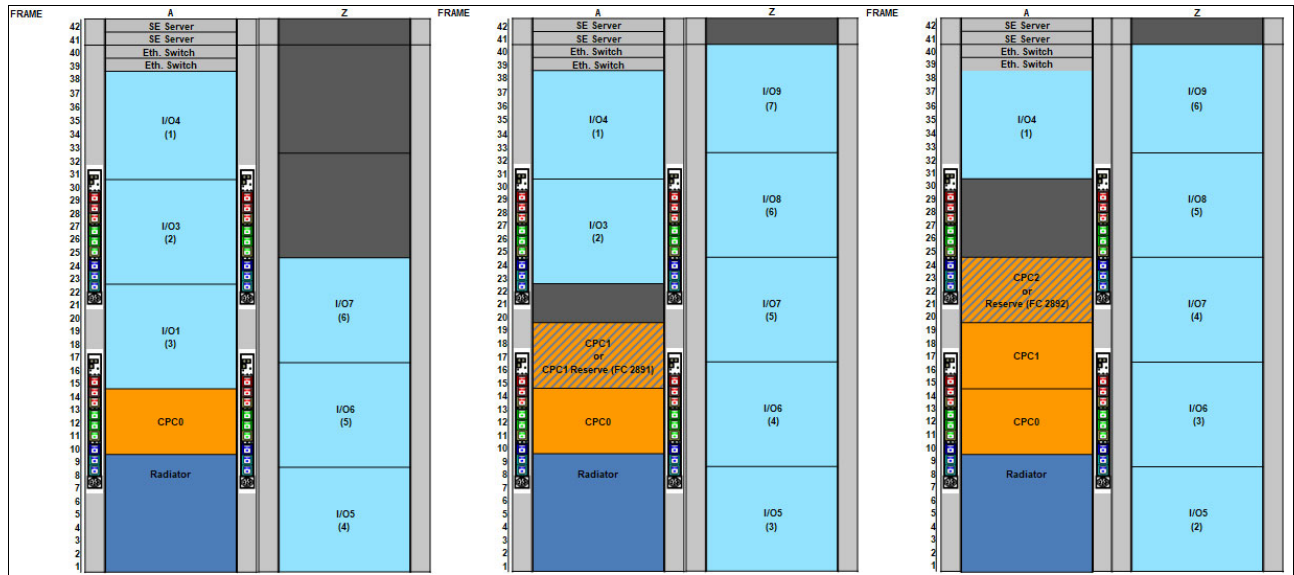


Figure E-2 Two frames AZ



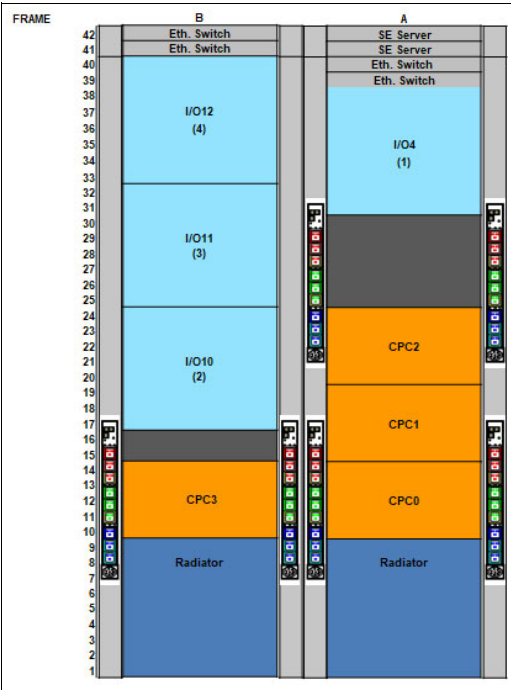


Figure E-3 Two frames AB

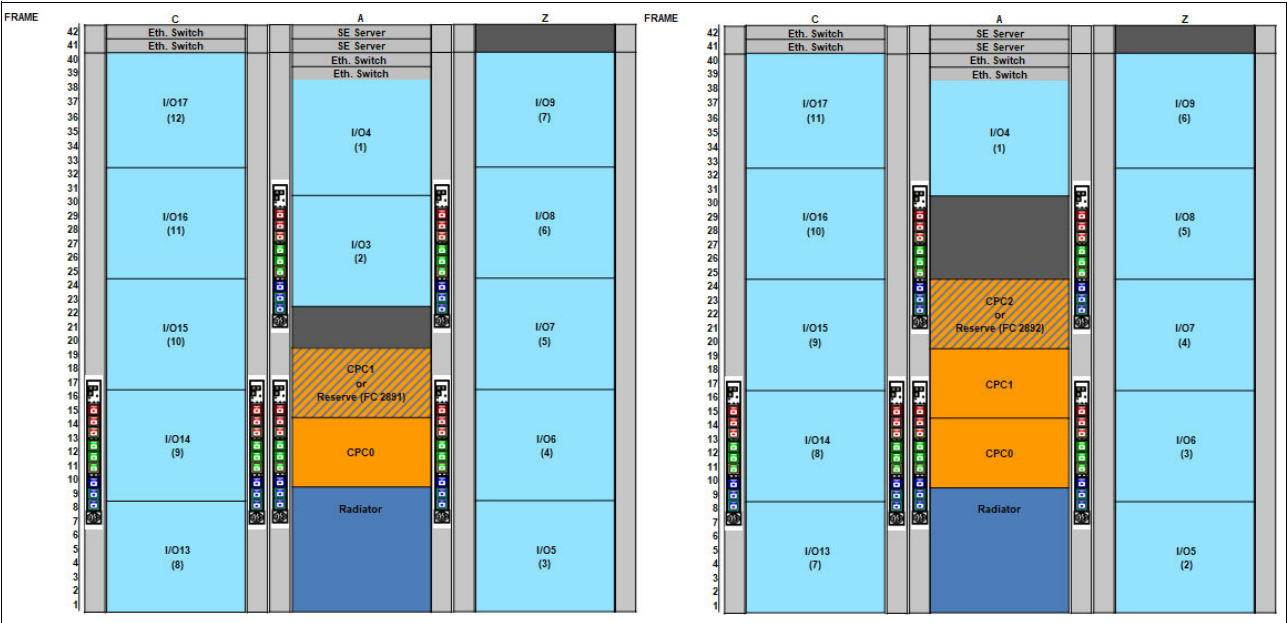


Figure E-4 Three frames ZAC



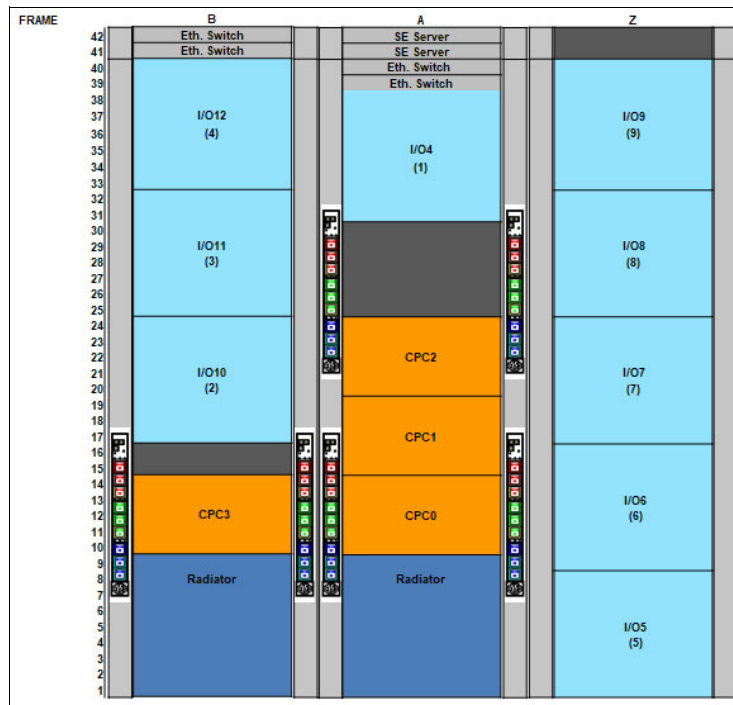


Figure E-5 Three frames ZAB

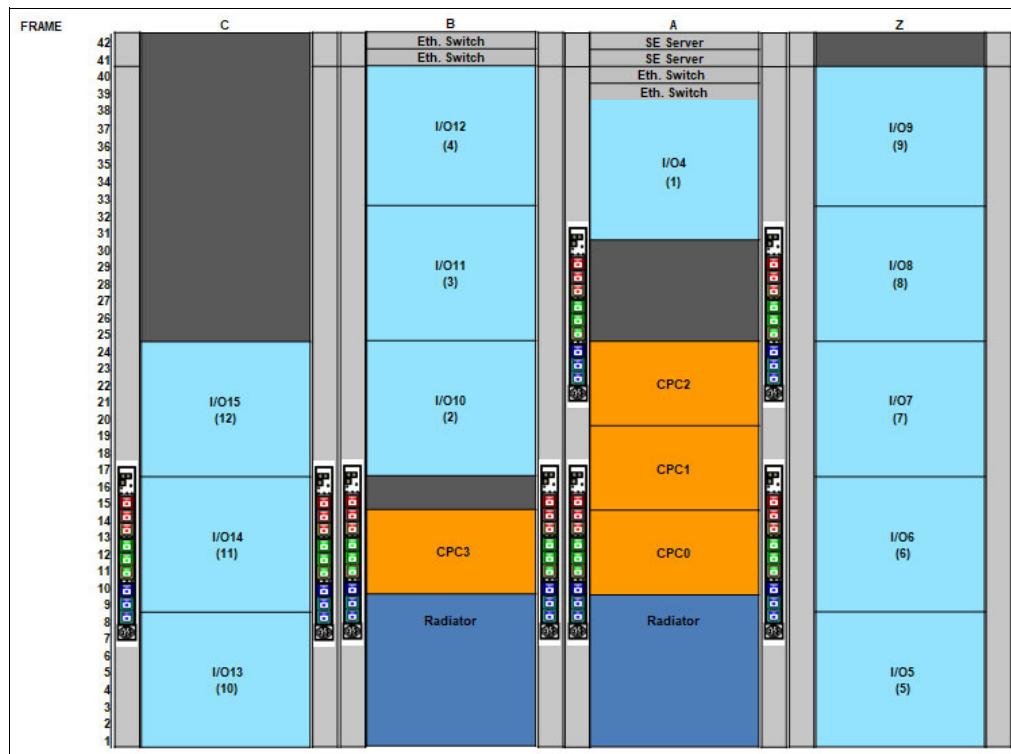


Figure E-6 Four frames ZABC, PDU



## E.2 New PCIe+ I/O drawers / PCHIDs location Options

There is a feature/option, FC 0352, to fill the Z Frame first. Z-First FC allows a customer to plan for a frame roll upgrade to a future IBM Z server installation.

The first I/O drawer will be located at Z01B in Z frame rather than A31B in A frame. And the Z frame will be filled with five I/O drawers before using A31B for the 6th I/O drawer. Then for the IBM Z server upgrade to a future server, the Z frame is reused as-is, except for the removal of unsupported cards and addition of new cards that are ordered.

For this feature, the configuration needs to have at least two frames (ZA).

Due to Z-First, PCHID numbers are now assigned to fixed I/O drawer locations rather than being associated with the logical plug sequence of I/O drawers. This allows a Z-First and non-Z-First config to have the same PCHIDs for the same I/O drawers.

Table E-1 Z-First and non-zFirst Feature Code PCHID Locations

PCHID Range	I/O Drawer Locations				
	9175 Max43 1 CPC w/o Z-FIRST	9175 Max43 1 CPC w/ Z-FIRST	9175 Max90 2 CPCs	9175 Max136 CPCs	9175 Max183 Max208 4 CPCs
100-13F	Z01B	Z01B	Z01B	Z01B	Z01B
140-17F	Z09B	Z09B	Z09B	Z09B	Z09B
180-1BF	Z17B	Z17B	Z17B	Z17B	Z17B
1C0-1FF	-	Z25B	Z25B	Z25B	Z25B
200-23F	-	z33B	z33B	z33B	z33B
240-27F	A31B	A31B	A31B	A31B	A31B
280-2BF	A23B	-	A23B	-	B33B
2C0-2FF	A15B	-	C01B	C01B	C01B
300-33F	-	-	C09B	C09B	C09B
340-37F	-	-	C17B	C17B	C17B
380-3BF	-	-	C25B	C25B	B25B
3C0-3FF	-	-	C33B	C33B	B17B







**F**

# Sustainability

This section discusses sustainability aspects. It comprises information on:

- ▶ Sustainability improvements
- ▶ Sustainability instrumentation

## F.1 Sustainability Improvements

Sustainability improvements are focused on

- ▶ Improved performance per kilowatt
- ▶ Reduced shipping impacts, floor space savings, and simplification

## F.2 Improved Performance Per Kilowatt

IBM Z has increased the total capacity per kilowatt by more than 125 times over the last 15 generations. z17 continues this trend.

The main contributors to increasing performance per unit of power are

- ▶ Improved I/O density with the on-chip DPU enabling eg 4-port FICON cards
- ▶ VCL or Dynamic Voltage Control reducing processor power consumption
- ▶ Moving from 7nm process to 5nm
- ▶ Adding the Spyre accelerator within the same overall power footprint as z16

VCL is a co-optimization between hardware and firmware to dynamically optimize and micro-tune the voltage applied to each processor chip for the current operating conditions. This reduces the net leakage current and power consumption of the machine. (On prior generations of hardware this voltage level was set at a fixed margin for all parts shipped to the field.)



The VCL algorithm monitors power and thermal sensors within each processor core to detect when margins are approached on the voltage level. This might result in rare instances of processor pipeline throttling and adjusts the voltage level to alleviate said conditions.

## F.3 Reduced Shipping Impacts, Floor Space Savings, And Simplification

Other sustainability improvements include

- ▶ Optimized & reduced packaging for frame doors and panels
- ▶ Enabling future I/O expansion frame carry forward
- ▶ Enabling future reusable frame door bases
- ▶ Pre-filled coolant and removal of coolant disposal

## F.4 Sustainability Instrumentation

z17 provides power consumption information to LPARs through Diagnose 324. z/OS, z/VM and Linux on Z provide software support for this information. Diagnose 324 enables an LPAR to see its own power consumption, as well as that of the machine.

This new support provides real time reporting in a way that

- ▶ Allows for trending
- ▶ Relates power consumption to workload variations

This instrumentation is ideal to enable Performance people to talk to such business people as their organisation's Chief Sustainability Officer.

For machines with multiple LPARs a comprehensive break down of power usage requires collecting data from each LPAR and aggregating. (Individual LPARs' consumption can be subtracted from the "partition power" consumption of the machine to provide power usage by those LPARs not enabled for data collection, such as Internal Coupling Facility LPARs.)

Similarly, new for z17 is the addition of System and Partition event monitors.

Events are triggered when power levels exceed customer defined thresholds over a specified period of time.

The system monitor supports total system power, total partition power, infrastructure power, and unassigned power.

The partition monitor supports partition power.

### F.4.1 z/OS

For z/OS RMF support adds fields to the SMF 70 Subtype 1 CPU Control Section in four categories:

- ▶ Total machine power



- ▶ Partition power
- ▶ Infrastructure power - for components (including infrastructure switches, SE/HMAs, and PDUs) that should not be accounted to individual partitions.
- ▶ Unassigned power for unused I/O adapters and components that are not assigned to any partition (including standby components).

Total machine power is, of course, the sum of the other three.

Machine power and partition power is further broken down into CPU, memory, and I/O.

It's important to enable SMF 70-1 on all z/OS LPARs on a machine - so each can report its own power consumption.

RMF support doesn't explicitly provide for workload-level reporting at the Service Class or Report Class level. You can pro-rate it from SMF Type 72 Subtype 3 records and the LPAR's consumption, as reported in SMF 70 Subtype 1.

If the record level (SMF70SRL in the RMF Product Section) is X'93' or higher the following fields are added to the CPU Control Section:

Offset (Dec)	Offset (Hex)	Name	Description
432	1B0	SMF70_CPUPower	Accumulated microwatts readings taken for all CPUs of the LPAR during the interval. Divide by SMF70_PowerReadCount to retrieve the average power measurement of the interval.
440	1B8	SMF70_StoragePower	Accumulated microwatts readings taken for storage of the LPAR during the interval. Divide by SMF70_PowerReadCount to retrieve the average power measurement of the interval.
448	1C0	SMF70_IOPower	Accumulated microwatts readings for I/O of the LPAR during the interval. Divide by SMF70_PowerReadCount to retrieve the average power measurement of the interval.
456	1C8	SMF70_CPCTotalPower	Accumulated microwatts readings for all electrical and mechanical components in the CPC. Divide by SMF70_PowerReadCount to retrieve the average power measurement of the interval.
464	1D0	SMF70_CPCUnassResPower	Accumulated microwatts readings for all types of resources in the standby or reserved state. Divide by SMF70_PowerReadCount to retrieve the average power measurement of the interval.
472	1D8	SMF70_CPCInfraPower	Accumulated microwatts readings for all subsystems in the CPC which do not provide CPU, storage, or I/O resources to logical partitions. These include service elements, cooling systems, power distribution, and network switches, among others. Divide by SMF70_PowerReadCount to retrieve the average power measurement of the interval.
480	1E0	SMF70_PowerReadCount	Number of power readings for the LPAR during the interval. (2 byte integer)



Offset (Dec)	Offset (Hex)	Name	Description
482	1E2		Reserved. (6 bytes)
488	1E8	SMF70_PowerPartitionName	The name of the LPAR to which the LPAR-specific power fields apply. (8 EBCDIC characters)

**Note:** All fields are 8-byte binary - except where noted.

## F.4.2 z/VM

The z/VM Performance Data Pump will include power usage for the z/VM host for z/VM 7.3 or 7.4. The z/VM Performance Data Pump Power Monitor Grafana Dashboard will include

- ▶ LPAR level CPU, memory, and I/O power information
- ▶ Guest level apportionment approximation details
- ▶ CPC level information details when global performance data has been enabled on at least one LPAR

## F.4.3 Linux on Z

Linux will obtain CEC and LPAR power consumption readings for the CEC and LPAR in which a Linux image is running, and expose this information in binary, raw (i.e., unmodified) format through a file in sysfs.

In addition, a tool will be provided through our s/390-tools package to display in human readable format.

Information to be provided includes LPAR level CPU, memory, and I/O power information.

Guest level apportionment will not be provided, but instructions will be provided regarding how to complete the calculation.

Information will also include CPC level details when global performance data has been enabled on the respective LPAR.

The code is planned to be included in the following Linux distributions

Ubuntu 25.04 (expected at General Availability)

SLES 15 SP7 (Expected for June 2025)

SLES 16 SP1 (Expected for October 2026)

RHEL 10.1 (Expected for November 2025)



# Related publications

The publications that are listed in this section are considered particularly suitable for a more detailed discussion of the topics that are covered in this book.

## IBM Redbooks

The following IBM Redbooks publications provide more information about the topic in this document. Note that some publications that are referenced in this list might be available in softcopy only:

- ▶ *IBM z16 Technical Introduction*, SG24-8950
- ▶ *IBM Z Connectivity Handbook*, SG24-5444
- ▶ *IBM z16 (3931) Configuration Setup*, SG24-8960

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

[ibm.com/redbooks](http://ibm.com/redbooks)

## Other publications

The publication *IBM Z 8561 Installation Manual for Physical Planning*, GC28-7002, also is relevant as another information source.

## Online resources

The following online resources are available:

- ▶ The IBM Resource Link for documentation and tools website:  
<http://www.ibm.com/servers/resourceLink>
- ▶ IBM Telum II Processor: The next-gen microprocessor for IBM Z and IBM LinuxONE:  
<https://www.ibm.com/blogs/systems/ibm-telum-processor-the-next-gen-microprocessor-for-ibm-z-and-ibm-linuxone/>
- ▶ Leveraging ONNX Models on IBM Z and LinuxONE:  
<https://community.ibm.com/community/user/ibmz-and-linuxone/blogs/andrew-sica/2021/10/29/leveraging-onnx-models-on-ibm-z-and-linuxone>
- ▶ Jump starting your experience with AI on IBM Z:  
<https://blog.share.org/Article/jump-starting-your-experience-with-ai-on-ibm-z>



## Help from IBM

IBM Support and downloads

[ibm.com/support](https://ibm.com/support)

IBM Global Services

[ibm.com/services](https://ibm.com/services)



# **IBM z17 (9175) Technical Guide**











SG24-8579-00

ISBN 0738460788