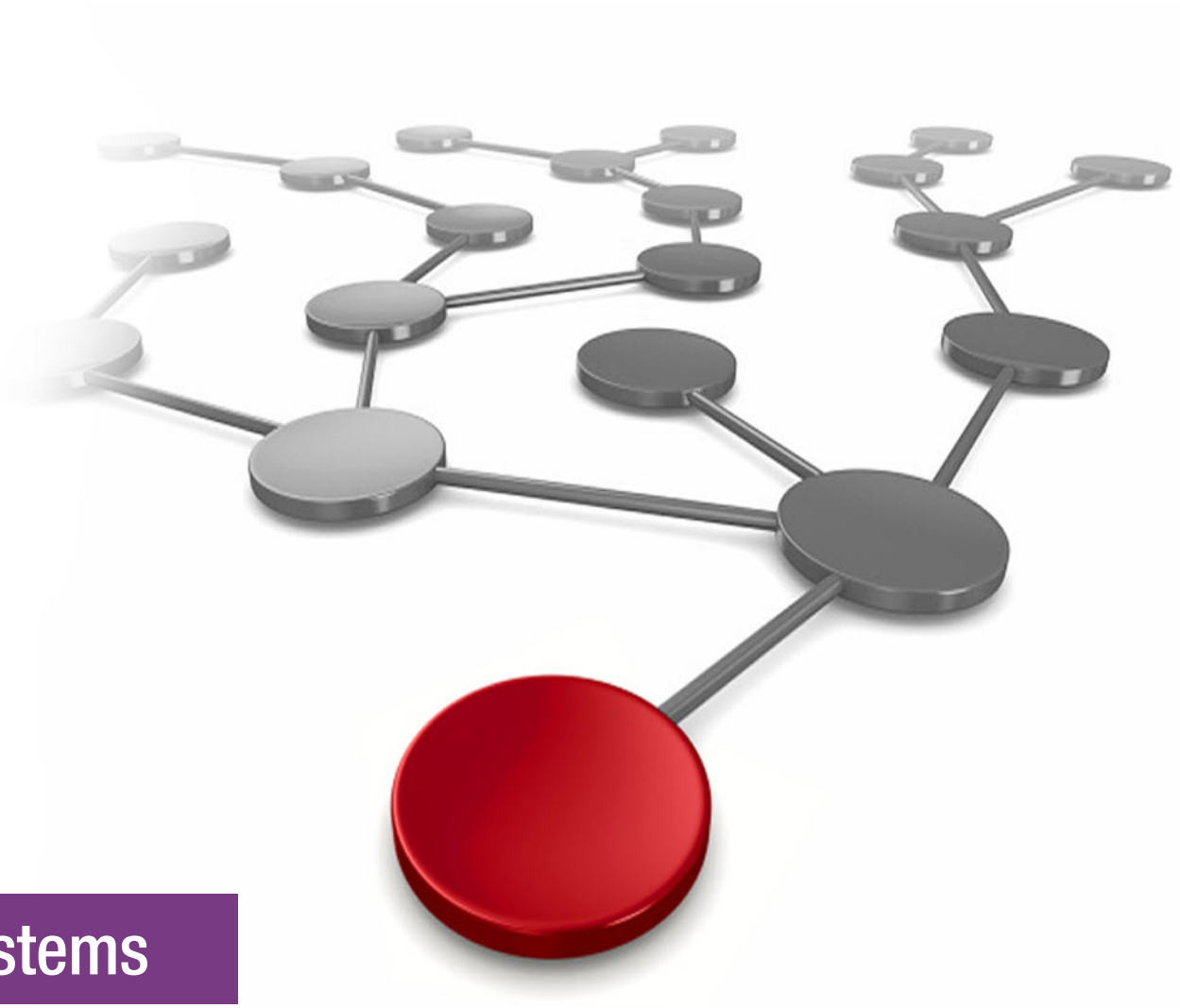
Draft Document for Review April 4, 2023 11:58 am SG24-5444-22

Redbooks
ibm.com/redbooks

IBM zSystems Connectivity Handbook

Octavian Lascu
Ewerson Palacio
Bill White



IBM zSystems

IBM

Redbooks

IBM Redbooks

IBM zSystems Connectivity Handbook

April 2023

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

Twenty-third Edition (April 2023)

This edition applies to connectivity options available on the IBM z16 A01,IBM z16 A02, IBM z16 AGZ, IBM z15 T01, IBM z15 T02, IBM z14 M0x, and IBM z14 Model ZR1.

Contents

Notices ix
Trademarks x

Preface xi
Authors xii
Now you can become a published author, too! xiii
Comments welcome xiii
Stay connected to IBM Redbooks xiii

Chapter 1. Introduction 1
1.1 I/O channel overview 2
1.1.1 I/O hardware infrastructure 2
1.1.2 I/O connectivity features 3
1.2 FICON Express 4
1.3 zHyperLink Express 5
1.4 Open Systems Adapter-Express 6
1.5 HiperSockets 7
1.6 Parallel Sysplex and coupling links 8
1.7 Shared Memory Communications 10
1.8 I/O feature support 11
1.9 Special-purpose feature support 14
1.9.1 Crypto Express features 14
1.9.2 Virtual Flash Memory 14
1.9.3 zEDC Express feature 15

Chapter 2. Channel subsystem overview 17
2.1 CSS description 18
2.1.1 CSS elements 18
2.1.2 Multiple CSSs 19
2.1.3 Multiple CSS structures 19
2.1.4 Physical channel ID 20
2.1.5 Adapter ID 22
2.1.6 Multiple CSS construct examples 24
2.1.7 Channel spanning 25
2.1.8 Multiple subchannel sets 27
2.1.9 Summary 28
2.2 I/O configuration management 29
2.2.1 Hardware Configuration Definition 29
2.2.2 CHPID Mapping Tool 29
2.3 I/O configuration planning 30
2.3.1 I/O configuration rules 30
2.4 References 32

Chapter 3. Fibre Channel connectivity 33
3.1 FICON Express description 34
3.1.1 FICON modes and topologies 34
3.1.2 FCP Channel 36
3.1.3 FCP and FICON mode characteristics 41
3.2 FICON elements 45

3.2.1 FICON channel . . . 45
3.2.2 High-Performance FICON for IBM zSystems . . . 48
3.2.3 Platform and name server registration in FICON channel . . . 50
3.2.4 Open exchanges . . . 51
3.2.5 Spanned channels . . . 59
3.2.6 Control unit port . . . 59
3.2.7 IBM z/OS Discovery and Automatic Configuration . . . 59
3.3 Connectivity . . . 60
3.3.1 FICON Express32S . . . 62
3.3.2 FICON Express16SA . . . 62
3.3.3 IBM Fibre Channel Endpoint Security . . . 63
3.3.4 FICON Express16S+ . . . 64
3.3.5 FICON Express16S . . . 64
3.3.6 FICON Express8S . . . 65
3.3.7 Qualified FICON and FCP products . . . 66
3.3.8 Software support . . . 66
3.3.9 Resource Measurement Facility . . . 66
3.4 References . . . 66

Chapter 4. IBM zHyperLink Express . . . 69
4.1 Description . . . 70
4.2 zHyperLink elements . . . 70
4.3 Connectivity . . . 71
4.4 References . . . 72

Chapter 5. IBM Open Systems Adapter Express . . . 73
5.1 Functional description . . . 74
5.1.1 Standard Ethernet support . . . 74
5.1.2 Operating modes . . . 74
5.1.3 Non-QDIO mode (CHPID OSE) . . . 76
5.1.4 QDIO mode (CHPID type OSD) . . . 76
5.1.5 OSA addressing support . . . 79
5.1.6 OSA/SF support . . . 80
5.2 OSA capabilities . . . 81
5.2.1 Virtual IP address . . . 81
5.2.2 Primary/secondary router function . . . 81
5.2.3 IPv6 support . . . 82
5.2.4 Large send for IP network traffic . . . 82
5.2.5 VLAN support . . . 83
5.2.6 SNMP support for z/OS and Linux on IBM Z . . . 85
5.2.7 IP network multicast and broadcast support . . . 86
5.2.8 Address Resolution Protocol cache management . . . 86
5.2.9 IP network availability . . . 87
5.2.10 Checksum offload support for z/OS and Linux on IBM Z . . . 87
5.2.11 Dynamic LAN idle for z/OS . . . 88
5.2.12 QDIO optimized latency mode . . . 88
5.2.13 Layer 2 support . . . 89
5.2.14 QDIO data connection isolation for z/VM . . . 90
5.2.15 QDIO interface isolation for z/OS . . . 92
5.2.16 Layer 3 VMAC for z/OS . . . 93
5.2.17 Enterprise Extender . . . 94
5.2.18 TN3270E server . . . 94
5.2.19 Adapter interruptions for QDIO . . . 95

5.2.20 Inbound workload queuing 95
5.2.21 Network management: Query and display OSA configuration 96
5.3 Connectivity 97
5.3.1 OSA-Express features 97
5.3.2 OSA function support 103
5.3.3 Software support 104
5.3.4 Resource Measurement Facility 105
5.4 Summary 105
5.5 References 106

Chapter 6. Console Communications (OSA-ICC) 107
6.1 Description of the OSA-ICC 108
6.2 Connectivity 110
6.3 Software support 111
6.3.1 TN3270E emulation 111
6.4 Summary 111
6.5 References 112

Chapter 7. Shared Memory Communications 113
7.1 SMC overview 114
7.1.1 Remote Direct Memory Access 115
7.1.2 Direct Memory Access 115
7.2 SMC over Remote Direct Memory Access 116
7.2.1 SMC-R connectivity 117
7.3 SMC over Direct Memory Access (intra-CPC) 120
7.4 Software support 121
7.4.1 SMC-R (Version1 and Version 2) 121
7.4.2 SMC-D (Version 1 and Version 2) 122
7.4.3 Shared Memory Communication Version 2 123
7.5 Reference material 125

Chapter 8. HiperSockets 127
8.1 Overview 128
8.1.1 HiperSockets benefits 128
8.1.2 Server integration with HiperSockets 129
8.1.3 HiperSockets function 130
8.1.4 Supported functions 132
8.2 Connectivity 137
8.3 Summary 140
8.4 References 140

Chapter 9. Coupling links and common time 141
9.1 IBM zSystems Parallel Sysplex 142
9.1.1 Coupling links and STP 143
9.1.2 Multi-site Parallel Sysplex considerations 144
9.2 Connectivity options 144
9.2.1 Coupling link options 144
9.2.2 Internal Coupling link 146
9.2.3 Integrated Coupling Adapter Short Range 147
9.2.4 Coupling Express Long Reach 147
9.2.5 InfiniBand coupling links (IBM z14 M0x only) 148
9.2.6 Dynamic I/O reconfiguration for stand-alone CF CPCs 148
9.3 Time functions 149
9.3.1 Server Time Protocol 149

9.3.2 Dynamic Split and Merge for Coordinated Timing Network 153
9.3.3 Operating system support 154
9.4 References 154

Chapter 10. Extended distance solutions 155
10.1 Unrepeated distances 156
10.2 Fibre Channel connection 158
10.2.1 FICON unrepeated distance 158
10.2.2 FICON repeated distance solutions 159
10.3 Coupling links 160
10.4 Wavelength-division multiplexing 161
10.4.1 GDPS qualification 162
10.4.2 IBM zSystems qualified WDM vendor products 163
10.5 References 163

Appendix A. Cryptographic solutions 165
Overview 166
Crypto Express8S features (1 HSM and 2 HSM) 167
Crypto Express7S (1 port or 2 port) 168
Crypto Express6S 169
References 170

Appendix B. Channel conversion options 171
Conversion solutions 172

Appendix C. Channel feature attributes 175
Cable types and attributes 176

Appendix D. Fiber optic cables 183
Description 184
Connector types for fiber cables 185
Mode-conditioning patch cables 185
zHyperLink Express and ICA SR cables 187
Conversion kits 188
References 189

Abbreviations and acronyms 191

Related publications 195
IBM Redbooks 195
Other publications 195
Online resources 196
Help from IBM 196

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 client 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.

AIX®
CICS®
Cognos®
Db2®
DS8000®
FICON®
GDPS®
Global Technology Services®
Guardium®
IBM®
IBM Security™
IBM Z®
IBM z13®
IBM z13s®
IBM z14®
Parallel Sysplex®
RACF®
Redbooks®
Redbooks (logo) ®
Resource Link®
S/390®
System z®
Tivoli®
VIA®
VTAM®
WebSphere®
z/Architecture®
z/OS®
z/VM®
z/VSE®
z13®
z13s®
z15™
zEnterprise®

The following terms are trademarks of other companies:

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.

OpenShift, 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.

VMware, and the VMware logo are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

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

Preface

This IBM® Redbooks® publication describes the connectivity options that are available for use within and beyond the data center for the IBM zSystems® family of mainframes, which includes:

- IBM z16 A01
- IBM z16 A02
- IBM z16 AGZ[1]
- IBM z15 T01
- IBM z15 T02
- IBM z14 M01 - M05
- IBM z14 ZR1

This book highlights the hardware and software components, functions, typical uses, coexistence, and relative merits of these connectivity features. It helps readers understand the connectivity alternatives that are available when planning and designing their data center infrastructures.

The changes to this edition are based on the IBM zSystems hardware announcement dated April 04, 2023.

This book is intended for data center planners, IT professionals, systems engineers, and network planners who are involved in the planning of connectivity solutions for IBM mainframes.

[1] The IBM z16 AGZ indicates a rack mount configuration that allows the core compute, I/O and networking features to be installed into and powered by a client-designated rack with power distribution units (PDUs), respectively. The rack mount configuration options are under a combined AGZ warranty umbrella.

Authors

This book was produced by a team of specialists from around the world working with IBM Redbooks, Poughkeepsie Center.

Octavian Lascu is an IBM Redbooks Project Leader and a Senior IT Consultant with over 25 years of experience. He specializes in designing, implementing, and supporting complex IT infrastructure environments (systems, storage, and networking), including high availability and disaster recovery (HADR) solutions and high-performance computing deployments. He has developed materials for and taught over 50 workshops for technical audiences around the world. He is the author of several IBM publications.

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 zSystems Redbooks publications, and created and presented ITSO seminars around the globe.

Bill White is an IBM Redbooks Project Leader and Senior Infrastructure Specialist at IBM Redbooks, Poughkeepsie Center.

A special thanks to the authors of the previous edition of this IBM Redbooks publication:

Markus Ertl, Jannie Houlbjerg, Hervey Kamga, Gerard Laumay, Slav Martinski, Kazuhiro Nakajima, Martijn Raave, Paul Schouten, Anna Shugol, André Spahni, John Troy, Roman Vogt, and Bo Xu

A thank you to following people for their contributions to this project:

Robert Haimowitz
IBM Redbooks, Poughkeepsie Center

Dave Surman, Darelle Gent, David Hutton, Eysha Powers, Patty Driever, Brian Valentine, Purvi Patel, Les Geer III, Bill Bitner, Christine Smith, Barbara Weiler, Dean St Piere, Tom Morris, Marna Walle, Franco Pinto, Walter Niklaus, Martin Recktenwald, Ron Geiger, Lisa Schloemer, Richard Gagnon, Susan Farrell, Les Geer III, Yamil Rivera.
IBM

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

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

- Send your comments in an email to:

  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

1

Introduction

This chapter gives a brief overview of the input/output (I/O) channel architecture and introduces the connectivity options that are available on IBM zSystems platforms.

This chapter includes the following topics:

- 1.1, “I/O channel overview” on page 2
- 1.2, “FICON Express” on page 4
- 1.3, “zHyperLink Express” on page 5
- 1.4, “Open Systems Adapter-Express” on page 6
- 1.5, “HiperSockets” on page 7
- 1.6, “Parallel Sysplex and coupling links” on page 8
- 1.7, “Shared Memory Communications” on page 10
- 1.8, “I/O feature support” on page 11
- 1.9, “Special-purpose feature support” on page 14

Note: The *link data rates* that are described throughout this book do not represent the actual performance of the links. The actual performance depends on many factors, which include latency through the adapters and switches, cable lengths, and the type of workload that uses the connection.

1.1 I/O channel overview

I/O channels are components of the IBM zSystems architecture (IBM z/Architecture®). They provide a pipeline through which data is exchanged between systems or between a system and external devices (in storage or on the network). z/Architecture channel connections, referred to as *channel paths,* have been a standard attribute of all IBM zSystems platforms back to the IBM S/360. Over the years, numerous extensions have been made to the z/Architecture to improve I/O throughput, reliability, availability, and scalability.

One of the many key strengths of the IBM zSystems platform is its ability to deal with large volumes of simultaneous I/O operations. The channel subsystem (CSS) provides the function for IBM zSystems platforms to communicate with external I/O and network devices and manage the flow of data between those external devices and system memory. This goal is achieved by using a system assist processor (SAP) that connects the CSS to the external devices.

The SAP uses the I/O configuration definitions that are loaded in the hardware system area (HSA) of the system to identify the external devices and the protocol they support. The SAP also monitors the queue of I/O operations that are passed to the CSS by the operating system.

Using an SAP, the processing units (PUs) are relieved of the task of communicating directly with the devices, so data processing can proceed concurrently with I/O processing.

Increased system performance demands higher I/O and network bandwidth, speed, and flexibility, so the CSS evolved with the advances in scalability of the IBM zSystems platforms. z/Architecture provides functions for scalability in the form of multiple CSSs that can be configured within the same IBM zSystems platform, for example:

- IBM z16 A01, IBM z15 T01, and IBM z14 M0x support up to six CSSs
- IBM z16 A02, IBM z16 AGZ[1], IBM z15 T02, and IBM z14 ZR1 support up to three CSSs

All of these IBM zSystems platforms deliver a significant increase in I/O throughput. For more information, see Chapter 2, “Channel subsystem overview” on page 17.

1.1.1 I/O hardware infrastructure

The I/O infrastructure on IBM zSystems adopted the Peripheral Component Interconnect Express (PCIe) standard. The latest IBM zSystems I/O hardware connectivity is implemented with PCIe Generation 3 standards. The I/O features are housed in PCIe I/O drawers or PCIe+ I/O drawers.

[1] The IBM z16 AGZ indicates a rack mount configuration that allows the core compute, I/O and networking features to be installed into and powered by a client-designated rack with power distribution units (PDUs), respectively. The rack mount configuration options are under a combined AGZ warranty umbrella.

PCIe+ I/O drawers and PCIe I/O drawers provide more I/O granularity and capacity flexibility than older IBM zSystems I/O infrastructures. The PCIe+ I/O drawers and PCIe I/O drawers connect to the CPC drawer fanout cards. The interconnection speed is 16 GBps (PCIe Generation 3):

- PCIe+ I/O drawer (Feature Code (FC) 4023)

  With IBM z16, an updated PCIe+ I/O drawer is introduced. The PCIe+ I/O drawer is fitted in a 19-inch format that allows the installation of 16 PCIe features. The PCIe+ I/O drawer is attached to the PCIe+ Gen3 (dual-port) fanouts (installed in the CPC drawer) with an interconnection speed of 16 GBps per port. PCIe+ I/O drawers can be installed and repaired concurrently in the field.

  Older PCIe+ I/O drawers cannot be carried forward to IBM z16. For an MES upgrade, new PCIe+ I/O drawers are provided to hold carried-forward PCIe features.

- PCIe+ I/O drawer (FC 4021 and FC 4001)

  With IBM z14 Model ZR1, a PCIe+ I/O drawer (FC 4001) was introduced. IBM z15 follows with a feature code (FC 4021) in the same format. The PCIe+ I/O drawer is fitted in a 19-inch format that allows the installation of 16 PCIe features (features that can be installed in the PCIe I/O drawers as well). The PCIe+ I/O drawer is attached to the PCIe+ Gen3 (dual-port for IBM z15) or PCIe Gen3 (single-port for IBM z14 ZR1) fanouts (installed in the CPC drawer) with an interconnection speed of 16 GBps per port. PCIe+ I/O drawers can be installed and repaired concurrently in the field.

- PCIe I/O drawer (FC 4013 and FC 4032)

  PCIe I/O drawers allow a higher number of features (four times more than I/O drawers in older IBM zSystem platforms) and increased port granularity. Each PCIe I/O drawer can accommodate up to 32 PCIe features in any combination. They are organized in four hardware domains per drawer, with eight features per domain.

  The PCIe I/O drawer is attached to a PCIe fanout in the CPC drawer, with an interconnection speed of 8 GBps with PCIe Gen2 and 16 GBps with PCIe Gen3. PCIe I/O drawers can be installed and repaired concurrently in the field.

1.1.2 I/O connectivity features

The most common attachment to a IBM zSystems I/O channel is a storage control unit (CU), which can be accessed through a Fibre Channel connection (IBM FICON) channel, for example. The CU controls I/O devices such as disk and tape drives. SCSI over FCP for disks and tape libraries is also supported.

System-to-system communications are typically implemented by using the IBM Integrated Coupling Adapter (ICA SR), Coupling Express Long Reach (CE LR), InfiniBand coupling links[2] (IBM z14 M0x only), Shared Memory Communications, and FICON channel-to-channel (FCTC) connections.

The Internal Coupling (IC) channel, IBM HiperSockets, and Shared Memory Communications can be used for communications between logical partitions (LPARs) within the IBM zSystems platform.

The Open Systems Adapter (OSA) features provide direct, industry-standard Ethernet connectivity and communications.

[2] IBM z16, IBM z15, and IBM z14 ZR1 do not support InfiniBand coupling links. Careful planning is required if IBM z14 M0x using Infiniband coupling/timing links is part of a Sysplex/CTN configuration.

The 25GbE RoCE Express3 and 2.x and 10GbE RoCE Express3 and 2.x features provide a high-speed, low-latency networking fabric for IBM Remote Direct Memory Access (RDMA) communications (IBM z/OS to z/OS, z/OS to Linux on IBM Z, and Linux on IBM Z to Linux on IBM Z).

As part of system planning activities, decisions are made about where to locate the equipment (for distance reasons), how it will be operated and managed, and the business continuity requirements for disaster recovery (DR), tape vaulting, and so on. The types of software (operating systems and applications) that will be used must support the features and devices on the IBM zSystems platform.

From a hardware point of view, all the features in the PCIe+ I/O drawers are managed by the IBM zSystems Support Elements. This function applies to installing and updating Licensed Internal Code (LIC) to features and other operational tasks.

Many features have an integrated processor that handles the adaptation layer functions that are required to present the necessary features to the rest of the system in a uniform manner. Therefore, all the operating systems have the same interface with the I/O subsystem.

The IBM zSystems platform supports industry-standard PCIe adapters that are called *native PCIe adapters*. For native PCIe adapter features, there is no adaptation layer, but the device driver is present in the operating system. The adapter management functions (such as diagnostics and firmware updates) are provided by Resource Groups.

Four Resource Groups are available on IBM z16, IBM z15, and IBM z14. The Resource Groups are managed by an integrated firmware processor (IFP) that is part of the system's base configuration.

The following sections briefly describe connectivity options for the I/O features that are available on the IBM zSystems platforms.

1.2 FICON Express

The Fibre Channel connection (FICON) Express features were originally designed to provide access to extended count key data (IBM ECKD) devices, and FICON channel-to-channel (CTC) connectivity. Then, came support for access to Small Computer System Interface (SCSI) devices (Fibre Channel Protocol (FCP)). This was followed by support for High-Performance FICON for IBM zSystems (zHPF) for OLTP I/O workloads that transfer small blocks of fixed-size data. These OLTP I/O workloads include IBM Db2® database, Virtual Storage Access Method (VSAM), partitioned data set extended (PDSE), and IBM z/OS File System (zFS).

IBM zSystems platforms build on this I/O architecture by offering high-speed FICON connectivity, as shown in Figure 1-1 on page 5.

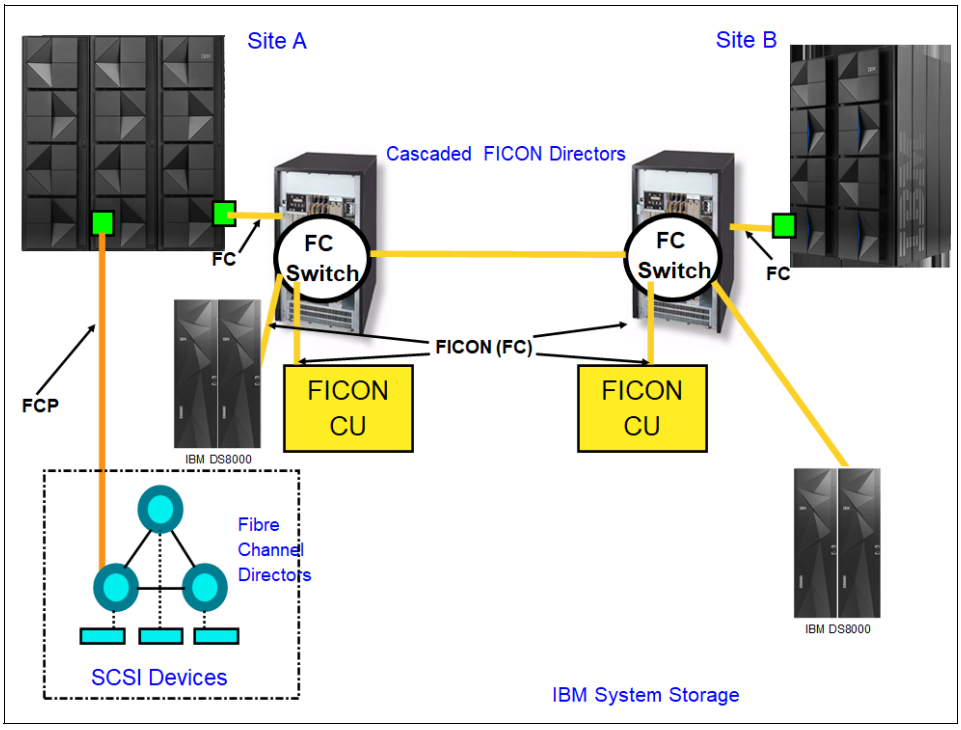


Figure 1-1 FICON connectivity

The FICON implementation enables full-duplex data transfer. In addition, multiple concurrent I/O operations can occur on a single FICON channel. FICON link distances can be extended by using various solutions. For more information, see 10.2.2, “FICON repeated distance solutions” on page 159.

The FICON features on IBM zSystems also support full fabric connectivity for the attachment of SCSI devices by using the Fibre Channel Protocol (FCP). Software support is provided by IBM z/VM, IBM z/VSE® (SCSI disk devices), Linux on IBM Z, and the KVM hypervisor.

IBM Fibre Channel Endpoint Security

An optional feature (feature on demand) is available for protecting the data in flight. Supported combinations of IBM z16 and FICON Express32S, IBM z15 T01[3] with FICON Express16SA, DS8000 storage, and IBM Security™ Guardium® Key Lifecycle Management provides Fibre Channel Endpoint Authentication and Encryption of Data-in-Flight between IBM zSystems and select DS8000 storage. IBM z16 platform continues to support FC 1146 with the new FICON Express32S adapters. For more information, see this IBM announcement.

1.3 zHyperLink Express

IBM zHyperLink is a technology that provides a low-latency point-to-point connection from IBM z16, IBM z15, and IBM z14 to an IBM DS8880 and newer. The transport protocol is defined for reading and writing ECKD data records. It provides a 5-fold reduction in I/O services time for *I/O* requests.

The zHyperLink Express feature is a PCIe adapter and can be shared by multiple LPARs.

[3] Not supported with the IBM z15 T02.

The zHyperLink Express feature is installed in the PCIe I/O drawer. On the IBM DS8000 side, the fiber optic cable connects to a zHyperLink PCIe interface in I/O bay.

The zHyperLink Express feature has the same qualities of service, as do all IBM zSystems I/O channel features.

Figure 1-2 shows a point-to-point zHyperLink connection with an IBM z16.

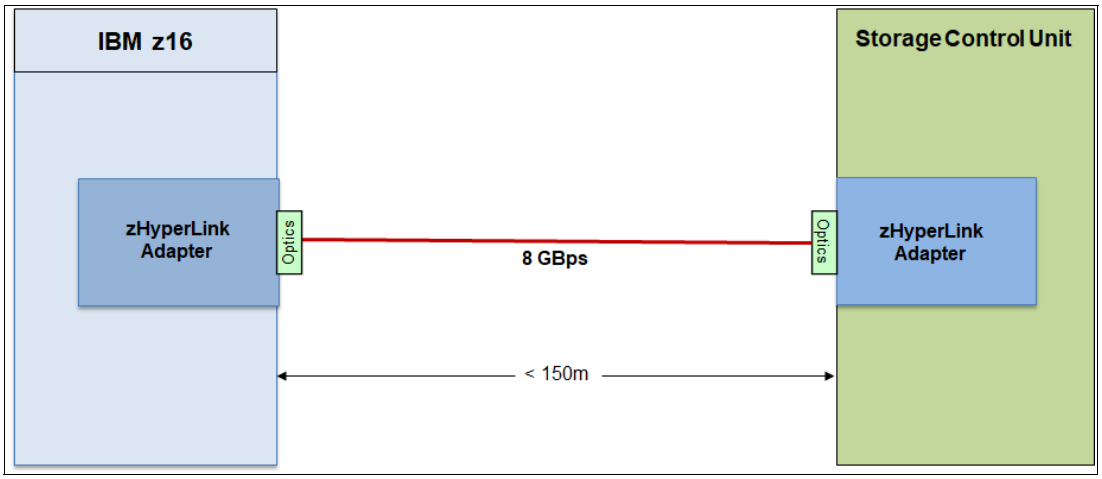


Figure 1-2 zHyperLink physical connectivity

1.4 Open Systems Adapter-Express

The Open Systems Adapter-Express (OSA-Express) features are the pipeline through which data is exchanged between the IBM zSystems platforms and devices in the network. OSA-Express6S and OSA-Express7S features provide direct, industry-standard LAN connectivity in a networking infrastructure, as shown in Figure 1-3.

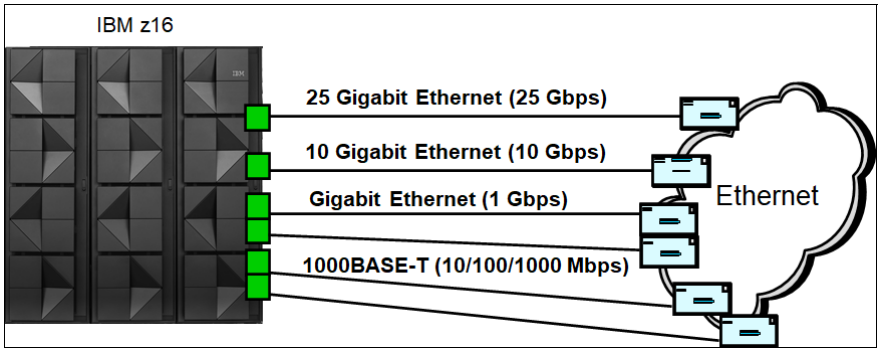


Figure 1-3 OSA-Express connectivity for IBM zSystems platforms

The OSA-Express features bring the strengths of the IBM zSystems family, such as security, availability, and enterprise-wide access to data to the LAN environment. OSA-Express provides connectivity for the following LAN types:

- 1000BASE-T Ethernet (100/1000 Mbps[4])
- 1 Gbps Ethernet SR and LR

[4] OSA Express7s 1000BASE-T and OSA Express7S 1.2 1000BASE-T support only 1000 Mbps full duplex.

- 10 Gbps Ethernet SR and LR
- 25 Gbps Ethernet SR and LR (The LR option is new with IBM z16).

Removal of support for OSA-Express 1000BASE-T hardware adapters:[a] IBM z16 will be the last IBM zSystems system to support OSA-Express 1000BASE-T hardware adapters (#0426, #0446, and #0458). Definition of all valid OSA CHPID types will be allowed only on OSA-Express GbE adapters, and potentially higher bandwidth fiber Ethernet adapters, on future servers.

a. IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remain at our sole discretion.

1.5 HiperSockets

IBM HiperSockets technology provides seamless network connectivity to consolidate virtual servers in an advanced infrastructure intraserver network. HiperSockets creates multiple independent and integrated virtual local area networks (VLANs) within an IBM zSystems platform.

This technology provides high-speed connectivity between combinations of LPARs or virtual servers. It eliminates the need for any physical cabling or external networking connection between these virtual servers. This *network within the box* concept minimizes network latency and maximizes bandwidth capabilities between z/VM, Linux on IBM Z and the KVM hypervisor, IBM z/VSE, and z/OS images, or combinations of them. HiperSockets use is also possible under the IBM z/VM operating system, which enables establishing internal networks between guest operating systems, such as multiple Linux servers.

The z/VM virtual switch can 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 directly communicate with other guest virtual machines on the virtual switch and with external network hosts through the virtual switch OSA UPLINK port.

Figure 1-4 shows an example of HiperSockets connectivity with multiple LPARs (LPARs) and virtual servers.

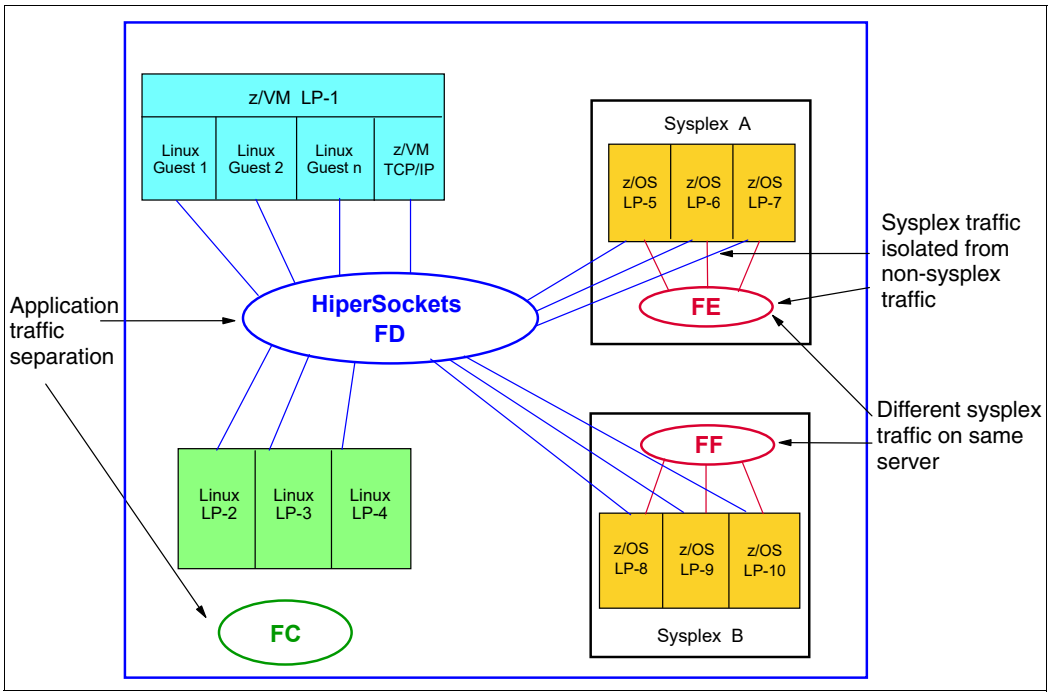


Figure 1-4 HiperSockets connectivity: multiple LPARs and virtual servers (FC, FD, FE, FF are CHPIDs)

HiperSockets technology is implemented by IBM zSystems Licensed Internal Code (LIC), with the communication path in system memory and the transfer information between the virtual servers at memory speed.

1.6 Parallel Sysplex and coupling links

IBM Parallel Sysplex is a clustering technology that represents a synergy between hardware and software. It consists of the following components:

- Parallel Sysplex-capable servers
- A coupling facility
- Coupling links (CS5, CL5, IC, InfiniBand[5])
- Server Time Protocol (STP)
- A shared direct access storage device (DASD)
- Software, both system and subsystem

[5] IBM z14 M0x is the last server that supports InfiniBand coupling links. Coupling links to other IBM z15 and IBM z16 can use only CS5 and CL5 CHPID types.

These components are all designed for parallel processing, as shown in Figure 1-5[6].

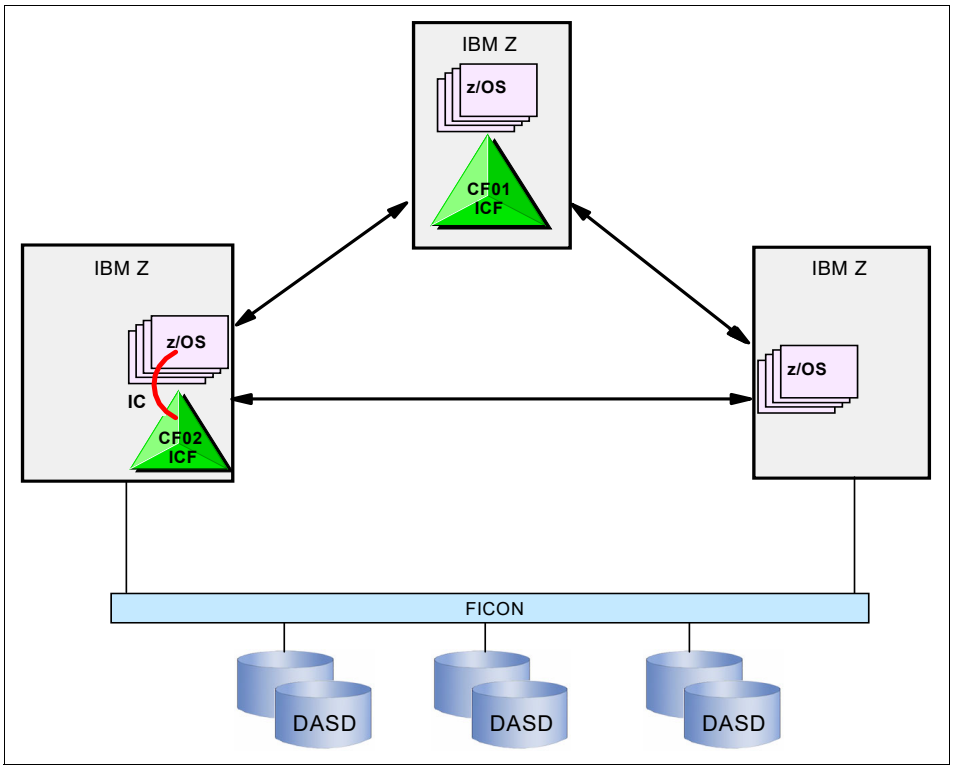


Figure 1-5 Sample configuration for Parallel Sysplex connectivity

Parallel Sysplex cluster technology is a highly advanced, clustered, commercial processing system. It supports high-performance, multisystem, read/write data sharing, which enables the aggregate capacity of multiple z/OS systems to be applied against common workloads.

The systems in a Parallel Sysplex configuration are linked and can fully share devices and run the same applications. This feature enables you to harness the power of multiple IBM zSystems platforms as though they are a single logical computing system.

The architecture is centered around the implementation of a coupling facility (CF) that runs the Coupling Facility Control Code (CFCC) and high-speed coupling connections for intersystem and intrasystem communications. The CF provides high-speed data sharing with data integrity across multiple IBM zSystems platforms.

Parallel Sysplex technology provides high availability for business-critical applications. The design is further enhanced with the introduction of System-Managed Coupling Facility Structure Duplexing, which provides the following extra benefits:

- Availability: Structures do not need to be rebuilt if a coupling facility fails.
- Manageability and usability: A consistent procedure is established to manage structure recovery across users.
- Configuration benefits: A sysplex can be configured with internal CFs *only*.

[6] There are many configurations possible for Parallel Sysplex depending on the available hardware and clustering requirements.

Attention: Parallel Sysplex technology and an STP CTN network support connectivity between systems that differ by up to two generations (n-2). For example, an IBM z16 can participate in Parallel Sysplex cluster with IBM z15, and IBM z14.

The IBM z16 does *not* support InfiniBand. You can set up connectivity by using only PCIe-based coupling, such as the Integrated Coupling Adapter Short Reach (ICA SR) and Coupling Express Long Reach (CE LR) features.

1.7 Shared Memory Communications

Shared Memory Communications (SMC) on IBM zSystems platforms is a technology that can improve throughput by accessing data faster with less latency. SMC reduces CPU resource consumption compared to traditional TCP/IP communications. Furthermore, applications do not need to be modified to gain the performance benefits of SMC.

SMC allows two peers to send and receive data by using system memory buffers that each peer allocates for its partner's use. Two types of SMC protocols are available on the IBM zSystems platform:

- SMC-Remote Direct Memory Access (SMC-R) Version 1 (SMC-Rv1) and Version 2 (SMC-Rv2).

  SMC-R is a protocol for Remote Direct Memory Access (RDMA) communication between TCP socket endpoints in LPARs in different systems. SMC-R runs over networks that support RDMA over Converged Ethernet (RoCE). It allows existing TCP applications to benefit from RDMA without requiring modifications. SMC-R provides dynamic discovery of the RDMA capabilities of TCP peers and automatic setup of RDMA connections that those peers can use.

  The 25GbE RoCE Express3, 25GbE RoCE Express2.1, 25GbE RoCE Express2, 10GbE RoCE Express3, 10GbE RoCE Express2.1 and 10GbE RoCE Express2 features provide the RoCE support that is needed for LPAR-to-LPAR communication across IBM zSystems platforms.

  While SMC-Rv1 does not support routing (communication across multiple subnets), SMC-Rv2 updates to the SMC-R protocol allowing SMC-Rv2 enabled hosts to connect and communicate across multiple IP subnets.

- SMC-Direct Memory Access (SMC-D)

  SMC-D implements the same SMC protocol that is used with SMC-R to provide highly optimized intra-system communications. Where SMC-R uses RoCE for communicating between TCP socket endpoints in separate systems, SMC-D uses Internal Shared Memory (ISM) technology for communicating between TCP socket endpoints in the same IBM zSystems platform.

  ISM provides adapter virtualization (virtual functions (VFs)) to facilitate the intra-system communications. Hence, SMC-D does not require any additional physical hardware (no adapters, switches, fabric management, or PCIe infrastructure).

  Therefore, significant cost savings can be achieved when using the ISM for LPAR-to-LPAR communication within the same IBM zSystems platform.

Both SMC protocols use shared memory-architectural concepts, eliminating TCP/IP processing in the data path, yet preserving TCP/IP quality of service (QoS) for connection management purposes.

Figure 1-6 shows the connectivity for SMC-D and SMC-R configurations.

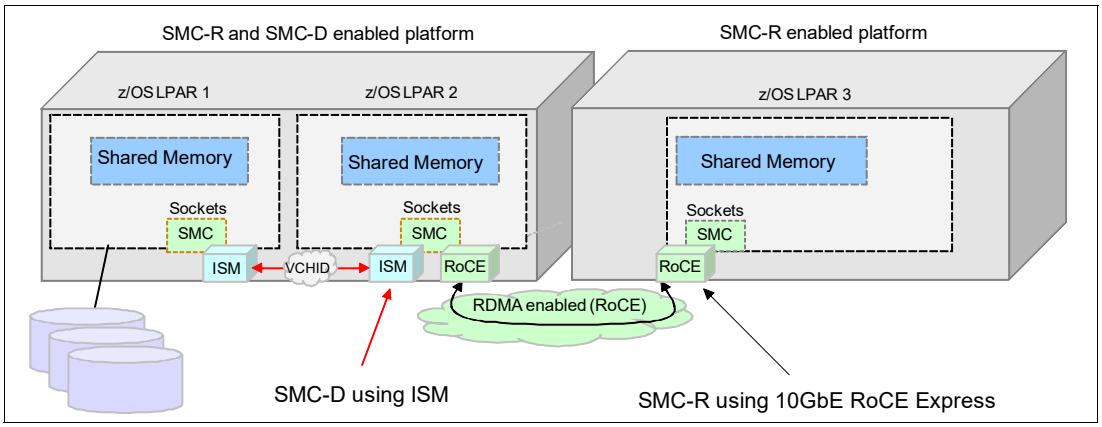


Figure 1-6 Connectivity for SMC-D and SMC-R configurations

1.8 I/O feature support

Table 1-1 lists the I/O features that are available on IBM zSystems platforms. Not all I/O features can be ordered on all systems, and certain features are available only for a system upgrade. Depending on the type and version of IBM Z, there might be further restrictions (for example, on the maximum number of supported ports).

Table 1-1 IBM zSystems I/O features

I/O feature	Feature code	Maximum number of ports						Ports per feature
		IBM z16 A01	IBM z16 A02/AGZ	IBM z15 T01	IBM z15 T02	IBM z14 M0x	IBM z14 ZR1	
zHyperLink Express		**Chapter 4, “IBM zHyperLink Express” on page 69**						
zHyperLink 1.1	0451	32	16	32	16	N/A	N/A	2
zHyperLink	0431	32	16	32	16	16	16	2
FICON Express		**Chapter 3, “Fibre Channel connectivity” on page 33**						
FICON Express32S LX	0461	384	96	N/A	N/A	N/A	N/A	2[a]
FICON Express32S SX	0462	384	96	N/A	N/A	N/A	N/A	2[a]
FICON Express16SA LX	0436	384	N/A	384	N/A	N/A	N/A	2[a]
FICON Express16SA SX	0437	384	N/A	384	N/A	N/A	N/A	2[a]
FICON Express16S+ LX	0427	320	96	320	128	320	128	2[a]
FICON Express16S+ SX	0428	320	96	320	128	320	128	2[a]
FICON Express16S LX	0418	N/A	N/A	320	128	320	128	2
FICON Espress16S SX	0419	N/A	N/A	320	128	320	128	2
FICON Express8S 10 KM LX	0409	N/A	N/A	128	128	320	128	2
FICON Express8S SX	0410	N/A	N/A	128	128	320	128	2
FICON Express8 10 KM LX	3325	N/A	N/A	N/A	N/A	N/A	64	4

I/O feature	Feature code	Maximum number of ports						Ports per feature
		IBM z16 A01	IBM z16 A02/AGZ	IBM z15 T01	IBM z15 T02	IBM z14 M0x	IBM z14 ZR1	
FICON Express8 SX	3326	N/A	N/A	N/A	N/A	N/A	64	4
OSA-Express		**Chapter 5, “IBM Open Systems Adapter Express” on page 73 Chapter 6, “Console Communications (OSA-ICC)” on page 107**						
OSA-Express7S 1.2 25GbE LR	0459	48	48	N/A	N/A	N/A	N/A	1
OSA-Express7S 1.2 25GbE SR	0460	48	48	N/A	N/A	N/A	N/A	1
OSA-Express7S 25GbE SR1.1	0449	48	N/A	48	N/A	N/A	N/A	1
OSA-Express7S 25GbE SR	0429	N/A	N/A	48	48	48	48	1
OSA-Express7S 1.2 10 GbE LR	0456	48	48	N/A	N/A	N/A	N/A	1
OSA-Express7S 1.2 10 GbE SR	0457	48	48	N/A	N/A	N/A	N/A	1
OSA-Express7S 10 GbE LR	0444	48	N/A	48	N/A	N/A	N/A	1
OSA-Express7S 10 GbE SR	0445	48	N/A	48	N/A	N/A	N/A	1
OSA-Express6S 10 GbE LR	0424	48	48	48	48	48	N/A	1
OSA-Express6S 10 GbE SR	0425	48	48	48	48	48	N/A	1
OSA-Express5S 10 GbE LR	0415	N/A	N/A	48	48	48	48	1
OSA-Express5S 10 GbE SR	0416	N/A	N/A	48	48	48	48	1
OSA-Express4S 10 GbE LR	0406	N/A	N/A	N/A	N/A	48	48	1
OSA-Express4S 10 GbE SR	0407	N/A	N/A	N/A	N/A	48	48	1
OSA-Express7S GbE 1.2 LX	0454	96	96	N/A	N/A	N/A	N/A	2[b]
OSA-Express7S GbE 1.2 SX	0455	96	96	N/A	N/A	N/A	N/A	2[b]
OSA-Express7S GbE LX	0442	96	N/A	96	N/A	N/A	N/A	2[b]
OSA-Express7S GbE SX	0443	96	N/A	96	N/A	N/A	N/A	2[b]
OSA-Express6S GbE LX	0422	96	96	96	96	96	96	2[b]
OSA-Express6S GbE SX	0423	96	96	96	96	96	96	2[b]
OSA-Express5S GbE LX	0413	N/A	N/A	96	96	96	96	2[b]
OSA-Express5S GbE SX	0414	N/A	N/A	96	96	96	96	2[b]
OSA-Express4S GbE LX	0404	N/A	N/A	N/A	N/A	96	96	2[b]
OSA-Express4S GbE SX	0405	N/A	N/A	N/A	N/A	96	96	2[b]
OSA Express7S 1.2 1000BASE-T	0458	96	96	N/A	N/A	N/A	N/A	2[b]
OSA-Express7S 1000BASE-T	0446	96	N/A	96	N/A	N/A	N/A	2[b]

I/O feature	Feature code	Maximum number of ports						Ports per feature
		IBM z16 A01	IBM z16 A02/AGZ	IBM z15 T01	IBM z15 T02	IBM z14 M0x	IBM z14 ZR1	
OSA-Express6S 1000BASE-T	0426	96	96	96	96	96	96	2[b]
OSA-Express5S 1000BASE-T	0417	N/A	N/A	N/A	96	96	96	2[b]
OSA-Express4S 1000BASE-T	0408	N/A	N/A	N/A	N/A	96	N/A	2[b]
RoCE Express		**Chapter 7, “Shared Memory Communications” on page 113**						
25GbE RoCE Express3 SR	0452	32	16	N/A	N/A	N/A	N/A	2[c]
25GbE RoCE Express3 LR	0453	32	16	N/A	N/A	N/A	N/A	2[c]
25GbE RoCE Express2.1	0450	32	16	32	16	N/A	N/A	2[c]
25GbE RoCE Express2	0430	32	16	16	16	8	4	2[c]
10GbE RoCE Express3 SR	0440	32	16	N/A	N/A	N/A	N/A	2
10GbE RoCE Express3 LR	0441	32	16	N/A	N/A	N/A	N/A	2
10GbE RoCE Express2.1	0432	32	16	32	16	N/A	N/A	2
10GbE RoCE Express2	0412	32	16	16	16	8	4	2
10GbE RoCE Express	0411	N/A	N/A	16	16	8	4	2
HiperSockets		**Chapter 8, “HiperSockets” on page 127**						
HiperSockets	N/A	32	32	32	32	32	32	N/A
Coupling links		**Chapter 9, “Coupling links and common time” on page 141**						
IC	N/A	64	64	64	64	32	32	N/A
Coupling Express2 LR	0434	64	64	N/A	N/A	N/A	N/A	N/A
Coupling Express LR	0433	N/A	N/A	64	64	64	32	2
ICA SR 1.1	0176	96	48	96	48	N/A	N/A	2
ICA SR	0172	96	48	96	48	80	16	2
HCA3-O (12x InfiniBand or 12x InfiniBand3)	0171	N/A	N/A	N/A	N/A	32	N/A	2
HCA3-O LR (1x InfiniBand)	0170	N/A	N/A	N/A	N/A	64	N/A	4

a. One feature, two ports, with one CHPID per port. *Both* ports must be the same CHPIT type (either Fibre Channel (FC) or FCP).

b. Both ports are on the same CHPID.

c. 25GbE RoCE Express2.x and Express3 supports 25GbE link only, and must not be used in the same SMC-R Link Group.

1.9 Special-purpose feature support

In addition to the I/O connectivity features, several special purpose features are available that can be installed in the PCIe+ I/O drawers such as:

- Crypto Express
- zEDC (IBM z14 only)

1.9.1 Crypto Express features

Integrated cryptographic features provide leading cryptographic performance and functions. The cryptographic features are designed for the highest level of security certifications that are available at the time of the design, and they provide reliability, availability, and serviceability support that is unmatched in the industry.

The new Crypto Express8S is the first adapter that supports the new Quantum Safe function, and it is also the first one that is designed for FIPS 140-3 Level 4.

Quantum-safe cryptography refers to efforts to identify algorithms that are resistant to attacks by both classical and quantum computers to keep information assets secure even after a large-scale quantum computer is built.

Crypto Express8S, Crypto Express7S, and Crypto Express6S are tamper-sensing and tamper-responding programmable cryptographic features that provide a secure cryptographic environment. Each adapter contains a tamper-resistant Hardware Security Module (HSM).

The HSM can be configured as a secure IBM Common Cryptographic Architecture (CCA) coprocessor, as a secure IBM Enterprise PKCS #11 (EP11) coprocessor, or as an accelerator.

1.9.2 Virtual Flash Memory

The Flash Express card was phased out and replaced with the IBM z14 M0X generation. The replacement for the card is Virtual Flash Memory (VFM), which is in the physical memory of the machine. The VFM is available on IBM z14 M0X, IBM z14 ZR1, IBM z15 T01, and T02, as IBM z16 A01.

Replacement of Flash Express

Starting with the IBM z14, Flash Express (FC 0402 and FC 0403) was replaced by Virtual Flash Memory (VFM). VFM implements an Extended Asynchronous Data Mover (EADM) architecture by using HSA-like memory instead of flash card pairs.

The VFM features are available in the following sizes:

- For IBM z16 A01, one VFM feature is 512 GB (FC 0644). Up to 12 features per system can be ordered.
- For IBM z16 A02 and IBM z16 AGZ, one VFM feature is 512 GB (FC 0644). Up to four (4) features per system can be ordered.
- For IBM z15 T01, one VFM feature is 512 GB (FC 0643). Up to 12 features per system can be ordered.
- For IBM z15 T02, one VFM feature is 512 GB (FC 0643). Up to four features per system can be ordered.

- For IBM z14 M0x, one VFM feature is 1.5 TB (FC 0604). Up to four features per system can be ordered.
- For IBM z14 ZR1, one VFM feature is 512 GB (FC 0614). Up to four features per system can be ordered.

1.9.3 zEDC Express feature

IBM zEnterprise® Data Compression is a hardware feature that is implemented on the IBM zSystems platform. Optional zEDC software is required for certain z/OS data set compression operations. IBM z14 was the last machine that supported the optional zEDC Express *PCIe features (cards)*. The zEDC Express PCIe feature cannot be carried forward to IBM z15 or IBM z16.

In IBM z15 and IBM z16, the compression capability is integrated into the processor chip. The Integrated Accelerator for zEDC provides hardware-based acceleration of data compression and decompression, which improves cross-platform data exchange, reduces processor use, and saves disk space. The Integrated Accelerator for zEDC implements compression as defined by RFC1951 (DEFLATE). For more information about the DEFLATE Compress Data Format Specification, see RFC for DEFLATE Compressed Data Format Specification Version 1.3.

On IBM z14, a zEDC Express PCIe feature can be shared by up to 15 LPARs. z/OS Version 2.2 and later supports the zEDC Express feature. z/VM 7.1 with PTFs and later provides guest exploitation. With IBM z15 and newer systems, the Integrated Accelerator for zEDC is integrated onto the processor chip, so there is no virtualization requirement.

Table 1-2 on page 15 lists the special-purpose features that are available on IBM zSystems platforms. Not all special-purpose features can be ordered on all systems, and certain features are available only with a system upgrade. All special-purpose features are installed in the PCIe I/O drawer or the PCIe+ I/O drawer.

Table 1-2 IBM zSystems special-purpose features

Special-purpose feature	Feature codes	Maximum number of features					
		IBM z16 A01	IBM z16 A02/AGZ	IBM z15 T01	IBM z15 T02	IBM z14 M0x	IBM z14 ZR1
Crypto Express		**Appendix A, “Cryptographic solutions” on page 165**					
Crypto Express 8S	0909 (1 HSM)[a]	30	20	N/A	N/A	N/A	N/A
	0908 (2 HSM)	16	16	N/A	N/A	N/A	N/A
Crypto Express7S	0899 (1 Port)[a]	30	20	30	20	N/A	N/A
	0898(2 Port)	16	16	16	16	N/A	N/A
Crypto Express6S	0893	16	16	16	16	16	16
zEDC Express							
zEDC Express	0420	N/A	N/A	N/A	N/A	8	8

a. The Crypto Express8S (2 HSM) and Crypto Express7S (2 Port) have two IBM PCIe Cryptographic Coprocessors (PCIeCCs). The PCIeCC is a Hardware Security Module. The other Crypto Express features have only one HSM. The maximum number of combined HSMs is 60 for IBM z16 A01 and IBM z15 Model T01, and 40 for IBM z16 A02. IBM z16 AGZ and IBM z15 Model T02.

2

Channel subsystem overview

This chapter describes the channel subsystem (CSS), which handles all the system input/output (I/O) operations for IBM zSystems platform. The role of the CSS is to control communication between internal and external channels, control units and devices.

This chapter includes the following topics:

- 2.1, “CSS description” on page 18
- 2.2, “I/O configuration management” on page 29
- 2.3, “I/O configuration planning” on page 30
- 2.4, “References” on page 32

2.1 CSS description

CSS enables communication from system memory to peripheral components by using channel connections. The channels in the CSS allow transfer of data between memory and I/O devices or other servers under the control of a channel program. The CSS allows channel I/O operations to continue independently of other operations in the system, which allows other functions to resume after an I/O operation is initiated. The CSS also provides internal channels for communication between logical partitions (LPARs) in a physical system.

2.1.1 CSS elements

CSS includes the elements that are described in this subsection, they include:

- Channel path
- Subchannels
- Channel path identifier
- Control units
- I/O devices

Channel path

A channel path is a single interface between a system and one or more control units. Commands and data are sent across a channel path to process I/O requests. A CSS can have up to 256 channel paths.

Subchannels

A subchannel provides the logical representation of a device to the program and contains the information that is required to sustain a single I/O operation. One subchannel is assigned for each device that is defined to the LPAR. Subchannel set availability per platform is shown in the following list:

- Four subchannel sets are available on IBM z16 A01, IBM z15 T01, and IBM z14 M0x.
- Three subchannel sets are available on IBM z16 A02, IBM z16 AGZ, IBM z15 T02, and IBM z14 ZR1.

Channel path identifier

The channel subsystem communicates with I/O devices through channel paths between the channel subsystem and control units. Each channel path is assigned a channel path identifier (CHPID) value that uniquely identifies that path.

With IBM zSystems platforms, a CHPID number is assigned to a physical location (slot or port) by the user through either the HCD or the Input/Output Configuration Program (IOCP).

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 it can be physically and logically integrated with the I/O device, the channel subsystem, or in the system itself.

I/O devices

An I/O device provides external storage, or 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 channel path.

2.1.2 Multiple CSSs

The IBM zSystems design offers considerable processing power, memory size, and I/O connectivity. The CSS concept has been scaled up to support the larger I/O capability. IBM zSystems implement the multiple CSS concept, which provides relief for the number of supported LPARs, channels, and devices that are available to the system. Up to six Logical Channel Subsystems (LCSSs), each with four subchannel sets (SS) and up to 256 channels, are supported, depending on the IBM zSystems machine type.

Table 2-1 lists the maximum number of CSSs and LPARs that are supported by IBM zSystems platforms. CSSs are numbered 0 - 5 or 0 - 2 (system dependent), with the numbers referred to as *CSS image IDs*.

Table 2-1 Maximum number of CSSs and LPARs that are supported by IBM zSystems

Systems	Maximum number of CSSs	Maximum number of LPARs
IBM z16 A01	6	85
IBM z16 A02	3	40
IBM z16 AGZ	3	40
IBM z15 T01	6	85
IBM z15 T02	3	40
IBM z14	6	85
IBM z14 ZR1	3	40

2.1.3 Multiple CSS structures

The structure of multiple CSSs provides channel connectivity to the defined LPARs in a manner that is apparent to subsystems and application programs. IBM zSystems platforms enable you to define more than 256 CHPIDs in the system through the multiple CSSs. As previously noted, the CSS defines CHPIDs, control units, subchannels, and so on. This feature enables you to define a balanced configuration for the processor and I/O capabilities.

For easier management, consider using the Hardware Configuration Definitions (HCD) tool to build and control the IBM zSystems input/output configuration definitions. HCD support for multiple channel subsystems is available with IBM z/VM and IBM z/OS. HCD provides the capability to make dynamic I/O configuration changes.

LPARs cannot be added until at least one CSS is defined. LPARs are defined for a CSS, not for a system. A LPAR is associated with only one CSS. CHPID numbers are unique within a CSS, but the same CHPID number can be reused within all CSSs.

All channel subsystems are defined within a single I/O configuration data set (IOCDS). The IOCDS is loaded into the hardware system area (HSA) and initialized during power-on reset.

On the IBM zSystems platform, the HSA has a fixed size, which is not included in the purchased memory. Table 2-2 lists the HSA sizes

Table 2-2 HSA size

	IBM z16 A01	IBM z16 A02/AGZ	IBM z15 T01	IBM z15 T02	IBM z14 M0x	IBM z14 ZR1
HSA size	256 GB	160 GB	256 GB	160 GB	192 GB	64 GB

2.1.4 Physical channel ID

A physical channel ID (PCHID) reflects the physical location of a channel-type interface. A PCHID number is based on the PCIe+ I/O drawer, PCIe I/O drawer, or I/O drawer location, the channel feature slot number, and the port number of the channel feature. A CHPID does not directly correspond to a hardware channel port, but is assigned to a PCHID in HCD or IOCP.

CHPIDs are not preassigned on IBM zSystems platform. You must assign the CHPID numbers by using the CHPID Mapping Tool or directly by using HCD/IOCP. Assigning CHPIDs means that the CHPID number is associated with a physical channel port location (PCHID) or the adapter ID (AID) and a CSS. The CHPID number range is from 00 to FF and must be unique in a CSS. Any CHPID that is not connected to a PCHID fails validation when an attempt is made to build a production IODF or an IOCDS.

Example 2-1 shows a portion of a sample PCHID REPORT for an IBM z16 A01 system.

Example 2-1 PCHID REPORT for an IBM z16 A01

```
CHPIDSTART
  31463036                      PCHID REPORT                        Nov 10,2021
  Machine: 3931-A01  SN1
  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  Source           Drwr  Slot  F/C    PCHID/Ports or AID              Comment
  A10/LG06         A10B  LG06  0176   AID=05
  A15/LG06         A15B  LG06  0176   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    0440   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  17    0458   130/D1D2
  A20/LG12/J02     Z01B  18    0457   134/D1
  A20/LG12/J02     Z01B  19    0457   138/D1
  A20/LG12/J02     Z01B  20    0458   13C/D1D2
  A10/LG12/J02     Z09B  02    0458   140/D1D2
  A10/LG12/J02     Z09B  03    0461   144/D1 145/D2
  A10/LG12/J02     Z09B  04    0457   148/D1
  A10/LG12/J02     Z09B  07    0451   150/D1D2
  A10/LG12/J02     Z09B  09    0452   158/D1D2             RG1
  A10/LG12/J02     Z09B  10    0457   15C/D1
  A20/LG12/J01     Z09B  12    0461   160/D1 161/D2
  A20/LG12/J01     Z09B  14    0462   168/D1 169/D2
  A20/LG12/J01     Z09B  17    0457   170/D1
  A20/LG12/J01     Z09B  18    0457   174/D1
  A20/LG12/J01     Z09B  19    0459   178/D1
```

```
A20/LG12/J01     Z09B  20    0908   17C/P00 17D/P01
A20/LG09/J02     Z17B  02    0462   180/D1 181/D2
A20/LG09/J02     Z17B  03    0461   184/D1 185/D2
A20/LG09/J02     Z17B  05    0457   18C/D1
A20/LG09/J02     Z17B  08    0457   194/D1
A20/LG09/J02     Z17B  09    0908   198/P00 199/P01
A20/LG09/J02     Z17B  10    0451   19C/D1D2
A15/LG12/J01     Z17B  12    0458   1A0/D1D2
A15/LG12/J01     Z17B  13    0461   1A4/D1 1A5/D2
A15/LG12/J01     Z17B  14    0457   1A8/D1
A15/LG12/J01     Z17B  17    0440   1B0/D1D2             RG2
A15/LG12/J01     Z17B  18    0457   1B4/D1
A15/LG12/J01     Z17B  19    0434   1B8/D1D2             RG2
A15/LG12/J01     Z17B  20    0459   1BC/D1
A10/LG12/J01     Z25B  02    0461   1C0/D1 1C1/D2
A10/LG12/J01     Z25B  03    0462   1C4/D1 1C5/D2
A10/LG12/J01     Z25B  05    0458   1CC/D1D2
A10/LG12/J01     Z25B  07    0457   1D0/D1
A10/LG12/J01     Z25B  08    0459   1D4/D1
A10/LG12/J01     Z25B  09    0457   1D8/D1
A10/LG12/J01     Z25B  10    0434   1DC/D1D2             RG3
A20/LG09/J01     Z25B  13    0461   1E4/D1 1E5/D2
A20/LG09/J01     Z25B  15    0458   1EC/D1D2
A20/LG09/J01     Z25B  17    0457   1F0/D1
A20/LG09/J01     Z25B  18    0457   1F4/D1
A20/LG09/J01     Z25B  19    0908   1F8/P00 1F9/P01
A20/LG09/J01     Z25B  20    0452   1FC/D1D2             RG4
Legend:
Source  Book Slot/Fanout Slot/Jack
A20B    CEC Drawer 3 in A frame
A15B    CEC Drawer 2 in A frame
A10B    CEC Drawer 1 in A frame
Z01B    PCIe Drawer 1 in IBM Z frame
Z09B    PCIe Drawer 2 in IBM Z frame
Z17B    PCIe Drawer 3 in IBM Z frame
Z25B    PCIe Drawer 4 in IBM Z frame
0461    FICON Exp 32S LX 2 Ports
0462    FICON Exp 32S SX 2 Ports
0458    OSA Express7S 1.2 1000BASE T 2 Ports
0457    OSA Express7S 1.2 10 GbE SR 1 Ports
0459    OSA Express7S 1.2 25 SR 1 Ports
RG3     Resource Group 3
0434    Coupling Express 10G LR
0908    Crypto Express8S 2 Ports
RG4     Resource Group 4
0452    25GbE RoCE Express3 SR
0451    zHyperLink Express1.1
RG2     Resource Group 2
0440    10GbE RoCE Express3 SR
RG1     Resource Group 1
0176    ICA SR1.1 2 Links
```

Refer to the System Overview publication for your processor or the IOCP Users Guide for recommendations on configuring control units and devices to best exploit the RAS and

performance of the processor. The CHPID Mapping tool, available from IBM Resource Link®, may also be used to aid in configuring channel to control units and devices.

You must assign CHPID numbers to your new channels by using IOCP or HCD. The PCHID and AID values in the above report will be needed to complete the definition. You might find the CHPID Mapping Tool, which is available from Resource Link, helpful in this effort.

You can get a PCHID report from the IBM account team when ordering.

2.1.5 Adapter ID

The adapter ID (AID) number assigns a CHPID to a port by using HCD/IOCP for IBM Parallel Sysplex cluster technology.

On the IBM zSystems platform, the AID is bound to the serial number of the fanout. If the fanout is moved, the AID moves with it. No IOCDS update is required if adapters are moved to a new physical location.

AIDs are included in the PCHID report that IBM provides for new build systems and upgrades. Example 2-2 shows an AID in a PCHID report.

Example 2-2 AID assignment in IBM z16 A01 PCHID report

```
CHPIDSTART
  23760121                      PCHID REPORT
  Machine: 3931 A01
  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Source Drwr  Slot    F/C PCHID /Ports or AID Comment
A15/LG12/J02 Z01B 10 0440 11C/D1D2
A10/LG12/J02 Z09B 09 0452 158/D1D2
A15/LG12/J01 Z17B 17 0440 1B0/D1D2
A15/LG12/J01 Z17B 19 0434 1B8/D1D2
A10/LG12/J01 Z25B 10 0434 1DC/D1D2
A20/LG09/J01 Z25B 20 0452 1FC/D1D2
.....< snippet >.................
```

Table 2-3 shows the adapter ID numbers for an IBM z14 M0x.

Table 2-3 Fanout AID numbers for IBM z14 M0x

Drawer[a] (DX)[b]	Location	Fanout slot	AIDs
First (D3)	A15A	LG03-LG12 (PCIe)	2E-37
		LG13-LG16 (InfiniBand)	0C-0F
Second (D2)	A19A	LG03-LG12 (PCIe)	24-2D
		LG13-LG16 (InfiniBand)	08-0B
Third (D1)	A23A	LG03-LG12 (PCIe)	1A-23
		LG13-LG16 (InfiniBand)	04-07

Drawer[a] (DX)[b]	Location	Fanout slot	AIDs
Fourth (D0)	A27A	LG03-LG12 (PCIe)	10-19
		LG13-LG16 (InfiniBand)	00-03

a. Indicates the IBM z14 physical CPC drawer installation order.
b. The designation between the parenthesis indicates the logical CPC drawer number.

Table 2-4 lists the adapter ID numbers for an IBM z14 ZR1.

Table 2-4 AID number assignment for IBM z14 ZR1

Fanout location	CPC drawer location	AIDs
LG01 - LG04	A09B	10 - 13
LG07 - LG10		14 - 17

Table 2-5 lists the adapter ID numbers for an IBM z15 T01.

Table 2-5 AID number assignment for IBM z15 T01

Fanout locations	CPC0 drawer location A10 AID (Hex)	CPC1 drawer location A15 AID (Hex)	CPC2 drawer location A20 AID (Hex)	CPC3 drawer location B10 AID (Hex)	CPC4 drawer location B15 AID (Hex)
LG01	00	0C	18	24	30
LG02	01	0D	19	25	31
LG03	02	0E	1A	26	32
LG04	03	0F	1B	27	33
LG05	04	10	1C	28	34
LG06	05	11	1D	29	35
LG07	06	12	1E	2A	36
LG08	07	13	1F	2B	37
LG09	08	14	20	2C	38
LG10	09	15	21	2D	39
LG11	0A	16	22	2E	3A
LG12	0B	17	23	2F	3B

Table 2-6 lists the adapter ID numbers for an IBM z15 T02.

Table 2-6 AID number assignment for IBM z15 T02

Fanout location	CPC drawer location (number)	AID (Hex)
LG01 - LG12	A10B (CPC0)	00 - 0B
LG01 - LG12	A15B (CPC1)	0C - 17

Table 2-7 lists the adapter ID numbers for an IBM z16 A01.

Table 2-7 AID number assignment for IBM z16 A01

Fanout locations	CPC0 drawer location A10 AID (Hex)	CPC1 drawer location A15 AID (Hex)	CPC2 drawer location A20 AID (Hex)	CPC3 drawer 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

Table 2-8 lists the adapter ID numbers for an IBM z16 A02.

Table 2-8 AID number assignment for IBM z16 A02

Fanout location	CPC drawer location (number)[a]	AID (Hex)
LG01 - LG12	A10B (CPC0)	00 - 0B
LG01 - LG12	A15B (CPC1)	0C - 17

a. For IBM z16 AGZ, location codes are determined by the symbolic name of the component (ACP0, ACP1, AIO1, AIO2, AIO3).

2.1.6 Multiple CSS construct examples

In this example, each CSS has three LPARs with their associated MIF image identifiers. In each CSS, the CHPIDs are shared across all LPARs. The CHPIDs are assigned to their designated PCHIDs.

Figure 2-1 on page 25 shows two CSSs defined as CSS0 and CSS1.

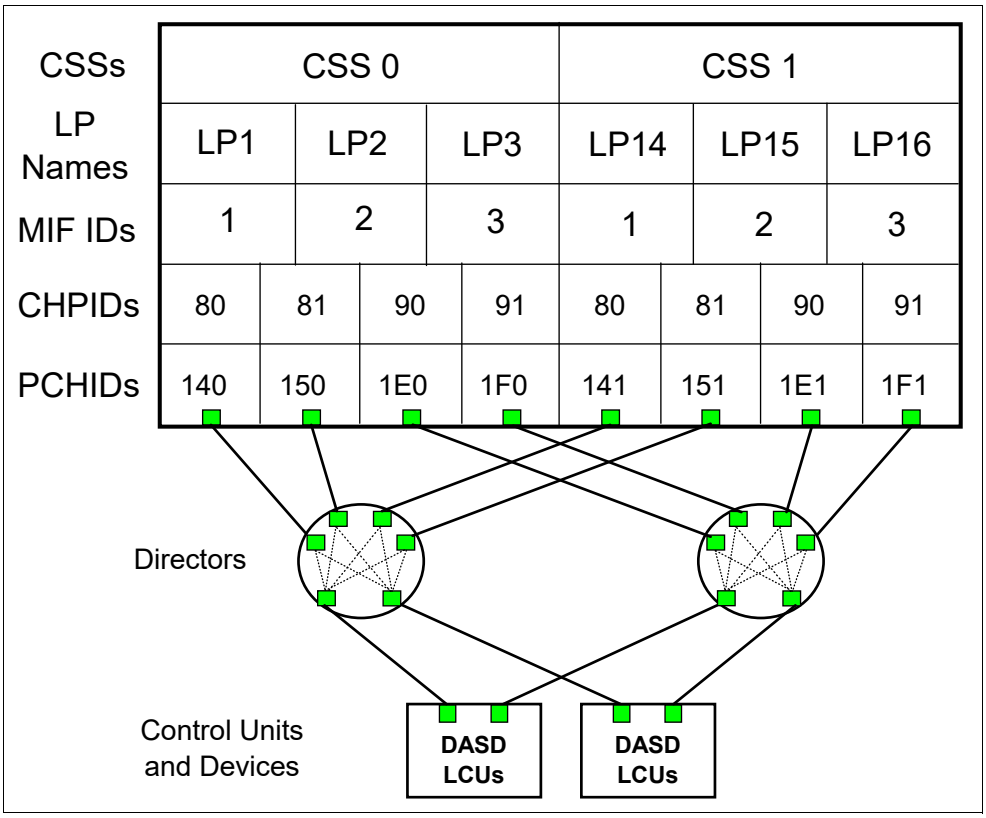


Figure 2-1 Multiple CSS construct

The CHPIDs are mapped to their designated PCHIDs by using the IBM CHPID Mapping Tool or manually by using HCD or IOCP. The output of the CHPID Mapping Tool is used as input to HCD or the IOCP to establish the CHPID-to-PCHID assignments.

2.1.7 Channel spanning

Channel spanning extends the MIF concept of sharing channels across LPARs to sharing channels across channel subsystems and LPARs.

Spanning is the ability for the channel to be configured to multiple channel subsystems. When so defined, the channels can be transparently shared by any or all of the configured LPARs, regardless of the channel subsystem to which the LPAR is configured.

A channel is considered a spanned channel if the same CHPID number in different CSSs is assigned to the same PCHID in the IOCP or is defined as *spanned* in HCD. The same applies to internal channels such as IBM HiperSockets technology, but there is no PCHID association. Internal channels are defined with the same CHPID number in multiple CSSs.

CHPIDs that span CSSs reduce the total number of channels that are available on IBM zSystems. The total is reduced because no CSS can have more than 256 CHPIDs. For an IBM z16 A01, IBM z15 T01, and IBM z14 M0x with six CSSs, a total of 1536 CHPIDs are supported, while for IBM z16 A02, IBM z16 AGZ, IBM z15 T02, and IBM z14 ZR1 with three CSSs, a total of 768 CHPIDs are supported.

If all CHPIDs are spanned across multiple CSSs, only 256 channels can be supported. Table 2-9 on page 26 shows which channel types can be defined as shared or spanned channels.

Table 2-9 Spanned and shared channel

Channel type		CHPID definition	Shared channels	Spanned channels
FICON Express8	External	Fibre Channel (FC), FCP	Yes	Yes
FICON Express8S	External	FC, FCP	Yes	Yes
FICON Express16S	External	FC, FCP	Yes	Yes
FICON Express16S+	External	FC, FCP	Yes	Yes
FICON Express16SA	External	FC, FCP	Yes	Yes
FICON Express32S	External	FC, FCP	Yes	Yes
zHyperLink Express	External	N/A	Yes	Yes
zHyperLink Express 1.1	External	N/A	Yes	Yes
OSA-Express4S[a]	External	OSC, OSD, OSE, OSM, OSN, OSX	Yes	Yes
OSA-Express5S[a]	External	OSC, OSD, OSE, OSM[d], OSN[b], OSX[b]	Yes	Yes
OSA-Express6S	External	OSC, OSD, OSE, OSM[d], OSN[b], OSX[b]	Yes	Yes
OSA-Express7S[c]	External	OSC, OSD, OSE, OSM[d,e]	Yes	Yes
OSA-Express7S 1.1	External	OSD	Yes	Yes
OSA-Express7S 1.2	External	OSD, OSE, OSC	Yes	Yes
ICA SR	External	CS5	Yes	Yes
ICA SR1.1	External	CS5	Yes	Yes
CE2 LR[f]	External	CL5	Yes	Yes
CE LR	External	CL5	Yes	Yes
InfiniBand[g]	External	CIB	Yes	Yes
ISC-3	External	CFP	Yes	Yes
IC	Internal	ICP	Yes	Yes
HiperSockets[h]	Internal	IQD	Yes	Yes

a. Not every CHPID is supported by the different OSA-Express features. For more information, see Chapter 5, “Open Systems Adapter-Express” on page 71.
b. OSN and OSX are NOT supported on 14 and IBM z15 and IBM z16.
c. OSA-Express7S 25GbE SR
d. The OSM CHPID type cannot be defined for user configuration on an IBM z15. OSM is used in DPM mode for internal management only.
e. OSM CHPID is not supported on IBM z16
f. For new build IBM z16 systems, CE LR cannot be carried forward to IBM z16 (IBM z16 supports Coupling Express2 LR).
g. InfiniBand coupling/timing links ar not supported on IBM z16, IBM z15 and IBM z14 ZR1.
h. The CHPID statement of HiperSockets devices requires the keyword VCHID. VCHID specifies the virtual channel identification number that is associated with the channel path. The valid range is 7C0 - 7FF.

2.1.8 Multiple subchannel sets

Do not confuse the multiple *subchannel sets* (MSS) functions with multiple *channel subsystems*. In most cases, a *subchannel* represents an *addressable device*. For example, a disk control unit with 30 drives uses 30 subchannels. An addressable device is associated with a device number.

Subchannel numbers, including their implied path information to devices, are limited to four hexadecimal digits by hardware and software architectures. Four hexadecimal digits provide 64 thousand addresses, which are known as a *set*. IBM reserved 256 subchannels, leaving 63,750 subchannels for general use with IBM zSystems platform. Parallel access volumes (PAVs) make this limitation of subchannels a challenge for larger installations. A single-disk drive (with PAVs) often uses at least four subchannels[1].

It was difficult to remove this constraint because the use of four hexadecimal digits for subchannels and device numbers that correspond to subchannels is specified in several places. Expanding the field would break too many programs.

The solution allows *sets* of subchannels (*addresses*) with a current implementation of four sets with IBM z16 A01, IBM z15 T01, IBM z14 M0x, and three sets with IBM z16 A02, IBM z16 AGZ, IBM z15 T02 and IBM z14 ZR1. Each set provides 64K (64,536) addresses minus one. Subchannel set 0, the first set, reserves 256 subchannels for use by IBM (65,280 devices (64K-256 or 63.75K)). Subchannel sets 1 - 3 provide a full range of 64K minus one subchannel on the IBM zSystems platform (65,535 devices (64K-1)).

The first subchannel set (SS0) allows definitions of any type of device that is supported, for example, bases, aliases, secondaries, and devices, other than disks that do not implement the concept of associated aliases or secondaries.

The second, third, and fourth subchannel sets (SS1, SS2, and SS3) are designated for use for disk alias devices, both primary and secondary devices, and for IBM Metro Mirror secondary devices only.

There is no required correspondence between addresses in the three sets. For example, it is possible to have device number 8000 in subchannel set 0 and device number 8000 in subchannel sets 1 or 2, and they can refer to different devices. Likewise, device number 1234, subchannel set 0, and device number 4321, subchannel set 1, might be the base and an alias for the same device.

There is no *required* correspondence between the device numbers that are used in the four subchannel sets. Each channel subsystem can have multiple subchannel sets, as shown in Figure 2-2 on page 28. The number of subchannels sets in the figure apply to IBM z16 A01, IBM z15 T01, and IBM z14 M0x.

[1] Four subchannel sets are mostly used with PAV. They represent base addresses and three alias addresses.

IBM z16 Model A01					
HSA = 256 GB					
LCSS 0	LCSS 1	LCSS 2	LCSS 3	LCSS 4	LCSS 5
Up to 15 Logical Partitions	**Up to 15 Logical Partitions**	**Up to 15 Logical Partitions**	**Up to 15 Logical Partitions**	**Up to 15 Logical Partitions**	**Up to 10 Logical Partitions**
Subchannel Sets: SS 0 – 63.75 k SS 1 – (64 k -1) SS 2 – (64 k -1) SS 3 – (64 k -1)	**Subchannel Sets: SS 0 – 63.75 k SS 1 – (64 k -1) SS 2 – (64 k -1) SS 3 – (64 k -1)**	**Subchannel Sets: SS 0 – 63.75 k SS 1 – (64 k -1) SS 2 – (64 k -1) SS 3 – (64 k -1)**	**Subchannel Sets: SS 0 – 63.75 k SS 1 – (64 k -1) SS 2 – (64 k -1) SS 3 – (64 k -1)**	**Subchannel Sets: SS 0 – 63.75 k SS 1 – (64 k -1) SS 2 – (64 k -1) SS 3 – (64 k -1)**	**Subchannel Sets: SS 0 – 63.75 k SS 1 – (64 k -1) SS 2 – (64 k -1) SS 3 – (64 k -1)**
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 2-2 Multiple CSSs and multiple subchannel sets

The appropriate subchannel set number must be included in IOCP definitions or in the HCD definitions that produce the IOCDS. The subchannel set number defaults to zero, and IOCP changes are needed only when using subchannel sets 1, 2, or 3.

Running an IPL from an alternative subchannel set

IBM z16 A01, IBM z15 T01, and IBM z14 M0x, support IPL from subchannel sets 0, 1, 2, and 3. IBM z16 A02, IBM z16 AGZ, IBM z15 T02, and IBM z14 ZR1, support running an IPL from subchannel sets 0, 1, and 2. Devices that are used early during IPL processing can now be accessed by using subchannel set 1, subchannel set 2, or subchannel set 3.

This feature allows the use of Metro Mirror (Peer-to-Peer Remote Copy, or PPRC) secondary devices. These devices are defined by using the same device number and a new device type in an alternative subchannel set to be used for IPL, IODF, and stand-alone memory dump volumes when needed.

Running an IPL from an alternative subchannel set is supported on the IBM zSystems platforms. It applies to the IBM Fibre Channel connection (FICON) and High-Performance FICON for IBM zSystems (zHPF) protocols.

2.1.9 Summary

Table 2-10 lists the maximum number of CSS elements that are supported per IBM zSystems platform.

Table 2-10 CSS elements

	IBM z14 ZR1, IBM z15 T02, IBM z16 A02, IBM z16 AGZ	**IBM z14 M0x, IBM z15 T01, IBM z16 A01**
CSSs	3 per system	6 per system
Partitions	15 for the first two CSSs and 10 for the third, up to 40 per system	15 for the first 5 CSSs and 10 for the sixth, up to 85 per system
CHPIDs	256 per CSS, up to 768 per system	256 per CSS, up to 1536 per system
Devices	65,280 (64K-256) on SS0 65,535 (64K-1) on SS1 and SS2	65,280 (64K-256) on SS0 65,535 (64K-1) on SS1, SS2, and SS3

2.2 I/O configuration management

Note: The information in this section applies to systems configured to run in PR/SM mode. For systems running in IBM zSystems Dynamic Partition Manager (DPM) mode, I/O configuration is managed trough the DPM Graphical User Interface (GUI) available on the Hardware Management Console.

CSS controls communication between a configured channel, the control unit, and the device. The IOCDS defines the channels, control units, and devices to the designated LPARs within the system. This communication is defined by using the IOCP.

The IOCP statements typically are built by using the HCD. An interactive dialog is used to generate the input/output definition file (IODF), start the IOCP program, and then build the production IOCDS. The IOCDS is loaded into the HSA and initialized during power-on reset. In prior IBM zSystems servers, the HSA storage was allocated based on the size of the IOCDS, partitions, channels, control units, and devices. Extra storage was reserved for dynamic I/O reconfiguration, if enabled.

The HSA sizes are listed in Table 2-2 on page 20.

With IBM zSystems platform, CHPIDs are mapped to PCHIDs or AIDs by using the configuration build process through HCD or IOCP.

The sections that follow describe tools that are used to maintain and optimize the I/O configuration on IBM zSystems platform.

2.2.1 Hardware Configuration Definition

HCD supplies an interactive dialog to generate the IODF and then the IOCDS. Consider using the HCD to generate the IOCDS, as opposed to writing IOCP statements. The validation checking that HCD does as data is entered helps eliminate errors before the I/O configuration is implemented.

2.2.2 CHPID Mapping Tool

The CHPID Mapping Tool helps with IBM zSystems requirements. It provides a mechanism to map CHPIDS to PCHIDS and identify the best availability options for installed features and defined configuration.

Consider using the mapping tool for all new builds of IBM zSystems family or when upgrading from one system to another. You can also use it as part of standard hardware changes (for example, MES[2]) for an existing IBM zSystems.

The mapping tool takes input from two sources:

- The Configuration Report file (CFreport) that is produced by the IBM order tool and provided by the IBM account team or produced by IBM manufacturing and obtained from IBM Resource Link.
- An IOCP statement file.

[2] Miscellaneous Equipment Specification - process of upgrading/changing system features

The mapping tool produces the following outputs:

- Tailored reports

  Save all reports for reference. Supply the port report that is sorted by CHPID number and location to the IBM hardware service representative for IBM zSystems installations.

- An IOCP input file with PCHIDs mapped to CHPIDs

  This IOCP input file can then be migrated back into HCD and used to build a production IODF.

The mapping tool does not automatically map CS5 or CIB CHPIDs to AIDs and ports. This process must be done manually, either in HCD, IOCP, or the mapping tool. The mapping tool provides availability intersects for defined CIB CHPIDs. For more information about the CHPID Mapping Tool, see IBM Documentation.

2.3 I/O configuration planning

I/O configuration planning for IBM zSystems requires the availability of physical resources, and must comply with the specific rules of the logical definitions. The following physical resources are the minimum that is required for connectivity:

- Platform frame
- PCIe+ I/O drawer, PCIe I/O drawer, I/O drawer, in a frame
- I/O slot in a PCIe+ I/O drawer, PCIe I/O drawer, or I/O drawer
- Channel feature in a slot of a PCIe+ I/O drawer, PCIe I/O drawer, or I/O drawer
- Port on a channel feature

For a system configuration, the IBM zSystems configurator build process includes all physical resources that are required for a particular I/O configuration, based on the supplied channel type and quantity requirements.

The definition phase starts after the physical resources are ordered. The channels must be defined according to the architecture's rules, the system's implementation, and the order.

The operational characteristics of a particular channel type, along with the addressing capabilities, can affect configuration complexity, topology, infrastructure, and performance.

2.3.1 I/O configuration rules

The following sections briefly describe the IBM zSystems configuration rules, which are identified through the architecture and the specific system that are implemented and enforced by using the HCD and IOCP.

All control units and I/O devices that attach to the system must be defined to a CSS. Specify the following items when defining the I/O configuration for the system through HCD/IOCP:

- LPARs (LPAR name, CSS ID, and MIF ID, where appropriate)
- Channel paths on the system, their assignment to CSSs, and LPARs
- FICON Directors (where appropriate)
- Control units that are attached to the channel paths
- I/O devices that are assigned to the control units

Cryptographic features: The cryptographic features on IBM zSystems do not require a CHPID definition and are configured by using the HMC or SE.

Certain definition characteristics that must be considered for the I/O configuration are found in the architecture (z/Architecture). Other definition characteristics are specified only for the system. Table 2-11 lists general IOCP rules for IBM Z.

Table 2-11 IBM zSystems general IOCP rules

Constant machine attributes		**IBM z16, IBM z15, IBM z14**
Maximum configurable physical control units (PCUs)	PCUs per OSD	16
	PCUs per OSE, OSC, or OSN[a]	1
	PCUs per OSM[b] or OSX[c]	16
	PCUs per CFP or ICP	1
	PCUs or link addresses per FC	256
	PCUs per FCP	1
	CUs per IQD	64
Maximum configurable devices	Per CIB (12x InfiniBand) or CS5 (ICA SR)	8
	CIB (1x InfiniBand) and CL5	32
	Per CNC	1024
	Per CTC	512
	Per OSC[d]	253
	Per OSD	1920
	Per OSE	254
	Per OSM[b] or OSX[c]	1920
	Per OSN[a]	480
	Per FCP	480
	Per FC	32K
	For all IQD channel paths	12K

a. The OSN CHPID type is *not* supported on IBM z16, IBM z15, or IBM z14.
b. OSM CHPID cannot be defined for user configurations on IBM z15 running in PR/SM mode. On IBM z15, OSM CHPID is used in DPM mode for internal management only. OSM is not supported on IBM z16.
c. OSX CHPID is NOT supported with IBM z16 or IBM z15
d. A limit of 120 clear sessions and 48 encrypted sessions can be defined at the HMC/SE.

For more information about CHPID types and channel configuration rules, see *Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7177.

2.4 References

The following publications include information that is related to the topics that are covered in this chapter:

- *z/OS Hardware Configuration Definition User's Guide*, SC34-2669
- *z/OS Hardware Configuration Definition Planning*, GA32-0907
- *Hardware Configuration Definition: User's Guide*, SC34-2699
- *Hardware Configuration Definition Planning*, GA32-0907
- *Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7177
- *IBM z16 (3931) Technical Guide,* SG24-8951
- *IBM z16 (3932) Technical Guide*, SG24-8952
- *IBM z16 Technical Introduction*, SG24-8950
- *IBM z15 (8561) Technical Guide*, SG24-8851
- *IBM z15 Technical Introduction*, SG24-8850
- *IBM z14 Technical Introduction*, SG24-8450
- *IBM z14 (3906) Technical Guide*, SG24-8451
- *IBM z14 Model ZR1 Technical Introduction*, SG24-8550
- *IBM z14 ZR1 Technical Guide*, SG24-8651

3

Fibre Channel connectivity

This chapter describes the Fibre Channel connection (FICON) Express features and the protocols that they support on IBM zSystems platforms.

This chapter includes the following topics:

- 3.1, “FICON Express description” on page 34
- 3.2, “FICON elements” on page 45
- 3.3, “Connectivity” on page 60
- 3.4, “References” on page 66

3.1 FICON Express description

FICON provides a non-drop distance of up to 100 km. FICON supports a link data rate of 2, 4, 8,16, or 32 Gbps, depending on the FICON features:

- FICON Express32S features automatically negotiate to 8, 16, or 32 Gbps.
- FICON Express16SA features automatically negotiate to 8 or 16 Gbps.
- FICON Express16S+ and FICON Express16S features automatically negotiate to 4, 8, or 16 Gbps.
- FICON Express8S features automatically negotiate to 2, 4, or 8 Gbps.

Note: Not all FICON Express features are supported on all IBM zSystems platforms. Support for each feature is listed in Table 3-2 on page 60.

The FICON implementation enables full-duplex data transfer, so data travels in both directions simultaneously. FICON also enables multiple concurrent I/O operations.

Terminology: Throughout this chapter, *FICON* refers to the FICON Express8S, FICON Express16S, FICON Express16S+, FICON Express16SA, and FICON Express32S features, except when the function that is described is applicable to a specific feature.

The FICON channel matches data storage and access requirements with the latest technology in servers, control units, and storage devices. FICON channels allow faster and more efficient data transfer while allowing you to use their currently installed single mode and multimode fiber optic cabling plant.

FICON uses long wavelength (LX) and short wavelength (SX) transceivers with single mode and multi-mode fiber optic media for data transmission.

3.1.1 FICON modes and topologies

IBM zSystems support the operation of FICON channels in one of two modes:

- FICON native mode (CHPID type Fibre Channel (FC))
- Fibre Channel Protocol (CHPID type FCP)

FICON native mode (FC)

As shown in Figure 3-1 on page 35, a FICON channel in FICON native mode (CHPID type FC) can access FICON native interface control units by using the following topologies:

- Point-to-point (direct connection)
- Switched point to point (through a Fibre Channel switch)
- Cascaded FICON Directors (through *two* Fibre Channel switches)

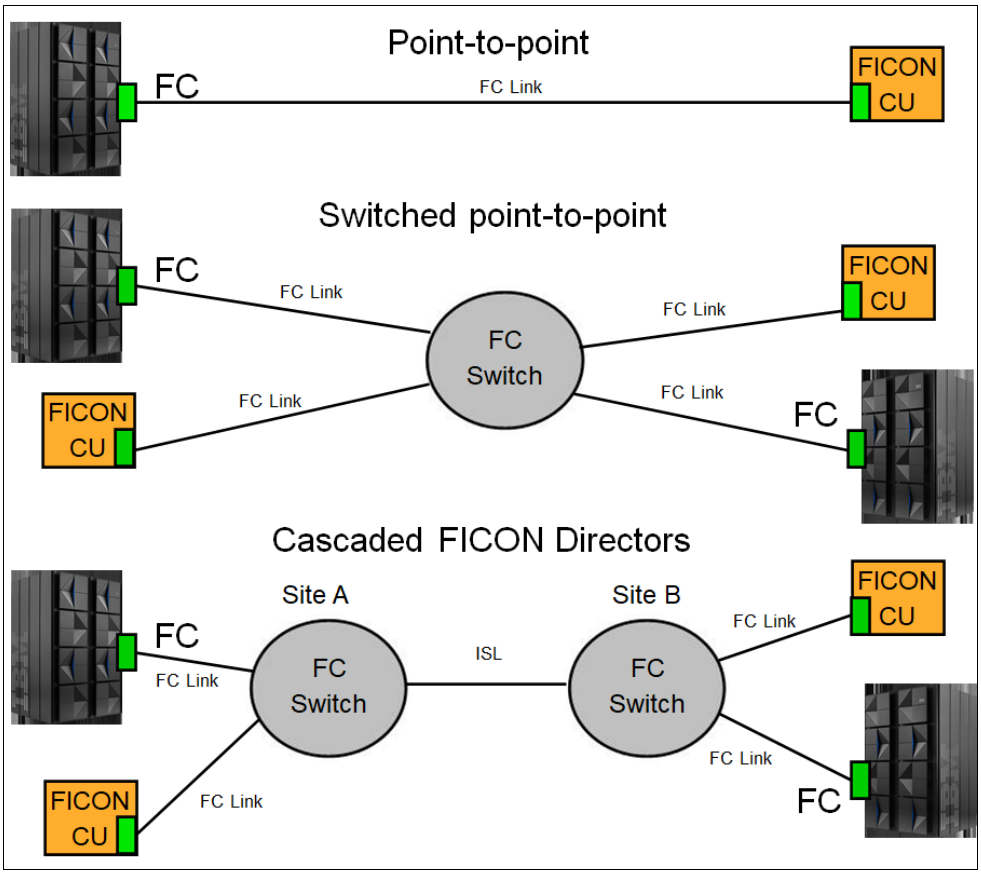


Figure 3-1 Native FICON supported topologies

A FICON native channel also supports channel-to-channel communications. The FICON channel at each end of the FICON CTC connection, which supports the FCTC control units, can also communicate with other FICON native control units, such as disk storage devices and tape. At least one end of the FICON CTC connection must be an IBM zSystems installation.

Fibre Channel Protocol (FCP) mode

A FICON channel in Fibre Channel Protocol mode (CHPID type FCP) can access FCP devices in either of the following ways:

- A FICON channel in FCP mode through a single Fibre Channel switch or multiple switches to a Small Computer System Interface (SCSI) device
- A FICON channel in FCP mode through a single Fibre Channel switch or multiple switches to a Fibre Channel-to-SCSI bridge.
- FCP-attached tape libraries (such as TS3310) because customers use them with FCP

The FICON features support Fibre Channel and SCSI devices in IBM z/VM, IBM z/VSE, Linux on IBM Z and the KVM hypervisor, as shown in Figure 3-2.

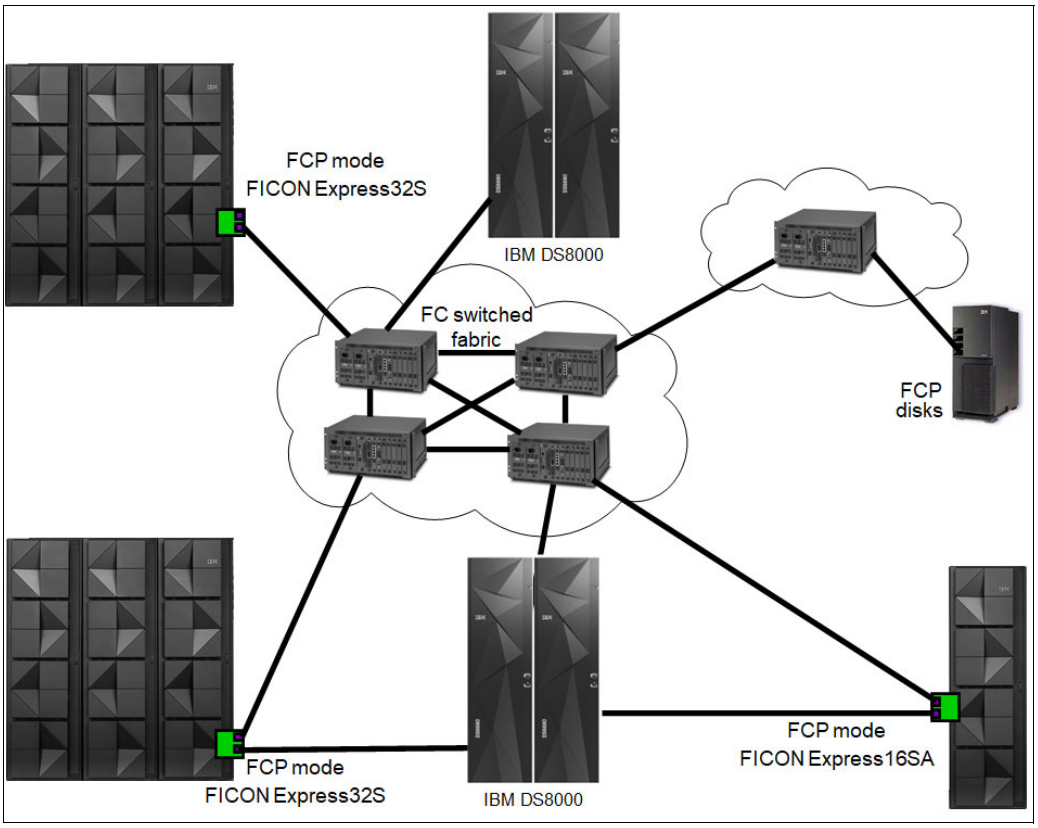


Figure 3-2 IBM zSystems FCP example topology

With IBM zSystems, point-to-point connections can be used to access data that is stored on devices without using a Fibre Channel switch. In addition, an operating system or other stand-alone program can undergo an IPL through a point-to-point connection by using the SCSI IPL feature. N_Port ID Virtualization is not supported by FCP point to point. For more information, see “Worldwide port name tool” on page 39.

The FCP support allows z/VM, Linux on IBM Z and the KVM hypervisor, and z/VSE operating systems to access industry-standard SCSI devices. For disk applications, these FCP storage devices use fixed block sectors rather than the extended count key data (ECKD) format.

3.1.2 FCP Channel

The FC-FCP standard was developed by the International Committee of Information Technology Standards (INCITS) and published as an ANSI standard. The IBM zSystems FCP I/O architecture conforms to the FC standards specified by the INCITS. For more information about the FC standards, see the INCITS Technical Committee T11 website and their page for SCSI Storage Interfaces (this committee within INCITS is responsible for the Fibre Channel Interface).

FICON channels in FCP mode provide full fabric attachment of SCSI devices to the operating system images by using the Fibre Channel Protocol and provide point-to-point attachment of SCSI devices. This technique allows z/VM, Linux on IBM Z and the KVM hypervisor, and z/VSE to access industry-standard SCSI storage controllers and devices.

FCP channel full fabric support means that multiple numbers of directors or switches can be placed between the IBM zSystems platform and the SCSI device. This technique allows many *hops* through a storage area network (SAN) and provides improved use of intersite connected resources and infrastructure. This expanded ability to attach storage devices provides more choices for storage solutions and the ability to use existing storage devices. This configuration can facilitate the consolidation of UNIX server farms onto IBM zSystems platform, protecting investments in SCSI-based storage.

For a list of switches, storage controllers, and devices that are verified to work in a Fibre Channel network that is attached to FCP channel, and for software requirements to support FCP and SCSI controllers or devices, see the I/O Connectivity website.

FICON channels in FCP mode are based on the Fibre Channel standards that are defined by INCITS and published as ANSI standards. FCP is an upper-layer Fibre Channel mapping of SCSI on a common stack of Fibre Channel physical and logical communication layers.

SCSI is supported by a wide range of controllers and devices, complementing the classical storage attachment capability through FICON channels. FCP is the base for industry-standard Fibre Channel networks or SANs.

Fibre Channel networks consist of servers, storage controllers, and devices as end nodes, which are interconnected by Fibre Channel switches, directors, and hubs. Switches and directors are used to build Fibre Channel networks or fabrics. Fibre Channel Arbitrated Loops (FC-ALs) can be constructed by using Fibre Channel hubs. In addition, different types of bridges and routers can be used to connect devices with different interfaces, such as parallel SCSI. All of these interconnections can be combined in the same network.

SCSI is implemented by many vendors in many different types of storage controllers and devices. These controllers and devices are widely accepted in the marketplace and have proven to be able to meet the reliability, availability, and serviceability (RAS) requirements of many environments.

FICON channels in FCP mode use the queued direct input/output (QDIO) architecture for communication with the operating system. The QDIO architecture for FCP channels derives from the QDIO architecture, which was defined initially for the OSA-Express features and for HiperSockets communications.

FCP channels do not use control devices. Instead, data devices that represent QDIO queue pairs are defined, consisting of a request queue and a response queue. Each queue pair represents a communication path between an operating system and the FCP channel. It allows an operating system to send FCP requests to the FCP channel through the request queue. The FCP channel uses the response queue to pass completion indications and unsolicited status indications to the operating system. Figure 3-3 on page 38 illustrates this process.

HCD or Input/Output Configuration Program (IOCP) is used to define the FCP channel type and QDIO data devices. However, there is no definition requirement for the Fibre Channel storage controllers and devices, or for the Fibre Channel interconnect units, such as switches, directors, and bridges. The FCP industry standard architecture requires that the Fibre Channel devices (end nodes) in a Fibre Channel network are addressed by using worldwide names (WWNs), Fibre Channel Identifiers (IDs), and logical unit numbers (LUNs).

These addresses are configured on an operating system level and passed to the FCP channel together with the corresponding Fibre Channel I/O or service request through a logical QDIO device (queue).

Figure 3-3 shows the necessary FCP I/O definitions and compares them to FICON I/O definitions.

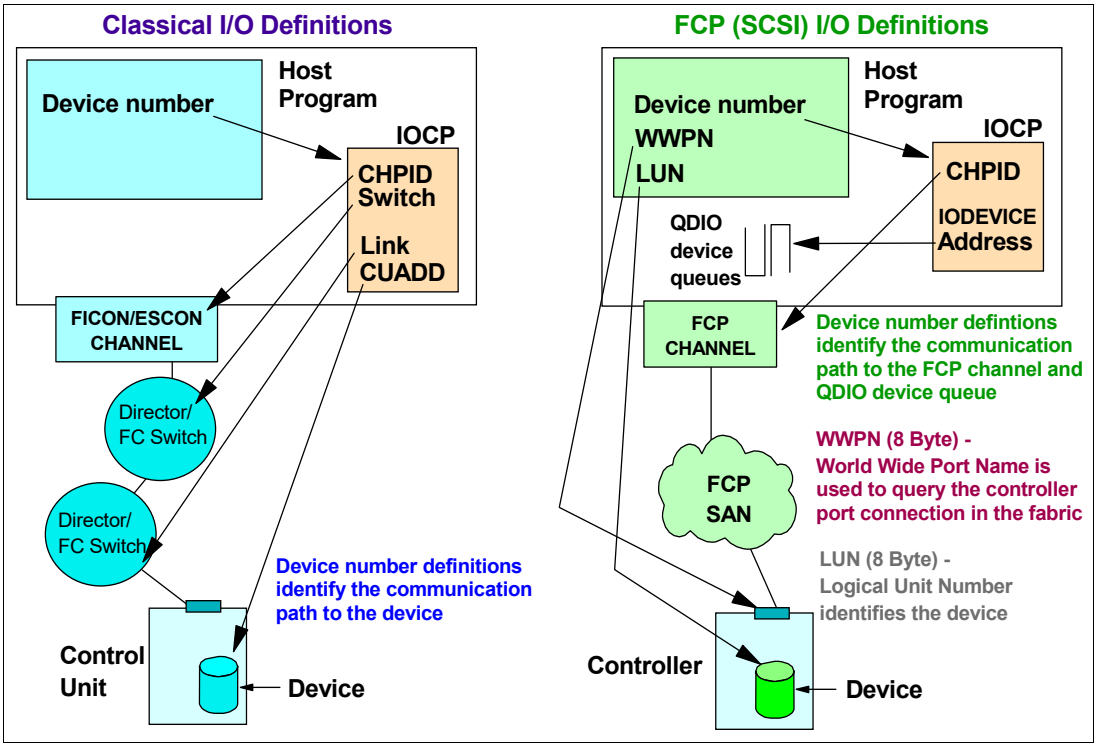


Figure 3-3 I/O definition comparison (FCP to FICON)

Channel and device sharing

An FCP channel can be shared among multiple Linux operating systems, each running in a logical partition (LPAR) or as a guest operating system under z/VM. To access the FCP channel, each operating system needs its own QDIO queue pair, which is defined as a data device on an FCP channel in the HCD/IOCP.

Each FCP channel can support up to 480 QDIO queue pairs with the IBM zSystems platform. This support allows each FCP channel to be shared among 480 operating system instances (with a maximum of 252 guests per LPAR).

Host operating systems that share access to an FCP channel can establish up to 2048 concurrent connections to up to 512 different remote Fibre Channel ports that are associated with Fibre Channel controllers. The total number of concurrent connections to end devices, which are identified by LUNs, must not exceed 4096.

Although multiple operating systems can concurrently access the same remote Fibre Channel port through a single FCP channel, Fibre Channel devices, which are identified by their LUNs, can be reused only serially. For two or more unique operating system instances to share concurrent access to a single Fibre Channel or SCSI device (LUN), each of these operating systems must access this device through a different FCP channel.

If two or more unique operating system instances attempt to share concurrent access to a single Fibre Channel or SCSI device (LUN) over the same FCP channel, a LUN sharing conflict occurs and errors result. A way to alleviate this sharing conflict on IBM zSystems platform is to use *N_Port ID Virtualization*.

Worldwide port name tool

The worldwide port name tool assigns WWPNs to each FCP channel or port by using the same WWPN assignment algorithms that a system uses when you are assigning WWPNs for channels that use NPIV. Therefore, the SAN can be set up in advance, allowing operations to proceed much faster after the system is installed.

WWPN tool: The WWPN tool is supported in PR/SM mode. DPM uses a different assignment algorithm for generating and assigning NPIV WWPNs.

The WWPN tool can calculate and show WWPNs for both virtual and physical ports ahead of system installation. This feature means that the SAN configuration can be retained rather than altered by assigning a WWPN to physical FCP ports when a FICON feature is replaced.

The WWPN tool takes an adhesive 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 (`.csv`) file can either be created manually or exported from the Hardware Configuration Definition or Hardware Configuration Manager (HCD or HCM).

The WWPN tool can be downloaded from the Tools section of IBM Resource Link (requires registration).

FCP SCSI IPL feature enabler

This function runs an IPL of an operating system from an FCP channel-attached disk either in a LPAR or as a guest operating system under z/VM. In particular, SCSI IPL can directly IPL a Linux operating system that was installed previously on a SCSI disk. Therefore, there is no need for a classical channel-attached device (FICON), such as an ECKD disk control unit, to install and IPL a Linux operating system.

The IPL device is identified by its SAN address, which consists of the WWPN of the disk controller and the LUN of the IPL device.

Important: If a z/VM system second level undergoes an IPL from an FCP SCSI LUN, a minimum virtual memory is required, which depends on the model of the processor on which the z/VM system is running. To ensure success on all processor models, you should define at least 768 MB of virtual storage.

SCSI IPL is supported in the following conditions:

- FCP access control
- Point-to-point connections
- N-Port ID Virtualization (NPIV)

A *stand-alone-dump program* can also undergo an IPL from an FCP channel that is attached to a SCSI disk. The stand-alone-dump program can also store the generated dumped data on a SCSI disk. z/VM support of SCSI IPL allows Linux and other guest operating systems that support this feature to undergo an IPL from an FCP-attached SCSI disk when z/VM is running on an IBM zSystems platform. Therefore, Linux guests can be started and run completely from an FCP channel-attached disk.

FCP multipathing concept

Multipath access to disk subsystems is a basic feature with the channel subsystem on IBM zSystems platform. FICON connections support multiple hardware paths to any physical disk device. The IBM zSystems platform handles multipathing invisibly through the operating system. With FICON channels, the operating system is presented a single device for I/O operations, and multipathing happens under channel subsystem control.

Multipathing over FCP is different. With FCP multipathing on Linux on IBM Z, each path to each LUN appears to the operating system as a separate device. For example, if there are four paths to five LUNs, the Linux kernel defines 20 SCSI devices.

Currently, supported distributions use device-mapper multipath in the Linux kernel along with multipath-tools in user space. For more information, see the corresponding distribution documentation and How to use FC-attached SCSI devices.

FCP access control

The ability to control access to nodes and devices is provided as a function in switches and controllers, and is called *LUN masking* and *zoning*. LUN masking and zoning can be used to prevent systems from accessing storage that they are not permitted to access:

LUN masking A LUN represents a portion of a controller, such as a disk device. With the use of LUNs, a controller can be logically divided into independent elements or groups of elements. Access to these LUNs can be restricted to distinctive WWPNs as part of the controller configuration. This method is known as *LUN masking*.

Zoning Segmentation of a switched fabric is achieved through *zoning*. It should be used to fence off certain portions of the switched fabric, allowing only the members of a zone to communicate within that zone. All others that attempt to access from outside of that zone are rejected.

I/O devices

The IBM zSystems FCP channel implements the Fibre Channel Protocol standard as defined by the INCITS Fibre Channel Protocol for SCSI (FC-FCP) and Fibre Channel Protocol for SCSI Second Version (FCP-2), and the relevant protocols for the SCSI-2 and SCSI-3 protocol suites. Theoretically, each device that conforms to these protocols works when attached to an IBM zSystems FCP channel. However, experience shows that there are small deviations in the implementations of these protocols.

Also, for certain types of FCP and SCSI controllers and devices, specific drivers in the operating system might be required to use all of the capabilities of the controller or device. The drivers might also be required to cope with unique characteristics or deficiencies of the device.

Note: Do appropriate conformance and interoperability testing to verify that a particular storage controller or device can be attached to an IBM zSystems FCP channel in a particular configuration. For example, test that it can be attached through a particular type of Fibre Channel switch, director, or point-to-point connection.

Hardware assists for z/VM guests

A complementary virtualization technology is available for the IBM zSystems platform. The technology includes these capabilities:

- QDIO Enhanced Buffer-State Management (QEBSM), with two hardware instructions that are designed to eliminate the overhead of hypervisor interception.
- Host Page-Management Assist (HPMA), which is an interface to the z/VM main storage management function that allows the hardware to assign, lock, and unlock page frames without z/VM hypervisor assistance.

These hardware assists allow a cooperating guest operating system to start QDIO operations directly to the applicable channel, without interception by z/VM, providing more performance improvements. This support is integrated into the IBM zSystems platform. Consult the appropriate Preventive Service Planning (PSP) buckets (3931DEVICE, 8561DEVICE, 3906DEVICE, 3932DEVICE, 8562DEVICE, or 3907DEVICE) before implementation.

Support of T10-DIF for enhanced reliability

Because high reliability is important for maintaining the availability of business-critical applications, the IBM zSystems FCP supports the ANSI T10 Data Integrity Field (DIF) standard. Data integrity protection fields are generated by the operating system and propagated through the SAN. IBM zSystems helps to provide added end-to-end data protection between the operating system and the storage device.

An extension to the standard, Data Integrity Extensions (DIX), provides checksum protection from the application layer through the host bus adapter (HBA), where cyclical redundancy check (CRC) protection is implemented.

T10-DIF support by the FICON features, when defined as CHPID type FCP, is available on the IBM zSystems platform. Using the T10-DIF standard requires support by the operating system and the storage device.

3.1.3 FCP and FICON mode characteristics

The single largest difference between the FICON channel (FC) and FCP channel mode types is the treatment of data access control and security. FICON channels rely on a multiple image facility (MIF) to address concerns about shared channels and devices. MIF provides ultra-high access control and security of data so that one operating system image and its data requests cannot interfere with another operating system's data requests. With the introduction of IBM zSystems, MIF continues this ultra-high level of access control and security across channel subsystems (CSSs).

FCP and MIF

Linux guest operating systems under z/VM can have access to an FCP channel defined to the z/VM operating system. Using MIF, an FCP channel can also be shared between Linux LPARs and z/VM LPARs with Linux guests.

The FCP industry-standard architecture does not use the data access control and security functions of MIF. As a result, FCP has the following limitations:

- Channel sharing

  When NPIV is not implemented, and if multiple Linux images share an FCP channel and all of the Linux images have connectivity to all of the devices that are connected to the FCP fabric, all of the Linux images use the same worldwide port name (WWPN). They use this name to enter the fabric and are indistinguishable from each other within the fabric. Therefore, the use of zoning in switches and LUN masking in controllers is not effective in creating appropriate access controls among the Linux images.

  By using NPIV, each operating system that shares an FCP channel is assigned a unique WWPN. The WWPN can be used for *device-level* access control in storage controllers (LUN masking) and in *switch-level* access control (zoning).

- Device sharing

  Without using NPIV, an FCP channel prevents logical units from being opened by more than one Linux image at a time. Access is on a first-come, first-served basis. This system prevents problems with concurrent access from Linux images that share an FCP channel and the same WWPN. This serialization means that one Linux image can block other Linux images from accessing the data on one or more logical units unless the sharing images (z/VM guests) are not in contention.

FICON versus FCP

FICON and FCP have other significant differences. Certain differences are fundamental to the IBM zSystems family, and others are fundamental to the two channel architectures. Still others depend on the operating system and the device being attached. Without taking the operating system and the storage device into consideration, I/O connectivity through IBM zSystems FCP and FICON channels has the following differences:

- Direct connection

  With all FICON features on the IBM zSystems platform, storage controllers can be directly connected to the channel by using point-to-point attachment when in FCP mode. There is no need for a director or switch between the FCP channel and storage controllers.

  Note: N_Port ID Virtualization (NPIV) is supported in switched topology, and FCP with NPIV is **not** supported in a point-to-point topology.

- Switch topology

  FCP channels support full fabric connectivity, meaning that several directors or switches can be used between a IBM zSystems platform and the device. With the FICON cascaded director support, the FICON storage network topology is limited to a two-director, single-hop configuration.

- Enterprise fabric

  The use of cascaded FICON Directors ensures the implementation of a high-integrity fabric. For FCP, a high-integrity fabric solution is not mandatory, although it must be considered. For example, if an FCP inter-switch link (ISL) must be moved, data potentially might be sent to the wrong path without notification. This type of error does not happen on an enterprise fabric with FICON.

- Transmission data checking

  When a transmission is sent through an FCP channel, because of its full fabric capability, FCP checks data for each leg of that transmission. FICON also checks intermediate data.

- Serviceability:
  - Licensed Internal Code (LIC) updates and the IBM zSystems platform itself allow FCP concurrent patches. FICON channels, when configured as CHPID type FCP, support concurrent patches, allowing the application of a LIC without requiring a configuration of off/on. This exclusive FCP availability feature is available with all FICON features.
  - The FICON features have Small Form-factor Pluggable (SFP) optics to permit each channel to be individually serviced during a fiber optic module failure. The traffic on the other channels on the same feature can continue to flow if a channel requires servicing.
- Problem determination:
  - Request Node Identification (RNID)

    RNID assists with the isolation of FICON detected cabling errors. Resolving fiber optic cabling problems can be a challenge in a fiber optic environment with extended distances. To facilitate resolution, the operating system can request the RNID data for each device or control unit that is attached to native FICON channels. Then, you can display the RNID data by using an operator command. RNID is available to the IBM zSystems platform and is supported by all FICON features (CHPID type FC) and IBM z/OS.
  - Link incident reporting

    Link incident reporting is integral to the FICON architecture. When a problem on a link occurs, this mechanism identifies the two connected nodes between which the problem occurred, leading to faster problem determination and service. For FCP, link incident reporting is not a requirement for the architecture, although it might be offered as an optional switch function. Therefore, important problem determination information might not be available if a problem occurs on an FCP link.

    IBM zSystems allows z/OS to register for all FICON link incident records. This feature improves your ability to capture data for link incident analysis across multiple systems.
  - Simplified problem determination

    To more quickly detect fiber optic cabling problems in a storage area network, all FICON channel error information is forwarded to the HMC. This function facilitates detection and reporting trends and thresholds for the channels with aggregate views, including data from multiple systems.

    Problem determination can be simplified by using the HMC to pinpoint fiber optic cabling issues in your SAN fabric without involving IBM service personnel.

    All FICON channel error information is forwarded to the HMC. In the HMC, it is analyzed to detect and report the trends and thresholds for all FICON channels on the IBM zSystems platform. The Fibre Channel Analyzer task on the HMC can be used to display analyzed information about errors on FICON channels (CHPID type FC) of attached Support Elements. Data includes information about the PCHID, CHPID, channel type, source link address, and destination link address where the error occurred. This report shows an aggregate view of the data and can span multiple systems.

    Starting with z13, similar FICON problem determination tools have been implemented for FCP channels. These channel problem determination tools for FCP channels include functions such as analyze channel information, subchannel data, device status, serial link status, and link error statistic block. In addition to the analyze functions, fabric status login and SAN explorer functions are also available. These FCP problem determination tools are accessed from the HMC in the same way as for the FICON channels.

- FICON purge path extended

  The purge path extended function enhances FICON problem determination. FICON purge path error recovery function is extended so that it transfers error-related data and statistics between the channel and entry switch and the control unit and its entry switch to the host operating system. FICON purge path extended use requires a switch or device that supports this function. The purge path extended function for FC channels is available on IBM z16, IBM z15, and IBM z14.

- FICON error recovery

  IBM zSystems platform, z/OS, and I/O recovery processing are designed to allow the system to detect switch or director fabric problems that might cause FICON links to fail and recover multiple times in a short time.

  This feature allows the system to detect these conditions and keep an affected path offline until an operator action is taken. This process is expected to limit the performance impacts of switch or director fabric problems. The FICON error recovery function is available in z/OS.

Forward Error Correction

Forward Error Correction (FEC) is a technique that is used for controlling errors in data transmission over unreliable or noisy communication channels. By adding redundancy and error-correcting code (ECC) to the transmitted information, the receiver detects and corrects a limited number of errors in the information instead of requesting a retransmission. This process improves the reliability and bandwidth utilization of a connection by saving retransmissions due to bit errors. This advantage is true especially for connections across long distances, such as an Inter-Switch Link (ISL) in an IBM Geographically Dispersed Parallel Sysplex (IBM GDPS) Metro Mirror environment.

FICON Express16SA, FICON Express16S+, and FICON Express16S support FEC coding on top of their 64 b/66 b data encoding for 16 Gbps connections. Their FEC design can correct up to 11 bit errors per 2112 bits that are transmitted, and FICON Express32G uses 256b/257b encoding and can correct up to 20 bit errors per 5140 bits that are transmitted. Thus, when connected to devices that support FEC at 16 or 32 Gbps connections, the FEC design allows FICON Express channels to operate at higher speeds over longer distances and with reduced power and higher throughput. Concurrently, the FEC design maintains the same reliability and robustness that FICON channels are known for.

With IBM DS8870 or later, the IBM z16, IBM z15, IBM z14, and IBM z14 ZR1 can extend the use of FEC to the fabric N_Ports[1] for a completed end-to-end coverage of 16 or 32 Gbps FC links. For more information, see *IBM DS8000 and IBM Z Synergy DS8000: Release 9.2 and z/OS 2.5*, REDP-5186.

FICON dynamic routing

With the IBM z16, IBM z15, IBM z14 and IBM z14 ZR1, FICON channels are no longer restricted to the use of static SAN routing policies for ISLs for cascaded FICON directors. IBM zSystems now support dynamic routing in the SAN with the FICON Dynamic Routing (FIDR) feature. It supports the dynamic routing policies that are provided by the FICON director manufacturers, such as Brocade Exchange Based Routing 7 (EBR 7) and Cisco Open Exchange ID Routing (OxID).

[1] Node Port

With FIDR, IBM z16, IBM z15, IBM z14, and IBM z14 ZR1 have advantages for performance and management in configurations with ISL and cascaded FICON directors:

- Support sharing of ISLs between FICON and FCP (PPRC or distributed)
- Better balanced I/O traffic between all available ISLs
- Improved utilization of the FICON director and ISL
- Easier management with a predicable 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 has implemented a health check function for FICON dynamic routing.

FICON performance

See the latest FICON and FCP performance information at the IBM server connectivity web page.

3.2 FICON elements

FICON enables multiple concurrent I/O operations to occur simultaneously to multiple control units. FICON channels also permit intermixing of large and small I/O operations on the same link. The data center I/O configuration now has increased flexibility for connectivity because of the increased I/O rate, increased bandwidth, and multiplexing of mixed workloads.

3.2.1 FICON channel

FICON channel architecture is compatible with the following protocols:

- Fibre Channel Physical and Signaling standard (FC-FS)
- Fibre Channel Switch Fabric and Switch Control Requirements (FC-SW)
- Fibre Channel Single-Byte-3 (FC-SB-3) and Fibre Channel Single-Byte-4 (FC-SB-4) standards

Cabling specifications are defined by the Fibre Channel - Physical Interface - 4 (FC-PI-4) standard and used by IBM zSystems FICON features. Table 3-1 identifies cabling types and link data rates that are supported in the FICON environment, including their allowable maximum distances and link loss budget. The link loss budget is derived from the channel insertion loss budget that is defined by the FC-PI-4 standard (Revision 8.00).

Table 3-1 Fiber optic cabling for FICON: Maximum distances and link loss budget

FC-PI-4 Fiber core	Cable Type	2 Gbps	4 Gbps	8 Gbps	16 Gbps	32 Gbps	10 Gbps ISL[a]
		Distance / Link-loss budget (dB)	Distance / Link-loss budget (dB)	Distance / Link-loss budget (dB)	Distance / Link-loss budget (dB)	Distance / Link-loss budget (dB)	Distance / Link-loss budget (dB)
9 µm SM	OS1/OS2	10 km / 7.8	10 km / 7.8	10 km / 6.4	10 km / 6.4	10 km / 6.34	10 km / 6.4
9 µm SM	OS1/OS2	4 km / 4.8	4 km / 4.8	N/A	N/A	N/A	N/A
50 µm MM	OM4	500 m / 3.31	400 m / 2.95	190 m / 2.19	125 m / 1.95	100 m / 1.86	N/A
50 µm MM	OM3	500 m / 3.31	380 m / 2.88	150 m / 2.04	100 m / 1.86	70 m / 1.75	300 m / 2.6
50 µm MM	OM2	300 m / 2.62	150 m / 2.06	50 m / 1.68	35 m / 1.63	20 m / 1.57	82 m / 2.3
62.5 µm MM	OM1	150 m / 2.1	70 m / 1.78	21 m / 1.58	N/A	N/A	N/A

a. ISL between two FICON Directors.

Note: IBM does not support a mix of 50 µm and 62.5 µm fiber optic cabling in the same physical link.

When an application performs an I/O operation to a device represented by a unit control block (UCB), it initiates an I/O request by using macros or Supervisor Call to the Input/Output Supervisor (IOS). The application or access method also provides the channel program (channel command words (CCWs)) and extra parameters in the operation request block (ORB). This request is queued on the UCB. The IOS services the request from the UCB on a priority basis.

The IOS then issues a start subchannel (SSCH) instruction, with the subsystem identifier (SSID) that represents the device and the ORB as operands. The CSS is signaled to perform the operation. This flow is shown in Figure 3-4 on page 47.

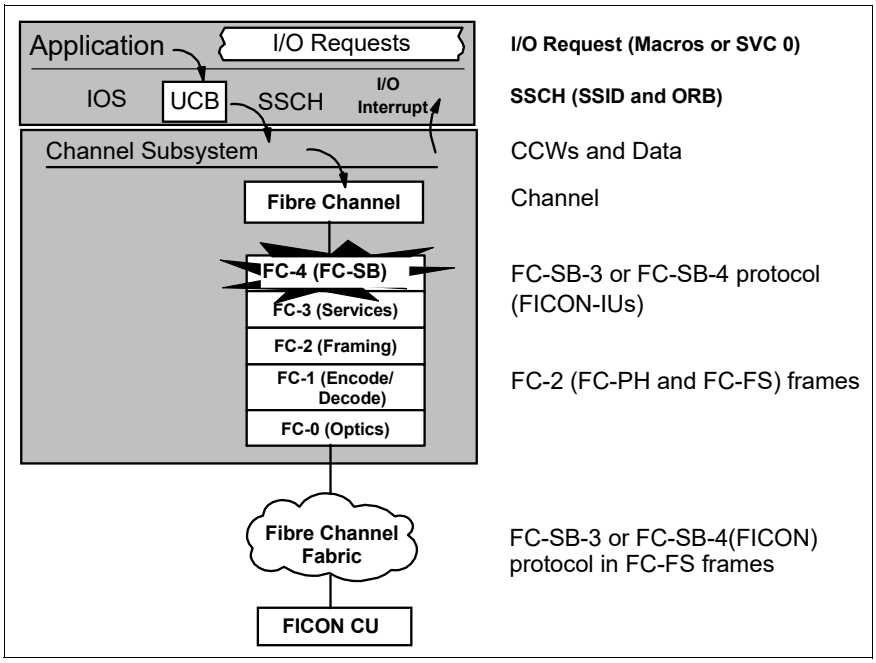


Figure 3-4 FICON channel operation flow

The most appropriate FICON channel is selected by the CSS. The FICON channel fetches channel programs (CCWs) prepared by the application, fetches data from memory (write) or stores data into memory (read), and presents the status of the operation to the application (I/O interrupt).

The z/Architecture channel commands, data, and status are packaged by the FICON channel into FC-SB-3 or FC-SB-4 (FC-4 layer) Information Units (IUs). IUs from several different I/O operations to the same or different control units and devices are multiplexed or demultiplexed by the FC-2 layer (framing). These FC-2 frames, with encapsulated FC-SB-3 or FC-SB-4 IUs, are encoded or decoded by the FC-1 layer (encode or decode) and sent to or received from the FC-0 fiber optic medium (optics).

On a FICON channel, CCWs are transferred to the control unit without waiting for the first command response from the control unit or for a CE/DE after each CCW execution. The device presents a logical *end* to the control unit after each CCW execution. If the last CCW of the CCW chain has been run by the device, the control unit presents CE/DE to the channel. Figure 3-5 shows a CCW operation on a FICON channel that uses CCW chaining.

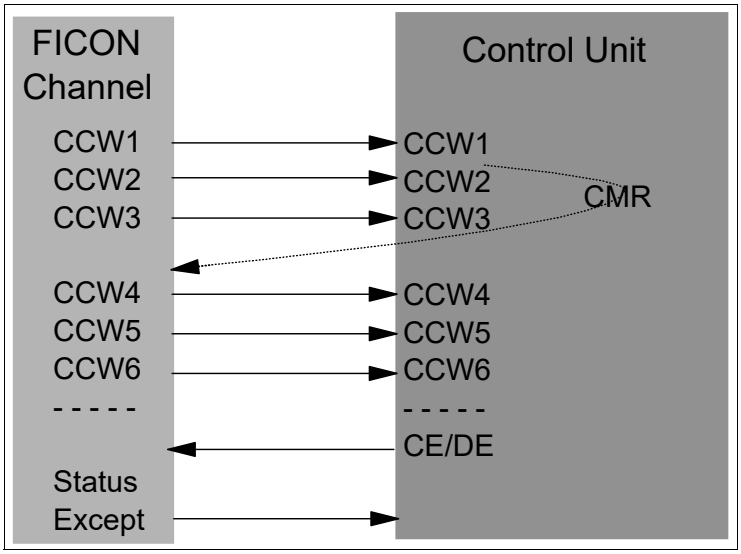


Figure 3-5 CCW chaining

3.2.2 High-Performance FICON for IBM zSystems

High-Performance FICON for IBM zSystems (zHPF) is an enhancement of the FICON channel architecture that is compatible with these protocols:

- Fibre Channel Physical and Signaling standard (FC-FS)
- Fibre Channel Switch Fabric and Switch Control Requirements (FC-SW)
- Fibre Channel Single-Byte-4 (FC-SB-4) standards

Using zHPF with the FICON channel, the z/OS operating system, and the control unit reduces the FICON channel overhead. This goal is achieved by protocol optimization and reducing the number of IUs processed, resulting in more efficient use of the fiber link.

The FICON Express32S, FICON Express16SA, FICON Express16S+, FICON Express16S, and FICON Express8S features support both the existing FICON architecture and the zHPF architecture. From the z/OS point of view, the existing FICON architecture is called *command mode*, and the zHPF architecture is called *transport mode*. A parameter in the ORB is used to determine whether the FICON channel is running in command or transport mode.

The mode that is used for an I/O operation depends on the control unit that is supporting zHPF and the settings in the z/OS operating system. An IECIOSxx parameter and SETIOS commands in z/OS can enable or disable zHPF. Support is also added for the D IOS, ZHPF system command to indicate whether zHPF is enabled, disabled, or not supported on the system.

During link initialization, both the channel and the control unit indicate whether they support zHPF. The Process Login (PRLI) support indicator is presented in response to the RNID Extended Link Services (ELS). If PRLI is supported, the channel sends a PRLI ELS. The PRLI response then indicates that zHPF is supported by the CU.

Similar to the existing FICON channel architecture described previously, the application or access method provides the channel program (CCWs) and parameters in the ORB. Bit 13 in word 1 of the ORB specifies how to handle the channel program in either command mode or transport mode.

The way that zHPF transport mode manages CCW operation is significantly different from the CCW operation for the existing FICON architecture command mode, as shown in Figure 3-6. In command mode, each single CCW is sent to the control unit for execution. In transport mode, all CCWs are sent over the link in one single frame to the control unit. Certain complex CCW chains are not supported by zHPF. Figure 3-6 shows an example of the optimization by a zHPF transport mode read operation.

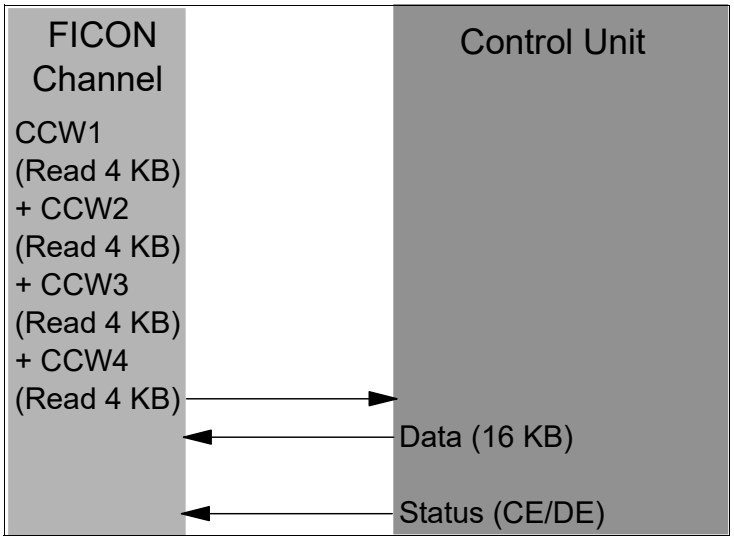


Figure 3-6 High-performance FICON read operation

The channel sends all the required CCWs and read operations of 4 KB of data in one single frame to the control unit. The control unit transfers the requested data over the link to the channel, followed by a CE/DE if the operation was successful. Less overhead is generated compared with the existing FICON architecture.

Figure 3-7 shows the same reduction of frames and open exchanges for a zHPF transport mode write operation.

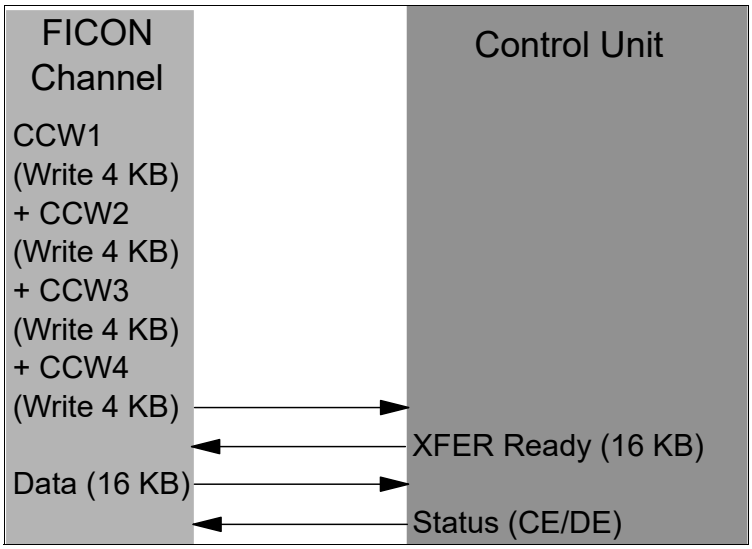


Figure 3-7 High-performance FICON write operation

The channel sends all the required CCWs and write operations of 4 KB of data in one frame to the control unit. The control unit responds with XFER when it is ready to receive the data. The channel then sends the 16 KB of data to the control unit. If the control unit successfully receives the data and finishes the write operation, the CE/DE status is sent by the control unit to indicate the completion of the write operation.

zHPF supports multitrack operations. It allows the channel to operate at rates that fully use the bandwidth of a FICON Express channel. The zHPF fully supports multiple tracks of data that can be transferred in a single operation.

For more information about the FICON channel protocol and zHPF, see *FICON Planning and Implementation Guide*, SG24-6497.

3.2.3 Platform and name server registration in FICON channel

All FICON features on the IBM zSystems platform support platform and name server registration to the fabric. That support exists only if the FICON feature is defined as channel path identifier (CHPID) type FC.

Information about the channels that are connected to a fabric, if registered, allow other nodes or SAN managers to query the name server to determine what is connected to the fabric. The following attributes are registered for the IBM zSystems platform:

- Platform information:
  - Worldwide node name (WWNN). This name is the node name of the platform and is the same for all channels that belong to the platform.
  - Platform type.
  - Host computer type.
  - Platform name. The platform name includes vendor ID, product ID, and vendor-specific data from the node descriptor.
- Channel information.

- Worldwide port name (WWPN).
- Port type (N_Port_ID).
- FC-4 types 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.

3.2.4 Open exchanges

An *open exchange*, part of FICON and FC terminology, represents an I/O operation in progress over the channel. Many I/O operations can be in progress over FICON channels at any one time. For example, a disk I/O operation might temporarily disconnect from the channel when performing a seek operation or while waiting for a disk rotation. During this disconnect time, other I/O operations can be managed as follows:

- Command mode open exchanges

  In command mode, the number of open exchanges is limited by the FICON Express feature. FICON Express32S, FICON Express16SA, FICON Express16S+, FICON Express16S, and FICON Express8S allow up to 64 open exchanges. One open exchange (an exchange pair) in command mode is the same as one I/O operation in progress.

- Transport mode open exchanges

  In transport mode, one exchange is sent from the channel to the control unit. Then, the same exchange ID is sent back from the control unit to the channel to complete the I/O operation. The maximum number of simultaneous exchanges that the channel can have open with the CU is 750 exchanges. The CU sets the maximum number of exchanges in the status area of the transport mode response IU. The default number is 64, which can be increased or decreased.

In addition, FICON channels can multiplex data transfer for several devices at the same time. This feature also allows workloads with low to moderate control unit cache hit ratios to achieve higher levels of activity rates per channel.

If the open exchange limit is reached, more I/O operations are refused by the channel, which can result in queues and retries by the operating system.

Extended distances

Degradation of performance at extended distances can be avoided by implementing an enhancement to the industry standard FICON architecture (FC-SB-3). This enhancement is a protocol for persistent IU pacing. Control units that use the architecture can increase the pace count, meaning the number of information units that are allowed to be underway between the channel and the control unit. Extended distance FICON channels remember the last pacing information and use this information for subsequent operations. This feature avoids performance degradation at the start of a new operation.

Information unit pacing helps to optimize the link use and simplifies the requirements for channel extension equipment because more commands can be in flight. Extended distance FICON is apparent to the operating systems and is applicable to all FICON features defined with CHPID type FC.

Modified Indirect Data Address Word

On IBM zSystems, the Modified Indirect Data Address Word (MIDAW) provides alternatives to using CCW data chaining in channel programs. The MIDAW facility has been added to z/Architecture and can coexist with the current CCW IDAW facility.

MIDAW decreases channel, fabric, and control unit overhead by reducing the number of CCWs and frames that are processed and allows scattering of data in memory for non-contiguous real pages. Although the CCW IDAW function requires all but the first and last IDAW in a list to deal with complete 2 thousand or 4 thousand units of data, the MIDAW facility allows page boundary crossing on either 2 thousand or 4 thousand boundaries. This feature allows access to data buffers anywhere in a 64-bit buffer space.

Figure 3-8 shows an example of MIDAW use.

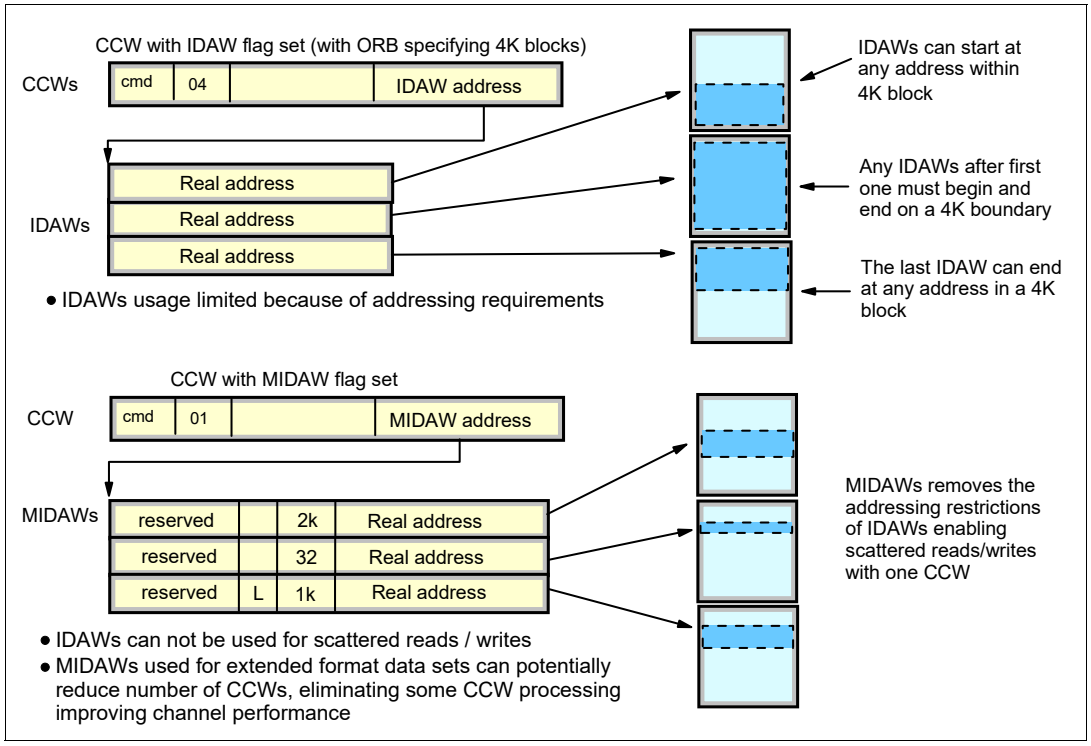


Figure 3-8 IDAW and MIDAW use

The use of MIDAWs is indicated by the MIDAW bit (flag bit 7) in the CCW. The last MIDAW in the list has a *last* flag set, indicated by L in Figure 3-8. MIDAW provides significant performance benefits, especially when processing extended format data sets with FICON channels.

FICON link

The transmission medium for the FICON interface is a fiber optic cable. Physically, it is a pair of optical fibers that provide two dedicated, unidirectional, serial-bit transmission lines. Information in a single optical fiber flows, bit by bit, in one direction. At any link interface, one optical fiber is used to receive data, and the other is used to transmit data. Full-duplex capabilities are used for data transfer. The Fibre Channel Standard (FCS) protocol specifies that for normal I/O operations frames flow serially in both directions, allowing several concurrent read and write I/O operations on the same link.

These are the link data rates:

- 2, 4, or 8 Gbps for FICON Express8S channels
- 4, 8, or 16 Gbps for FICON Express16S, FICON Express16S+ channels
- 8 or 16 Gbps for FICON Express16SA channels
- 8, 16, or 32 Gbps for FICON Express32S channels

Whether these link data rates can be achieved depends on many factors, such as transfer sizes and access methods used. The link speed capability is automatically negotiated between the system and FICON Director, and the director and CUs, and is apparent to the operating system and the application.

In general, the FICON channels and the FICON Director or CU communicate and agree on either a 2, 4, 8,16, or 32 Gbps (that is, 200 MBps, 400 MBps, 800 MBps, 1600 MBps, or 3200 MBps) link speed. This speed determination is based on the supported speeds in the system feature, FICON Director, and CU.

Note: The link speed is the theoretical maximum unidirectional bandwidth capability of the fiber optic link. The actual link data rate, whether it is measured in I/O operations per second or MBps, depends on the type of workload, fiber infrastructure, and storage devices in place.

FICON LX features use LX transceivers and 9 µm single mode fiber optic media (cables and trunks) for data transmission. FICON SX features use SX transceivers and 50 or 62.5 µm multimode fiber optic media (cables and trunks) for data transmission. A *transceiver* is a transmitter and receiver. The transmitter converts electrical signals to optical signals to be sent over the fiber optic media. The receiver converts optical signals to electrical signals to be sent through the system, director, or CU.

FICON to ESCON conversion

For more information about the requirements for connecting to ESCON devices, see Appendix B, “Channel conversion options” on page 171.

FICON and Fibre Channel switch

The Fibre Channel switch fabric (FC-SW) supports *packet-switching*. It allows up to 64 simultaneous concurrent I/O operations (read and write) from multiple FICON-capable systems to multiple FICON control units.

Figure 3-9 shows a conceptual view of frame processing in a switched point-to-point configuration for multi-system and multi-control unit environments.

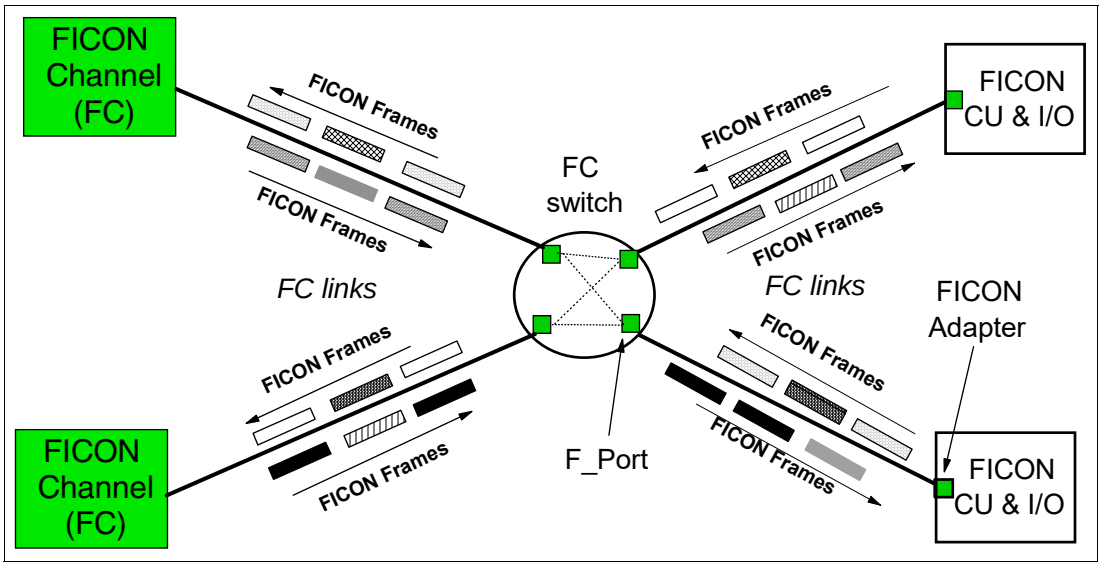


Figure 3-9 FICON switch function

FICON support of cascaded Fibre Channel Directors

FICON native channels on IBM zSystems platform support cascaded FICON Directors. This support is for a two-director configuration only. With cascading, a FICON native channel, or a FICON native channel with the channel-to-channel (FCTC) function, you can connect a system to a device or other system through two native connected Fibre Channel Directors. This cascaded director support is for all native FICON channels that are implemented on IBM zSystems platform.

FICON support of cascaded Fibre Channel Directors, sometimes referred to as *cascaded switching* or *two-switch cascaded fabric*, is for single-vendor fabrics only.

Cascaded support is important for disaster recovery and business continuity solutions, as shown in Figure 3-10 on page 55. It can provide high availability connectivity and the potential for fiber infrastructure cost savings for extended storage networks. FICON two-director cascaded technology allows for shared links, so it improves use of connected site resources and infrastructure. Solutions such as IBM GDPS can benefit from the reduced intersite configuration complexity that is provided by FICON support of cascaded directors.

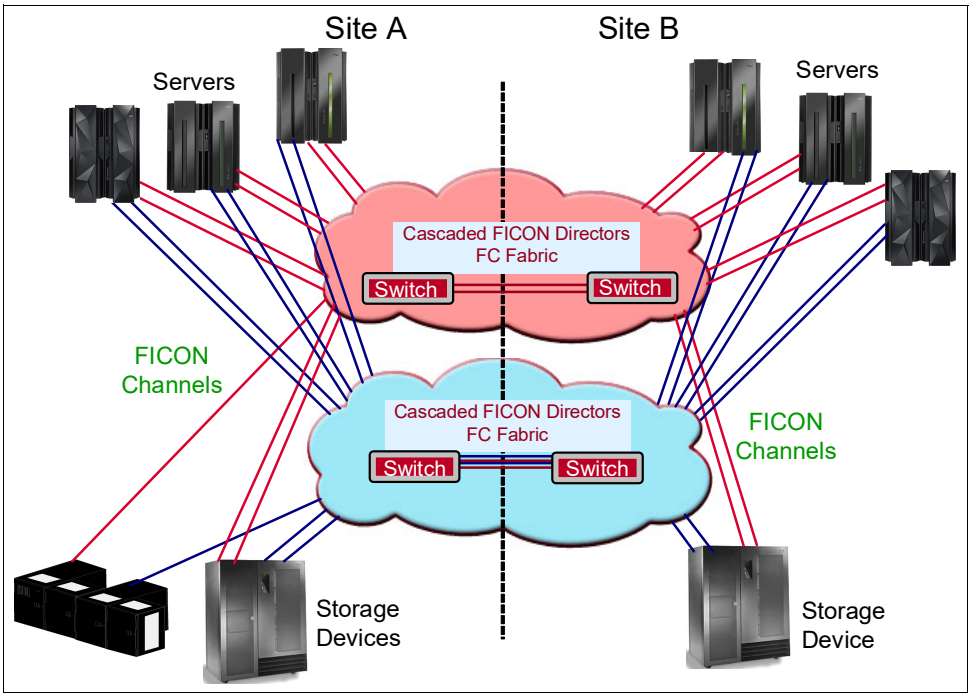


Figure 3-10 Two-site FICON connectivity with cascaded directors

Generally, organizations that have data centers that are separated between two sites can reduce the number of cross-site connections by using cascaded directors. Specific cost savings vary depending on the infrastructure, workloads, and size of data transfers. Further savings can be realized by reducing the number of channels and switch ports. Another important value of FICON support of cascaded directors is its ability to provide high-integrity data paths. The high-integrity function is an integral component of the FICON architecture when you are configuring FICON channel paths through a cascaded fabric.

To support the introduction of FICON cascaded switching, IBM worked with Fibre Channel Director vendors. IBM and the vendors worked to ensure that robustness in the channel-to-control unit path is maintained to the same high standard of error detection, recovery, and data integrity as the initial implementation of FICON.

End-to-end data integrity is maintained through the cascaded director fabric. Data integrity helps ensure that any changes to the data streams are always detected and that the data frames (data streams) are delivered to the correct endpoint. The endpoint is a FICON channel port or a FICON control unit port. For FICON channels, CRCs and longitudinal redundancy checks (LRCs) are bit patterns that are added to the data streams to allow for detection of any bit changes in the data stream. With FICON support of cascaded switching, integrity features are introduced within the FICON channel and the FICON cascaded switch fabric. This feature helps to ensure the detection and reporting of any incorrect cabling actions that occur within the fabric during operation that might cause a frame to be delivered to the wrong endpoint.

A FICON channel, when configured to operate with a cascaded switch fabric, requires that the switch fabric supports high integrity. During initialization, the FICON channel queries the switch fabric to determine that it supports high integrity. If it does, the channel completes the initialization process, allowing the channel to operate with the fabric.

After a FICON switched fabric is customized to support FICON cascaded directors and the required WWNN and domain IDs are added in the fabric membership list, the director checks that its ISLs are attached to the correct director before they are operational. If an accidental cable swap occurs, the director starts logical path testing, reporting, isolation, and recovery. The high integrity fabric feature for cascaded FICON Directors protects against miscabling and misdirecting of data streams, as shown in Figure 3-11. The checking process follows this sequence:

1. Channel initialization completes.
2. Later, miscabling occurs (for example, cables are swapped at a patch panel).
3. The director port enters `invalid attachment state` and sends a `state change` back to the IBM zSystems platform.
4. The IBM zSystems platform starts the channel logical path testing, reporting, isolation, and error recovery.
5. Any I/O requests to the invalid route are discarded until the error is corrected.
6. Data is protected. Channel initialization completes.

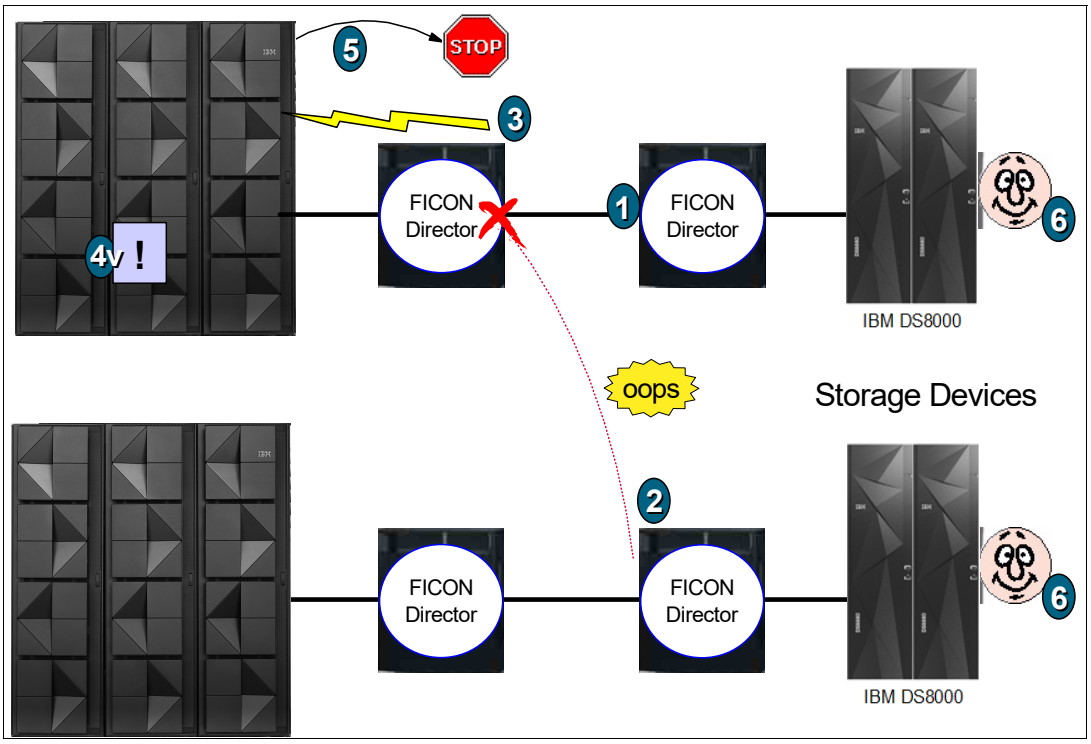


Figure 3-11 High-integrity fabric feature

High-integrity fabric architecture support includes the following capabilities:

- Fabric binding support

  The ability of the fabric to prevent a switch from being added to the fabric that is not configured to support the high integrity fabric. For example, all switches in the required fabric must be defined to the fabric by using a fabric membership list.

- Insistent domain IDs support

  This support does not allow a switch address to be automatically changed when a duplicate switch address is added to the enterprise fabric. It requires operator action to change a switch address.

FICON Dynamic Routing

FICON Dynamic Routing is a feature on the IBM zSystems platform that enables exploitation of SAN dynamic routing policies in the fabric to lower cost and improve performance for supporting I/O devices.

A static SAN routing policy typically 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).

Port-based routing (PBR) assigns the ISL routes statically based on first come, first served when a port starts a fabric login (FLOGI) to a destination domain. The ISL is round robin for assignment. Thus, I/O flow from the same incoming port to the same destination domain is always assigned the same ISL route, regardless of the destination port of each I/O. This process can result in some ISLs being overloaded while others are under utilized. The ISL routing table is changed every time that a IBM zSystems server undergoes a power-on-reset (POR). Because of this POR, the ISL assignment is unpredictable.

Device-base routing (DBR) assigns the ISL routes statically based on a hash of the source and destination port. That I/O flow from the same incoming port to the same destination is assigned with the same ISL route. Compared to PBR, DBR can better spread load across ISLs for I/O flow from the same incoming port to different destination ports within the same destination domain.

When using a static SAN routing policy, the FICON director has limited capability to assign ISL routes based on workload. There are also chances of ISL overloaded or under-utilization.

With dynamic routing, ISL routes are dynamically changed based on the Fibre Channel exchange ID, which is unique for each I/O operation. The ISL is assigned at I/O request time so that different I/Os from the same incoming port to the same destination port are assigned different ISLs. FICON dynamic routing is shown in Figure 3-12.

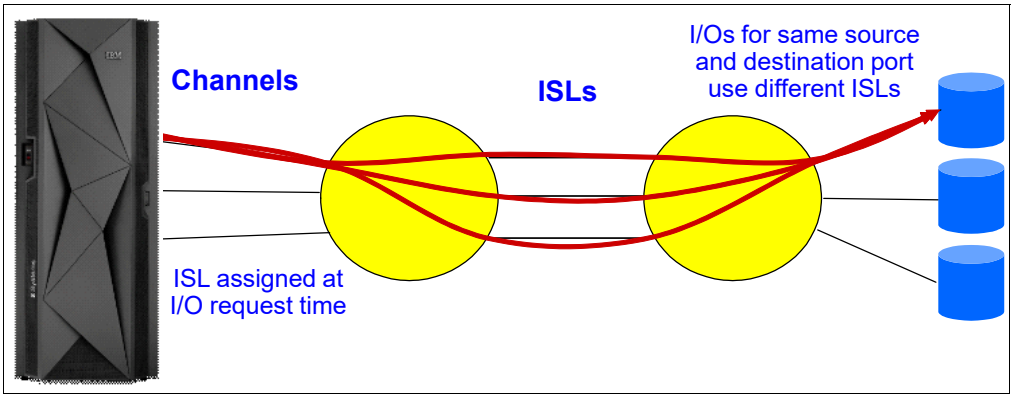


Figure 3-12 FICON dynamic routing

Dynamic routing (Brocade EBR or Cisco OxID) dynamically changes the routing between the channel and control unit based on the *Fibre Channel Exchange ID*. Each I/O operation has a unique exchange ID. The following are some of the benefits of FIDR:

- Reduces cost by allowing sharing of ISLs between FICON and FCP (PPRC or distributed)
- Provides better balanced I/O traffic between all available ISLs
- Improves utilization of switch and ISL hardware
- Provides easier management
- Allows for easier capacity planning for ISL bandwidth requirements
- Provides predictable, repeatable I/O performance

FICON control units and devices

The control unit receives commands from the channel, receives data from and transmits data to the channel, and controls execution of commands and data transfer at associated devices. The control unit attaches to one or multiple links through port interfaces.

FICON does not allow multiple physical control units, or control unit link interfaces, to be on the same link. A control unit can contain multiple images with dedicated or shared physical control unit facilities. The FICON I/O architecture provides addressing for these multiple CU images.

Channel device-addressing support

The FC device addressing is 32,000 devices[2] for the IBM zSystems platform.

Multiple image facility

MIF enables sharing FICON channels among LPARs that are running on IBM zSystems platform. For more information, see Chapter 2, “Channel subsystem overview” on page 17.

[2] Applies to the FICON Express32S, FICON Express16SA, FICON Express16S+, FICON Express16S, and FICON Express8S features that are supported by z/OS, z/VM, and Linux on IBM Z.

3.2.5 Spanned channels

Spanning is the ability to configure channels to multiple channel subsystems. When so defined, the channels can be transparently shared by any or all of the configured LPARs, regardless of the channel subsystem to which the LPAR is configured.

FICON channels can be spanned across multiple CSSs in the IBM zSystems platform. For more information about MIF and spanned channels, see 2.1.2, “Multiple CSSs” on page 19.

3.2.6 Control unit port

The CUP function allows z/OS to manage a FICON Director with the greater level of control and security. Host communication includes control functions like blocking or unblocking ports, monitoring, and error-reporting functions.

IBM Tivoli® System Automation for z/OS (SA for z/OS) includes support for FICON channels and FICON Directors. You can find additional information, updates, extensions, tools, and demonstrations at the IBM DevOps web page.

Before using SA for z/OS in your FICON environment, check the latest maintenance recommendations in the appropriate z/OS subset of the appropriate Preventive Service Planning (PSP) buckets (3931DEVICE, 8561DEVICE, 3906DEVICE, 3932DEVICE, 8562DEVICE, or 3907DEVICE) before implementation.

3.2.7 IBM z/OS Discovery and Automatic Configuration

IBM z/OS Discovery and Automatic Configuration (zDAC) is a function that is supported on the IBM zSystems platform. It automatically performs several I/O configuration definition tasks for new and changed disk and tape controllers that are connected to a FICON Director. It helps simplify I/O configurations of IBM zSystems CPCs running z/OS and helps reduce complexity and setup time.

The zDAC function is integrated into the existing HCD tool. A policy can be defined in the HCD according to the availability and bandwidth requirements, including parallel access volume (PAV) definitions, control unit numbers, and device number ranges. The zDAC proposed configurations are created as work input/output definition files (IODFs) that can be converted to production IODF and activated.

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 LCUs and devices, zDAC compares the discovered controller information with the current system configuration to determine delta changes to the configuration for a proposed configuration.

All new added or changed logical control units and devices are added to the proposed configuration with proposed control unit and device numbers and channel paths based on the defined policy. zDAC uses a channel path chosen algorithm to minimize single point of failure. zDAC applies to all FICON features that are supported on the IBM zSystems platform when configured as CHPID type FC.

3.3 Connectivity

The connectivity options for the FICON I/O interface environment are discussed in this section. Table 1-1 on page 11 lists the maximum number of FICON channels that are supported, based on each system.

Table 3-2 lists the available FICON features and their respective specifications. All FICON features use LC duplex connectors. For long wavelength FICON features that can use a data rate of 1 Gbps, mode-conditioning patch cables (MCP), either 50 or 62.5 MM, can be used. The maximum distance for this connection is reduced to 550 m at a link data rate of 1 Gbps. Details for each feature follow the table.

Table 3-2 IBM zSystems channel feature support

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Platform
FICON Express8S 10KM LX	0409	2, 4, or 8 Gbps	SM 9 µm	10 km	IBM z15[b], IBM z14[b]
FICON Express8S SX	0410	8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000)	IBM z15[b], IBM z14[b]
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000)	
		2 Gbps	MM 62.5 µm MM 50 µm	150 m (200) 300 m (500) 500 m (2000)	
FICON Express16S LX	0418	4, 8, or 16 Gbps	SM 9 µm	10 km	IBM z15[b], IBM z14[b]
FICON Express16S SX	0419	16 Gbps	MM 62.5 µm MM 50 µm	15 m (200) 35 m (500) 100 m (2000) 125 m (4700)	IBM z15[b], IBM z14[b]
		8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)	
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)	
FICON Express16S+ LX	0427	4, 8, or 16 Gbps	SM 9 µm	10 km	IBM z16[b], IBM z15[b], IBM z14

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Platform
FICON Express16S+ SX	0428	16 Gbps	MM 62.5 µm MM 50 µm	15 m (200) 35 m (500) 100 m (2000) 125 m (4700)	IBM z16[b], IBM z15[b], IBM z14
		8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)	
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)	
FICON Express16SA LX	0436	8 or 16 Gbps	SM 9 µm	10 km	IBM z16 A01[b] and IBM z15 T01
FICON Express16SA SX	0437	16 Gbps	MM 62.5 µm MM 50 µm	15 m (200) 35 m (500) 100 m (2000) 125 m (4700)	IBM z16 A01[b] and IBM z15 T01
		8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)	
FICON Express32S LX	0461	8, 16, or 32 Gbps	SM 9 µm	10 km[c]	IBM z16
FICON Express32S SX	0462	32 Gbps	MM 50 µm	20 m (500) 70 m (2000) 100 m (4700)	IBM z16
		16 Gbps	MM 62.5 µm MM 50 µm	15 m (200) 35 m (500) 100 m (2000) 125 m (4700)	
		8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)	

a. Minimum fiber bandwidths in MHz km for multimode fiber optic links are included in parentheses where applicable
b. Carry forward only.
c. FICON Express32S LX supports up to 5 km @ 32 Gbps for point-to-point connections.

The ports on a single FICON Express16S and FICON Express8S feature can be configured individually and can be defined in different channel modes (FC and FCP).

FICON Express32S, FICON Express16SA, and FICON Express16S+: With FICON Express32S, FICON Express16SA, and FICON Express16S+, both ports must be defined as the *same* channel type (either FC or FCP). Mixing channel types is not supported.

For more information about extended distances for FICON, see Chapter 10, “Extended distance solutions” on page 155.

3.3.1 FICON Express32S

FICON Express32S supports a link data rate of 32 Gbps with auto-negotiation to 16 Gbps and 8 Gbps for synergy with current-generation switches, directors, and storage devices. It works with your existing fiber optic cabling environment, both single-mode and multi-mode optical cables. The FICON Express32S feature running at end-to-end 32 Gbps link speeds provides reduced latency for large read and write operations and increased bandwidth compared to the FICON Express16SA and FICON Express16S+ features.

The FICON Express32S features have two independent ports. Each feature occupies a single I/O slot by using one CHPID per channel. Each channel supports 8 Gbps, 16 Gbps, or 32 Gbps link data rates with auto-negotiation. The FICON Express32S is supported on IBM z16.

FICON Express32S helps absorb large application and transaction spikes driven by large unpredictable AI and hybrid cloud workloads, and it is a key component of the IBM Fibre Channel Endpoint Security solution.

Each FICON Express32S port has a PCHID and can be defined as CHPID type FC or FCP. However, both ports must be same CHPID type. FICON Express32S CHPIDs can be defined as a spanned channel and can be shared among LPARs within and across CSSs.

The FICON Express32S features are installed in the PCIe+ I/O drawer and use Small Form-factor Pluggable (SFP) optics to permit each channel to be individually serviced during a fiber optic module failure. Traffic on the other channels on the same feature can continue to flow if a channel requires servicing.

The FICON Express32S features are ordered in two channel increments and are added concurrently. This concurrent update capability allows you to continue to run workloads through other channels when the FICON Express32S features are being added.

The FICON Express32S features are designed for connectivity to systems, switches, directors, disks, tapes, and printers, and can be defined in two ways:

- CHPID type FC:
  - Native FICON, zHPF, and FICON channel-to-channel (CTC) traffic
  - Supported in z/OS, IBM z/VM, IBM z/VSE V6.2 with PTFs, IBM z/TPF V1.1 with PTFs, and Linux on IBM Z[3]
- CHPID type FCP:
  - Fibre Channel Protocol traffic for communication with SCSI devices
  - Supported in IBM z/VM, z/VSE V6.2 with PTFs, and Linux on IBM Z

3.3.2 FICON Express16SA

The FICON Express16SA features have two independent ports. Each feature occupies a single I/O slot by using one CHPID per channel. Each channel supports 8 Gbps and 16 Gbps link data rates with auto-negotiation. The FICON Express16SA is supported on IBM z15 T01 and IBM z16 A01.

Each FICON Express16SA feature has two ports. Each port has a PCHID and can be defined as FC or FCP type. However, both ports must be same CHPID type. FICON Express16SA

[3] The support statements for Linux on IBM zSystems server also cover the KVM hypervisor on distribution levels that have KVM support.

CHPIDs can be defined as a spanned channel and can be shared among LPARs within and across CSSs.

The FICON Express16SA features can be placed only in the PCIe+ I/O drawer, and they use Small Form-factor Pluggable (SFP) optics to permit each channel to be individually serviced during a fiber optic module failure. Traffic on the other channels on the same feature can continue to flow if a channel requires servicing.

The FICON Express16SA features are ordered in two channel increments and are added concurrently. This concurrent update capability allows you to continue to run workloads through other channels when the FICON Express16SA features are being added.

The FICON Express16SA features are designed for connectivity to systems, switches, directors, disks, tapes, and printers, and can be defined in two ways:

- CHPID type FC:
  - Native FICON, zHPF, and FICON channel-to-channel (CTC) traffic
  - Supported in the z/OS, IBM z/VM V7.1 with corresponding APARs, IBM z/VSE V6.2 with PTFs, IBM z/TPF V1.1 with PTFs, and Linux on IBM Z[4]
- CHPID type FCP:
  - Fibre Channel Protocol traffic for communication with SCSI devices
  - Supported in IBM z/VM V6.4 and V7.1, z/VSE V6.2 with PTFs, and Linux on IBM Z

3.3.3 IBM Fibre Channel Endpoint Security

FICON Express32S (FC 0461 and FC 0462) and FICON Express16SA (FC 0436 and FC 0437) support Fibre Channel Endpoint Authentication and Encryption of data in-flight.

Based closely on Fibre Channel–Security Protocol-2 (FC-SP-2) 1 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 existing IBM solutions for key server infrastructure in the Storage System (for data at-rest encryption).

IBM Security Key Lifecycle Manager server provides shared secret key generation in a master and subordinate relationship between a Fibre Channel initiator (IBM zSystems) and the Storage target. The solution implements authentication and key management called IBM Secure Key Exchange (SKE).

Data that is in-flight (from or to IBM zSystems and IBM Storage) is encrypted when it leaves either endpoint (source) and decrypted at the destination. Encryption and decryption are done at the Fibre Channel adapter level. The operating system that is running on the host (IBM zSystems) is not involved in Endpoint Security-related operations. Tools are provided at the operating system level for displaying information about encryption status.

IBM Fibre Channel Endpoint Security is an orderable feature (FC 1146) for IBM z16 and IBM z15 T01[5], and it requires CPACF enablement (FC 3863), specific storage (DS8900), and FICON Express32S or FICON Express16SA features.

For more information and implementation details, see *IBM Fibre Channel Endpoint Security for IBM DS8900F and IBM Z*, SG24-8455, and this announcement letter.

[4] The support statements for Linux on IBM zSystems server also cover the KVM hypervisor on distribution levels that have KVM support.
[5] Not supported with IBM z15 T02.

3.3.4 FICON Express16S+

The FICON Express16S+ features have two independent ports. Each feature occupies a single I/O slot by using one CHPID per channel. Each channel supports 4 Gbps, 8 Gbps, and 16 Gbps link data rates with auto-negotiation.

These features can be carried forward during an upgrade or MES to IBM z16 A01, IBM z16 A02, IBM z16 AGZ and IBM z15 T01. They also can be ordered on a new build for IBM z15 T02, IBM z14 M0x, and IBM z14 ZR1. FICON Express16S+ features increase performance compared to FICON Express16S.

Each FICON Express16S+ feature has two ports. Each port has a PCHID and can be defined as FC or FCP type. However, both ports must be same CHPID type. FICON Express16S+ CHPIDs can be defined as a spanned channel and can be shared among LPARs within and across CSSs.

All FICON Express16S+ features are in the PCIe I/O drawer or PCIe+ I/O drawer[6] and use Small Form-factor Pluggable (SFP) optics to permit each channel to be individually serviced during a fiber optic module failure. Traffic on the other channels on the same feature can continue to flow if a channel requires servicing.

The FICON Express16S+ features are ordered in two channel increments and are added concurrently. This concurrent update capability allows you to continue to run workloads through other channels when the FICON Express16S+ features are being added.

The FICON Express16S+ features are designed for connectivity to systems, switches, directors, disks, tapes, and printers, and can be defined in two ways:

- CHPID type FC:
  - Native FICON, zHPF, and FICON channel-to-channel (CTC) traffic
  - Supported in the z/OS, IBM z/VM hypervisor, IBM z/VSE V6.2 (earlier z/VSE versions have no zHPF support), IBM z/Transaction Processing Facility (z/TPF), and Linux on IBM Z and the KVM hypervisor environments
- CHPID type FCP:
  - Fibre Channel Protocol traffic for communication with SCSI devices
  - Supported in IBM z/VM, z/VSE, and Linux on IBM Z and the KVM hypervisor environments

3.3.5 FICON Express16S

The FICON Express16S features have two independent ports. Each feature occupies a single I/O slot by using one CHPID per channel. Each channel supports 4 Gbps, 8 Gbps, and 16 Gbps link data rates with auto-negotiation.

These features are supported on IBM z15, IBM z14, and IBM z14 ZR1 with carry forward, and they are designed to deliver increased performance in comparison to FICON Express8S features. The FICON Express16S features include half the number of ports per feature in comparison with the FICON Express8 features. This design facilitates purchasing the correct number of ports to help satisfy your application requirements and to better optimize for redundancy.

All FICON Express16S features are in the PCIe I/O drawer or PCIe+ I/O drawer, and they use SFP optics to permit each channel to be individually serviced during a fiber optic module

[6] PCIe+ I/O drawer (FC 4023 on IBM z16, FC 4021 on IBM z15, and FC 4001 on IBM z14 ZR1) replaces the PCIe I/O drawer. All PCIe features that can be carried forward during an upgrade can be installed in the PCIe+ I/O drawer. PCIe I/O drawer (FC 4032) *cannot* be carried forward during an upgrade.

failure. Traffic on the other channels on the same feature can continue to flow if a channel requires servicing.

The FICON Express16S features are ordered in two channel increments and are added concurrently. This concurrent update capability allows you to continue to run workloads through other channels when the FICON Express16S features are being added.

FICON Express16S CHPIDs can be defined as a spanned channel and can be shared among LPARs within and across CSSs.

The FICON Express16S features are designed for connectivity to systems, switches, directors, disks, tapes, and printers and can be defined in two ways:

- CHPID type FC:
  - Native FICON, zHPF, and FICON CTC traffic
  - Supported in the z/OS, IBM z/VM hypervisor, IBM z/VSE V6.2 (earlier z/VSE versions have no zHPF support), IBM z/TPF, and Linux on IBM Z and the KVM hypervisor environments
- CHPID type FCP:
  - Fibre Channel Protocol traffic for communication with SCSI devices
  - Supported in IBM z/VM, z/VSE, and Linux on IBM Z and the KVM hypervisor environments

3.3.6 FICON Express8S

The FICON Express8S features have two independent ports. Each feature occupies a single I/O slot by using one CHPID per channel. Each channel supports 2 Gbps, 4 Gbps, and 8 Gbps link data rates with auto-negotiation.

These features are supported on the IBM zSystems platform (carry forward only on IBM z15, IBM z14, and IBM z14 ZR1), and are designed to deliver increased performance in comparison to FICON Express8 features. The FICON Express8S features include half the number of ports per feature in comparison with the FICON Express8 features. This design facilitates purchasing the correct number of ports to help satisfy your application requirements and to better optimize for redundancy.

All FICON Express8S features are in the PCIe I/O drawer or PCIe+ I/O drawer and use SFP optics to permit each channel to be individually serviced during a fiber optic module failure. Traffic on the other channels on the same feature can continue to flow if a channel requires servicing.

The FICON Express8S features are ordered in two channel increments and are added concurrently. This concurrent update capability allows you to continue to run workloads through other channels when the FICON Express8S features are being added.

FICON Express8S CHPIDs can be defined as a spanned channel and can be shared among LPARs within and across CSSs.

The FICON Express8S features are designed for connectivity to systems, switches, directors, disks, tapes, and printers, and can be defined in two ways:

- CHPID type FC:
  - Native FICON, zHPF, and FICON CTC traffic
  - Supported in the z/OS, IBM z/VM hypervisor, IBM z/VSE V6.2 (earlier z/VSE versions have no zHPF support), IBM z/TPF, and Linux on IBM Z and the KVM hypervisor environments

- CHPID type FCP:
  - Fibre Channel Protocol traffic for communication with SCSI devices
  - Supported in z/VM, z/VSE, and Linux on IBM Z and the KVM hypervisor environments

3.3.7 Qualified FICON and FCP products

For more information about IBM zSystems qualified FICON and FCP products, and products that support intermixing of FICON and FCP within the same physical FC switch or FICON Director (requires registration), see IBM Resource Link and the IBM System Storage Interoperability center.

On the left side of the web page, select **Library** and locate the listing for “Hardware products for servers” in the middle of the web page. Then, select **Switches and directors that are qualified for IBM zSystems FICON and FCP channels**.

3.3.8 Software support

For more information about the operating systems that are supported on IBM zSystems platform, see this website.

Note: Certain functions might require specific levels of an operating system, program temporary fixes (PTFs), or both. That information is provided when necessary within this chapter.

See the appropriate Preventive Service Planning (PSP) buckets (3931DEVICE, 8561DEVICE, 3906DEVICE, 3932DEVICE, 8562DEVICE, or 3907DEVICE) before implementation.

3.3.9 Resource Measurement Facility

IBM Resource Measurement Facility (IBM RMF) reporting is available for all FICON features. This application enables you to capture performance data through the following reports:

- Channel path activity report (of primary interest)
- Device activity report
- FICON Director activity report
- I/O queuing activity report

With these reports, you can analyze possible bandwidth bottlenecks to determine the cause. For more information about performance, see the IBM zSystems I/O connectivity web page.

3.4 References

The following publications contain information that is related to the topics covered in this chapter:

- *FICON I/O Interface Physical Layer,* SA24-7172
- *Planning for Fiber Optic Links,* GA23-1409
- *Linux on System z: Fibre Channel Protocol Implementation Guide*, SG24-6344
- *FICON Planning and Implementation Guide*, SG24-6497

For more information about Fibre Channel Standard publications, see the Technical Committee T11 website.

For more information about the SCSI Storage Interface standards, see the Technical Committee T10 website.

4

IBM zHyperLink Express

IBM zHyperLink is a technology that can provide up to a 5x reduction in I/O latency through an enhanced Synchronous I/O model. This goal is achieved by using a direct connection between an IBM zSystems platform and IBM Storage.

This chapter describes the zHyperLink connectivity option, which is offered with the IBM z16, IBM z15, and IBM z14, in conjunction with the IBM Storage DS888x and DS8900.

Note: The zHyperLink Express1.1 (FC 0451) is a technology refresh that includes the same functional and software requirements as the zHyperLink Express (FC 0431). Throughout this chapter, both features are referred to as zHyperLink or zHyperLink Express, unless otherwise specified.

This chapter includes the following topics:

- 4.1, “Description” on page 70
- 4.2, “zHyperLink elements” on page 70
- 4.3, “Connectivity” on page 71
- 4.4, “References” on page 72

4.1 Description

The zHyperLink technology was created to provide fast access to data through a low latency, short-distance direct connection between the IBM zSystems platform and IBM DS8000 storage system. This connection is intended to speed up Db2 for IBM z/OS transaction processing and improve activelog throughput, and improve VSAM read I/O requests as well.

zHyperLink Express is designed for up to 5X lower read latency than High-Performance FICON.[1] Working with the FICON SAN Infrastructure, zHyperLink can improve application response time by cutting I/O-sensitive workload response time by up to 50% without requiring application changes.

The zHyperLink Express feature in the IBM z16, IBM z15, IBM z14, and IBM z14 ZR1 allows you to make synchronous I/O requests for data that is in the cache of the IBM DS888x or newer model storage. This task is done by directly connecting a zHyperLink Express port in the IBM z16, IBM z15, IBM z14, or IBM z14 ZR1 to an I/O Bay zHyperLink port of the DS888x (or newer).

To better plan your zHyperLink implementation, use the PC-based tool IBM zBNA (IBM Z Batch Network Analyzer), which provides graphical and text reports. For more information, see this IBM Support web page.

Note: zHyperLink connections work in conjunction with FICON channels, they do not replace them (a FICON channel is required to “drive” the zHyperLink). Only z/OS and Extended Count Key Data (ECKD) are supported, and the z/OS image must run in an LPAR, not as a guest under IBM z/VM.

4.2 zHyperLink elements

Synchronous I/O is an I/O model type that is used as part of the zHyperLink technology. It allows the operating system to read data records synchronously, thus avoiding the scheduling and interrupt overhead associated with asynchronous operations.

A new Synchronous I/O command was defined to z/OS to synchronously read one or more data records. In addition, an option is provided to allow z/OS to initiate a Synchronous I/O command and return control to perform other processing requests.

When a traditional I/O operation is requested by a task to start the Input/Output Supervisor (IOS), the I/O operation is not handled by the central processor (CP) assigned to the z/OS LPAR. There are specialized components to do the job: SAPs and channel programs.

A SAP is a special processing unit (PU) that helps set up the I/O operation. The SAP finds an available channel path, but is not responsible for the data movement between the storage control unit and z/OS LPAR memory. The channel program communicates with control unit to manage the data movement. After the I/O operation completes, an I/O interrupt notifies the CP so that IOS can be run again.

For Synchronous I/O, the CP directly issues the I/O request to the storage control unit through a zHyperLink connection with a new z/OS I/O command and new hardware capabilities (the zHyperLink Express feature, IBM zSystems firmware, and a DS888x (or newer) I/O system

[1] Results observed in IBM internal lab. Actual performance can vary depending on the workload.

board). The SAP and channel subsystem are bypassed by using the Synchronous I/O model. I/O interrupts and I/O path-lengths are minimized, which results in improved performance.

4.3 Connectivity

The IBM zHyperLink Express takes up one slot in the PCIe I/O drawer or the PCIe+ I/O drawer, and has two ports. Both ports are on a single PCHID. The zHyperLink Express uses PCIe Gen3 technology, with x16 lanes that are bifurcated into x8 lanes. It is designed to support distances up to 150 meters at a link data rate of 8 GBps.

The following zHyperLink Express features are available:

- zHyperLink Express1.1 - Feature Code (FC) 0451: The feature is available with the IBM z15 and on IBM z16 for new build systems and carry forward.
- zHyperLink Express - FC 0431: Available on IBM z14 (can be carried forward to IBM z15 and IBM z16).

Figure 4-1 shows a zHyperLink Express feature in an IBM z15 connecting to a zHyperLink feature in the I/O Bay of a DS888x (or later). This configuration is valid for the IBM z14 and IBM z16.

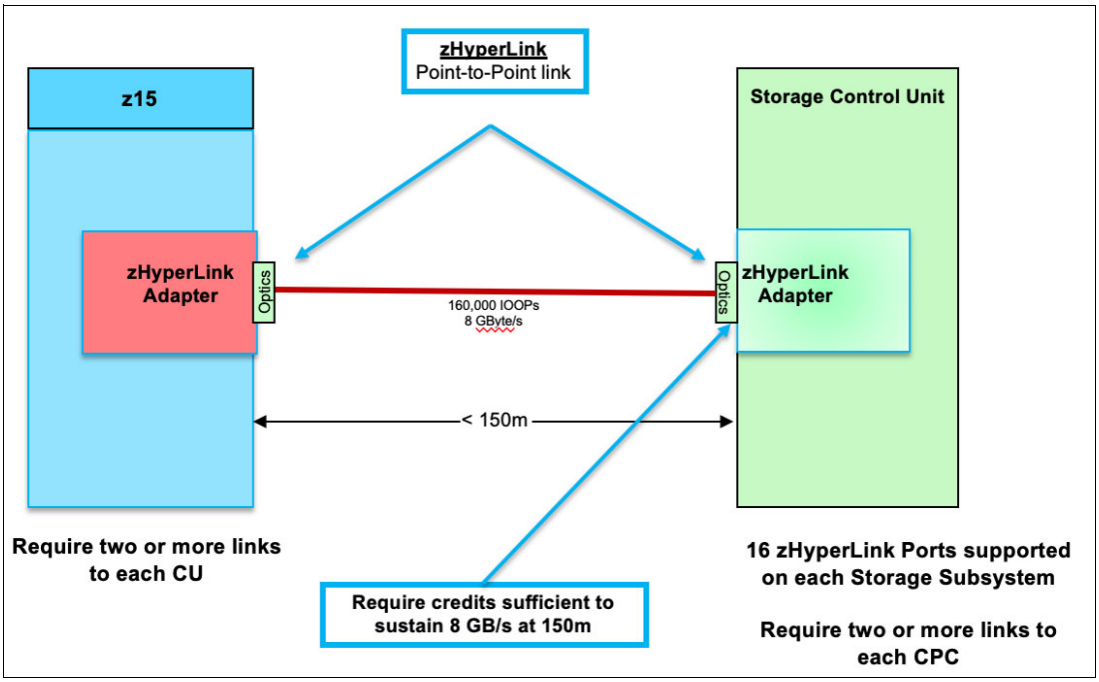


Figure 4-1 zHyperLink point-to-point connectivity

The zHyperLink Express requires 24x MTP-MTP cables to connect to a DS888x (or newer) and z/OS V2.1 or later.

The zHyperLink Express feature is an IBM-designed PCIe adapter. It is managed by using native z/PCI commands, with enhancements for Synchronous I/O.

Up to 16 zHyperLink Express adapters can be installed in an IBM z16, IBM z15, and IBM z14 for up to 32 links.

The zHyperLink Express feature works as native PCIe adapter that can be shared by multiple LPARs. Each port can support up to 127 Virtual Functions (VFs), with one or more VFs or PFIDs being assigned to each LPAR. This configuration supports a maximum of 254 VFs per adapter.

At a minimum, a DS888x with firmware R8.3 is required, with the I/O Bay system board updated to support the zHyperLink interface.

zHyperLink cabling requirements

A 24x MTP-MTP cable is required for each port of the zHyperLink Express feature. It is a single 24-fiber cable with Multi-fiber Termination Push-on (MTP) connectors. Internally, the single cable houses 12 fibers for transmit and 12 fibers for receive.

Two fiber options are available with specifications that support different distances for the zHyperLink Express features:

- Up to 150 m: OM4 50/125 micrometer multimode fiber optic cable with a fiber bandwidth wavelength of 4.7 GHz-km @ 850 nm.
- Up to 100 m: OM3 50/125 micrometer multimode fiber optic cable with a fiber bandwidth wavelength of 2.0 GHz-km @ 850 nm.

z/OS support for zHyperLink

Table 4-1 shows z/OS and Db2 minimum requirements to have READS and WRITES support with zHyperLink.

Table 4-1 z/OS Support for zHyperLink

Version	Reads support requires at a minimum	Writes support requires at a minimum
zHyperLink 1.1 zHyperLink	► z/OS V2.5 ► z/OS V2.4 with PTFs ► z/OS V2.3 with PTFs ► z/OS V2.2 with PTFs ► Db2 V11 plus PTFs	► z/OS V2.5 ► z/OS V2.4 with PTFs ► z/OS V2.3 with PTFs ► Db2 V12 plus PTFs

4.4 References

The following publications contain information that is related to the topics covered in this chapter:

- *IBM DS8000 and IBM Z Synergy DS8000: Release 9.2 and z/OS 2.5*, REDP-5186
- *IBM z16 (3931) Technical Guide*, SG24-8951
- *IBM z16 A02 and IBM AGZ Technical Guide*, SG24-8952
- *Getting Started with IBM zHyperLink for z/OS*, REDP-5493

5

IBM Open Systems Adapter Express

This chapter describes the IBM Open Systems Adapter (OSA) Express features that are available on IBM zSystems platforms. These features provide connectivity in 1000BASE-T Ethernet (100 and 1000 Mbps), gigabit Ethernet (GbE), 10-gigabit Ethernet, and 25-gigabit Ethernet environments.

Terminology: The terms *OSA* and *OSA-Express* are used throughout this chapter to simplify descriptions and discussion related to common functions and support across all OSA-Express features.

The chapter includes the following topics:

- 5.1, “Functional description” on page 74
- 5.2, “OSA capabilities” on page 81
- 5.3, “Connectivity” on page 97
- 5.4, “Summary” on page 105
- 5.5, “References” on page 106

5.1 Functional description

The OSA-Express features integrate network interface hardware and support many networking transport protocols. Every OSA-Express feature provides capabilities in the areas of:

- Standard Ethernet support
- Operating modes
- Connectivity options (bandwidth and data throughput)
- Availability, reliability, and serviceability

5.1.1 Standard Ethernet support

The following Ethernet standards are applicable for OSA-Express features (standard transmission schemes):

- IEEE 802.3 and IEEE 802.1
- Ethernet V2.0
- DIX V2

5.1.2 Operating modes

With IBM zSystems platform, the integration of a channel path with network ports makes the OSA a unique channel or CHPID type, which is recognized by the hardware I/O configuration on a port-by-port basis. The following CHPID types are discussed in further detail in subsequent sections:

- Queued Direct Input/Output (QDIO) (OSD)
- Non-Queued Direct Input/Output (OSE)[1]
- OSA Integrated Console Controller (OSC)[2]

Table 5-1 provides an overview of the types of traffic that are supported by IBM zSystems platforms for each CHPID type. It also indicates whether the Open Systems Adapter Support Facility (OSA/SF) is required to configure the OSA-Express ports that are based on the supported modes of operation (CHPID types).

Table 5-1 Supported CHPID types for OSA-Express features

CHPID type	SNA, APPN, and HPR traffic	IP traffic	TN3270E traffic	OSA/SF
OSD	No[a,b]	Yes	No	Optional
OSE	Yes	Yes	No	Required
OSC[c]	No	No	Yes	N/A

a. SNA over IP with the use of Enterprise Extender or TN3270. See 5.2.17, “Enterprise Extender” on page 94 and 5.2.18, “TN3270E server” on page 94.
b. Layer 2 support allows for non-IP protocols, such as SNA. See 5.2.13, “Layer 2 support” on page 89.
c. OSA-ICC (OSC Channel) supports Secure Sockets Layer for IBM zSystems platforms.

[1] IBM z16 is planned to be the last IBM zSystems Server to support OSE networking channels. IBM zSystems support for the System Network Architecture (SNA) protocol being transported natively out of the server by using OSA-Express 100BASE-T adapters configured as channel type “OSE” will be eliminated after IBM z16.
[2] See Chapter 6, “Console Communications (OSA-ICC)” on page 129 for details

Not all features support all CHPID types, see Table 5-3 on page 100, Table 5-4 on page 101 and Table 5-5 on page 102 for more information.

Open Systems Adapter Support Facility

OSA/SF is a host-based tool that is used to customize and manage OSA-Express features:

- OSA/SF is not required for the OSA feature that is configured for the QDIO mode or the default IP pass-through non-QDIO mode. However, it can be used for problem determination purposes.
- OSA/SF is a required facility when the OSA feature is being configured for shared non-QDIO mode and where SNA definitions are involved.
- With the IBM zSystems platform and OSA/SF, the HMC is enhanced to provide configuration, validation, activation, and display support for the OSA-Express7S, OSA-Express7S 1.2, OSA-Express6S features:
  - OSA/SF on the HMC is required for OSA-Express7S 1.2, OSA-Express7S, and OSA-Express6S features.
  - One OSA/SF application can communicate with all OSA features in an IBM zSystems platform.
  - OSA/SF communicates with an OSA feature through a device (type OSD) that is defined by using HCD or the Input/Output Configuration Program (IOCP). For more information, see 5.1.6, “OSA/SF support” on page 80.

QDIO versus non-QDIO

Figure 5-1 shows the I/O process when in QDIO mode rather than non-QDIO mode. I/O interrupts and I/O path-lengths are minimized, which result in improved performance versus non-QDIO mode, reduced system assist processor (SAP) use, improved response time, and reduced system use.

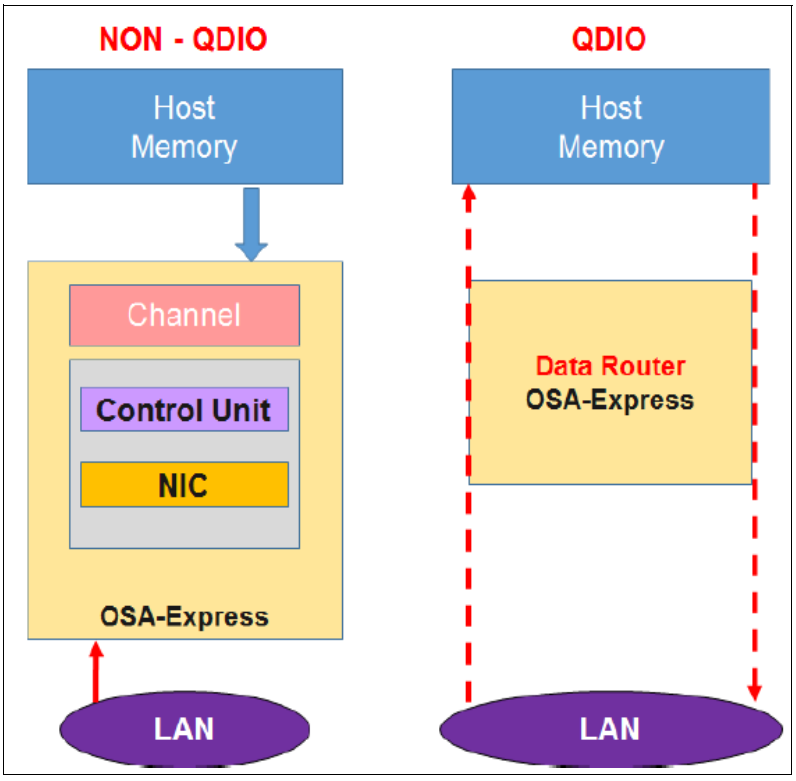


Figure 5-1 Non-QDIO data path versus QDIO data path

5.1.3 Non-QDIO mode (CHPID OSE)

Similar to any other channel-attached control unit and device, an OSA port can run channel programs (CCW chains) and present I/O interrupts to the issuing applications. For non-QDIO mode, the OSA ports are defined as channel type OSE. The non-QDIO mode requires the use of the OSA/SF for setup and customization of the OSA features.

The OSA-Express 1000BASE-T features support non-QDIO mode. This mode supports SNA, APPN, HPR, and IP network traffic simultaneously through the OSA port. This section describes the non-QDIO mode types.

IP network pass-through

In IP network pass-through mode, an OSA feature transfers data between an IP network stack to which it is defined and clients on an Ethernet 100/1000[3] Mbps LAN. The LAN is attached to the port on a 1000BASE-T feature and supports one of the following frame protocols:

- Ethernet II that uses the DEC Ethernet V 2.0 envelope
- Ethernet 802.3 that uses the 802.2 envelope with SNAP

For IP network pass-through mode, the default OSA Address Table (OAT) can be used. In that case, no configuration or setup is required.

SNA/APPN/HPR support

In this mode, an OSA feature acts as an SNA pass-through agent to clients that use the SNA protocol on the LAN that is directly attached to the OSA. If an OSA feature is running in the SNA mode, it is viewed by IBM VTAM® as an external communication adapter (XCA) that can use either switched or non-switched lines of communication.

5.1.4 QDIO mode (CHPID type OSD)

QDIO is a highly efficient data transfer mechanism that dramatically reduces system overhead and improves throughput by using system memory queues and a signaling protocol to directly exchange data between the OSA microprocessor and network software.

OSA-Express features use direct memory access (DMA) and a data router model to eliminate store and forward delays.

Also in QDIO mode, all OSA features dynamically receive configuration information from the host. This process reduces configuration and setup time, eliminates duplicate data entry, and reduces the possibility of data entry errors and incompatible definitions.

QDIO is the interface between the operating system and the OSA hardware.

The following components make up QDIO:

- Direct memory access (DMA)
- Data router
- Priority queuing
- Dynamic OSA address table building
- LPAR-to-LPAR communication
- Internet Protocol (IP) Assist functions

[3] OSA-Express7S and newer support only 1000 Mbps (no negotiation at lower speed).

QDIO supports IP and non-IP traffic with the OSA-Express features. The following features support two transport modes:

- Layer 2 (Link Layer) for IP (IPv4, IPv6) and non-IP (AppleTalk DECnet, IPX, NetBIOS, or SNA) traffic
- Layer 3 (Network Layer) for IP traffic only

For more information about the Layer 2 support, see 5.2.13, “Layer 2 support” on page 89.

Direct memory access

OSA and the operating system share a common storage area for memory-to-memory communication, reducing system overhead, and improving performance. Data can move directly from the OSA microprocessor to system memory and vice versa by using a store-and-forward technique in DMA. There are no read or write channel programs for data exchange. For write processing, no I/O interrupts must be handled. For read processing, the number of I/O interrupts is minimized.

Data router

The IBM Application Specific Integrated Circuit (ASIC) processor of the OSA feature handles packet construction, inspection, and routing. This feature allows packets to flow between host memory and the LAN at line speed without firmware intervention. With the data router, the store and forward technique in DMA is no longer used, which enables a direct host memory-to-LAN flow and avoids a hop. It reduces latency and increases throughput for standard frames (1492 bytes) and jumbo frames (8992 bytes).

Priority queuing

Priority queuing is supported by the QDIO architecture and introduced with the Service Policy Server (for z/OS environments only). It sorts outgoing IP message traffic according to the service policy that is set up for the specific priority that is assigned in the IP header.

This capability is an alternative to the best-effort priority assigned to all traffic in most IP networks. Priority queuing allows the definition of four different priority levels for IP network traffic through the OSA features defined for QDIO. For example, the highest priority can be granted to interactive communications, and the lowest can be granted to batch traffic, with two more categories in between based on particular user groups or projects.

QDIO uses four write (outbound) queues and one read (inbound) queue for each IP network stack that is sharing the OSA feature.

OSA signals the z/OS Communications Server when there is work to do. z/OS Communications Server puts outbound packets in one of the four queues, based on priority settings.

At a certain time, z/OS Communications Server signals the OSA feature that there is work to do. The OSA feature searches the four possible outbound queues by priority and sends the packets to the network, giving more priority to queues 1 and 2, and less priority to queues 3 and 4. For example, if there is data on every queue, queue 1 is served first, then portions of queue 2, then fewer portions of queue 3, then even fewer portions of queue 4, and then back to queue 1. This process means that four transactions are running across the four queues. Over time, queue 1 finishes first, queue 2 finishes second, and so on.

Note: With OSA-Express, priority queuing is enabled by default. This feature reduces the total number of supported IP network stacks and devices. For more information, see “Maximum IP network stacks and subchannels” on page 80.

Dynamic OSA Address Table update

With QDIO, the dynamic OSA Address Table (OAT) update process simplifies installation and configuration tasks. The definition of IP addresses is done in one place, the IP network profile, thus removing the requirement to enter the information into the OAT by using the OSA/SF.

The OAT entries are dynamically built when the corresponding IP device in the IP network stack is started.

At device activation, all IP addresses contained in the IP network stack's IP HOME list are downloaded to the OSA port. Corresponding entries are built in the OAT. Subsequent changes to these IP addresses cause an update of the OAT.

LPAR-to-LPAR communication

Access to an OSA port can be shared among the system images that are running in the logical partitions (LPARs) to which the channel path is defined to be shared. Also, access to a port can be shared concurrently among IP network stacks in the same LPAR or in different LPARs.

When sharing ports, an OSA port operating in QDIO mode can send and receive IP traffic between LPARs without sending the IP packets to the LAN and then back to the destination LPAR.

For outbound IP packets, the OSA port uses the next-hop IP address within the packet to determine where it is sent. If the next-hop IP address was registered by another IP network stack that is sharing the OSA port, the packet is sent directly to that IP network stack, not onto the LAN. This feature makes the forwarding of IP packets possible within the same host system.

Internet Protocol Assist functions

OSA QDIO helps IP processing and offloads the IP network stack functions for the following processes:

- Multicast support (see 5.2.7, "IP network multicast and broadcast support" on page 86)
- Broadcast filtering (see 5.2.7, "IP network multicast and broadcast support" on page 86)
- Building MAC and LLC headers
- Address Resolution Protocol (ARP) processing (see 5.2.8, "Address Resolution Protocol cache management" on page 86)
- Checksum offload (see 5.2.10, "Checksum offload support for z/OS and Linux on IBM Z" on page 87)

QDIO functions

The QDIO functions that are described in this section are supported on the IBM zSystems platform.

IP network functions

The following IP network functions are available:

- Large send for IP network traffic for OSA-Express (for more information, see 5.2.4, "Large send for IP network traffic" on page 82)
- 640 IP network stacks (for more information, see "Maximum IP network stacks and subchannels" on page 80)

Hardware assists for z/VM guests

Complementary virtualization technology is available that includes these capabilities:

- QDIO Enhanced Buffer-State Management (QEBSM). Two hardware instructions help eliminate the overhead of hypervisor interception.
- Host Page-Management Assist (HPMA). An interface to the IBM z/VM operating system main storage management function to allow the hardware to assign, lock, and unlock page frames without z/VM hypervisor assistance.

These hardware assists allow a cooperating guest operating system to start QDIO operations directly to the applicable channel, without interception by z/VM, which helps improve performance. Support is integrated in IBM zSystems Licensed Internal Code.

QDIO Diagnostic Synchronization for z/OS

QDIO Diagnostic Synchronization is exclusive to IBM zSystems and the OSA-Express features when configured as CHPID type OSD (QDIO). It provides the system programmer and network administrator with the ability to coordinate and simultaneously capture both operating system (software) and OSA (hardware) traces at the same instance of a system event.

This process allows the host operating system to signal the OSA-Express features to stop traces and capture the current trace records. By using existing tools (traps) and commands, the operator can capture both hardware and software traces at the same time and then correlate the records during post processing.

OSA-Express Network Traffic Analyzer for z/OS

The OSA-Express Network Traffic Analyzer (ENTA) is exclusive to IBM zSystems and the OSA-Express features when configured as CHPID type OSD (QDIO). It allows trace records to be sent to the host operating system to improve the capability to capture data for both the system programmer and the network administrator. This function allows the operating system to control the sniffer trace for the LAN and capture the records in host memory and storage. It uses existing host operating system tools to format, edit, and process the sniffer records.

5.1.5 OSA addressing support

This section describes the maximum number IP addresses, MAC addresses, and subchannels that are supported by the OSA features.

Maximum IP addresses per OAT

The OAT is a component of an OSA feature's configuration. An OAT entry defines the data path between an OSA feature port and an LPAR and device unit address. That is, it manages traffic through the OSA CHPID.

OSA-Express features support up to 4096 IP addresses per port.

When the OSA port is defined in QDIO mode, the OAT entries are built and updated dynamically.

Maximum number of MAC addresses

When configured as an OSD CHPID type, up to 4096 (IBM z14 and later) Media Access Control (MAC) or virtual MAC (VMAC) addresses are supported for each port of the OSA feature. Included in the maximum number of MAC addresses is the "burned-in" MAC address of the OSA port.

The MAC or VMAC addresses are added to the Layer 2 table of the OAT when the IP network stacks (in which the addresses are defined) are started.

For more information, see 5.2.13, “Layer 2 support” on page 89.

Maximum IP network stacks and subchannels

A subchannel is a logical representation of a device. One subchannel is assigned for each device that is defined to the LPAR. Therefore, if an OSA CHPID is being shared across 15 LPARs and one device is defined, that device uses 15 subchannels.

The following maximum number of IP network stacks and subchannels are supported:

- OSA port in non-QDIO mode (CHPID type OSE)

  An OSA port in non-QDIO mode can support up to 120 IP network stacks and 240 subchannels for all IBM zSystems platforms.

- OSA port in QDIO mode (CHPID type OSD)

  The OSA features support 640 IP network stack connections per dedicated CHPID, or 640 total stacks across multiple LPARs when defined as a shared or spanned CHPID. The maximum number of subchannels that are allowed is 1920 (1920 subchannels / 3 = 640 stacks).

Note: By default, OSA-Express features have multiple priorities for outbound queues enabled (four QDIO priorities). Therefore, the maximum number of supported subchannels is reduced to 480 (1920 subchannels / 4 = 480 subchannels), which reduces the total number of supported IP network stacks to 160 (480 subchannels / 3 = 160 stacks). Priority queues can be disabled through HCD or IOCP. For example, in IOCP use the CHPARM=02 value to disable priority queuing.

5.1.6 OSA/SF support

OSA/SF includes a graphical user interface (GUI) that is based on Java to support the client application. The Java GUI is independent of any operating system or server, and is expected to operate wherever the current Java runtime support is available.

Use of the GUI is optional. A REXX command interface is also included with OSA/SF. OSA/SF is integrated in z/OS, z/VM, and z/VSE, and runs as a host application. For OSA/SF, Java GUI communication is supported through IP networks only. This version of OSA/SF is not being offered as a separately licensed product.

The HMC has been enhanced to use OSA/SF function for the OSA-Express features. OSA/SF on the HMC or the OSA/SF in the operating system component can be used for the OSA-Express features(6S, 7S, 7S 1.2). For the OSA-Express7S 1.2, OSA-Express7S and OSA-Express6S features, OSA/SF on the HMC is required.

OSA/SF is used primarily for the following purposes:

- Manage all OSA ports.
- Configure all OSA non-QDIO ports.
- Configure local MAC addresses.
- Display registered IPv4 addresses (in use and not in use). This display is supported on IBM zSystems platforms for QDIO ports.

- Display registered IPv4 or IPv6 Virtual MAC and VLAN ID associated with all OSA Ethernet features configured as QDIO Layer 2.
- Provide status information about an OSA port and its shared or exclusive use state.

This support is applicable to all OSA features on IBM zSystems platforms.

OSA/SF is not always required to customize an OSA feature. However, it can be used to gather operational information, to help with problem determination to monitor and control ports. The OSA/SF Query function provides performance information about the OSA CHPIDs. As shown in Table 5-1 on page 74, OSA/SF is not required to configure the OSA features in any operating modes except OSE.

5.2 OSA capabilities

This section describes the capabilities that use the OSA-Express features.

5.2.1 Virtual IP address

In the IP network environment, virtual IP address (VIPA) frees IP network hosts from dependence on a particular network attachment, allowing the establishment of primary and secondary paths through the network. VIPA is supported by all of the OSA features.

An IP address traditionally ties to a physical link at one end of a connection. If the associated physical link goes down, it is unreachable. The virtual IP address exists only in software and has no association to any physical link. The IP network stack is the destination IP address rather than the network attachment.

VIPA provides for multiple IP addresses to be defined to an IP network stack, allowing fault-tolerant, redundant backup paths to be established. Applications become insensitive to the condition of the network because the VIPAs are always active. This configuration enables users to route around intermediate points of failure in the network.

VIPA takeover and takeback

Because a VIPA is associated with an IP network stack and not a physical network attachment, it can be moved to any IP network stack within its network. If the IP network stack that the VIPA is on fails (because of an outage), the same VIPA can be brought up automatically on another IP network stack (VIPA takeover). This process allows users to reach the backup server and applications. The original session between the user and original server is not disrupted. After the failed IP network stack is restored, the same VIPA can be moved back automatically (VIPA takeback).

5.2.2 Primary/secondary router function

The primary/secondary router function enables an OSA port to forward packets with unknown IP addresses to an IP network stack for routing through another IP network interface, such as IBM HiperSockets or another OSA feature.

For an OSA port to forward IP packets to a particular IP network stack for routing to its destination, the PRIRouter must be defined on the DEVICE statement in the IP network profile.

If the IP network stack that has an OSA port that is defined as PRIRouter becomes unavailable, the following process occurs: A second IP network stack that is defined as the secondary router (SECRouter on the DEVICE statement in the IP network profile) receives the packets for unknown IP addresses.

For enhanced availability, the definition of one primary router and multiple secondary routers for devices on an OSD-type CHPID is supported. However, only one secondary router is supported for devices on an OSE-type CHPID.

Important Sharing a single OSA port (with a single MAC address) can fail in load-balancing solutions. A workaround is to use generic routing encapsulation (GRE) or network address translation (NAT), which can have a negative effect on performance. A Layer 3 virtual MAC is a function that is available on IBM zSystems platforms with OSA features that allows multiple MAC addresses on a single OSA port. For more information, see 5.2.16, “Layer 3 VMAC for z/OS” on page 93.

5.2.3 IPv6 support

Internet Protocol version 6 (IPv6) is supported by the OSA features when configured in QDIO mode. IPv6 is the protocol that is designed by the Internet Engineering Task Force (IETF) to replace Internet Protocol version 4 (IPv4). IPv6 provides improved traffic management in the following areas:

- 128-bit addressing

  This improvement eliminates all practical limitations on global address ability. Private address space, along with the NATs used between a private intranet and the public internet, are no longer needed.

- Simplified header formats

  This improvement allows for more efficient packet handling and reduced bandwidth cost.

- Hierarchical addressing and routing

  This feature keeps routing tables small and backbone routing efficient by using address prefixes rather than address classes.

- Improved support for options

  This improvement changes the way IP header options are encoded, allowing more efficient forwarding and greater flexibility.

- Address auto-configuration

  This change allows stateless IP address configuration without a configuration server. In addition, IPv6 brings greater authentication and privacy capabilities through the definition of extensions, and integrated quality of service (QoS) through a traffic class byte in the header.

5.2.4 Large send for IP network traffic

Large send (also referred to as *TCP segmentation offload*) can improve performance by offloading TCP packet processing from the host to the OSA-Express features that are running in QDIO mode. Offload allows the host to send large blocks of data (up to 64 KB) directly to the OSA. The OSA then fragments those large blocks into standard Ethernet frames (1492 bytes) to be sent on the LAN (see Figure 5-2).

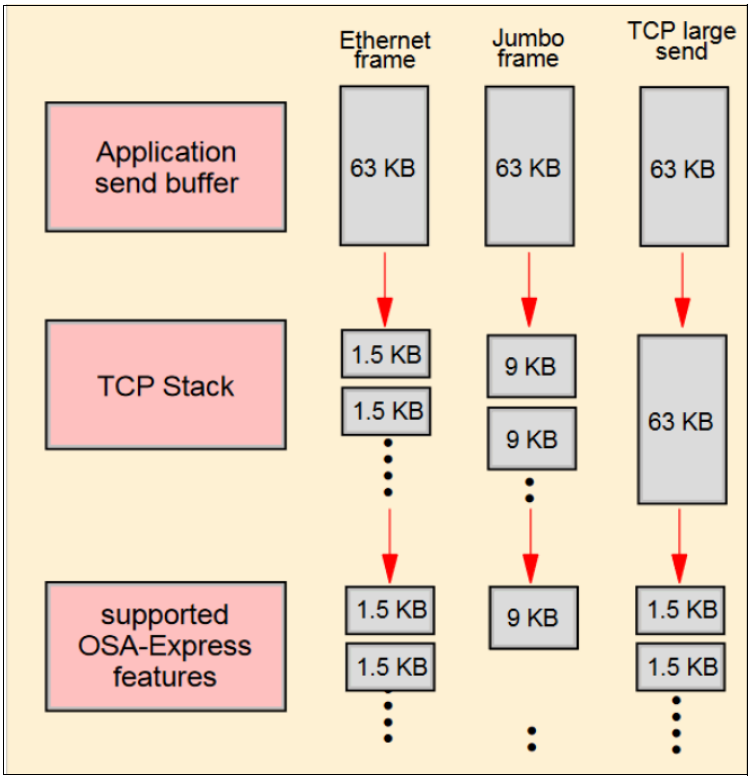


Figure 5-2 Large send versus standard Ethernet and jumbo frame sizes

Large send support reduces host processor use, so it returns processor cycles for application use and increases network efficiencies. For OSA-Express features, large send supports outbound IPv4 traffic only and applies solely to unicasts. Large send support for IPv6 packets applies to the OSA-Express features (CHPID types OSD) that are available on the IBM zSystems platform.

Note: Large send for IPv6 packets is not supported for LPAR-to-LPAR packets.

5.2.5 VLAN support

A virtual local area network (VLAN) is supported by the OSA-Express features when they are configured in QDIO mode. This support is applicable to z/OS, z/VM, and Linux on IBM Z[4] environments.

The IEEE standard 802.1q describes the operation of virtual bridged local area networks. A VLAN is defined to be a subset of the active topology of a local area network. The OSA features provide for the setting of multiple unique VLAN IDs per QDIO data device. They also provide for both tagged and untagged frames to flow from an OSA port. The number of VLANs supported is specific to the operating system.

VLANs facilitate easy administration of logical groups of stations, which can communicate as though they were on the same LAN. They also facilitate easier administration of moves, adds, and changes in members of these groups. VLANs are also designed to provide a degree of low-level security by restricting direct contact with a server to only the set of stations that comprise the VLAN.

[4] For more information about the KVM hypervisor, see:
https://www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/lnz_r_kvm.html

On IBM zSystems platforms where multiple IP network stacks exist and potentially share one or more OSA features, VLAN support provides a greater degree of isolation (see Figure 5-3).

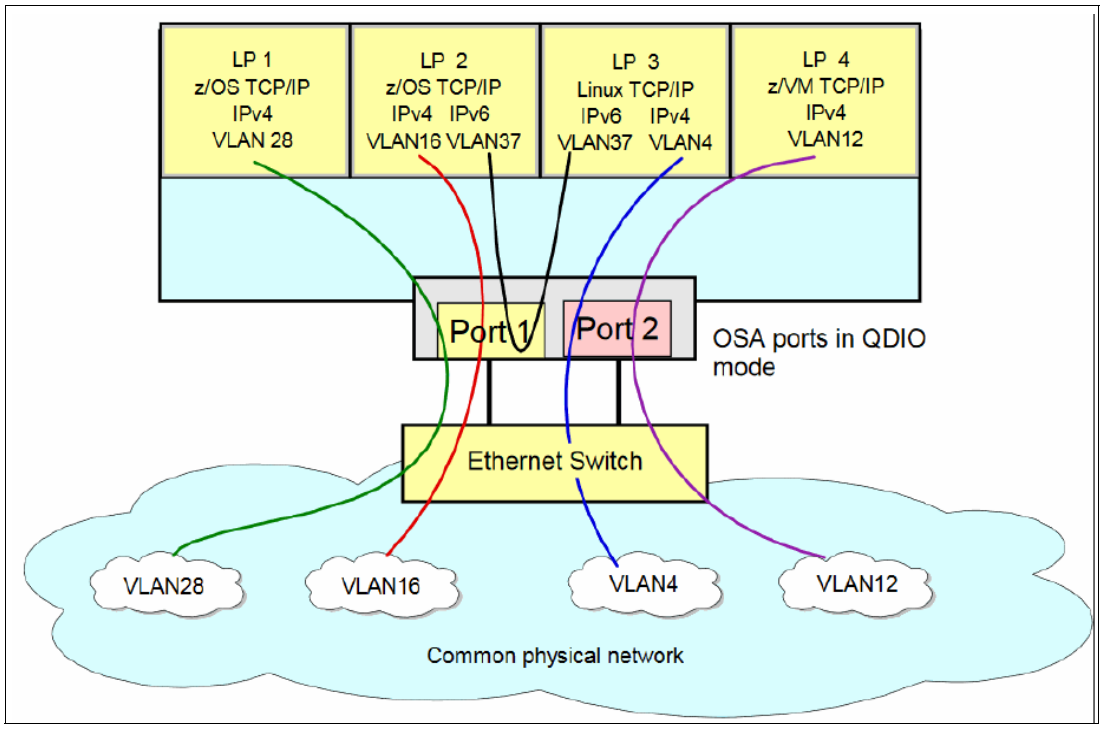


Figure 5-3 VLAN support

VLAN support for z/OS

Full VLAN support is offered for all OSA Ethernet features that are available on IBM zSystems platforms. The IBM z/OS Communications Server supports VLAN IDs. z/OS support is offered for up to 32 global VLAN IDs per OSA port for IPv4 and IPv6.

VLAN support for z/VM

z/VM uses the VLAN technology and conforms to the IEEE 802.1q standard. Support is offered for one global VLAN ID per OSA port for IPv4 and IPv6. Each port can be configured with a different VLAN ID.

VLAN support for Linux on IBM Z

VLAN support in a Linux on IBM Z environment is available for the OSA Ethernet features that operate in QDIO mode.

VLAN support of GARP VLAN Registration Protocol

The Generic Attribute Registration Protocol (GARP) (see Figure 5-4 on page 85) VLAN Registration Protocol (GVRP) is defined in the IEEE 802.1p standard for the control of IEEE 802.1q VLANs. It can be used to simplify networking administration and management of VLANs.

With GVRP support, an OSA-Express port can register or unregister its VLAN IDs with a GVRP-capable switch and dynamically update its table as the VLANs change. Support of GVRP is exclusive to IBM zSystems. It is applicable to all the OSA-Express features when in QDIO mode (CHPID type OSD), and it is supported by z/OS and z/VM.

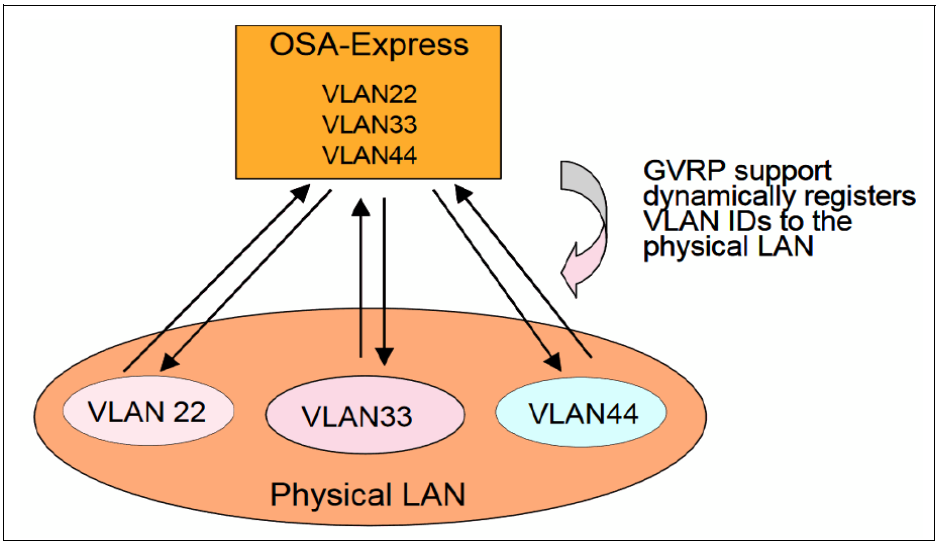


Figure 5-4 GVRP support

5.2.6 SNMP support for z/OS and Linux on IBM Z

Simple Network Management Protocol (SNMP) is supported for OSA features when configured in the QDIO mode (CHPID type OSD). The OSA features LIC includes the following support for the OSA SNMP subagents:

- Get and GetNext requests

  This support applies to all OSA features that are supported on IBM zSystems platform.

- dot3StatsTable

  Ethernet data for dot3StatsTable applies to all of the Ethernet features that are supported on IBM zSystems platform. It implements the SNMP EtherLike Management Information Base (MIB) module of RFC 2665, which provides statistics for Ethernet interfaces. These statistics can help in the analysis of network traffic congestion.

- Performance data

  This support applies to all of the OSA features that are supported on IBM zSystems platform. The performance data reflects the OSA usage.

- Traps and Set

  This support applies to all of the OSA features that are supported on IBM Z.

SNMP support for LAN channel station (LCS) applies to all of the OSA features that are supported on IBM Z, along with IP network applications only. It supports the same SNMP requests and alerts that are offered in QDIO mode (Get, GetNext, Trap, and Set) and is exclusive to the z/OS environment.

Tip: You can subscribe to the “OSA-Express Direct SNMP MIB module” document through Resource Link to receive email notifications of document changes.

OSA/SF is not required to manage SNMP data for the OSA features. An SNMP subagent exists on an OSA feature, which is part of a direct path between the z/OS or Linux on IBM Z master agent (IP network stacks) and an OSA-Express MIB.

The OSA features support an SNMP agent by providing data for use by an SNMP management application, such as IBM Tivoli NetView for z/OS. This data is organized into MIB tables that are defined in the IP network enterprise-specific MIB and standard RFCs. The data is supported by the SNMP IP network subagent (see Figure 5-5).

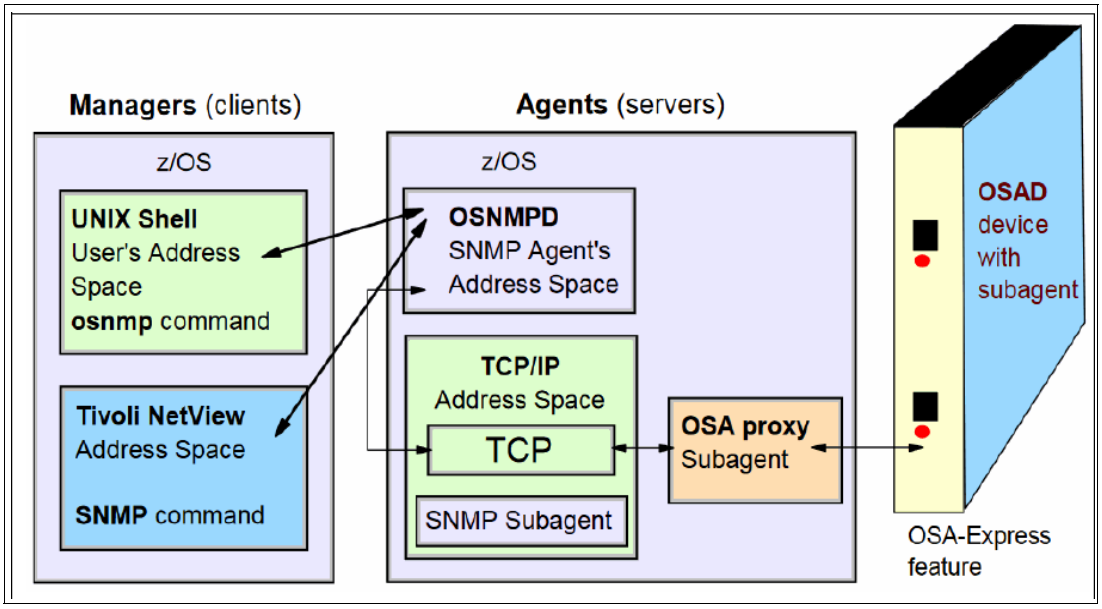


Figure 5-5 SNMP support: z/OS example

5.2.7 IP network multicast and broadcast support

Multicast and broadcast support are part of the Internet Protocol Assist (IPA) function of the OSA feature.

Multicast support

For sending data to multiple recipients, OSA features support IP multicast destinations only in QDIO or IP pass-through mode.

IP network broadcast support for z/OS, z/VM, and Linux on IBM Z

Broadcast support is included for all of the OSA features when configured in QDIO mode and supporting the Routing Information Protocol (RIP) version 1. Broadcast is also supported for all of the OSA features when carrying IP network traffic and configured in the non-QDIO mode (LAN Channel Station in LCS mode).

A broadcast simultaneously transmits data to more than one destination. Messages are transmitted to all stations in a network (for example, a warning message from a system operator). The broadcast frames can be propagated through an OSA feature to all IP network applications that require broadcast support, including applications that use RIP V1.

5.2.8 Address Resolution Protocol cache management

The query and purge Address Resolution Protocol (ARP) enhancements are supported for all OSA features when configured in QDIO mode. The OSA feature maintains a cache of recently acquired IP-to-physical address mappings (or *bindings*). When the binding is not found in the ARP cache, a broadcast (an ARP request: “How can I reach you?”) to find an address mapping is sent to all hosts on the same physical network. Because a cache is

maintained, ARP does not have to be used repeatedly, and the OSA feature does not have to keep a permanent record of bindings.

Query ARP table for IPv4 for Linux on IBM Z

The Query ARP table is supported by using Internet Protocol version 4 (IPv4). The IP network stack already has an awareness of Internet Protocol version 6 (IPv6) addresses.

Purging ARP entries in cache for IPv4 for z/OS and Linux on IBM Z

Purging of entries in the ARP cache is supported by using IPv4. The IP network stack already has an awareness of IPv6 addresses.

ARP takeover

ARP takeover provides the capability of switching OSA port operations from one OSA to another OSA running in the same mode in z/OS environments.

When a z/OS IP network is started in QDIO mode, it downloads all of the home IP addresses in the stack and stores them in each OSA feature to which it has a connection. This service is part of the QDIO architecture and occurs automatically only for OSD channels. For OSA ports that are set up as OSE channels (non-QDIO), multiple IP addresses must be defined in the OAT by using OSA/SF. The OSA then responds to ARP requests for its own IP address and for VIPAs.

If an OSA feature fails and a backup OSA available on the same network or subnetwork, IP network informs the backup OSA which IP addresses (real and VIPA) to take over, and the network connection is maintained. For this technique to work, multiple paths must be defined to the IP network stack. For example, MULTIPATH must be defined to the IPCONFIG statement of the IP network profile in z/OS.

ARP statistics

QDIO includes an IPA function, which gathers ARP data during the mapping of IP addresses to Media Access Control (MAC) addresses. CHPIDs that are defined as OSD maintain ARP cache information in the OSA feature (ARP offload). This data is useful in problem determination for the OSA feature.

5.2.9 IP network availability

There are several ways to ensure network availability if failure occurs at either the logical partition or the network connection level. Port sharing, redundant paths, and the use of primary and secondary ports all provide some measure of recovery. A combination of these items can ensure network availability regardless of the failing component.

5.2.10 Checksum offload support for z/OS and Linux on IBM Z

z/OS and Linux on IBM Z environments provide the capability of calculating and validating the Transmission Control Protocol/User Datagram Protocol (TCP/UDP) and IP header checksums. Checksums are used to verify the contents of files when transmitted over a network, such as these examples:

- OSA validates the TCP, UDP, and IP header checksums for inbound packets.
- OSA calculates the TCP, UDP, and IP header checksums for outbound packets.

Checksum offload is supported by all OSA Ethernet features when operating in QDIO mode. By offloading checksum processing to the supporting OSA features, host system cycles are reduced, which can result in improved performance for most IPv4 packets.

Note: Linux on IBM Z supports only inbound checksum offload (inbound packets).

When checksum is offloaded, the OSA feature calculates the checksum. When multiple IP stacks share an OSA port and an IP stack sends a packet to a next-hop IP address that is owned by another IP stack that shares an OSA port, OSA sends the IP packet directly to the other IP stack without placing it on the LAN. Checksum offload does not apply to such IP packets.

The checksum offload is performed for IPv6 packets and IPv4 packets. This process occurs regardless of whether the traffic goes to the LAN, comes in from the LAN, or flows from one LPAR to another LPAR through the OSA feature.

Checksum offload for IPv6 packets is supported by CHPID type OSD on the IBM zSystems platform. Checksum offload for LPAR-to-LPAR traffic in the z/OS environment is included in the OSA-Express design for both IPv4 and IPv6 packets.

Checksum offload support is available with z/OS and z/VM. For more information, see 5.3.3, “Software support” on page 104.

5.2.11 Dynamic LAN idle for z/OS

Dynamic LAN idle is exclusive to IBM zSystems platform and applies to the OSA-Express features (QDIO mode). It is supported by z/OS.

Dynamic LAN idle reduces latency and improves networking performance by dynamically adjusting the inbound blocking algorithm. When enabled, the z/OS IP network stack adjusts the inbound blocking algorithm to best match the application requirements.

For latency-sensitive applications, the blocking algorithm is modified to be latency sensitive. For streaming (throughput sensitive) applications, the blocking algorithm is adjusted to maximize throughput. In all cases, the z/OS IP network stack dynamically detects the application requirements and makes the necessary adjustments to the blocking algorithm. The monitoring of the application and the blocking algorithm adjustments are made in real time, so they dynamically adjust the application’s LAN performance.

System administrators can authorize the z/OS IP network stack to enable a dynamic setting, which was previously a static setting. The z/OS IP network stack dynamically determines the best setting for the current running application, based on system configuration, system, inbound workload volume, processor use, traffic patterns, and related items.

5.2.12 QDIO optimized latency mode

QDIO optimized latency mode (OLM) can help improve performance for applications that have a critical requirement to minimize response times for inbound and outbound data. OLM optimizes the interrupt processing according to direction:

- For inbound processing, the IP network stack looks more frequently for available data to process to ensure that any new data is read from the OSA feature without requiring more program-controlled interrupts (PCIs).

- For outbound processing, the OSA features also look more frequently for available data to process from the IP network stack. Therefore, a Signal Adapter instruction (SIGA) is not required to determine whether more data is available.

OLM is supported by the Communications Server for z/OS with any OSA feature on the IBM zSystems platform.

5.2.13 Layer 2 support

The OSA Ethernet features on IBM zSystems platform can support two transport modes of the OSI model:

- Layer 2 (Link Layer or MAC Layer)
- Layer 3 (Network Layer)

The Layer 2 transport mode allows for communication with IP and non-IP protocols. OSA works with either a z/VM IP network or Linux on IBM Z Layer 2 support that is running in an LPAR or as a z/VM guest.

The z/VM virtual switch can also be used to enable the Layer 2 function for guest systems (see Figure 5-6).

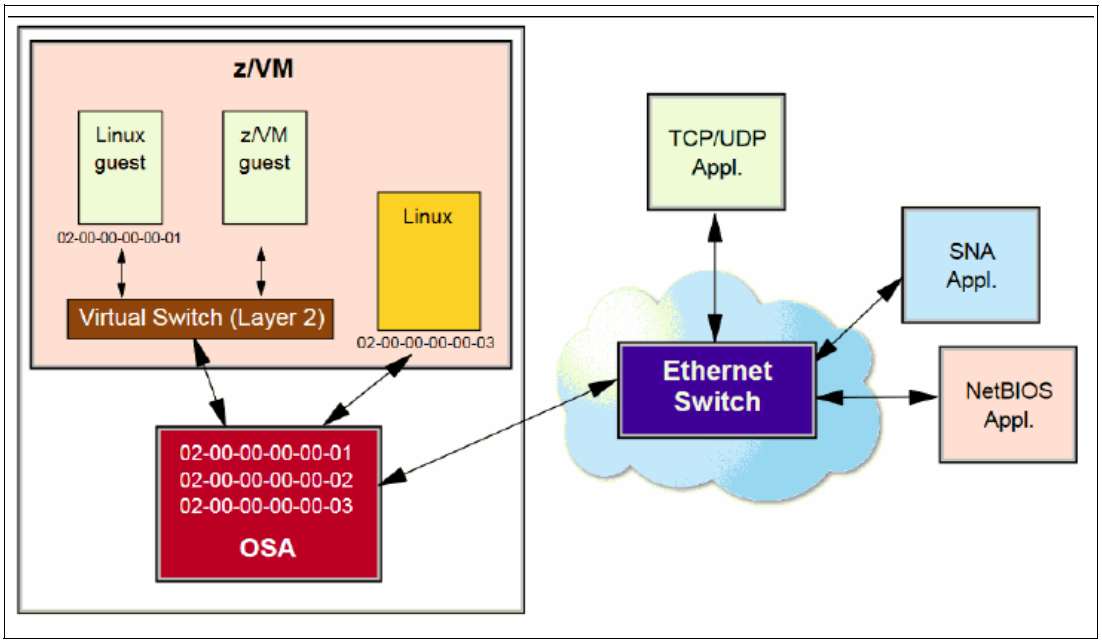


Figure 5-6 Layer 2 support for OSA

The virtual switch uses both Layer 2 and Layer 3 support in the z/VM Control Program. For Layer 2 support, the z/VM Control Program owns the connection to the OSA feature and manages the MAC addresses and VLAN connectivity of the attached guest systems. The virtual switch automatically generates the MAC address and assignment to ensure that the z/VM guest systems are unique. MAC addresses can also be locally administered.

The virtual switch uses each guest system's unique MAC address to forward frames. Data is transported and delivered within Ethernet frames. This process transports both IP and non-IP frames (for example, NetBIOS and SNA) through the fabric that the virtual switch supports. Through the address-resolution process, each guest system's MAC address becomes known to hosts on the physical side of the LAN segment. All inbound or outbound frames that pass

through the OSA port have the guest system's corresponding MAC address as the source or destination address.

The OSA Ethernet features can filter inbound frames by virtual local area network identification (VLAN ID, IEEE 802.1q), the Ethernet destination MAC address, or both. Filtering can reduce the amount of inbound traffic that is being processed by the operating system, which reduces processor use. Filtering by VLAN ID or MAC address can also enable you to isolate portions of your environment that have sensitive data to provide a degree of low-level security.

Link aggregation for z/VM in Layer 2 mode

Link aggregation is exclusive to IBM zSystems and is applicable to the OSA-Express features in Layer 2 mode when configured as CHPID type OSD (QDIO). Link aggregation is supported by z/VM.

z/VM virtual switch-controlled (VSWITCH-controlled) link aggregation (IEEE 802.3ad) allows you to do the following when the port is participating in an aggregated group and is configured in Layer 2 mode:

- Aggregated links are viewed as one logical trunk and contain all of the VLANs that are required by the LAN segment.
- Link aggregation is between a VSWITCH and the physical network switch.
- Load-balance communications are across multiple links in a trunk to prevent a single link from being overrun.
- Up to eight OSA-Express ports are supported in one aggregated link.
- OSA ports can be added or removed to provide on-demand bandwidth.
- It operates in full-duplex mode (send and receive) for increased throughput.

Note: Target links for aggregation must be of the same type (for example, gigabit Ethernet to gigabit Ethernet).

5.2.14 QDIO data connection isolation for z/VM

The QDIO data connection isolation function provides a higher level of security when sharing a OSA-Express port and virtual switch (VSWITCH) across multiple z/VM guest systems. The VSWITCH is a virtual network device that provides switching between OSA-Express ports and the connected guest systems.

Two modes allow for connection isolation: *Port isolation* and *Virtual Ethernet Port Aggregator (VEPA)*. Port isolation and VEPA are mutually exclusive.

Port isolation

z/VM allows you to disable guest system-to-guest system communication through a virtual switch, preserving each guest system's ability to communicate with hosts or routers in the external network. When a virtual switch is defined to be *isolated*, all communication between the guest systems' ports on that virtual switch (VSWITCH) are disabled.

An isolated virtual switch is unable to communicate directly with other LPARs that share the OSA-Express port. Communications must be relayed through another network-based device, such as a router or firewall, in the external network.

As shown in Figure 5-7, when in *isolation mode*, data traffic that is destined for a guest system port in the VSWITCH is blocked. However, traffic that is going to an external network device is sent to the OSA-Express port for delivery. The isolation options (ON or OFF) can be set by using the **SET VSWITCH ISOLATION** command.

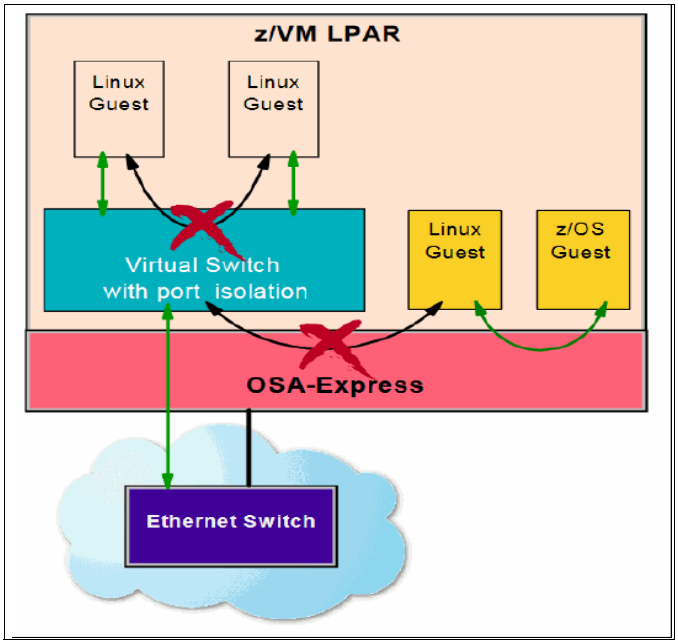


Figure 5-7 VSWITCH port isolation

Virtual Ethernet Port Aggregation

VEPA is part of the IEEE 802.1Qbg standard. It provides the capability of sending all attached guest systems' data traffic to an external Ethernet switch for further processing. This mode does not allow any direct guest system-to-guest system communications through the VSWITCH. In tandem with the VSWITCH, the OSA-Express feature prevents any data traffic between the VSWITCH and any connected systems that share that OSA-Express port. Hence, the isolation is provided within both the VSWITCH and the OSA-Express feature. However, VEPA mode does allow data traffic from a guest system to be sent to a router or similar network-based devices and come back through the same VSWITCH to another guest system, which is known as hair pinning.

For a VSWITCH to enter VEPA mode, the external Ethernet switch must be in Reflective Relay mode.

As shown in Figure 5-8, when in *VEPA mode*, data traffic that is destined for a guest system in the VSWITCH is forced to go to an external Ethernet switch through an OSA-Express port for further processing. VEPA mode (ON or OFF) can be set by using the SET VSWITCH VEPA command.

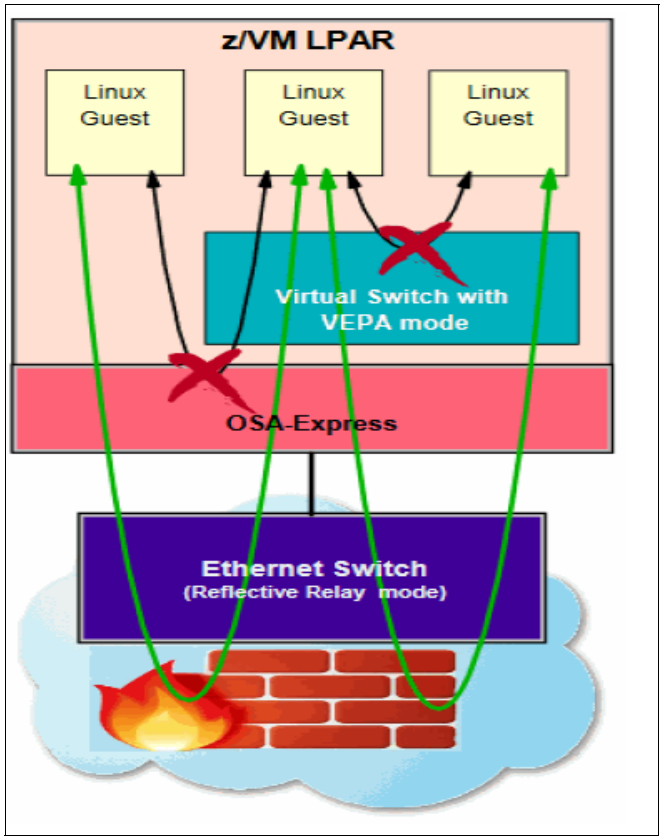


Figure 5-8 VSWITCH in VEPA mode

QDIO data connection isolation is supported by all OSA-Express7S 1.2, OSA-Express7S, and OSA-Express6S features on IBM z16, all OSA-Express7S (except for 25 GbE) and OSA-Express6S features on IBM z15, and all OSA-Express6S features on IBM z14.

5.2.15 QDIO interface isolation for z/OS

Some environments require strict controls for routing data traffic between systems or nodes. In certain cases, the LPAR-to-LPAR capability of a shared OSA port 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 OAT as *isolated*.

For example, as shown in Figure 5-9 on page 93, interface isolation is enabled for LPAR 1. Therefore, data traffic from LPAR 2 that is destined for LPAR 1 is dropped by the OSA.

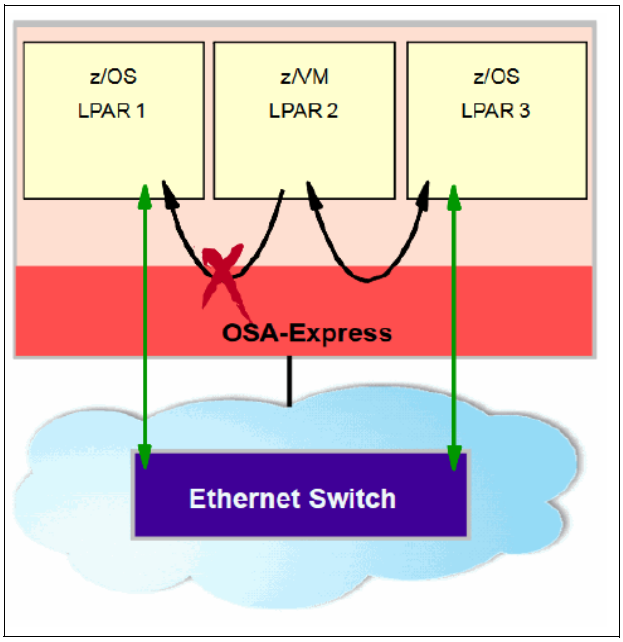


Figure 5-9 QDIO interface isolation

QDIO interface isolation is supported by the Communications Server for z/OS with OSA features on the IBM zSystems platform.

5.2.16 Layer 3 VMAC for z/OS

To help simplify the infrastructure and provide load balancing when an LPAR is sharing an OSA MAC address with another LPAR, each operating system instance can now have its own unique VMAC address. All IP addresses associated with an IP network stack are accessible by using their own VMAC address rather than sharing the MAC address of an OSA port. This situation applies to Layer 3 mode and to an OSA port shared among LPARs.

This support features the following benefits:

- Improves IP workload balancing.
- Dedicates a Layer 3 VMAC to a single IP network stack.
- Removes the dependency on Generic Routing Encapsulation (GRE) tunnels.
- Improves outbound routing.
- Simplifies configuration setup.
- Allows z/OS to use a standard interface ID for IPv6 addresses.
- Allows IBM WebSphere® Application Server content-based routing to support an IPv6 network.
- Eliminates the need for PRIROUTER/SECROUTER function in z/OS.

OSA Layer 3 VMAC for z/OS is applicable to OSA Ethernet features when configured as CHPID type OSD (QDIO).

5.2.17 Enterprise Extender

The Enterprise Extender (EE) function of z/OS Communications Server allows you to run SNA applications and data on IP networks and IP-attached clients. It can be used with any OSA feature that is running IP traffic. EE is a simple set of extensions to the open High-Performance Routing technology that integrates HPR frames into UDP/IP packets, which provide these advantages:

- SNA application connectivity by using an IP backbone support for:
  - SNA-style priority
  - SNA Parallel Sysplex
- Improved throughput and response times
- Compatible support for TCP and UDP traffic on the IP portion of the application traffic path (SNA and HPR and UDP and IP traffic can coexist on an EE connection.)

The EE function is an IP network encapsulation technology. It carries SNA traffic from an endpoint over an IP network (for example, through the OSA port to the Communications Server) to another endpoint where it is de-encapsulated and presented to an SNA application.

EE requires APPN/HPR at the endpoints. To enable EE, you must configure the IP network stack with a virtual IP address and define an XCA major node. The XCA major node is used to define the PORT, GROUP, and LINE statements for the EE connections.

5.2.18 TN3270E server

The TN3270E Server is supported by z/OS. It allows desktop users to connect through an IP network directly to SNA applications.

The following support is provided:

- Secure access by using SSL and Client Authentication by using IBM Resource Access Control Facility (IBM RACF®).
- Over 64,000 sessions per server when using multiple ports.
- Over 2000 transactions per second with subsecond response time.
- Reconnect 16,000 sessions in less than a minute by using VIPA takeover support.
- IP QoS that uses a Service Policy server.
- Host Print support.
- Tivoli support provides IP visibility to IBM VTAM.
- Manage with your data center resources.
- More robust than external TN3270 servers.

z/OS Communications Server also supports a secure, RACF-based single sign-on logic called Express Logon Facility (ELF). ELF works with IBM TN3270 clients to securely authenticate the client, acquire a pass token, and then pass it on to the TN3270E server for replacement or submission to the application.

5.2.19 Adapter interruptions for QDIO

Linux on IBM Z and z/VM work together to provide performance improvements by using extensions to the QDIO architecture. Adapter interruptions, which were first added to IBM z/Architecture with HiperSockets, provide an efficient, high-performance technique for I/O interruptions. This technique reduces path lengths and overhead in both the host operating system and the adapter when using type OSD CHPID.

In extending the use of adapter interruptions to OSD (QDIO) channels, the programming overhead to process a traditional I/O interruption is reduced. This technique benefits OSA IP network support in Linux on IBM Z, z/VM, and z/VSE.

Adapter interruptions apply to all of the OSA features on the IBM zSystems platform when in QDIO mode (CHPID type OSD).

5.2.20 Inbound workload queuing

Inbound workload queuing (IWQ) helps reduce overhead and latency for inbound z/OS network data traffic and implements an efficient way for initiating parallel processing. This goal is achieved by using an OSA-Express feature in QDIO mode with multiple input queues and by processing network data traffic that is based on workload types.

Figure 5-10 shows the IWQ concept of using multiple inbound queues for different types of workloads (T1 through T4) compared to a single inbound queue.

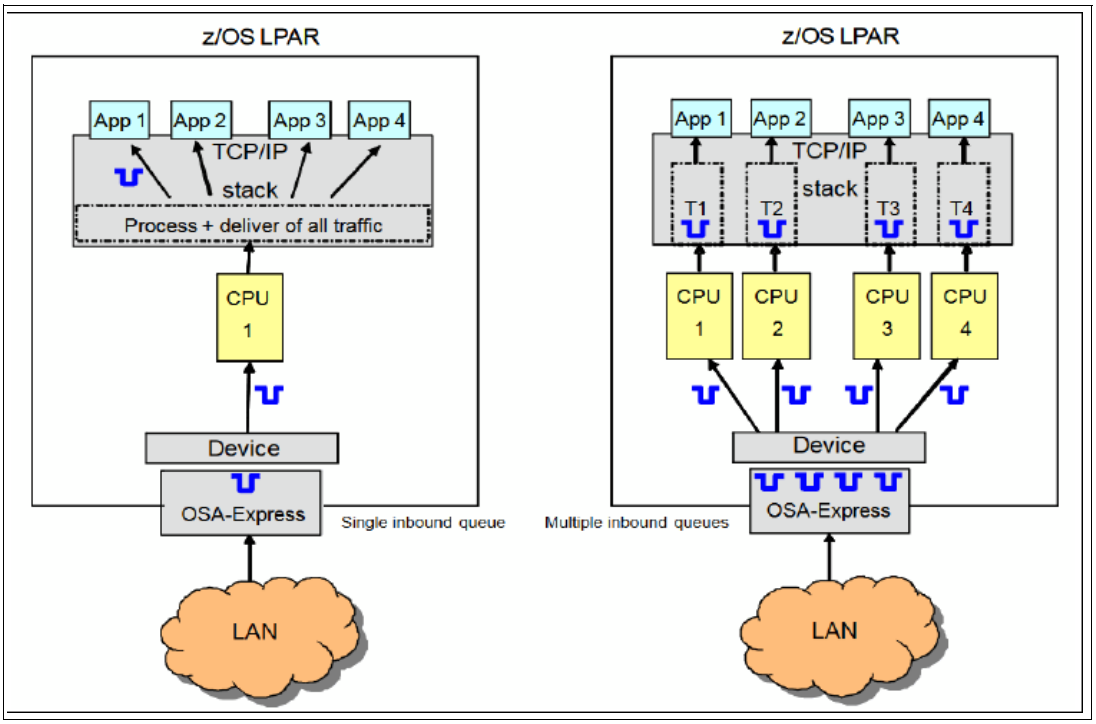


Figure 5-10 Single inbound queue versus multiple inbound queues

The data from a specific workload type is placed in one of four input queues, and a process is created and scheduled to run on one of multiple processors, independent from the three other queues. This process greatly improves performance because IWQ can use the symmetric multiprocessor (SMP) architecture of the IBM zSystems platform.

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. These 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 VIPA is assigned to a unique input queue, allowing 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 technique allows most of the data processing to be assigned the appropriate resources and isolated from critical interactive workloads.

- Enterprise Extender traffic

  Inbound Enterprise Extender traffic is differentiated and separated to a new input queue. This separation and processing provides improved scalability and performance for Enterprise Extender. It is supported on the IBM zSystems platform with OSA features.

IPsec traffic on OSA-Express7S 1.2, OSA-Express7S, and OSA-Express6S

OSA Express7S and OSA-Express6S provide new support for Inbound Workload Queuing for IPsec, which allows OSA to separate inbound IPsec packets and enables z/OS Communication Server to optimize the related software processing.

The supported IWQ IPsec traffic applies to the following protocols:

- Encapsulated Security Payload (ESP)
- Authentication Header (AH) protocol
- UDP protocol + port (NAT Traversal)

Users already using IWQ should be aware that each IWQ workload type that applies to their z/OS environment (for example, Sysplex distributor, Enterprise Extender) is automatically enabled when IWQ is enabled. Each unique input queue consumes extra fixed ECSA memory (approximately 4 MB per input queue per OSA interface).

5.2.21 Network management: Query and display OSA configuration

As complex functions have been added to OSA, the job of the system administrator to display, monitor, and verify the specific current OSA configuration unique to each operating system has become more complex. The operating system can directly query and display the current OSA configuration information (similar to OSA/SF). z/OS and z/VM use this OSA capability with an IP network operator command, which allows the operator to monitor and verify the current OSA configuration. This feature improves the overall management, serviceability, and usability of the OSA features. **`Display OSAINFO`** is possible on OSA-Express features.

5.3 Connectivity

The transmission medium and cabling requirements for the OSA ports depend on the OSA feature, OSA port type, and LAN environment. Acquiring cables and other connectivity items is the user's responsibility.

OSA devices

The different types of OSA channels (CHPID types) require the following device types:

- OSA devices for QDIO (OSD) and non-QDIO (OSE) CHPID types.
- 3270-X and 3287 devices (consoles) for the OSA-ICC (OSC) CHPID type.
- OSA/SF requires one device (defined through HCD) to be associated with the OSA CHPID as device type OSAD (UNITADD=FE). OSA/SF uses this device to communicate with the OSA feature.
- The OSA-Express Network Traffic Analyzer for z/OS requires one or more data paths devices for the OSAENTA trace interface, depending on the configuration.

Multiple image facility

The multiple image facility (MIF) enables OSA ports that are installed on IBM zSystems platform to be shared across LPARs. For more information, see Chapter 2, "Channel subsystem overview" on page 17.

Spanned channels

Spanning is the ability to configure channels to multiple channel subsystems. When defined that way, the channels can be transparently shared by any or all of the configured LPARs, regardless of the channel subsystem to which the LPAR is configured.

OSA ports can be spanned across multiple channel subsystems (CSSs) on IBM zSystems platform. For more information about spanned channels, see 2.1.2, "Multiple CSSs" on page 19.

5.3.1 OSA-Express features

Each OSA-Express feature occupies one slot in the I/O drawer of the IBM zSystems platform.

OSA-Express features are offered in two transmission medium types with different speeds:

- OSA-Express fiber Ethernet features use Small Form-factor Pluggable (SFP)[5] transceiver technology[6] with LC duplex receptacles and require either multi-mode (MM) at 62.5 µm or 50 µm fiber optic cables, or single mode (SM) at 9 µm fiber optic cables. OSA-Express fiber Ethernet features, known as either GbE, 10 GbE, or 25 GbE (based on the speed [gigabit per second] of the feature), do not support auto-negotiate to a lower speed.
- The OSA-Express copper Ethernet features use RJ45 jacks and require unshielded twisted pair (UTP) Cat5 or Cat6 cables. OSA-Express copper Ethernet features[7] (referred to as 1000BASE-T), support auto-negotiate at the following speeds:
  - 10 Mbps half-duplex or full-duplex[8]
  - 100 Mbps half-duplex or full-duplex
  - 1000 Mbps full-duplex

[5] SFP allows for a concurrent repair or replace action.
[6] Except for the OSA-Express4S features
[7] OSA-Express7S 1000BASE-T feature does not support auto-negotiate to a lower speed.
[8] OSA-Express6S and OSA-Express5S 1000BASE-T features do not support 10 Mbps.

If you are not using auto-negotiate, the OSA port attempts to join the LAN at the specified speed. If this speed does not match the speed and duplex mode of the signal on the cable, the OSA port will not connect.

LAN speed can be set explicitly by using OSA/SF or the OSA Advanced Facilities function of the HMC. The explicit settings override the OSA feature port's ability to auto-negotiate with its attached Ethernet switch.

Table 5-2 lists the OSA-Express features that are supported on the respective IBM zSystems platform with the maximum unrepeated distance and cable type.

Table 5-2 IBM zSystems OSA-Express features

Feature name	Feature code	Cable type	Maximum unrepeated distance[a]	System
OSA-Express7S 1.2 25GbE SR	0459	MM 50 µm	70 m (2000) 100 m (4700)	IBM z16
OSA-Express7S 1.2 25GbE LR	0460	SM 9 µm	10 km (6.2 miles)	IBM z16
OSA-Express7S 1.2 10GbE LR	0456	SM 9 µm	10 km (6.2 miles)	IBM z16
OSA-Express7S 1.2 10GbE SR	0457	MM 62.5 µm MM 50 µm	33 m (200) 82 m (500) 300 m (2000)	IBM z16
OSA-Express7S 1.2 GbE LX	0454	SM 9 µm	5 km (3.1 miles)	IBM z16
OSA-Express7S 1.2 GbE SX	0455	MM 62.5 µm MM 50 µm	275 m (200) 550 m (500)	IBM z16
OSA-Express7S 1.2 1000BASE-T	0458	UTP Cat5 or Cat6	100 m	IBM z16
OSA-Express7S 25GbE SR 1.1	0449	MM 50 µm	70 m (2000) 100 m (4700)	IBM z16 A01[b] and IBM z15 T01
OSA-Express7S 25GbE SR	0429	MM 50 µm	70 m (2000) 100 m (4700)	IBM z15 T01[b], IBM z15 T02, and IBM z14
OSA-Express7S 10 GbE LR	0444	SM 9 µm	10 km (6.2 miles)	IBM z16 A01[b] and IBM z15 T01
OSA-Express7S 10 GbE SR	0445	MM 62.5 µm	33 m (200) 82 m (500) 300 m (2000)	IBM z16 A01[b] and IBM z15 T01
OSA-Express7S GbE LX	0442	SM 9 µm	5 km (3.1 miles)	IBM z16 A01[b] and IBM z15 T01
OSA-Express7S GbE SX	0443	MM 62.5 µm MM 50 µm	275 m (200) 550 m (500)	IBM z16 A01[b] and IBM z15 T01
OSA-Express7S 1000BASE-T	0446	UTP Cat5 or Cat6	100 m	IBM z16 A01[b] and IBM z15 T01

Feature name	Feature code	Cable type	Maximum unrepeated distance[a]	System
OSA-Express6S 10 GbE LR	0424	SM 9 µm	10 km	IBM z16[b], IBM z15 T01[b], IBM z15 T02, and IBM z14
OSA-Express6S 10 GbE SR	0425	MM 50 µm	550 m (500)	IBM z16[b], IBM z15 T01[b], IBM z15 T02, and IBM z14
		MM 62.5 µm	275 m (200) 220 m (160)	
OSA-Express6S GbE LX	0422	SM 9 µm	10 km	IBM z16[b], IBM z15 T01[b], IBM z15 T02, and IBM z14
OSA-Express6S GbE SX	0423	MM 50 µm	550 m (500)	IBM z16[b], IBM z15 T01[b], IBM z15 T02, and IBM z14
		MM 62.5 µm	275 m (200) 220 m (160)	
OSA-Express6S 1000BASE-T	0426	UTP Cat5 or Cat6	100 m	IBM z16[b], IBM z15 T01[b], IBM z15 T02, and IBM z14
OSA-Express5S 10 GbE LR[c]	0415	SM 9 µm	10 km	IBM z15 T01[b] and IBM z14[b]
OSA-Express5S 10 GbE SR[c]	0416	MM 50 µm	550 m (500)	IBM z15 T01[b] and IBM z14[b]
		MM 62.5 µm	275 m (200) 220 m (160)	
OSA-Express5S GbE LX[c]	0413	SM 9 µm	5 km	IBM z15 T01[b] and IBM z14[b]
		MCP	550 m (500)	
OSA-Express5S GbE SX[c]	0414	MM 50 µm	550 m (500)	IBM z15 T01[b] and IBM z14[b]
		MM 62.5 µm	275 m (200) 220 m (160)	
OSA-Express5S 1000BASE-T[c]	0417	UTP Cat5 or Cat6	100 m	IBM z15 T01[b] and IBM z14
OSA-Express4S 10 GbE LR[c]	0406	SM 9 µm	10 km	IBM z15 T01[b] and IBM z14[b]
OSA-Express4S 10 GbE SR[c]	0407	MM 50 µm	550 m (500)	IBM z15 T01[b] and IBM z14[b]
		MM 62.5 µm	275 m (200) 220 m (160)	
OSA-Express4S GbE LX[c]	0404	SM 9 µm	5 km	IBM z14 ZR1[b]
		MCP	550 m (500)	
OSA-Express4S GbE SX[c]	0405	MM 50 µm	550 m (500)	IBM z14 ZR1[b]
		MM 62.5 µm	275 m (200) 220 m (160)	

Feature name	Feature code	Cable type	Maximum unrepeated distance[a]	System
OSA-Express4S 1000BASE-T[c]	0408	UTP Cat5 or Cat6	100 m	IBM z14[b]

a. Minimum fiber bandwidth in MHz*km for multimode fiber optic links is included in parentheses where applicable.
b. Available on carry-forward only.
c. *Not* supported on IBM z16.

Removal of support for OSA-Express 1000BASE-T features[a]: IBM z16 will be the last IBM zSystems platform to support OSA-Express 1000BASE-T features. In the future, valid OSA CHPID types will be supported only with OSA-Express GbE features, and potentially higher bandwidth fiber Ethernet adapters.

a. All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these SoDs is at the relying party's sole risk and will not create liability or obligation for IBM.

Maximum number of OSA features

The maximum number of OSA-Express features are as follows:

- On IBM z16, the maximum combined number of OSA-Express features is 48 (new and carry forward).
- On IBM z15, the maximum combined number of OSA-Express features is 48 (new and carry forward).
- On an IBM z14, the maximum combined number of OSA-Express features is 48 with any mix of OSA-Express6S, OSA-Express5S (carry-forward only), and supported OSA Express4S(carry-forward only) features.

Table 5-3 shows the CHPID types and number of OSA-Express features that are supported on IBM z16 platforms.

Table 5-3 IBM z16 supported OSA-Express features

I/O feature	Ports per feature	Ports per CHPID	Max. number of[a]		CHPID definition
			Ports	I/O slots	
OSA-Express7S 1.2 25GbE LR/SR[b]	1	1	48 / 48	48 / 48	OSD
OSA-Express7S 1.2 10GbE LR/SR[b]	1	1	48 / 48	48 / 48	OSD
OSA-Express7S 1.2 GbE LX/SX[b]	2	2	96 / 96	48 / 48	OSC and OSD
OSA-Express7S 1.2 1000BASE-T[b]	2	2	96 / 96	48 / 48	OSC, OSD, and OSE
OSA-Express7S 25GbE SR1.1[cd]	1	1	48	48	OSD
OSA-Express7S 25GbE SR[cd]	1	1	48	48	OSD

I/O feature	Ports per feature	Ports per CHPID	Max. number of[a]		CHPID definition
			Ports	I/O slots	
OSA-Express7S 10 GbE LR/SR[c]	1	1	48 / 48	48 / 48	OSD
OSA-Express7S GbE LX/SX[cd]	2	2	96 / 96	48/48	OSD and OSC
OSA-Express7S 1000BASE-T[cd]	2	2	96 / 96	48 / 48	OSC, OSD, and OSE
OSA-Express6S GbE LX/SX[c]	2	2	96 / 96	48 / 48	OSD
OSA-Express6S 10 GbE LR/SR[c]	1	1	48 / 48	48 / 48	OSD
OSA-Express6S 1000BASE-T[c]	2	2	96 / 96	48 / 48	OSC, OSD, and OSE

a. The maximum number OSA-Express features that are mixed is determined based on a maximum of 48 PCHIDs.
b. IBM z16 *only*.
c. Carry forward to IBM z16 A01.
d. Not available on IBM z16 A02 and IBM z16 AGZ

Note: OSA-Express4S and OSA-Express5S are not supported on IBM z16.

Table 5-4 on page 101 shows the CHPID types and number of OSA-Express features that are supported on IBM z15 platforms.

Table 5-4 IBM z15 supported OSA-Express features

I/O feature	Ports per feature	Ports per CHPID	Max. number of[a]		CHPID definition
			Ports	I/O slots	
OSA-Express7S 25GbE SR1.1[b]	1	1	48	48	OSD
OSA-Express7S 25GbE SR[c]	1	1	48	48	OSD
OSA-Express7S 10 GbE LR/SR[c]	1	1	48 / 48	48 / 48	OSD
OSA-Express7S GbE LX/SX[b]	2	2	96 / 96	48/48	OSD and OSC
OSA-Express7S 1000BASE-T[bd]	2	2	96 / 96	48 / 48	OSC, OSD, and OSE
OSA-Express6S GbE LX/SX[c]	2	2	96 / 96	48 / 48	OSD
OSA-Express6S 10 GbE LR/SR[c]	1	1	48 / 48	48 / 48	OSD
OSA-Express6S 1000BASE-T[c]	2	2	96 / 96	48 / 48	OSC[e], OSD, and OSE

I/O feature	Ports per feature	Ports per CHPID	Max. number of[a]		CHPID definition
			Ports	I/O slots	
OSA-Express5S GbE LX/SX[f]	2	2	96 / 96	48 / 48	OSD
OSA-Express5S 10 GbE LR/SR[g]	1	1	48 / 48	48 / 48	OSD
OSA-Express5S 1000BASE-T[g]	2	2	96 / 96	48 / 48	OSC[b], OSD, and OSE

a. The maximum number OSA-Express features that are mixed is determined based on a maximum of 48 PCHIDs.
b. z15 T01 and carry forward to IBM z16.
c. Carry forward to z15 T01, and carry forward or new build for z15 T02.
d. OSA-Express7S 1000BASE-T supports only 1000 Mbps (no negotiation to lower speeds).
e. OSA-ICC (OSC Channel) now supports Secure Sockets Layer. Up to 48 secure sessions per CHPID are supported (the overall maximum of 120 connections is unchanged).
f. Carry forward only.

Table 5-5 lists the CHPID types and number of OSA-Express features that are supported on IBM z14 platforms.

Table 5-5 IBM z14 M0x and IBM z14 ZR1 supported OSA I/O features

I/O feature: z14 / z14 ZR1	Ports per feature	Ports per CHPID	Max. number of[a]		CHPID definition
			Ports	I/O slots	
OSA-Express7S 25GbE SR	1	1	48	48	OSD
OSA-Express6S GbE LX/SX	2	2	96 / 96	48 / 48	OSD
OSA-Express6S 10 GbE LR/SR	1	1	48 / 48	48 / 48	OSD
OSA-Express6S 1000BASE-T	2	2	96 / 96	48 / 48	OSC[b], OSD, and OSE
OSA-Express5S GbE LX/SX	2	2	96 / 96	48 / 48	OSD
OSA-Express5S 10 GbE LR/SR	1	1	48 / 48	48 / 48	OSD
OSA-Express5S 1000BASE-T	2	2	96 / 96	48 / 48	OSC[b], OSD, and OSE
OSA-Express4S2 GbE SX/LX	2	NA / 96	OSD / 96	NA / 48	OSD
OSA-Express4S 10 GbE SR/LR	1	1	NA / 48	NA / 48	OSD
OSA-Express4S 1000BASE-T	2	2	96 / NA	48 / NA	OSC[b], OSD, and OSE

a. The maximum number OSA-Express that are mixed is determined based on a maximum of 48 PCHIDs.
b. OSA-ICC (OSC Channel) now supports Secure Sockets Layer. Up to 48 secure sessions per CHPID are supported (the overall maximum of 120 connections is unchanged).

5.3.2 OSA function support

Table 5-6 lists the functions that are supported based on the OSA feature.

Table 5-6 OSA function support

Function	OSA-Express4S and OSA-Express5S			OSA-Express6S and OSA-Express7S			OSA-Express7S 25GbE SR and SR1.1	OSA-Express7S 1.2			
	10 GbE	GbE	1000 BASE-TGbE	10 GbE	GbE	1000BASE-T		25 GbE	10 GbE	GbE	1000BASE-T
Jumbo frame support (8992-byte frame size)[a]	x	x	x	x	x	x	x	x	x	x	x
Network Traffic Analyzer for z/OS	x	x	x	x	x	x	x	x	x	x	x
QDIO Diagnostic Synchronization for z/OS	x	x	x	x	x	x	x	x	x	x	x
640 IP network (with priority queues disabled)	x	x	x	x	x	x	x	x	x	x	x
Virtual IP address (VIPA)	x	x	x	x	x	x	x	x	x	x	x
Primary/secondary router function	x	x	x	x	x	x	x	x	x	x	x
Internet Protocol version 6 (IPv6)	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
Large send support for IPv4	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
Large send support for IPv6	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
VLAN (IEEE 802.1q)	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
VLAN support of GVRP (IEEE 802.1 p)[a]	x	x	x	x	x	x	x	x	x	x	x
SNMP support for z/OS and Linux on IBM Z[a]	x	x	x	x	x	x	x	x	x	x	x
Multicast and broadcast support	x	x	x	x	x	x	x	x	x	x	x
ARP cache management	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
ARP statistics[a]	x	x	x	x	x	x	x	x	x	x	x
ARP takeover	x	x	x	x	x	x	x	x	x	x	x
IP network availability	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
Checksum offload support for IPv4	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]

Function	OSA-Express4S and OSA-Express5S			OSA-Express6S and OSA-Express7S			OSA-Express7S 25GbE SR and SR1.1	OSA-Express7S 1.2			
	10 GbE	GbE	1000 BASE-TGbE	10 GbE	GbE	1000BASE-T		25 GbE	10 GbE	GbE	1000BASE-T
Checksum offload support for IPv6	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
Dynamic LAN Idle for z/OS[a]	x	x	x	x	x	x	x	x	x	x	x
QDIO optimized latency mode	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
Layer 2 support	x	x	x[a]	x	x	[a]	x	x	x	x	x[a]
Link aggregation for z/VM Layer 2 mode[a]	x	x	x	x	x	x	x	x	x	x	x
QDIO data connection isolation for z/VM	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
QDIO interface isolation for z/OS	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
Layer 3 VMAC for z/OS[a]	x	x	x	x	x	x	x	x	x	x	x
Enterprise Extender	x	x	x	x	x	x	x	x	x	x	x
TN3270E server for z/OS	x	x	x	x	x	x	x	x	x	x	x
Adapter interruptions for QDIO	x	x	x	x	x	x	x	x	x	x	x
Inbound workload queuing (IWQ)	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]
Query and display OSA configuration	x	x	x[a]	x	x	x[a]	x	x	x	x	x[a]

a. Only in QDIO mode (CHPID type: OSD).

5.3.3 Software support

Certain functions might require specific levels of an operating system, program temporary fixes (PTFs), or both. That information is provided when necessary within this chapter.

Consult the appropriate Preventive Service Planning (PSP) buckets (3931DEVICE, 3932DEVICE, 8561DEVICE, 8562DEVICE, 3906DEVICE, 3907DEVICE) before implementation.

Consider the following points:

- Not every operating system supports all features. The operating system support information is provided for IBM z/OS, IBM z/VM, IBM z/VSE, IBM z/TPF, and Linux on IBM Z (supported distributions).

- KVM is included with Linux distributions. For more information about the latest support, contact your Linux distribution providers. The KVM hypervisor is supported with the following minimum Linux distributions:
  - SUSE Linux Enterprise Server 15 SP1 with service, SUSE Linux Enterprise Server 12 SP4 with service, and SUSE Linux Enterprise Server 11 SP4 with service.
  - RHEL 9.0, RHEL 8.0 with service, RHEL 7.7 with service, and RHEL 6.10 with service.
  - Ubuntu 22.04, 20.04, and 18.04, 16.04.5 LTS with service.
- The support statements for Linux on IBM Z also cover the KVM hypervisor on distribution levels that have KVM support.

  For more information about the minimum required and recommended distribution levels, see this web page.

5.3.4 Resource Measurement Facility

Resource Measurement Facility (RMF) reporting (z/OS only) supports the OSA features. It can capture performance data for these features:

- Microprocessor use (per LPAR image, if it applies)
- Physical Peripheral Component Interconnect (PCI) bus use
- Bandwidth per port (both read and write directions) per LPAR image

For example, with this support, possible bandwidth bottlenecks and root causes can be analyzed.

5.4 Summary

The OSA features provide direct LAN connectivity as integrated features of the IBM zSystems platform. They bring the strengths of IBM zSystems and z/Architecture to the modern network environment.

Table 5-7 on page 105 summarizes the OSA features for the different modes of operation and maximum addressing ranges that are supported by IBM zSystems platform.

Table 5-7 OSA modes of operation and addressing support

Item	Value
Addresses	
IP addresses per channel path (IPv4, IPv6, VIPA)	4096
Multicast addresses (IPv4 + IPv6)ARP table size	16384
ARP table size	16384
MAC addresses	4096
Non-QDIO (OSE)[a,b]	
Subchannels per IP network link	2
IP network stacks per channel path	120
SNA PUs per port	4096
Subchannels per channel path	240

Item	Value
CUs per channel path	1
QDIO (OSD)	
Subchannels per IP network link	3
IP network stacks per channel path	640[c]
Subchannels per channel path	1920
CUs per channel path	16

a. 1000BASE-T feature only.

b. Removal of support for the OSE CHPID type: IBM z16 is planned to be the last IBM zSystems platform to support OSE networking channels. IBM zSystems support for the System Network Architecture (SNA) protocol being transported natively out of the server by using OSA-Express 100BASE-T adapters configured as channel type “OSE” will be eliminated after IBM z16. Client applications that rely on the SNA protocol and use OSE networking channels as the transport, as opposed to FICON CTC, must either migrate to TCP/IP, or the operating system configuration must be updated to use some form of SNA over IP technology if possible, such as z/OS Enterprise Extender.

c. If multiple priority for queues is enabled, the maximum value is reduced to 160 IP network stacks and 480 devices.

5.5 References

For more information about the OSA features and configuration, see the following publications:

- *Open Systems Adapter-Express Customer’s Guide and Reference*, SA22-7935
- *OSA-Express Implementation Guide*, SG24-5948
- *Communications Server: SNA Network Implementation Guide*, SC27-3672
- *Communications Server: IP Configuration*, SC27-3650
- *Resource Measurement Facility Report Analysis*, SC34-2665
- *Enterprise Extender Implementation Guide*, SG24-7359

6

Console Communications (OSA-ICC)

This chapter describes the IBM Open Systems Adapter-Express Integrated Console Controller (OSA-ICC) function of the OSA-Express7S 1.2, OSA-Express7S, OSA-Express6S, OSA-Express5S, and OSA-Express4S features.

The OSA-ICC function supports the TN3270 console (TN3270E), local non-SNA DFT 3270 emulation, 328x printer emulation, and 3215 console emulation for TPF. This emulation support for console session connections is integrated in IBM zSystems platforms.

This chapter includes the following topics:

- 6.1, “Description of the OSA-ICC” on page 108
- 6.2, “Connectivity” on page 110
- 6.3, “Software support” on page 111
- 6.4, “Summary” on page 111
- 6.5, “References” on page 112

6.1 Description of the OSA-ICC

The OSA-ICC support is a no-charge function that is included in Licensed Internal Code (LIC) on the IBM zSystems platform. Table 6-1 lists the support matrix for OSA-ICC.

Table 6-1 OSA-ICC support matrix

System	OSA-Express features that support OSA-ICC
IBM z16 A01	OSA-Express7S 1.2 GbE LX and SX (Feature Code (FC) 0454 and FC 0455) OSA-Express7S 1.2 1000BASE-T (FC 0458) OSA-Express7S GbE LX[a] and SX[a] (FC 0442 and FC 0443) OSA-Express7S 1000BASE-T[a] (FC 0446) OSA-Express6S 1000 BASE-T[a] (FC 0426)
IBM z16 A02 IBM z16 AGZ	OSA-Express7S 1.2 GbE LX and SX (Feature Code (FC) 0454 and FC 0455) OSA-Express7S 1.2 1000BASE-T (FC 0458) OSA-Express6S 1000 BASE-T[a] (FC 0426)
IBM z15 T01	OSA-Express7S GbE LX and SX (FC 0442 and FC 0443) OSA-Express7S 1000BASE-T (FC 0446) OSA-Express6S 1000 BASE-T[a] (FC 0426) OSA-Express5S 1000 BASE-T[a] (FC 0417)
IBM z15 T02	OSA-Express6S 1000 BASE-T[a] (FC 0426) OSA-Express5S 1000 BASE-T[a] (FC 0417)
IBM z14 M0x	OSA-Express6S 1000 BASE-T (FC 0426) OSA-Express5S 1000 BASE-T[a] (FC 0417) OSA-Express4S 1000 BASE-T[a] (FC 0408)
IBM z14 ZR1	OSA-Express6S 1000 BASE-T (FC 0426) OSA-Express5S 1000 BASE-T[a] (FC 0417)

a. Carry forward only.

Removal of support for OSA-Express 1000BASE-T hardware adapters[a]: IBM z16 will be the last IBM zSystems system to support OSA-Express 1000BASE-T hardware adapters (#0426, #0446, and #0458). Definition of all valid OSA CHPID types will be allowed only on OSA-Express GbE adapters, and potentially higher bandwidth fiber Ethernet adapters, on future servers.

a. Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM. Information regarding potential future products is intended to outline general product direction and should not be relied on in making a purchasing decision.

The OSA-ICC supports Ethernet-attached TN3270E consoles and provides a system console function at IPL time and operating systems support for multiple logical partitions (LPARs). Console support can be used by IBM z/OS, IBM z/VM, IBM z/VSE, and IBM z/Transaction Processing Facility (z/TPF) software.

The OSA-ICC also supports local non-SNA DFT 3270 and 328x printer emulation for IBM Time Sharing Option Extensions (TSO/E), IBM CICS®, IBM Information Management System (IMS), or any other 3270 application that communicates through VTAM. For IBM z16, IBM z15, and IBM z14 with FC 0034, 3215 emulation is provided for z/TPF 1.1. For previous server generations, 3215 emulation for z/TPF 1.1 is provided by RPQ 8P2339.

With the proper OSA-Express features, the OSA-ICC is configured on a port-by-port basis, by using the channel path identifier (CHPID) type OSC. When the CHPID shares two ports the following server definition rules apply:

- Each physical port is defined to a unique IP network port number.
- Different subnets are defined for each physical port host IP.
- There is a single defined common gateway router with an interface on each IP subnet. Only one of the IP network ports defines the IP default gateway.

Each CHPID can support up to 120 console session connections. These connections can be shared among LPARs by using the multiple image facility (MIF) and can be spanned across multiple channel subsystems (CSSs).

OSA-ICC (CHPID type OSC) SSL support

TLS/SSL with Certificate Authentication was added to the OSC CHPID to provide a secure and validated method for connecting clients to the IBM zSystems host. Up to 48 secure sessions per CHPID can be defined (the overall maximum of 120 connections unchanged).

Removal of TLS 1.0 for OSA, HMC, and SE: IBM z15 is the last IBM zSystems server to support the use of the Transport Layer Security protocol version 1.0 (TLS 1.0) for establishing secure connections to the Support Element (SE), Hardware Management Console (HMC), and OSA Integrated Console Controller (channel path type OSC).

OSA-ICC enhancements with HMC 2.14.1 (and later)

The following OSA-ICC enhancements were introduced with HMC 2.14.1:

- 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.
- The supported TLS protocol levels for the OSA-ICC 3270 client connection can be now specified. Supported protocol levels are: TLS 1.0, TLS 1.1, and TLS 1.2.
- In addition to a single certificate for all OSA-ICC PCHIDs in the system, separate and unique OSA-ICC 3270 certificates are now supported 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.

Figure 6-1 shows an example of the OSA-Express Integrated Console Controller in a single system configuration.

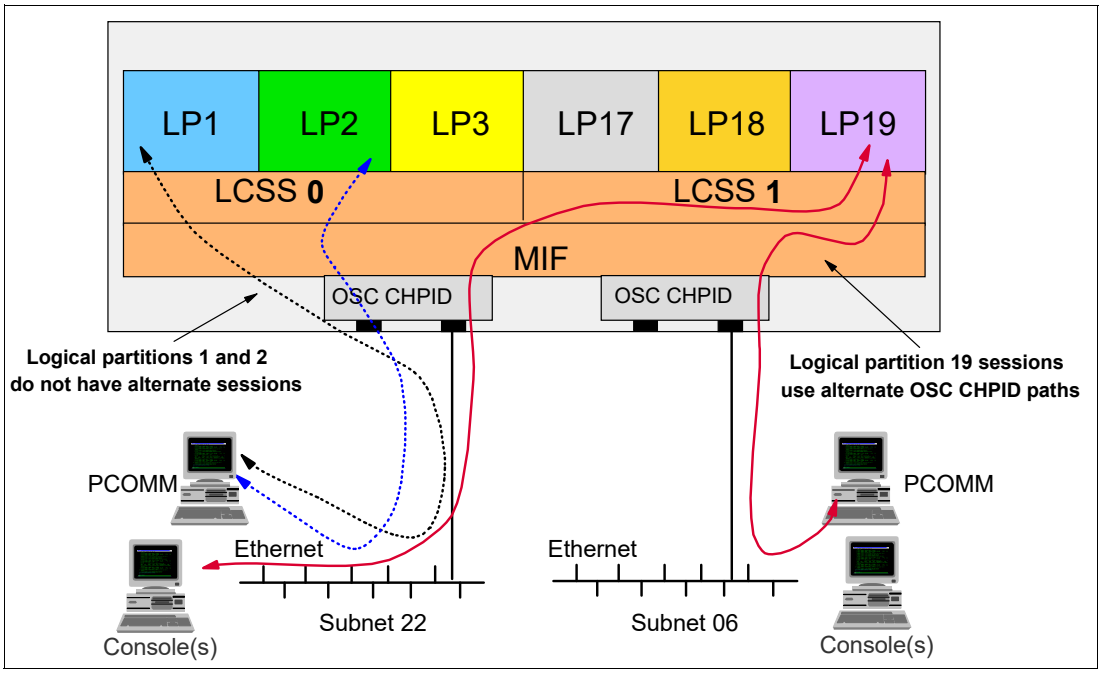


Figure 6-1 Example of an OSA-ICC configuration

The following features are included in base support for the OSA-ICC:

- Local and remote session connectivity, depending on the provided security and networking environment
- Local and remote connections for configuration changes by using security features of the IBM zSystems Hardware Management Console environment

6.2 Connectivity

IBM zSystems platform have base Licensed Internal Code support for the OSA-ICC function. At least one supported OSA-Express feature must be installed.

The Hardware Management Console (HMC) or Support Element (SE) is used to create the configuration source file for the OSA-ICC CHPID and for operation and diagnosis.

The OSA-Express7S 1.2-gigabit Ethernet LX and SX features (FC 0454 and FC 0466) support the OSC CHPID. These features are available on IBM z16 platform.

The OSA-Express7S gigabit Ethernet LX and SX features (FC 0442 and FC 0443) support the OSC CHPID. These features, which were introduced with the IBM z15 T01, can be carried forward to IBM z16 A01.

The OSA-Express6S 1000BASE-T (FC 0417) Ethernet features support the OSC CHPID. FC 0417 is supported on the IBM z14 systems and as a carry forward feature for IBM z15 and IBM z16.

6.3 Software support

For more information about the operating systems that are supported on IBM zSystems platform, see this website.

Note: Certain functions might require specific levels of an operating system, program temporary fixes (PTFs), or both. That information is provided when necessary within this chapter. FIXCAT categories information is available at IBM Fix Category Values and Descriptions.

Consult the appropriate Preventive Service Planning (PSP) buckets (3931DEVICE, 8561DEVICE, 8562DEVICE, 3906DEVICE, and 3907DEVICE) before implementation.

6.3.1 TN3270E emulation

The TN3270E emulation program that is used with the OSA-ICC must comply with the TN3270E (TN3270 Enhancements) protocol (RFC 2355), which allows for the following functions:

- The ability to request that a connection is associated with a specific 3270 LU or pool name.
- Access to the SNA bind information. Until it receives the bind information, the session runs at the SSCP-LU level.
- Handling of SNA positive and negative responses.
- Universal support of the 3270 ATTN and SYSREQ keys.
- Support for the emulation of the 328x class of printer.

IBM Personal Communications V14 supports TN3270 Extensions that are used for OSA-ICC support. Personal Communications is a component of the IBM Access Client Package for Multiplatforms program suite. For more information, see this web page.

6.4 Summary

Here are the key characteristics of the Open Systems Adapter-Express Integrated Console Controller (OSA-ICC) function:

- A no-charge function that is integrated into the IBM z16, IBM z15, and IBM z14 LIC.
- Each OSC CHPID can support up to 120 console session connections, be shared among LPARs, and be spanned across multiple CSSs.
- Support for TN3270 Enhancements (TN3270E) and local non-SNA DFT 3270 emulation by using Ethernet-attached workstations.
- Console support that can be used by z/OS, z/VM, z/VSE, and z/TPF.
- Local non-SNA DFT 3270 and 328x printer emulation support that can be used by TSO/E, CICS, IMS, or any other 3270 application that communicates through VTAM.
- TLS/SSL with Certificate Authentication was added to the OSC CHPID to provide a secure and validated method for connecting clients to the IBM zSystems host. Up to 48 secure sessions per CHPID can be defined (the overall maximum of 120 connections is unchanged).

The OSA-ICC function provides several potential benefits:

- Simplicity

  External controllers are no longer needed.
- Scalable capacity:
  - Facilitates the addition of partitions and operations support pools.
  - Can be shared by up to 85 LPARs on IBM z16 A01, IBM z15 T01, and IBM z14 M0x.
  - Can be shared by up to 40 partitions on IBM z16 A02, IBM z16 AGZ, IBM z15 T02 and IBM z14 ZR1.
- Improved availability:
  - Can enable *lights out* operation.
  - Hot-plug OSA availability characteristics.
  - OSA features are a highly integrated component of the IBM zSystems platform, with the reliability, availability, and serviceability characteristics inherent in IBM zSystems.
- Low operating cost versus external console controller:
  - Power, cooling, cabling, and floor space requirements are minimized.
  - No rack necessary.

6.5 References

For more information about the OSA-ICC function, see the following publications:

- *OSA-Express Integrated Console Controller Implementation Guide*, SG24-6364
- *Open Systems Adapter Integrated Console Controller User's Guide,* SC27-9003
- *Input/Output Configuration Program Users's Guide for ICP IOCP*, SB10-7177

7

Shared Memory Communications

Shared Memory Communications (SMC) on IBM zSystems platform is a technology that can improve throughput by accessing data faster with less latency while reducing CPU resource consumption compared to traditional TCP/IP communications. Furthermore, applications do not need to be modified to gain the performance benefits of SMC.

This chapter includes the following topics:

- 7.1, “SMC overview” on page 114
- 7.2, “SMC over Remote Direct Memory Access” on page 116
- 7.3, “SMC over Direct Memory Access (intra-CPC)” on page 120
- 7.4, “Software support” on page 121
- 7.5, “Reference material” on page 125

7.1 SMC overview

SMC is a technology that allows two peers to send and receive data by using system memory buffers that each peer allocates for its partner's use. Two SMC protocols are available on the IBM zSystems platform:

- SMC - Remote Direct Memory Access (SMC-R)

  SMC-R is a protocol for Remote Direct Memory Access (RDMA) communication between TCP socket endpoints in logical partitions (LPARs) in different systems. SMC-R runs over networks that support RDMA over Converged Ethernet (RoCE). It enables existing TCP applications to benefit from RDMA without requiring modifications. SMC-R provides dynamic discovery of the RDMA capabilities of TCP peers and automatic setup of RDMA connections that those peers can use.

- SMC - Direct Memory Access (SMC-D)

  SMC-D implements the same SMC protocol that is used with SMC-R to provide highly optimized intra-system communications. Where SMC-R uses RoCE for communicating between TCP socket endpoints in separate systems, SMC-D uses Internal Shared Memory (ISM) technology for communicating between TCP socket endpoints in the same system. ISM provides adapter virtualization (virtual functions (VFs)) to facilitate the intra-system communications. So, SMC-D does not require any additional physical hardware (no adapters, switches, fabric management, or PCIe infrastructure). Therefore, significant cost savings can be achieved when using the ISM for LPAR-to-LPAR communication within the same IBM zSystems platform.

Important: Both SMC-R and SMC-D require an existing TCP link between the images that are configured to use the SMC protocol. With the SMC Version 2 (SMCv2) you can configure shared memory communication over multiple subnets.

Both SMC protocols use shared memory-architectural concepts, eliminating TCP/IP processing in the data path, yet preserving TCP/IP quality of service (QoS) for connection management purposes.

RoCE Express is the strategic adapter for Linux on IBM Z: In the future[a], IBM plans to shift from OSA-Express to PCIe-based networking devices like RoCE Express as the target strategic adapter type for IBM zSystems direct access networking connections to Linux operating systems. MES updates between generations are planned to be supported. Linux on IBM Z clients that indirectly access the OSA-Express adapter family through the IBM z/VM Virtual Switch (VSwitch) will be unaffected by this change.

Linux on IBM Z networking currently supports two Ethernet networking connectivity options: the OSA-Express adapter family and the RoCE Express adapter family. Using PCIe-based networking devices as provided by the RoCE Express adapter family is aligned with the deployment model for Linux on other architectural platforms, facilitates use of broader existing Linux ecosystem tools, and eases the effort to enable exploitation of industry hardware optimizations and integrate into industry software-defined networking models and tools, including Red Hat OpenShift Container Platform (OCP).

Clients are encouraged to plan for their adoption of RoCE Express adapters for IBM zSystems networking connectivity. IBM plans to continue to work toward common networking adapters for all operating systems on IBM Z, IBM LinuxONE, and Linux on IBM Z.

a. All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these SoDs is at the relying party's sole risk and will not create liability or obligation for IBM.

7.1.1 Remote Direct Memory Access

Remote Direct Memory Access (RDMA) is primarily based on InfiniBand technology. It has been available in the industry for many years. In computing, RDMA is direct memory access from the memory of one computer to that of another without involving either one's operating system (kernel). This permits high-throughput, low-latency network transfer, which is commonly used in massively parallel computing clusters.

There are two key requirements for RDMA, as shown in Figure 7-1 on page 115:

- A reliable lossless[1] network fabric (LAN for Layer 2 in data center network distance)
- An RDMA-capable NIC and Ethernet fabric

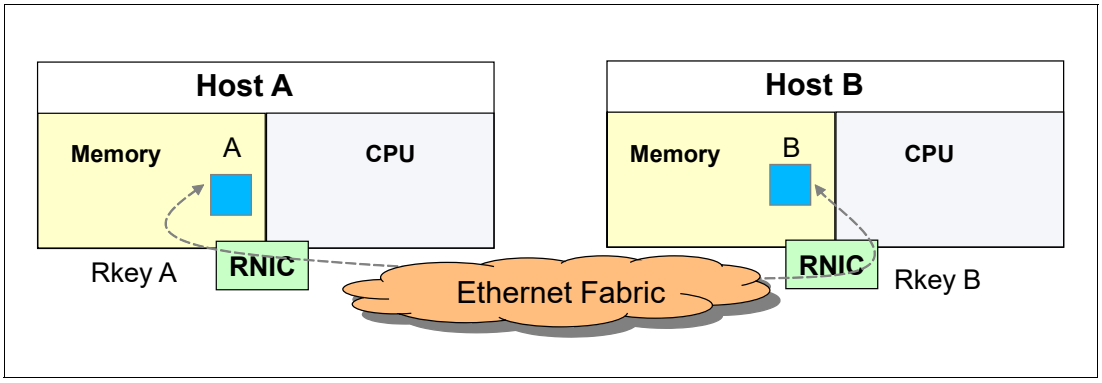


Figure 7-1 RDMA technology overview

RoCE uses existing Ethernet fabric (switches with Global Pause enabled as defined by the IEEE 802.3x port-based flow control) and requires RDMA-capable network interface cards (RNICs), such as 25 GbE RoCE Express3, 25 GbE RoCE Express2.1, 25 GbE RoCE Express2, 10 GbE RoCE Express3, 10 GbE RoCE Express2.1, and 10 GbE RoCE Express2 features.

7.1.2 Direct Memory Access

Intra-CPC Direct Memory Access (DMA) uses a virtual PCI adapter that is called ISM rather than an RNIC as with RDMA. The ISM interfaces are associated with IP interfaces (for example, HiperSockets or OSA-Express) and are dynamically created, automatically started and stopped, and are auto-discovered.

The ISM does not use queue pair (QP) technology like RDMA. Therefore, links and link groups based on QPs (or other hardware constructs) are not applicable to ISM. SMC-D protocol has a design concept of a *logical point-to-point connection* that is called an SMC-D link.

The ISM uses a virtual channel identifier (VCHID) similar to HiperSockets for addressing purposes. Figure 7-2 shows the SMC-D LPAR-to-LPAR communications concept.

[1] Network switching infrastructure requirements have relaxed with the new RoCE Express3 features.

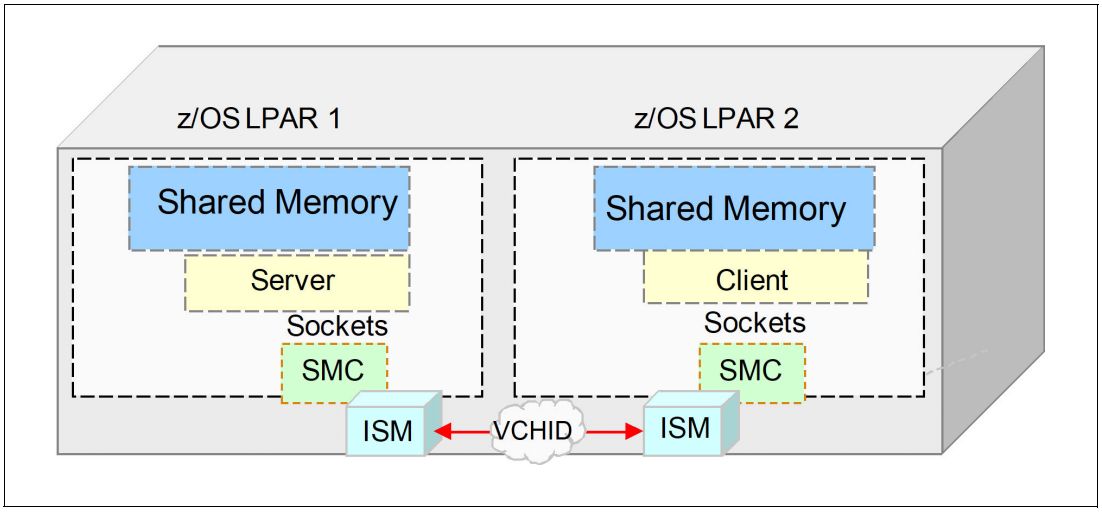


Figure 7-2 Connecting IBM z/OS LPARs in the same IBM zSystems platform by using ISMs

7.2 SMC over Remote Direct Memory Access

SMC-R is a protocol that allows TCP sockets applications to transparently use RDMA. SMC-R defines the concept of the SMC-R link, which is a logical point-to-point link that uses reliably connected queue pairs between TCP/IP stack peers over a RoCE fabric. An SMC-R link is bound to a specific hardware path, meaning a specific RNIC on each peer.

SMC-R is a hybrid solution (see Figure 7-3) for the following reasons:

- It uses TCP connection to establish the SMC-R connection.
- Switching from TCP to “out of band” SMC-R is controlled by a TCP option.
- SMC-R information is exchanged within the TCP data stream.
- Socket application data is exchanged through RDMA (write operations).
- TCP connection remains to control the SMC-R connection.
- This model preserves many critical existing operational and network management features of an IP network.

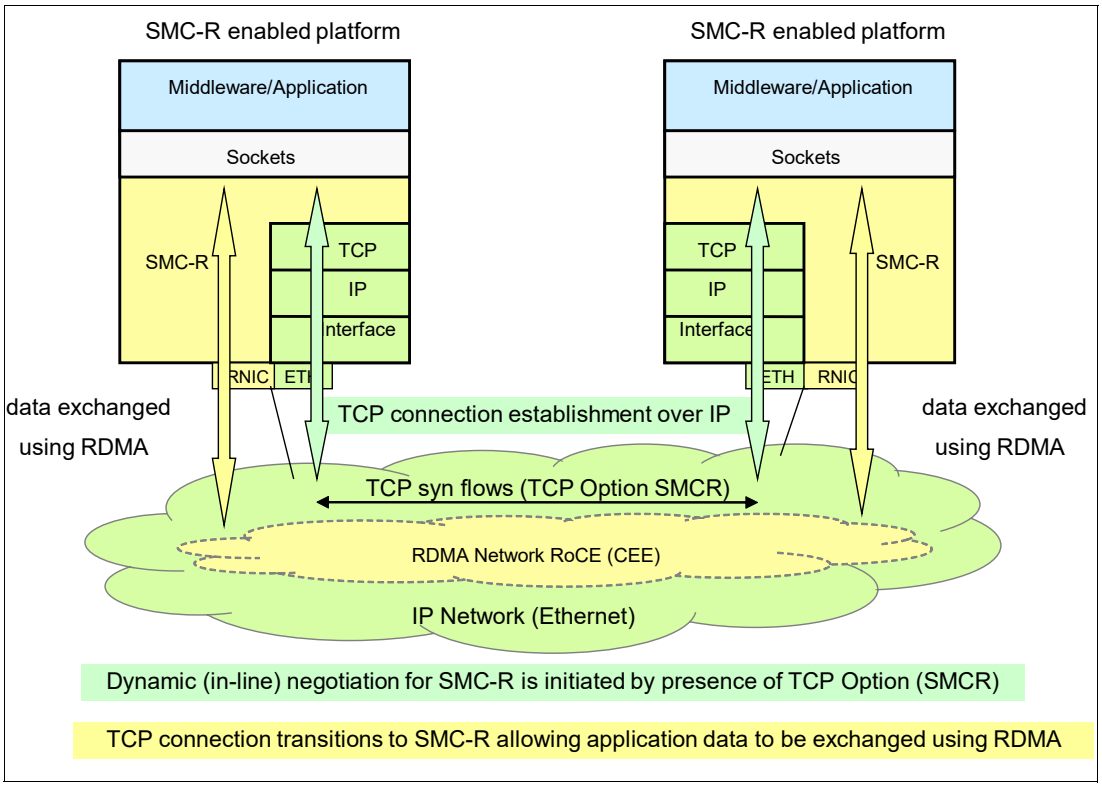


Figure 7-3 Dynamic transition from TCP to SMC-R

The hybrid model of SMC-R uses the following key existing attributes:

- Follows standard IP network connection setup.
- Dynamically switches to RDMA.
- TCP connection remains active (idle) and is used to control the SMC-R connection.
- Preserves critical operational and network management IP network features:
  - Minimal (or zero) IP topology changes.
  - Compatibility with TCP connection-level load balancers.
  - Preserves existing IP security model (for example, IP filters, policy, VLANs, or SSL).
  - Minimal network administrative or management changes.
- Changing host application software is not required, so all host application workloads can benefit immediately.

7.2.1 SMC-R connectivity

IBM introduced a new generation of RoCE that is called *RoCE Express3* with the IBM z16 A01, IBM z16 A02, and IBM z16 AGZ. This generation includes the 25GbE RoCE Express3 and 10GbE RoCE Express3 features.

With IBM z15, the 25GbE RoCE Express2.1 and 10GbE RoCE Express2.1 features were released. These features can be carried forward to IBM z16.

The previous generation, 25GbE RoCE Express2 and 10GbE RoCE Express2 are available with the IBM z14 and can be carried forward to IBM z16 and IBM z15.

The RoCE features are virtualized by the Resource Group LIC (firmware) that is running on a processor that is internally characterized as IFP. There is one IFP per system in IBM z15 and IBM z14, and two IFPs on IBM z16. The Resource Group provides VFs for each RoCE Express adapter that is installed in the IBM zSystems server.

RoCE Express2 and newer features support 126 VFs (63 per port). Older RoCE Express 10GbE supports only 31 VFs per feature and cannot be carried forward to IBM z16.

25GbE RoCE Express3 SR and LR

25GbE RoCE Express3 features are a technology update based on CX6 generation hardware, each provides two 25GbE physical ports. The features require 25GbE optics.

The 25GbE RoCE Express3 features are available in SR (FC 0452) and LR (FC 0453) versions.

IBM z16 A01 supports a maximum of 16 RoCE Express features (in any combination), while IBM z16 A02 and IBM z16 AGZ support a maximum of eight features. For 25GbE RoCE Express3, the PCI Function IDs (FIDs) are associated with a specific (single) physical port (that is port 0 or port 1).

25GbE RoCE Express2.1

IBM z15 introduced 25GbE RoCE Express2.1 (FC 0450), which is based on the existing RoCE Express2 generation hardware and provides two 25GbE physical ports. The features require 25GbE optics.

The following maximum number of features can be installed:

- Sixteen features in IBM z16 A01
- Sixteen features in IBM z15 T01
- Eight features in IBM z16 A02, IBM z16 AGZ, and IBM z15 T02

For 25GbE RoCE Express2.1, the PCI Function IDs (FIDs) are associated with a specific (single) physical port (that is port 0 or port 1).

25GbE RoCE Express2

The 25GbE RoCE Express2 (FC 0430) was introduced with IBM z14 and provides two 25GbE physical ports. The feature requires 25GbE optics. The following maximum number of features can be installed:

- Sixteen features in IBM z16 A01
- Sixteen features in IBM z15 T01
- Eight features in IBM z16 A02, IBM z16 AGZ, IBM z15 T02, and IBM z14 M0x
- Four features in IBM z14 ZR1

For 25GbE RoCE Express2, the PCI Function IDs (FIDs) are associated with a specific (single) physical port (that is port 0 or port 1).

10GbE RoCE Express3 SR and LR

IBM z16 introduces 10GbE RoCE Express3 SR and LR (FC 0440 and FC 0441), which is a technology update that is based on CX6 generation hardware and provides two 10 GbE physical ports. The features require 10 GbE optics.

IBM z16 A01 supports a maximum of 16 RoCE Express features (in any combination), while IBM z16 A02 and IBM z16 AGZ support a maximum of eight features. For 10GbE RoCE Express3, the PCI Function IDs (FIDs) are associated with a specific (single) physical port (that is port 0 or port 1).

10GbE RoCE Express2.1

The 10GbE RoCE Express2.1 is based on the existing RoCE Express2 generation hardware and provides two 10 GbE physical ports. The features require 10 GbE optics. The following maximum number of features can be installed:

- Sixteen features in IBM z16 A01
- Sixteen features in IBM z15 T01
- Eight features in IBM z16 A02, IBM z16 AGZ, and IBM z15 T02

Both ports are supported. For 10GbE RoCE Express2.1, the PCI Function IDs (FIDs) are associated with a specific (single) physical port (that is port 0 or port 1).

10GbE RoCE Express2

The 10GbE RoCE Express2 feature (FC 0412) are equipped with two physical 10 GbE ports. The following maximum number of features can be installed:

- Sixteen features in IBM z16 A01
- Sixteen features in IBM z15 T01
- Eight features in IBM z16 A02, IBM z16 AGZ, IBM z15 T02, and IBM z14 M0x
- Four features in IBM z14 ZR1

Both ports are supported. For 10GbE RoCE Express2, the PCI Function IDs (FIDs) are associated with a specific (single) physical port (that is port 0 or port 1).

Connectivity for 25GbE RoCE Express features

The 25GbE RoCE Express3 SR and LR and 25GbE RoCE Express2.x features require an Ethernet switch with 25GbE support. The switch port must support 25GbE (negotiation down to 10GbE is not supported).

A customer-supplied cable is required. The following types of cables can be used for connecting the port to the selected 25GbE switch or to another 25GbE RoCE Express feature on the attached IBM zSystems platform:

- 25GbE RoCE Express3 and 2.x SR features:
  - OM4 50-micron multimode fiber optic cable that is rated at 4700 MHz-km that is terminated with an LC duplex connector (up to 100 meters, or 328 feet).
  - OM3 50-micron multimode fiber optic cable that is rated at 2000 MHz-km that is terminated with an LC duplex connector (up to 70 meters, or 229.6 feet).
- 25GbE RoCE Express3 LR features: Single mode 9-micron fiber optic cable for a maximum connection distance of 10 km (6.2 miles).

Connectivity for 10GbE RoCE Express features

The 10GbE RoCE Express3 and 10GbE RoCE Express 2.x features are connected to 10 GbE switches (negotiation to another speed is not supported).

A customer-supplied cable is required. The following cable types can be used for connecting the port to the selected 10 GbE switch or to the RoCE Express feature on the attached IBM zSystems platform:

- 10GbE RoCE Express3 and 2.x SR features:
  - OM3 50-micron multimode fiber optic cable that is rated at 2000 MHz-km that is terminated with an LC duplex connector (up to 300 meters, or 984 feet).
  - OM2 50-micron multimode fiber optic cable that is rated at 500 MHz-km that is terminated with an LC duplex connector (up to 82 meters, or 269 feet).

  - OM1 62.5-micron multimode fiber optic cable that is rated at 200 MHz-km that is terminated with an LC duplex connector (up to 33 meters, or 108 feet).
- 10GbE RoCE Express3 LR features: Single mode 9-micron fiber optic cable for a maximum connection distance of 10 km (6.2 miles).

7.3 SMC over Direct Memory Access (intra-CPC)

SMC-D is a protocol that allows TCP socket applications to transparently use ISM. It is a hybrid solution (see Figure 7-4) and has the following features:

- It uses a TCP connection to establish the SMC-D connection.
- The TCP connection can be either through OSA Adapter or IQD HiperSockets.
- A TCP option (SMCD) controls switching from TCP to *out-of-band* SMC-D.
- The SMC-D information is exchanged within the TCP data stream.
- Socket application data is exchanged through ISM (write operations).
- The TCP connection remains to control the SMC-D connection.
- This model preserves many critical existing operational and network management features of TCP/IP.

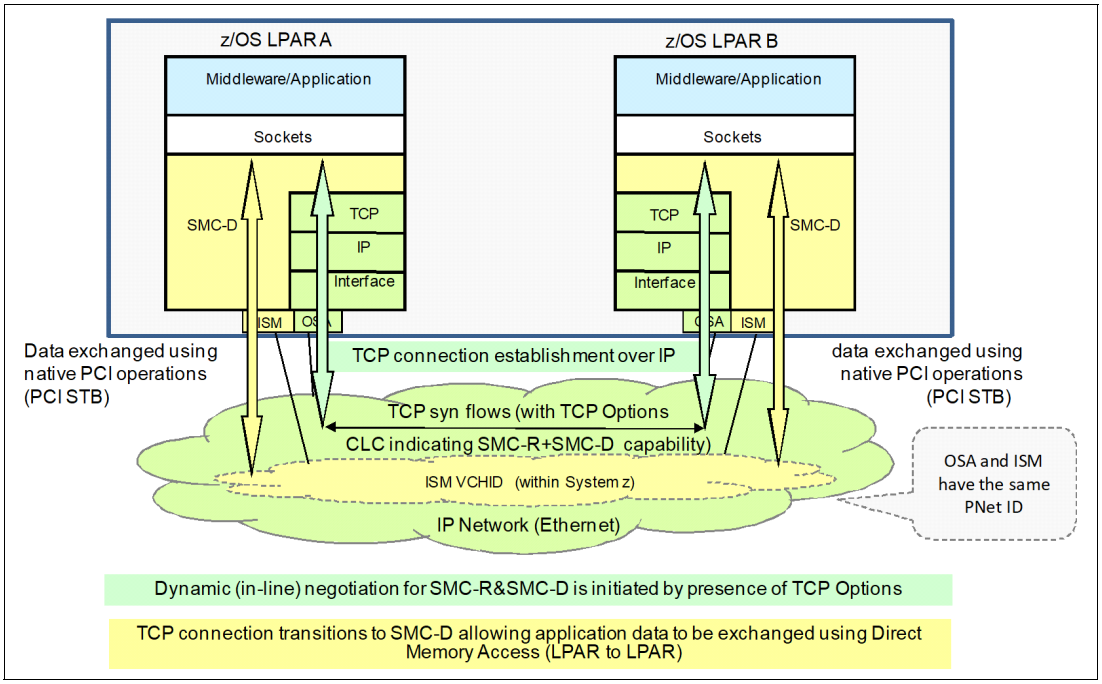


Figure 7-4 Dynamic transition from TCP to SMC-D

The hybrid model of SMC-D uses these key existing attributes:

- It follows the standard TCP/IP connection setup.
- The hybrid model switches to ISM (SMC-D) dynamically.
- The TCP connection remains active (idle) and is used to control the SMC-D connection.

- The hybrid model preserves the following critical operational and network management TCP/IP features:
  - Minimal (or zero) IP topology changes.
  - Compatibility with TCP connection-level load balancers.
  - Preservation of the existing IP security model, such as IP filters, policies, virtual local area networks (VLANs), and Secure Sockets Layer (SSL).
  - Minimal network administration and management changes.
- Host application software is not required to change, so all host application workloads can benefit immediately.
- The TCP path can be either through an OSA-Express port or HiperSockets connection.

7.4 Software support

This section provides software support information for SMC-R and SMC-D.

The SMC-R and SMC-D are hybrid protocols that have been designed and implemented to speed up TCP communication without changing applications.

SMC-R uses the RoCE Express features for providing (direct memory access) data transfer under the control of an established TCP/IP connection (OSA). SMC-D can use also an established TCP/IP connection over HiperSockets.

7.4.1 SMC-R (Version1 and Version 2)

This section describes the following topics:

- z/OS and IBM z/VM
- Linux on IBM Z

z/OS and IBM z/VM

SMC-R with RoCE provides high-speed communications performance across physical processors. It helps all TCP-based communications across z/OS LPARs that are in different central processor complexes (CPCs). It also can be used on supported Linux on IBM Z distribution for Linux to Linux or Linux to z/OS communications.

Here are some typical communication patterns:

- Optimized Sysplex Distributor intra-sysplex load balancing.
- IBM WebSphere Application Server Type 4 connections to remote IBM Db2, IBM Information Management System, and IBM CICS instances.
- IBM Cognos®-to-Db2 connectivity.
- CICS-to-CICS connectivity through Internet Protocol interconnectivity (IPIC).

IBM z/OS V2.1 or later with program temporary fixes (PTFs) supports the SMC-R protocol with RoCE. With IBM z/OS V2R4 and later support for SMC-Rv2 is also provided.

Also, consider the following factors:

- No rollback to previous z/OS releases.
- Requires Input/Output Configuration Program (IOCP) 3.4.0 or later.

- Extra PTFs are required to support 25GbE RoCE Express3 SR and LR, and 25GbE RoCE Express2.x.
- z/VM V7.1 and later support guest use of the RoCE Express feature of IBM zSystems. This feature allows guests to use RoCE for optimized networking. PTFs are required to support 25GbE RoCE Express3.

Linux on IBM Z

For Linux on IBM Z, the Linux distribution partners included SMC-D and SMC-R support. SMC running on Linux on IBM Z LPARs can be used also to communicate with LPARs running z/OS. The following minimum requirements must be fulfilled:

- RHEL 8
- SUSE Linux Enterprise Server 12 SP4 (kernel 4.12.14-95.13.1 or higher)
- SUSE Linux Enterprise Server 15 SP1
- Ubuntu 18.10

Linux has introduced also support for SMC-Rv2, available in Linux kernel 5.10 or later, and the following Linux distribution(s):

- Ubuntu 21.04
- RHEL 8.4
- SLES 15 SP3

SMC-Rv2 is available in Linux kernel 5.16, and requires smc-tools v1.17.

For more information about RoCE features support, check with the distribution owners.

IBM AIX® 7.2 Technology Level 2 is enhanced to provide SMC-R support and transparently use RDMA, which enables direct, high-speed, and low-latency communications.

7.4.2 SMC-D (Version 1 and Version 2)

This section describes the following topics:

- z/OS and z/VM
- Linux on IBM Z
- SMC-Dv2 Compatibility Support

z/OS and z/VM

SMC-D has the following prerequisites:

- IBM z16, IBM z15, and IBM z14 with HMC/SE for ISM vPCI functions.
- At least two z/OS V2.2 (or later) or supported Linux on IBM Z LPARs in the same IBM zSystems platform with the required services installed:
  - SMC-D can communicate with another z/OS V2.2 (or later) or supported Linux on IBM Z instances. The peer hosts must be in the same ISM PNet.
  - SMC-D requires an IP network with access that uses OSA-Express or HiperSockets, which have a defined PNet ID that matches the ISM PNet ID.
  - With IBM z/OS V2R4 and later support for SMCv2 is also provided.
- z/VM supports guest access to RoCE (for Guest Exploitation only).

Linux on IBM Z

For Linux on IBM Z, the Linux distribution partners included SMC-D and SMC-R support. SMC running on Linux on IBM Z LPARs can be used also to communicate with LPARs running z/OS. The following minimum requirements must be fulfilled:

- RHEL 8, RHEL 9
- SUSE Linux Enterprise Server 12 SP4 (kernel 4.12.14-95.13.1 or higher)
- SUSE Linux Enterprise Server 15 SP1
- Ubuntu 18.10, Ubuntu 20.04, Ubuntu 22.04

SMC-Dv2 Compatibility Support

z/OS announced SMC-Dv2 support. When SMC-Dv2 support is enabled in z/OS, a compatibility patch is required in Linux on IBM Z, which is available in the following minimum Linux distribution levels:

- RHEL 8.1, Linux kernel 4.18.0-147.27.1
- RHEL 8.2, Linux kernel 4.18.0-193.28.1
- RHEL 8.3, Linux kernel 4.18.0-228
- SUSE Linux Enterprise Server 12 SP5, Linux kernel 4.12.14-122.41.1
- SUSE Linux Enterprise Server 15 SP1, Linux kernel 4.12.14-197.61.1
- SUSE Linux Enterprise Server 15 SP2, Linux kernel 5.3.18-24.9.1
- Ubuntu 20.04, Linux kernel 5.4.0-45.49

Note: SMC (the existing architecture) cannot be used in the following circumstances:

- Peer hosts are not within the same IP subnet and VLAN.
- TCP traffic requires IPsec or the server uses FRCA.

7.4.3 Shared Memory Communication Version 2

Shared Memory Communications v2 (SMCv2) is available in z/OS V2R4 (with PTFs) and z/OS V2R5 (see Figure 7-5 on page 124).

The initial version of SMC (SMCv1) was limited to TCP/IP connections over the same Layer 2 network and was not routable across multiple IP subnets. The associated TCP/IP connection was limited to hosts within a single IP subnet, which 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.

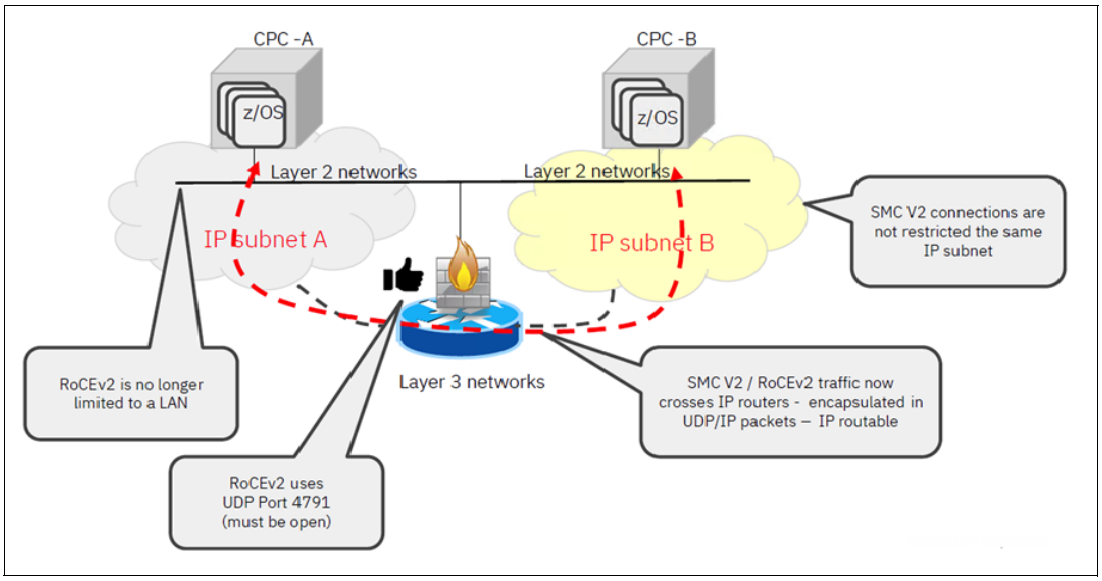


Figure 7-5 SMCv2 for z/OS logical diagram

SMCv2 supports SMC over multiple IP subnets for both SMC-D and SMC-R. It also 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 compatible with earlier versions so that SMCv2 hosts can continue to communicate with SMCv1 hosts.

Although SMCv2 changes the SMC connection protocol to enable multiple IP subnet support, SMCv2 does not change how the 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.

Shared Memory Communications v2: Linux support

SMC-Dv2 is available in Linux kernel 5.10 or later and the following Linux distributions:

- Ubuntu 21.04
- RHEL 8.4
- SUSE Linux Enterprise Server 15 SP3

SMC-Rv2 is available in Linux kernel 5.16, and it requires smc-tools V1.17.

When SMC-Dv2 support is enabled in z/OS, a compatibility patch is required in Linux on IBM Z, which is available at the following minimum Linux distribution levels:

- RHEL 8.1, Linux kernel 4.18.0-147.27.1
- RHEL 8.2, Linux kernel 4.18.0-193.28.1
- RHEL 8.3, Linux kernel 4.18.0-228
- SUSE Linux Enterprise Server 12 SP5, Linux kernel 4.12.14-122.41.1
- SUSE Linux Enterprise Server 15 SP1, Linux kernel 4.12.14-197.61.1
- SUSE Linux Enterprise Server 15 SP2, Linux kernel 5.3.18-24.9.1
- Ubuntu 20.04, Linux kernel 5.4.0-45.49

7.5 Reference material

For more information, see the following resources:

- IETF RFC for SMC-R
- Shared Memory Communications for Linux on IBM Z
- IBM z/OS Communications Server
- *IBM z16 (3931) Technical Guide*, SG24-8951
- *IBM z16 A02 and IBM AGZ Technical Guide*, SG24-8952
- *IBM z15 (8561) Technical Guide*, SG24-8851
- *IBM z14 (3906) Technical Guide*, SG24-8451
- *IBM z14 ZR1 Technical Guide*, SG24-8651

8

HiperSockets

This chapter presents a high-level overview of the IBM HiperSockets capabilities as they pertain to the IBM zSystems platform.

This chapter includes the following topics:

- 8.1, “Overview” on page 128
- 8.2, “Connectivity” on page 137
- 8.3, “Summary” on page 140
- 8.4, “References” on page 140

8.1 Overview

The HiperSockets function, also known as internal queued direct input/output (iQDIO) or internal QDIO, is an integrated function of the Licensed Internal Code (LIC) of the IBM zSystems platform. It provides an attachment to high-speed logical LANs with minimal system and network overhead.

HiperSockets provides internal virtual local area networks that act like IP networks within the IBM zSystems platform. Therefore, HiperSockets provides the fastest IP network communication between consolidated Linux, IBM z/VM, IBM z/VSE, and IBM z/OS virtual servers on a IBM zSystems platform.

The virtual servers form a virtual local area network. Using iQDIO, the communication between virtual servers is through I/O queues that are set up in the system memory of the IBM zSystems platform. Traffic between the virtual servers is passed at memory speeds. For more information about the number of HiperSockets that are available for each IBM zSystems platform, see 8.2, “Connectivity” on page 137.

This LIC function, which is coupled with supporting operating system device drivers, establishes a higher level of network availability, security, simplicity, performance, and cost effectiveness than is available when connecting single servers or logical partitions (LPARs) together by using an external IP network.

HiperSockets is supported by the following operating systems:

- All in-service z/OS releases
- All in-service z/VM releases
- All in service z/VSE releases
- Linux on IBM zSystems and the KVM hypervisor host

Note: Throughout this chapter, we provide operating system support information for the functions described. Not every operating system supports all features. The operating system support information is provided for IBM z/OS, IBM z/VM, IBM z/VSE, IBM z/TPF, and Linux on IBM Z (supported distributions).

Regarding KVM hypervisor, KVM support is provided by Linux distribution partners. For more information about KVM support for the IBM zSystems platform, see your distribution’s documentation.

8.1.1 HiperSockets benefits

Using the HiperSockets function has several benefits:

- HiperSockets eliminates the need to use I/O subsystem operations and the need to traverse an external network connection to communicate between LPARs in the same IBM zSystems platform.
- HiperSockets offers significant value in server consolidation, connecting many virtual servers in the same IBM zSystems CPC. It can be used rather than certain coupling link configurations in a Parallel Sysplex cluster. All of the consolidated hardware servers can be eliminated, along with the cost, complexity, and maintenance of the networking components that connect them.
- Consolidated servers that must access data on the IBM zSystems CPC can do so at memory speeds, bypassing all of the network overhead and delays.

- HiperSockets can be customized to accommodate varying traffic sizes. In contrast, LANs (such as Ethernet) have a maximum frame size that is predefined by their architecture. With HiperSockets, a maximum frame size can be defined according to the traffic characteristics transported for each of the possible HiperSockets virtual local area networks.
- Because there is no server-to-server traffic outside of the IBM zSystems platform, a much higher level of network availability, security, simplicity, performance, and cost-effectiveness is achieved, compared to servers that communicate across an external LAN. For example:
  - Because the HiperSockets feature has no external components, it provides a secure connection. For security purposes, servers can be connected to different HiperSockets. All security features, such as firewall filtering, are available for HiperSockets interfaces, as they are for other IP network interfaces.
  - HiperSockets looks like any other IP network interface. Therefore, it is apparent to applications and supported operating systems.
- HiperSockets can also improve IP network communications within a sysplex environment when the DYNAMICXCF facility is used.

8.1.2 Server integration with HiperSockets

Many data center environments are multi-tiered server applications, with various middle-tier servers that surround the IBM z16, IBM z15, and IBM z14 data and transaction servers. Interconnecting multiple servers affects the cost and complexity of many networking connections and components. The performance and availability of the interserver communication depends on the performance and stability of the set of connections. The more servers that are involved, the greater the number of network connections and complexity to install, administer, and maintain.

Figure 8-1 shows two configurations. The configuration on the left shows a server farm that surrounds an IBM zSystems platform, with its corporate data and transaction servers. This configuration has a great deal of complexity that is involved in the backup of the servers and network connections. This environment also results in high administration costs.

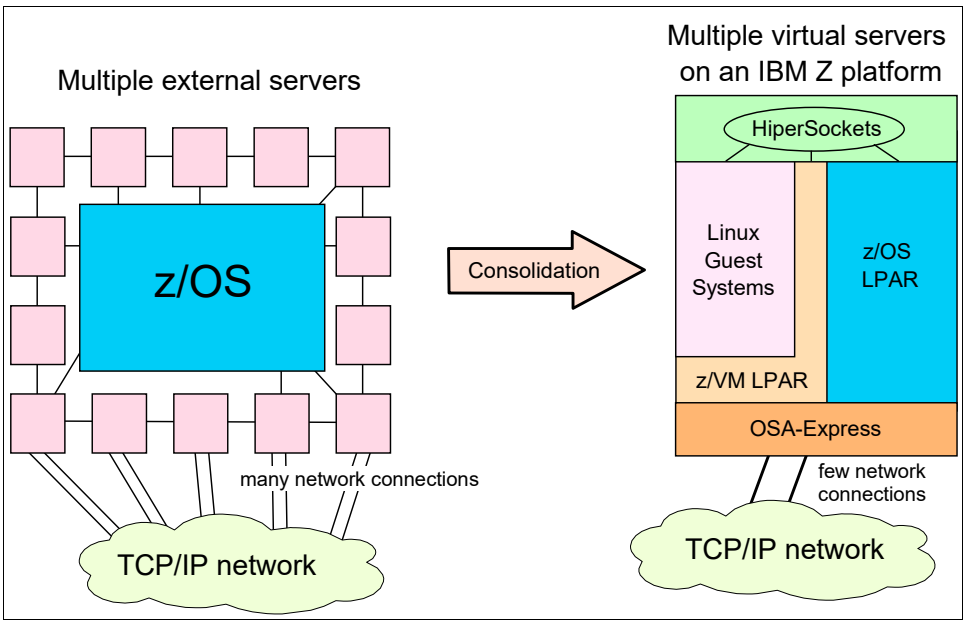


Figure 8-1 Server consolidation

Consolidating that mid-tier workload onto multiple Linux virtual servers that run on an IBM Z platform requires a reliable, high-speed network for those servers to communicate over. HiperSockets provides that. In addition, those consolidated servers also have direct high-speed access to database and transaction servers that are running under z/OS on the same IBM zSystems platform.

This configuration is shown on the right side in Figure 8-1 on page 129. Each consolidated server can communicate with others on the IBM zSystems platform through HiperSockets. In addition, the external network connection for all servers is concentrated over a few high-speed OSA-Express, and possibly RoCE interfaces.

8.1.3 HiperSockets function

HiperSockets implementation is based on the OSA-Express QDIO protocol. Therefore, HiperSockets is also called internal QDIO (iQDIO). The LIC emulates the link control layer of an OSA-Express QDIO interface.

Typically, before a packet can be transported on an external LAN, a LAN frame must be built. The MAC address of the destination host or router on that LAN must be then inserted into the frame. HiperSockets does not use LAN frames, destination hosts, or routers. IP network stacks are addressed by inbound data queue addresses rather than MAC addresses.

The IBM z16, IBM z15, and IBM z14 LICs maintain a lookup table of IP addresses for each HiperSockets function. This table represents a virtual local area network. At the time that an IP network stack starts a HiperSockets device, the device is registered in the IP address lookup table with its IP address and its input and output data queue pointers. If an IP network device is stopped, the entry for this device is deleted from the IP address lookup table.

HiperSockets copies data synchronously from the output queue of the sending IP network device to the input queue of the receiving IP network device by using the memory bus to copy the data through an I/O instruction.

The controlling operating system that performs I/O processing is identical to OSA-Express in QDIO mode. The data transfer time is similar to a cross-address space memory move, with hardware latency close to zero. For total elapsed time for a data move, the operating system I/O processing time must be added to the LIC data move time.

HiperSockets operations run on the processor where the I/O request is initiated by the operating system. HiperSockets starts write operations. The completion of a data move is indicated by the sending side to the receiving side with a signal adapter (SIGA) instruction. Optionally, the receiving side can use dispatcher polling rather than handling SIGA interrupts. The I/O processing is performed without using the system assist processor (SAP). This implementation is also called *thin interrupt*.

The data transfer is handled much like a cross-address space memory move that uses the memory bus, not the IBM zSystems I/O bus. Therefore, HiperSockets does not contend with other system I/O activity in the system. Figure 8-2 on page 131 shows the basic operation of HiperSockets.

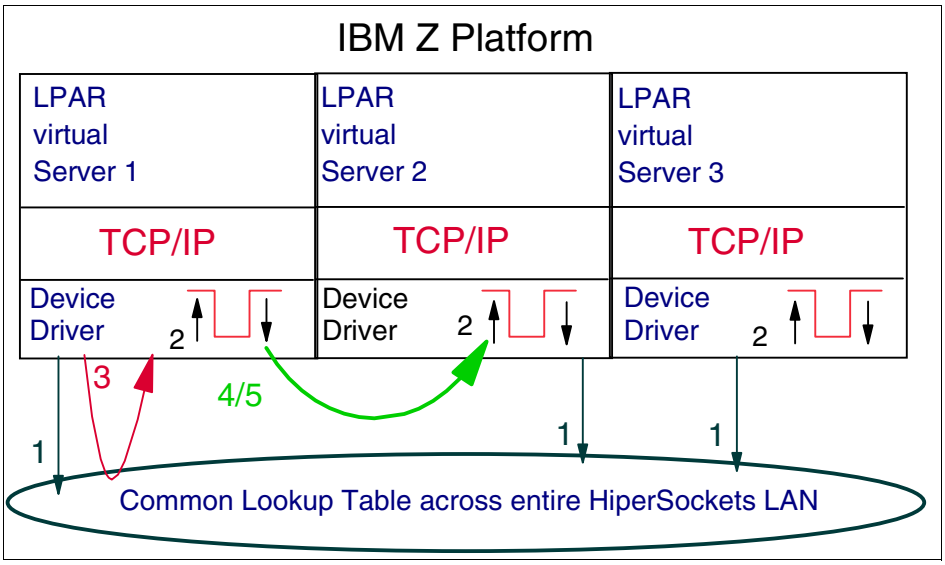


Figure 8-2 HiperSockets basic operation

The HiperSockets operational flow consists of five steps:

1. Each IP network stack registers its IP addresses in a server-wide common address lookup table. There is one lookup table for each HiperSockets virtual local area network. The scope of each LAN is the LPARs that are defined to share the HiperSockets IQD channel path identifier (CHPID).
2. The address of the IP network stack's receive buffers is appended to the HiperSockets queues.
3. When data is being transferred, the **send** operation of HiperSockets performs a table lookup for the addresses of the sending and receiving IP network stacks and their associated send and receive buffers.
4. The sending virtual server copies the data from its send buffers into the target virtual server's receive buffers (in system memory).
5. The sending virtual server optionally delivers an interrupt to the target IP network stack. This optional interrupt uses the *thin interrupt* support function of the IBM zSystems platform. This feature means that the receiving virtual server looks *ahead*, detecting and processing inbound data. This technique reduces the frequency of real I/O or external interrupts.

Hardware assists

A complementary virtualization technology that includes the following features is available for the IBM zSystems platform:

- QDIO Enhanced Buffer-State Management (QEBSM)

  Two hardware instructions that are designed to help eliminate the overhead of hypervisor interception.

- Host Page-Management Assist (HPMA)

  An interface to the z/VM main storage management function designed to allow the hardware to assign, lock, and unlock page frames without z/VM hypervisor assistance.

These hardware assists allow a cooperating guest operating system to start QDIO operations directly to the applicable channel, without interception by the z/VM operating system. This

process improves performance. Support is integrated in IBM zSystems. However, always check the suitable FIXCATs.

8.1.4 Supported functions

This section describes other functions that are supported by HiperSockets technology.

Broadcast support

Broadcasts are supported across HiperSockets on Internet Protocol version 4 (IPv4) for applications. Applications that use the broadcast function can propagate the broadcast frames to all IP network applications that are using HiperSockets. This support is applicable to Linux on IBM Z, z/OS, and z/VM environments.

Virtual local area network support

Virtual local area networks (VLANs), IEEE standard 802.1q, are supported by Linux on IBM Z and z/OS for HiperSockets. VLANs can reduce overhead by allowing networks to be organized by traffic patterns rather than physical location. This enhancement permits traffic flow on a VLAN connection both over HiperSockets and between HiperSockets and OSA-Express Ethernet features.

IPv6 support

HiperSockets supports Internet Protocol version 6 (IPv6). IPv6 is the protocol that was designed by the Internet Engineering Task Force (IETF) to replace IPv4 to help satisfy the demand for more IP addresses.

The support of IPv6 on HiperSockets (CHPID type IQD) is available on the IBM zSystems platform, and is supported by z/OS and z/VM. IPv6 support is available on the OSA-Express7S, OSA-Express6S, OSA-Express5S, and OSA-Express4 features in the z/OS, z/VM, and Linux on IBM Z environments.

Support of guests is expected to be apparent to z/VM if the device is directly connected to the guest (pass through).

HiperSockets Network Concentrator

Traffic between HiperSockets and OSA-Express can be transparently bridged by using the HiperSockets Network Concentrator. This technique does not require intervening network routing overhead, thus increasing performance and simplifying the network configuration. This goal is achieved by configuring a *connector* Linux system that has HiperSockets and OSA-Express connections that are defined to it.

The HiperSockets Network Concentrator registers itself with HiperSockets as a special network entity to receive data packets that are destined for an IP address on the external LAN through an OSA port. The HiperSockets Network Concentrator also registers IP addresses to the OSA feature on behalf of the IP network stacks by using HiperSockets, thus providing inbound and outbound connectivity.

HiperSockets Network Concentrator support uses the next-hop IP address in the QDIO header, rather than a Media Access Control (MAC) address. Therefore, VLANs in a switched Ethernet fabric are not supported by this support. IP network stacks that use only HiperSockets to communicate with no external network connection see no difference, so the HiperSockets support and networking characteristics are unchanged.

To use HiperSockets Network Concentrator unicast and multicast support, a Linux distribution is required. You also need s390-tools (tools for use with the IBM S/390® Linux kernel and device drivers), which are available at this page.

HiperSockets Layer 2 support

The IBM HiperSockets feature supports two transport modes on the IBM zSystems platform:

- Layer 2 (link layer)
- Layer 3 (Network and IP layer)

HiperSockets is protocol-independent and supports the following traffic types:

- Internet Protocol (IP) version 4 or version 6
- Non-IP (such as AppleTalk, DECnet, IPCX, NetBIOS, and SNA)

Each HiperSockets device has its own Layer 2 MAC address and allows the use of applications that depend on a Layer 2 address, such as DHCP servers and firewalls. LAN administrators can configure and maintain the mainframe environment in the same fashion as they do in other environments. This feature eases server consolidation and simplifies network configuration.

The HiperSockets device automatically generates a MAC address to ensure uniqueness within and across LPARs and servers. MAC addresses can be locally administered, and the use of group MAC addresses for multicast and broadcasts to all other Layer 2 devices on the same HiperSockets network is supported. Datagrams are delivered only between HiperSockets devices that use the same transport mode (for example, Layer 2 with Layer 2 and Layer 3 with Layer 3).

A HiperSockets device can filter inbound datagrams by VLAN identification, the Ethernet destination MAC address, or both. This feature reduces the amount of inbound traffic, which leads to lower processor use by the operating system.

As with Layer 3 functions, HiperSockets Layer 2 devices can be configured as primary or secondary connectors or multicast routers that enable high-performance and highly available Link Layer switches between the HiperSockets network and an external Ethernet.

HiperSockets Layer 2 support is available on the IBM zSystems platform with Linux on IBM Z and by z/VM guest use.

HiperSockets multiple write facility

HiperSockets performance has been increased by allowing streaming of bulk data over a HiperSockets link between LPARs. The receiving partition can process larger amounts of data per I/O interrupt. The improvement is apparent to the operating system in the receiving partition. Multiple writes with fewer I/O interrupts reduce processor use of both the sending and receiving LPARs and is supported in z/OS.

zIIP-Assisted HiperSockets for large messages

In z/OS, HiperSockets is enhanced for IBM zSystems Integrated Information Processor (zIIP) use. Specifically, the z/OS Communications Server allows the HiperSockets Multiple Write Facility processing of large outbound messages that originate from z/OS to be performed on zIIP.

z/OS application workloads that are based on XML, HTTP, SOAP, Java, and traditional file transfer can benefit from zIIP enablement by lowering general-purpose processor use.

When the workload is eligible, the HiperSockets device driver layer processing (write command) is redirected to a zIIP, which unblocks the sending application.

HiperSockets Network Traffic Analyzer

HiperSockets Network Traffic Analyzer (HS NTA) is a function that is available in the IBM LIC. It can make problem isolation and resolution simpler by allowing Layer 2 and Layer 3 tracing of HiperSockets network traffic.

With HiperSockets NTA, Linux on IBM Z can control the tracing of the internal VLAN. It captures records into host memory and storage (file systems) that can be analyzed by system programmers and network administrators by using Linux on IBM Z tools to format, edit, and process the trace records.

With a customized HiperSockets NTA rule, you can authorize an LPAR to trace messages only from LPARs that are eligible to be traced by the NTA on the selected IQD channel.

HS NTA rules can be set up on the Support Element (SE). There are four types of rules for the HS NTA:

- Tracing is unavailable for all IQD channels in the system (the default rule).
- Tracing is unavailable for a specific IQD channel.
- Tracing is allowed for a specific IQD channel. All LPARS can be set up for NTA, and all LPARs are eligible to be traced by an active Network Traffic Analyzer.
- Customized tracing is allowed for a specific IQD channel.

HiperSockets Completion Queue

The HiperSockets Completion Queue function enables HiperSockets to transfer data synchronously, if possible, and asynchronously if necessary. This process combines ultra-low latency with more tolerance for traffic peaks. With the asynchronous support, during high volume situations, data can be temporarily held until the receiver has buffers available in its inbound queue. This feature provides end-to-end performance improvement for LPAR to LPAR communication.

The HiperSockets Completion Queue function is supported on the z/OS, z/VSE, and Linux on IBM Z operating systems, and z/VM supports guest exploitation.

HiperSockets Completion Queue is supported by Linux on IBM Z through AF_IUCV socket communication. Fast Path to Linux in a Linux LPAR requires the HiperSockets Completion Queue function of the IBM zSystems platform.

HiperSockets virtual switch bridge support

The 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 either of these points:

- Other guest virtual machines on the virtual switch
- External network hosts, through the virtual switch OSA UPLINK port

Note: IBM z/VM 6.2[a] or later, IP network, and Performance Toolkit APARs are required for this support.

a. The earliest z/VM version in support at the time of writing is 7.1

A HiperSockets channel alone can provide only intra-CPC communications. The HiperSockets bridge port allows a virtual switch to connect IBM z/VM guests by using real HiperSockets devices. This feature can communicate with hosts that are external to the CPC. A single IP address and virtual machine network connection can be used to communicate over the internal and external segments of the LAN. The fact that any particular destination address might be on the local HiperSockets channel or outside of the CPC is apparent to the bridge-capable port.

Incorporating the HiperSockets channel into the flat Layer 2 broadcast domain through OSD adapters simplifies networking configuration and maintenance. The virtual switch HiperSockets bridge port eliminates the need to configure a separate next-hop router on the HiperSockets channel to provide connectivity to destinations that are outside of a HiperSockets channel. This configuration avoids the need to create routes for this internal route in all hosted servers and the extra hop of a router to provide the Layer 3 routing functions.

Figure 8-3 shows an example of a bridged HiperSockets configuration.

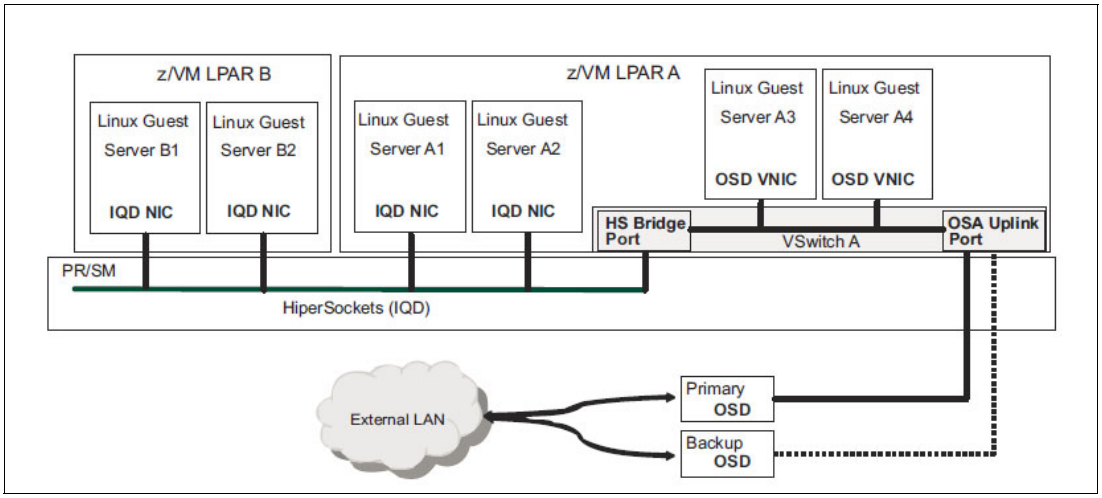


Figure 8-3 Bridge HiperSockets channels

The virtual switch HiperSockets bridge support expands the use cases and capabilities of the HiperSockets channel to include the following items:

- Full-function, industry-standard robust L2 bridging technology.
- Single NIC configuration, which simplifies network connectivity and management.
- No guest configuration changes are required for use (apparent to guest OS).
- Live Guest Relocation (LGR) of guests with real HiperSockets bridge-capable IQD connections within and between bridged CPCs.
- No limit on the number of z/VM LPARs that can participate in a bridged HiperSockets LAN.
- Ability to create a single broadcast domain across multiple CPCs (Cross CPC bridged HiperSockets channel network).
- Highly available network connection to the external network, provided by the z/VM virtual switch by default.

Figure 8-4 shows a sample Cross CPC HiperSockets LAN configuration. This configuration enables the creation of a single broadcast domain across CPCs and HiperSockets channels at the same time delivering a highly available configuration on both CPCs.

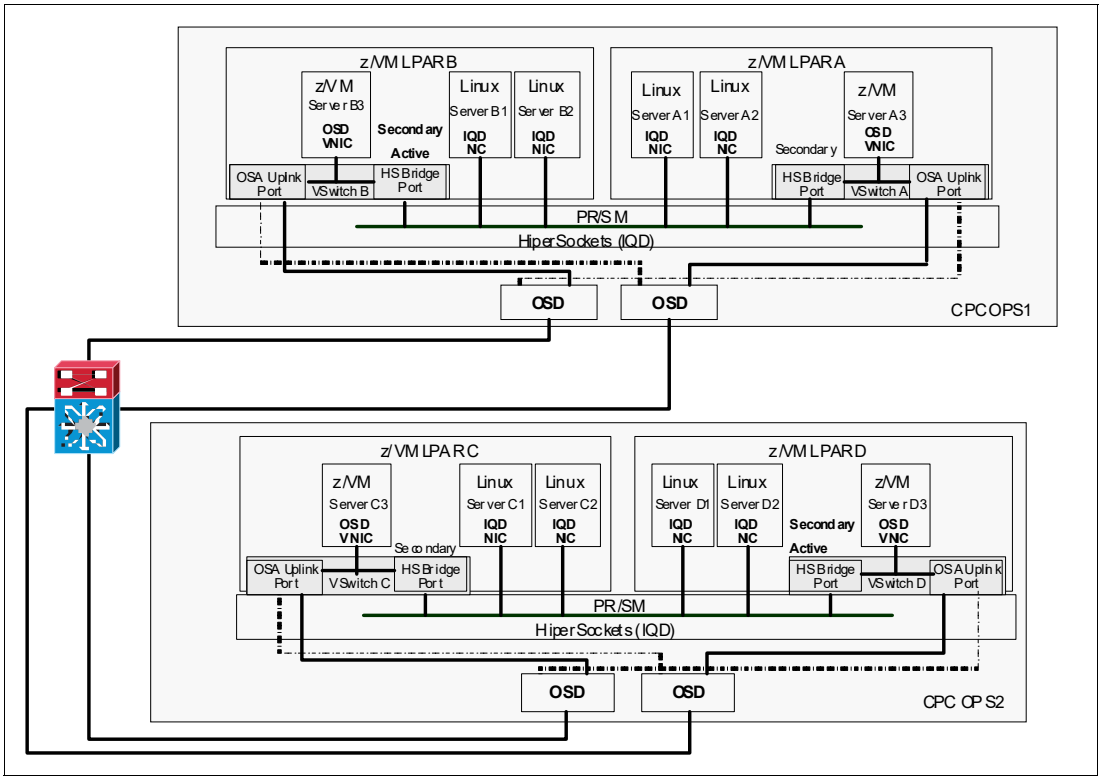


Figure 8-4 HiperSockets LAN spanning multiple CPCs

In this configuration, VSwitch B in LPAR B and VSwitch D in LPAR D are the active bridge ports that provide external connectivity between the external bridged IQD channel in CPC OPS1 and the external IQD channel in CPC OPS2. This flat Layer 2 LAN essentially joins or extends the HiperSockets LAN between CPCs across the external Ethernet network VSwitch UPLINK port and VSwitch D UPLINK port.

IBM z/VSE Fast Path to Linux

Fast Path to Linux allows z/VSE IP network applications to communicate with an IP network stack on Linux without using an IP network stack on z/VSE. Fast Path to Linux in an LPAR requires that the HiperSockets Completion Queue function is available on IBM zSystems platforms. The Fast Path to Linux function is supported starting with z/VSE 5.1.1[1] (version 5, release 1.1).

[1] Supported z/VSE at the time of writing is Version 6.2.

Figure 8-5 shows a sample configuration of z/VSE and the Fast Path to Linux function. z/VSE applications can directly communicate with Linux IP network through the HiperSockets without involving the IP network stack of z/VSE.

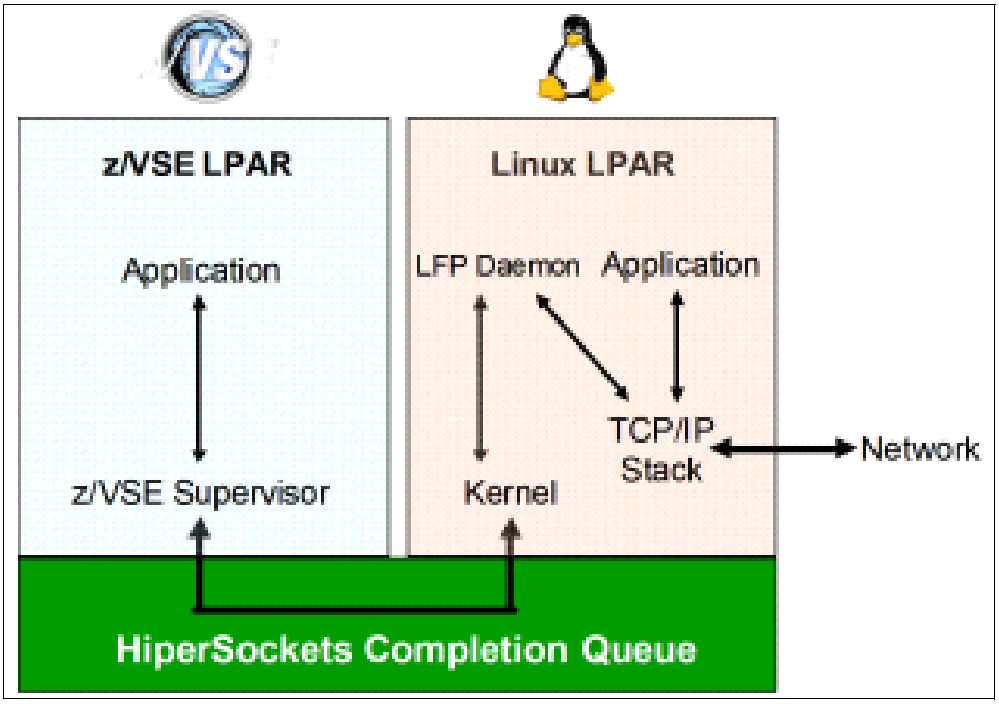


Figure 8-5 Fast Path to Linux on IBM Z in an LPAR

z/VSE z/VM IP Assist

The IBM zSystems platform processors provide an appliance that is called z/VSE z/VM IP Assist (IBM VIA®) that can be used by Linux Fast Path. With VIA, the processor provides a function to communicate to the Internet Protocol network without having a Linux distribution that is installed. This appliance is used for z/VSE running under z/VM.

8.2 Connectivity

HiperSockets has no external components or external network. There is no internal or external cabling. The HiperSockets data path does not go outside of one physical IBM zSystems server.

HiperSockets is not allocated a CHPID until it is defined. It does not occupy an I/O drawer, or a PCIe I/O drawer slot. HiperSockets cannot be enabled if all the available CHPIDs on the IBM zSystems platform are used. Therefore, HiperSockets must be included in the overall channel I/O planning.

HiperSockets IP network devices are configured similarly to OSA-Express QDIO devices. Each HiperSockets requires the definition of a CHPID like any other I/O interface. The CHPID type for HiperSockets is IQD, and the CHPID number must be in the range of hex 00 to hex FF. No other I/O interface can use a CHPID number that is defined for HiperSockets, even though HiperSockets does not occupy any physical I/O connection position.

Real LANs have a maximum frame size limit that is defined by their architecture. The maximum frame size for Ethernet is 1492 bytes. gigabit Ethernet has a jumbo frame option for a maximum frame size of 9 KB. The maximum frame size for a HiperSocket is assigned when the HiperSockets CHPID is defined. Frame sizes of 16 KB, 24 KB, 40 KB, and 64 KB can be selected. The default maximum frame size is 16 KB. The selection depends on the characteristics of the data that is transported over a HiperSockets and is also a tradeoff between performance and storage allocation.

The MTU size that is used by the IP network stack for the HiperSockets interface is also determined by the maximum frame size. Table 8-1 lists these values.

Table 8-1 Maximum frame size and MTU size

Maximum frame size	Maximum transmission unit size
16 KB	8 KB
24 KB	16 KB
40 KB	32 KB
64 KB	56 KB

The maximum frame size is defined in the hardware configuration (Input/Output Configuration Program (IOCP)) by using the CHPARM parameter of the CHPID statement.

z/OS allows the operation of multiple IP network stacks within a single image. The read control and write control I/O devices are required only once per image, and are controlled by VTAM. Each IP network stack within the same z/OS image requires one I/O device for data exchange.

Running one IP network stack per LPAR requires three I/O devices for z/OS (the same requirement as for z/VM and Linux on IBM Z). Each additional IP network stack in a z/OS LPAR requires only one more I/O device for data exchange. The I/O device addresses can be shared between z/OS systems that are running in different LPARs. Therefore, the number of I/O devices is not a limitation for z/OS.

An IP address is registered with its HiperSockets interface by the IP network stack when the IP network device is started. IP addresses are removed from an IP address lookup table when a HiperSockets device is stopped. Under operating system control, IP addresses can be reassigned to other HiperSockets interfaces on the same HiperSockets LAN. This feature allows flexible backup of IP network stacks.

Reassignment is only possible within the same HiperSockets LAN. A HiperSockets is *one network* or *subnetwork*. Reassignment is only possible for the same operating system type. For example, an IP address that is originally assigned to a Linux IP network stack can be reassigned only to another Linux IP network stack.

A z/OS dynamic VIPA can be reassigned only to another z/OS IP network stack, and a z/VM IP network VIPA can be reassigned only to another z/VM IP network stack. The LIC forces the reassignment. It is up to the operating system's IP network stack to control this change.

Enabling HiperSockets requires the CHPID to be defined as `type=IQD` by using HCD and IOCP. This CHPID is treated like any other CHPID and is counted as one of the available channels within the IBM zSystems platform.

HiperSockets definition: The IBM z16 IOCP definitions for HiperSockets devices require the keyword VCHID. VCHID specifies the virtual channel identification number that is associated with the channel path (type IQD). The valid range is 7C0 - 7FF.

The HiperSockets LIC on IBM zSystems supports the following features:

- Up to 32 independent HiperSockets.
- For z/OS, z/VM, Linux, and z/VSE the maximum number of IP network stacks or HiperSockets communication queues that can concurrently connect on a single IBM zSystems platform is 4096.
- Maximum total of 12288 I/O devices (valid subchannels) across all HiperSockets.
- Maximum total of 12288 IP addresses across all HiperSockets. These IP addresses include the HiperSockets interface and virtual IP addresses (VIPA) and dynamic VIPA that are defined for the IP network stack.

Sharing of HiperSockets is possible with the extension to the multiple image facility (MIF). HiperSockets channels can be configured to multiple channel subsystems (CSSs). They are transparently shared by any or all of the configured LPARs without regard for the CSS to which the partition is configured.

Figure 8-6 shows spanned HiperSockets that are defined on an IBM zSystems platform. For more information about spanning, see 2.1.7, “Channel spanning” on page 25.

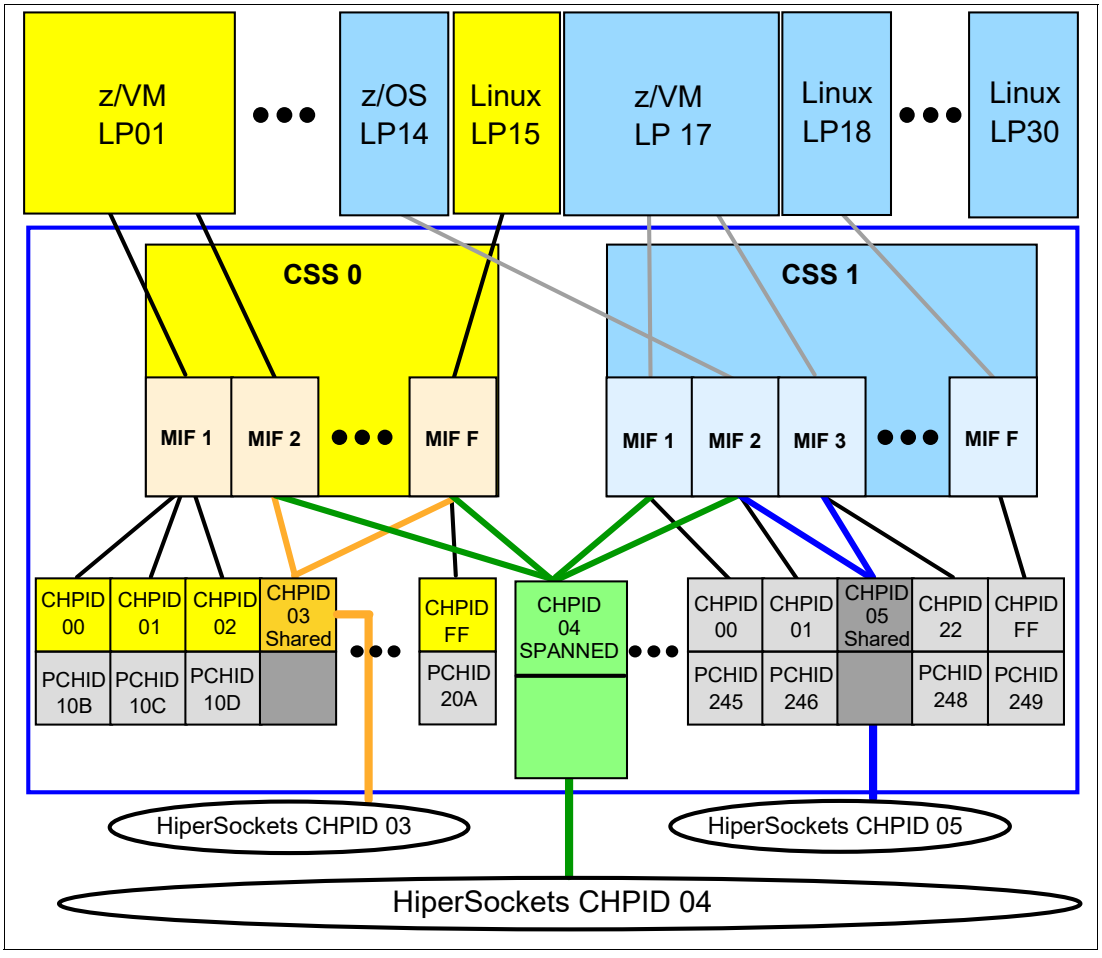


Figure 8-6 Spanned and non-spanned HiperSockets defined

8.3 Summary

HiperSockets is part of IBM z/Architecture technology and includes QDIO and advanced adapter interrupt handling. The data transfer is handled much like a cross-address space memory move, by using the memory bus. Therefore, HiperSockets does not contend with other I/O activity in the system.

HiperSockets can be defined to separate traffic between specific virtual servers and LPARs on one IBM zSystems server. Virtual private networks (VPNs) or network virtual local area networks across HiperSockets are supported to further isolate traffic as required. With integrated HiperSockets networking, there are no server-to-server traffic flows outside the IBM zSystems platform. The only way to probe these VLANs is by using the NTA function, and strict controls are required for that procedure.

The IBM zSystems platform supports up to 32 HiperSockets. Spanned channel support allows sharing of HiperSockets across multiple CSSs and LPARs.

8.4 References

For more information about the HiperSockets function and configuration, see *IBM HiperSockets Implementation Guide*, SG24-6816.

For more information about the HiperSockets virtual bridge support for z/VM, *see z/VM Connectivity,* SC24-6267.

9

Coupling links and common time

This chapter describes the connectivity options that support IBM Parallel Sysplex clustering technology and common time on IBM zSystems platform.

This chapter includes the following topics:

- 9.1, “IBM zSystems Parallel Sysplex” on page 142
- 9.2, “Connectivity options” on page 144
- 9.3, “Time functions” on page 149
- 9.4, “References” on page 154

9.1 IBM zSystems Parallel Sysplex

Parallel Sysplex brings the power of parallel processing to business-critical applications. A Parallel Sysplex cluster consists of up to 32 IBM z/OS images, which are connected to one or more coupling facilities (CFs) by using high-speed specialized links, called *coupling links*, for communication and time-keeping. The coupling facilities at the heart of the cluster enable high-speed record-level read/write data sharing among the images in a cluster.

Coupling links support communication between z/OS and coupling facilities. The coupling facility provides critical locking and serialization, data consistency, messaging, and queuing capabilities that allow the systems in the sysplex to coordinate and share data.

A configured cluster has no single point of failure and can provide users with near continuous application availability over planned and unplanned outages.

Figure 9-1 shows a possible Parallel Sysplex configuration with the Server Time Protocol (STP) feature. STP provides time synchronization for multiple servers and coupling facilities. The use of Network Time Protocol (NTP) servers as an External Time Source (ETS) is supported by STP.

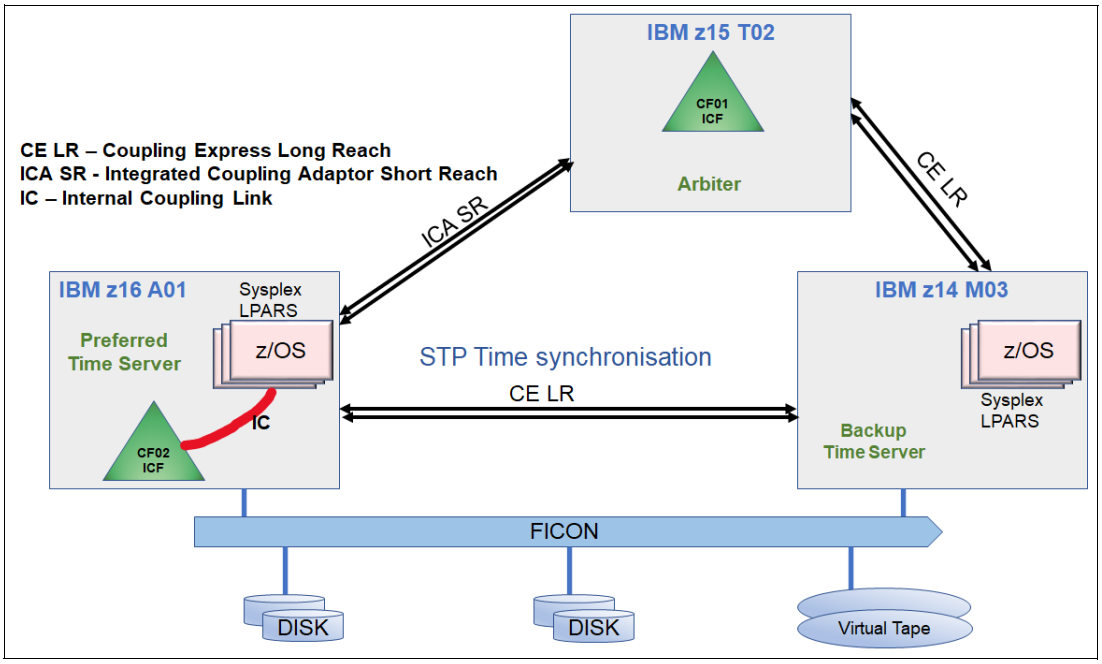


Figure 9-1 Sample Parallel Sysplex that uses stand-alone CFs and STP

IBM z16 supports improved time accuracy by connecting Precision Time Protocol (PTP) (IEEE 1588) and NTP time networks directly to the CPC drawers. Pulse Per Second (PPS) support is available for the highest accuracy to the external time reference for NTP time.

Note: STP is a mandatory hardware requirement for a Parallel Sysplex environment that consists of more than one IBM zSystems machine.

STP is a facility that is implemented in the Licensed Internal Code (LIC) that presents a single view of time to IBM Processor Resource/Systems Manager (IBM PR/SM).

The CPCs are configured with coupling links, such as the Integrated Coupling Adapter (ICA SR), Coupling Express Long Reach (CE LR), or Internal Coupling (IC) link. IC cannot be used for timing links (coupling links only).

Important: InfiniBand coupling or timing links should be planned carefully on supported systems (IBM z14 M0x) if an IBM z16 or IBM z15 are part of the IBM Parallel Sysplex or Coordinated Timing Network (CTN) configuration.

A CF runs the Coupling Facility Control Code (CFCC) that is loaded into main storage at the time the CF image is activated.

You can configure the coupling facilities both in the processor where z/OS images run and in a dedicated coupling-facility processor. The former is called an internal coupling facility, and the latter is called a stand-alone coupling facility.

If you use an internal coupling facility and the host CPC fails, both the z/OS images and the coupling facility fail simultaneously. In this situation, some structures might not be rebuilt in another coupling facility until the z/OS systems are recovered, which results in an application outage. Having a stand-alone coupling facility allows remaining systems to continue operating if a z/OS host fails. If the stand-alone CF machine fails, the z/OS systems can quickly rebuild the structures on the backup CF minimizing disruption to applications.

The alternative to using a stand-alone CF is to duplex CFs across two machines. There is an overhead to using this configuration, which is the tradeoff against having an extra machine footprint for the stand-alone CF.

Coupling link redundancy

There must be at least two coupling links between any two CPCs. This configuration provides redundancy to prevent the loss of a link causing sysplex communication failure between the CPCs.

9.1.1 Coupling links and STP

STP is a message-based protocol in which STP timekeeping information is passed over externally defined coupling links between CPCs.

Timing-only coupling links

For a CPC that is not part of a Parallel Sysplex, but required to be time-synchronized, coupling links must be configured for the CPC to be part of a CTN. Hardware Configuration Definition (HCD) supports the definition of *timing-only* links when there is no CF at either end of the coupling link. The timing-only links can be of type CL5 for CE LR, CS5 for ICA SR coupling links, or CIB for InfiniBand[1] coupling links (IBM z14 only). The control unit type is STP, and no devices are defined to it.

Note: CF messages cannot be transferred over timing-only links.

Coupling link redundancy for STP

There must be at least two coupling links between any two CPCs that are intended to exchange STP messages. This configuration provides redundancy to prevent the loss of a link causing STP communication failure between the CPCs.

[1] InfiniBand coupling and timing links are not allowed when an IBM z16 is a member in a Parallel Sysplex or Coordinated Timing Network configuration.

For more information, see *IBM zSystems Server Time Protocol Guide*, SG24-8480.

9.1.2 Multi-site Parallel Sysplex considerations

If a Parallel Sysplex is configured across two or more sites, you should plan to extend the Coupling Express LR beyond the distance that is supported without repeaters. IBM supports Wavelength Division Multiplexing (WDM) products that are qualified by IBM for use in multisite sysplex solutions, such as IBM Geographically Dispersed Parallel Sysplex (GDPS).

Note: Coupling Express LR (available on IBM z15 and older systems) is link-compatible with Coupling Express2 LR (available on IBM z16).

If any messages are to be transmitted across multiple sites through WDM products, ensure that only qualified WDM products and supported configurations are used.

For a list of WDM vendor products that are qualified for coupling links (transporting CF or STP messages) in a multi-site sysplex, see IBM Resource Link, which shows a list of qualified WDM products (requires a Resource Link ID).

9.2 Connectivity options

From a hardware perspective, when configuring a cluster, connectivity is a primary consideration, as are other hardware components, such as Server Time Protocol, and coupling links. In an IBM Parallel Sysplex, the objective is any-to-any connectivity and nondisruptive planned or unplanned outages. All channels, directors, and WDM equipment must be included in the high availability design.

For availability purposes, use at least two coupling links to connect each server to each CF in a Parallel Sysplex. Performance, availability, and distance requirements of a Parallel Sysplex are the key factors determining the appropriate connectivity options for a particular configuration.

9.2.1 Coupling link options

IBM z16, IBM z15, IBM z14 M0x, and IBM z14 ZR1 support the following coupling link types:

- Integrated Coupling Adapter Short Reach (ICA SR) links connect directly to the CPC drawer and are intended for short distances between CPCs of up to 150 meters.
- Coupling Express Long Reach (CE LR) adapters are located in the PCIe+ drawer and support unrepeated distances of up to 10 kms or up to 100 kms over qualified WDM services.
- Internal Coupling (IC) links are for internal links within a CPC.

Important: Parallel Sysplex supports connectivity between systems that differ by up to two generations (N-2). For example, an IBM z16 can participate in an IBM Parallel Sysplex cluster with other IBM z16, IBM z15, and IBM z14 systems.

Figure 9-2 shows the supported coupling link connections for the IBM z16, IBM z15, and IBM z14.

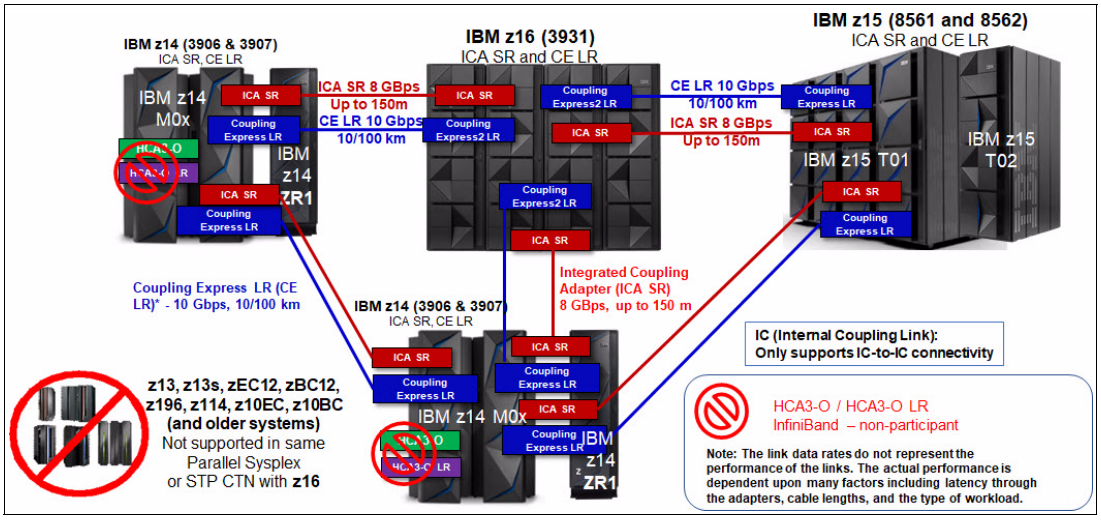


Figure 9-2 Parallel Sysplex connectivity options (IBM z16, IBM z15, and IBM z14)

Table 9-1 lists the coupling link support for each IBM zSystems platform. Restrictions on the maximum numbers can apply, depending on the configuration. Always check with your IBM support team for more information.

Table 9-1 Supported coupling link options

Type[a]	Description	Feature code	Link rate	Maximum unrepeated distance	Maximum number of physical links that are supported					
					IBM z16 A01	IBM z16 A02/ AGZ	IBM z15 T01	IBM z15 T02	IBM z14 M0x	IBM z14 ZR1
CE2 LR	Coupling Express2 LR	0434	10 Gbps	10 kms (6.2 miles)	64	46	-	-	-	-
CE LR	Coupling Express LR	0433	10 Gbps	10 kms (6.2 miles)	-	-	64	64	64	32
ICA SR1.1	Integrated Coupling Adapter	0176	8 GBps	150 meters (492 feet)	96*	48	96*	48*	N/A	N/A
ICA SR	Integrated Coupling Adapter	0172	8 GBps	150 meters (492 feet)	96*	48	96*	48*	80*	16
IC	Integrated Coupling Adapter	N/A	Internal speeds	N/A	64	64	64	64	32	32

Type[a]	Description	Feature code	Link rate	Maximum unrepeated distance	Maximum number of physical links that are supported					
					IBM z16 A01	IBM z16 A02/ AGZ	IBM z15 T01	IBM z15 T02	IBM z14 M0x	IBM z14 ZR1
HCA3-O LR[b]	InfiniBand Long Reach (1 x InfiniBand)	0170		10 kms (6.2 miles)	N/A	N/A	N/A	N/A	64*	N/A
HCA3-O[b]	InfiniBand (12 x InfiniBand)	0171		150 meters (492 feet)	N/A	N/A	N/A	N/A	32*	N/A

a. The maximum supported links depend on the IBM zSystems model or capacity feature code. These numbers are marked with an asterisk (*).
b. InfiniBand coupling or timing links are *not* supported for connecting to IBM z16, IBM z15 or IBM z14 ZR1.

Notes:

- The number of PCIe fanouts and ICA SR ports that are available on a server depends on the number of CPC drawers, and in the case of IBM z14 M/T 3907, the number of PU SCMs. Feature Code (FC) 0636 has one PU SCM and can place up to two PCIe I/O features in the CPC drawer. FC 0637 has two PU SCMs and can place up to four PCIe I/O features. FC 0638 and FC 0639 have four PU SCMs and can place up to eight PCIe I/O features.
- On an IBM z16 A01, for example, there are 12 PCIe+ Gen3 fanouts per CPC drawer, which means that there are a maximum of 24 ICA SR/ICA SR1.1 ports per drawer. So, a 4-drawer IBM z16 A01 server can support up to 96 ICA SR/ICA SR1.1 ports.

9.2.2 Internal Coupling link

IC links are Licensed Internal Code-defined links to connect a CF to a z/OS logical partition (LPAR) in the same CPC. These links are available on all IBM zSystems platforms. The IC link is an IBM zSystems coupling connectivity option that enables high-speed, efficient communication between a CF partition and one or more z/OS LPARs 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, effectively reducing 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 IBM z16 and IBM z15 servers, and IBM z14 supports up to 32 IC links per server.

9.2.3 Integrated Coupling Adapter Short Range

ICA SR (FC 0172) and ICA SR1.1 (FC 0176) are two-port, short-distance coupling features that allow the supported IBM zSystems servers to connect to each other. ICA SR and ICA SR1.1 use coupling channel (CHPID) type CS5. The ICA SR and ICA SR1.1 use PCIe Gen3 technology, with x16 lanes that are bifurcated into x8 lanes for coupling.

The ICA SR and SR1.1 are designed to drive distances up to 150 m and supports a link data rate of 8 GBps. It is designed to support up to four CHPIDs per port and seven subchannels (devices) per CHPID.

For more information, see *Planning for Fiber Optic Links*, GA23-1409. This publication is available at the Library section of Resource Link.

9.2.4 Coupling Express Long Reach

IBM z16 supports the new Coupling Express2 LR (FC 0434) technology update feature. Older Coupling Express LR (FC 0433) cards cannot be carried forward to IBM z16. Coupling Express2 LR and Coupling Express LR are two-port cards that occupy one slot in a PCIe+ I/O drawer or PCIe I/O drawer.[2] With these cards, the supported IBM zSystems servers can connect to each other over extended distances. Coupling Express features use coupling channel type CL5.

Note: Coupling Express2 Long Reach (FC 0434) is link-compatible and can connect to Coupling Express Long Reach (FC 0433).

Coupling Express LR uses 10GbE RoCE technology, can support distances up to 10 km unrepeated, and support a link data rate of 10 Gbps. For distance requirements greater than 10 km, clients must use wavelength-division multiplexing (WDM). The WDM vendor must be qualified by IBM zSystems.

Coupling Express features support up to four CHPIDs per port and 32 buffers (that is, 32 subchannels) per CHPID. The Coupling Express feature is in the PCIe+ I/O drawer on IBM z16, IBM z15, and IBM z14 ZR1, and in a PCIe I/O drawer in an IBM z14 M0x.

For more information, see *Planning for Fiber Optic Links*, GA23-1409. This publication is available at the Library section of Resource Link.

[2] PCIe+ I/O drawer (FC 4023 on IBM z16, FC 4021 on IBM z15, and FC 4001 with IBM z14 ZR1) is built in a 19” format with the PCIe cards oriented horizontally. All three features (FC 4023, FC 4021, and FC 4001) can hold up to 16 I/O PCIe features. The PCIe I/O drawer on an IBM z14 M0x cannot be carried forward to newer models.

9.2.5 InfiniBand coupling links (IBM z14 M0x only)

Notes:

- The IBM z14 M0x (M/T 3906) is the last IBM zSystems server to support InfiniBand coupling connectivity.
- InfiniBand coupling links are *not* supported on IBM z14 ZR1 (M/T 3907).

InfiniBand coupling links are high-speed links on IBM z14 M0x. The InfiniBand coupling links originate from two types of fanouts:

- HCA3-O (FC 0171) for 12x InfiniBand links
- HCA3-O LR (FC 0170) for 1x InfiniBand links

Each fanout that is used for coupling links has an assigned adapter ID number (AID) that must be used for definitions in IOCDS to have a relation between the physical fanout location and the CHPID number.

12x InfiniBand coupling links

The HCA3-O fanout support InfiniBand coupling links that operate at 6 GBps (12x InfiniBand). InfiniBand coupling links use a fiber optic cable that is connected to a HCA3-O fanout. The maximum distance for an InfiniBand link is 150 meters. The fiber cables are industry standard OM3 50/125 micrometer-multimode optical cables with Multifiber Push-On (MPO) connectors. 12x InfiniBand supports seven or 32 subchannels[3] per CHPID.

1x InfiniBand coupling links

The HCA3-O LR fanout supports 1x InfiniBand coupling links that operate at up to 5.0 Gbps. 1x InfiniBand coupling links use a fiber optic cable that is connected to a HCA3-O LR fanout. The maximum unrepeated distance for a 1x InfiniBand link is 10 km. When using repeaters, the maximum distance is up to 100 km. The fiber cables that are used for 1x InfiniBand links are standard 9 µm single mode fiber optic cables with an LC duplex connector. 1x InfiniBand supports seven subchannels per CHPID.

Fanout adapter ID

Unlike channels that are installed in a PCIe I/O drawer, which are identified by a PCHID number that is related to their physical location, InfiniBand coupling link fanouts and ports are identified by an AID. The adapter ID value depends on its physical location. The AID must be used to assign a CHPID to the fanout in the IOCDS definition. The CHPID assignment is done by associating the CHPID to an AID and port. For more information, see 2.1.5, “Adapter ID” on page 22.

9.2.6 Dynamic I/O reconfiguration for stand-alone CF CPCs

With z14 Driver Level 36 (z14 GA2) and newer, support for dynamic activation of a new or changed IODF on a stand-alone CF is supported without requiring a POR/IML of the stand-alone CF CPC and without requiring the presence of any z/OS or IBM z/VM image running an HCD instance on the same CPC.

The new hardware activation service is a function that is implemented in firmware and provides the HCD instance that is deployed on the stand-alone CF CPC. This firmware function is driven by an updated HCD and HCM that is running in a z/OS LPAR on a remote

[3] Depending on the version of the HCD, the default setting for the subchannels is set to 7 or 32.

IBM z16, IBM z15, or IBM z14 GA2 system and managed by an HMC V2.14.1 or later that is connected to the stand-alone CF CPC.

The dynamic I/O for a stand-alone CF CPC is shown in Figure 9-3.

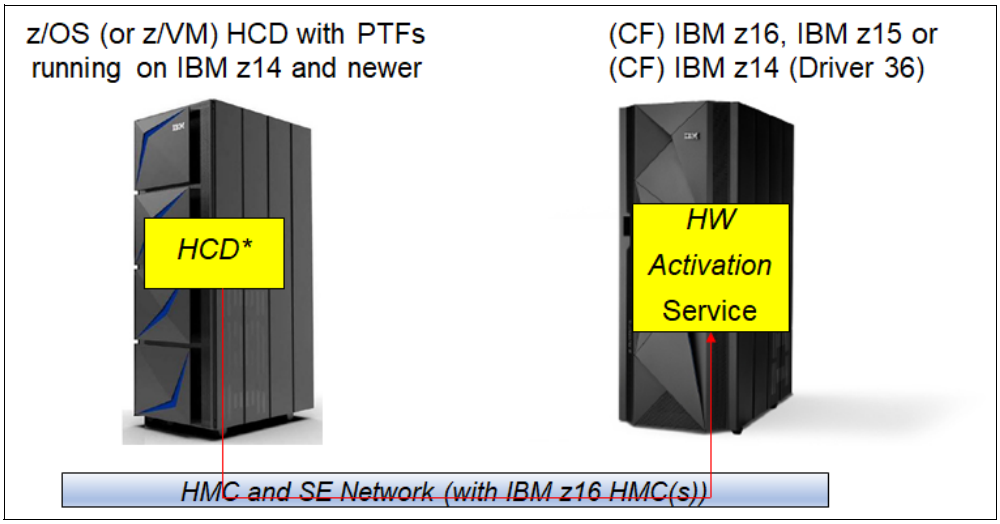


Figure 9-3 Dynamic I/O for a stand-alone CF CPC

9.3 Time functions

Business-critical applications require the accurate time and date to be set because they often are business and regulatory requirements. For example, the Financial Industry Regulatory Authority in the US and the EU markets in Financial Instruments Directive ii specify timing accuracy regulations for financial transactions. Time must be consistent across all the systems that process these transactions.

In the IBM z/Architecture, Server Time Protocol (STP) facilitates the synchronization of CPC time-of-day (TOD) clocks to ensure consistent time across multiple CPCs and operating systems. STP provides a way to synchronize TOD clocks in different CPCs with a centralized time reference. This setup, in turn, can be set accurately based on an international time standard (ETS). The architecture defines a time-signal protocol and a distribution network, which allows accurate setting, maintenance, and consistency of TOD clocks.

9.3.1 Server Time Protocol

Important: The “Sysplex Timer” menus on the Support Element were discontinued for IBM z15 and later systems. The Server Time Protocol (STP) can be configured and managed from the HMC task “Manage System Time”.

STP is a facility that is implemented in the Licensed Internal Code and presents a single view of time to Processor Resource/Systems Manager (PR/SM) across multiple CPCs. Any IBM zSystems system can be enabled for STP by installing the STP feature (FC 1021). Each CPC that is planned to be configured in a CTN must be STP enabled.

It is possible to manually set the hardware clock when a stand-alone machine is powered on, but for most customers it is not accurate enough. The systems in the CTN must obtain the time from an *ETS*.

The IBM zSystems platform has the following ETS support:

- NTP
- NTP with PPS
- PTP
- PTP with PPS

These protocols are described further in this section.

Tip: If you configured an STP CTN with three or more servers, see *Important Considerations for STP Server Role Assignments* in IBM Docs.

IBM z16 Oscillator hardware changes

IBM z16 configurations have 1 - 4 CPC drawers. Each CPC drawer has two combined Base Management Card (BMC) / OSC (oscillator) cards, each with one PPS port and one ETS port (RJ45 Ethernet for both PTP and NTP). This setup is a change from the IBM z15, which implemented a card that combined the FSP and OSC, and the NTP and PTP networks, which are connected through the Support Element.

For PPS signal redundancy, two PPS ports must be connected:

- For ETS redundancy for a single CPC drawer system, both PPS ports must be connected and configured.
- For redundancy for a system with two or more CPC drawers, PPS ports in the first and second CPC drawer must be used, and only two PPS ports (one in CPC 0 and one in CPC 1).

  For such configurations, PPS ports that are connected must be explicitly assigned in the STP menus.

  STP tracks the PPS signal to maintain time accuracy for the CTN. STP maintains an accuracy of 10 microseconds to the PPS input signal. Several variables, such as the cable that is used to connect the PPS signal and the accuracy of the NTP server to its time source (GPS or radio signals), determine the ultimate accuracy of STP to Coordinated Universal Time.

PTP requires network infrastructure support:

- For IBM z16, the PTP Ethernet cable must plug directly into the BMC and OSC customer port of CPC Drawer 0 and CPC Drawer 1. For single CPC drawer systems, PTP Ethernet cables must be connected to ETS1 and ETS2 ports on the CPC drawer.
- PPS is optional for PTP, but might still be required for NTP to meet financial regulations.

The IBM z16 implements a card that combines the BMC and OSC (see Figure 9-4).

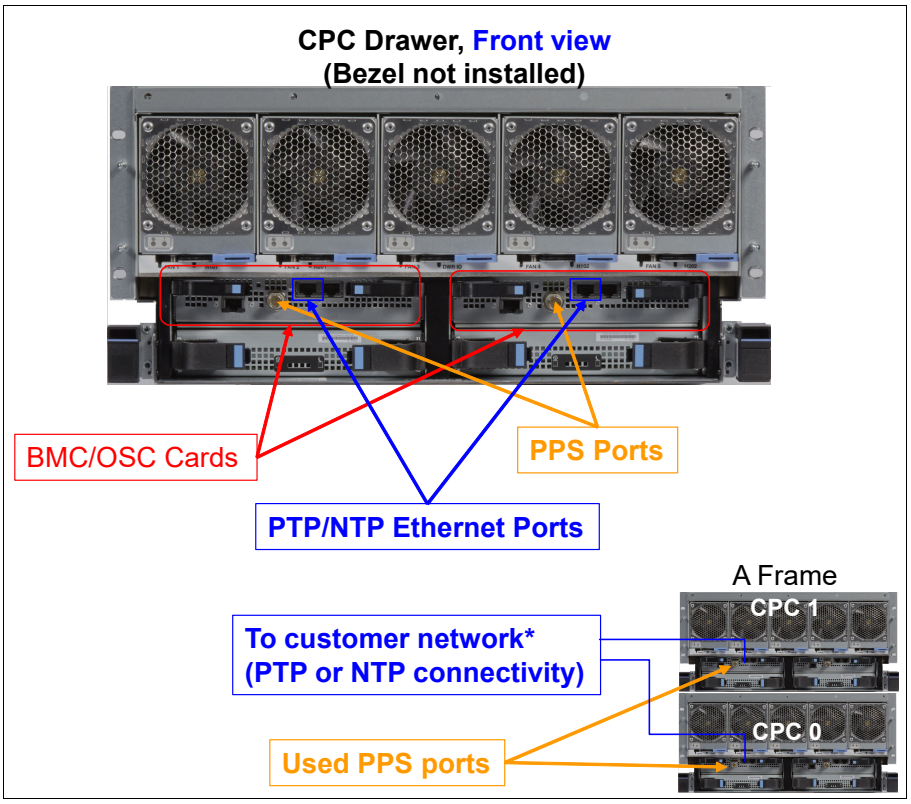


Figure 9-4 z16 PTP, NTP, and PPS ports

Network Time Protocol client support

Using NTP servers as an ETS usually fulfills the requirement for a time source or common time reference across heterogeneous platforms and for providing a higher time accuracy.

On the IBM z16, NTP client support is provided by a firmware function that is implemented in the CPC. The code interfaces with the NTP servers. This interaction allows an NTP server to become the single time source for the z16 CPC.

Precision Time Protocol

IBM z16 and IBM z15 servers support IEEE 1588 PTP as an ETS for an IBM zSystems CTN. The IBM z16 implementation is for PTP connectivity directly to the CPC (not to the Support Element as in previous generations of IBM zSystems servers). The implementation does not change the CTN implementation for time coordination other than providing the potential to use a PTP-based ETS for more accurate time reference to the external time.

Pulse Per Second

Each of the two oscillator cards in the IBM zSystems servers has a PPS port to provide connectivity to an NTP or PTP server with PPS support:

- On IBM z16, PPS is connected as shown in Figure 9-4 on page 151.
- On IBM z15, PPS is also connected to the CPC drawer, but to an FSP/OSC card instead.
- On IBM z14, PPS is connected to the oscillator cards in the back of frame A.
- On IBM z14 ZR1, the NTP server with PPS is connected to the oscillator cards that are in the CPC drawer (rear side of the CPC drawer).

External Time Source support

The IBM zSystems platform can be configured to obtain system time from an NTP or a PTP source (server). The server address is configured by using the Configure External Time function from the Manage System Time task on the HMC.

An ETS must be configured for the Current Time Server (CTS). Configuring the Preferred ETS and Secondary ETS servers for the PTS and BTS reduces the risk of an STP-only timing network losing its time source.

An ETS does not need to be configured for systems that are not the PTS or BTS. If an ETS is configured, its configuration is saved if the role of the system changes to the PTS or BTS.

For more information about how to configure ETS, see *IBM zSystems Server Time Protocol Guide*, SG24-8480.

Using an External Time Source server with PPS support

STP can use an ETS server that has a PPS output signal as its ETS. This type of ETS device is available worldwide from several vendors that provide network timing solutions.

When PPS is used, a network (Ethernet LAN) connection to the ETS server is also required.

For more information about ETS support, see the Parallel Sysplex advantages web page.

STP recovery enhancements

Here are the recovery enhancements to STP:

- N-mode power-imminent disruption signal for STP recovery.

  Important: N-mode power signal can be enabled if both PTS and BTS are on an the IBM z16.

  For more information, see *IBM zSystems Server Time Protocol Guide*, SG24-8480.

  N-mode power sensing allows automatic failover of the CTS. To enable this feature, there is a one-time setup step.

  In the HMC “Manage System Time” task, change the automatic switchover function from **CPC1** to **CPC2** for STP.

  - After the function is enabled, the power subsystem (both Bulk Power Assembly and intelligent Power Distribution Unit (iPDU)) detects any power source loss (at the power cord or power side level).
  - If there is a failure, a signal is generated from the CPC1 to CPC2 like for IBM z15 with Integrated Battery Facility. The generation of this signal can take up to 30 seconds depending on conditions.
  - If within 30 seconds CPC2 does not receive a signal that power is back to fully redundant on CPC1, CPC2 takes over as CTS.
  - After normal power is restored to CPC1, CPC1 can automatically return to the CTS role.

- Updated GUI for managing the STP environment.
- Enhanced Console Assisted Recovery.

  Enhanced Console Assisted Recovery (ECAR) speeds up the process of Backup Time Server (BTS) takeover:

  - When the Primary Time Server (PTS/CTS) encounters a checkstop condition, the CPC informs its Support Element (SE) and Hardware Management Console (HMC).
  - The PTS SE recognizes the checkstop-pending condition, and the PTS SE STP code is called.
  - The PTS SE sends an ECAR request by using the HMC to the BTS SE.
  - The BTS SE communicates with the BTS to start the takeover.
- Starting with IBM z14, an extra STP stratum level is supported (Stratum 4). This additional stratum level was implemented to alleviate the additional complexity and expense of system reconfiguration by using system upgrades. It should be used only as a temporary state during reconfiguration. Environments should not run with systems at Stratum level 4 for extended periods because of the lower quality of the time synchronization.
- "Going away" signal.

  The ICA SR, CE LR, and HCA3-O host channel adapters send a reliable, unambiguous "going away" signal to indicate that the server is about to enter a *Failed* state (check stopped). If the Preferred Time Server (PTS) is the CTS in an STP-only CTN and the going away signal is received by the Backup Time Server (BTS) from the CTS, the BTS can safely take over as the CTS without relying on the previous recovery methods of Offline Signal (OLS) in a two-server CTN or the Arbiter in a CTN with three or more servers.

  The available STP recovery design is still available for the cases when a "going away" signal is not received or for failures other than a server failure.

9.3.2 Dynamic Split and Merge for Coordinated Timing Network

Beginning with HMC 2.14.1, it is possible to dynamically split a CTN or merge two CTNs by using the STP GUI that is available on the HMC. STP understands the images that are present on each CPC in the source STP network and the sysplex affiliation of each of those z/OS images. STP can use this information for integrated checking, which operates with z/OS toleration for CTN change.

Attention: If any of the two CTNs being Joined/merged is restricted, the join operation will fail and can lead to outage. Therefore, for a successful join/merge operation, none of the two CTNs shall be restricted.

A restricted/bounded CTN does not allow any CEC to be added nor remove any CEC from the CTN.

A restricted CTN cannot be split.

CTN split

When splitting a CTN, STP checks to ensure that it does not result in a sysplex that "splits" across the two resultant CTNs. New STP roles are automatically assigned in the split-off CTN during the process. STP connectivity and max-stratum checking are performed in both split-off CTNs. Checking results are previewed and confirmed before proceeding.

CTN Merge

When merging two CTNs, STP roles are automatically maintained in the merged-into CTN while STP connectivity and max-stratum checking is performed in the merged CTN. Preview and confirm the results before proceeding.

9.3.3 Operating system support

Software requirements vary with the design and use of common time. All current IBM z/OS versions support Server Time Protocol (STP).

For more information about STP operating system support, see the STP tab on the Parallel Sysplex web page.

9.4 References

For more information about understanding, planning, and implementing a Parallel Sysplex cluster, see the Parallel Sysplex web page.

For more information about Parallel Sysplex, see the following publications:

- *Planning for Fiber Optic Links,* GA23-1409
- *IBM zSystems Server Time Protocol Guide*, SG24-8480
- *Getting the Most Out of a Parallel Sysplex*, SG24-2073

10

Extended distance solutions

This chapter describes architectural requirements and implementation for IBM zSystems platform connectivity over extended distances.

This chapter includes the following topics:

- 10.1, “Unrepeated distances” on page 156
- 10.2, “Fibre Channel connection” on page 158
- 10.3, “Coupling links” on page 160
- 10.4, “Wavelength-division multiplexing” on page 161
- 10.5, “References” on page 163

Note: Not all features discussed in this chapter are not supported on IBM zSystems platforms. For more information regarding a specific feature, refer to the appropriate chapter.

10.1 Unrepeated distances

This section lists the maximum unrepeated distance and link budget for each type of IBM Z fiber optic link. Longer distances are possible by using repeaters, switches, channel extenders, and wavelength division multiplexing (WDM).

In Table 10-1, a *link* is a physical connection over a transmission medium (fiber) that is used between an optical transmitter and an optical receiver. The maximum allowable link loss, or *link budget*, is the maximum amount of link attenuation (loss of light), expressed in decibels (dB), that can occur without causing a possible failure condition (bit errors). When you use multimode fiber, as the link data rate increases, the unrepeated distance and link budget decrease.

The link budget is derived from combining the channel insertion loss budget with the deallocated link margin budget. The link budget numbers are rounded to the nearest 10th of a dB.

Table 10-1 Fiber optic connections: Unrepeated distances

Feature type	Fiber type	Link data rate	Fiber bandwidth (MHz-km)	Maximum distance[a]	Link budget (dB)
FICON Express LX	SM 9 µm	8 Gbps	N/A	10 km	6.4
		16 Gbps	N/A	10 km	6.4
		32 Gbps	N/A	5 km	6.34
FICON Express SX	MM 62.5 µm	8 Gbps	200	21 m	1.58
	MM 50 µm	8 Gbps	500	50 m	1.68
			2000	150 m	2.04
			4700	190 m	2.19
	MM 50 µm	16 Gbps	500	35 m	1.63
			2000	100 m	1.86
			4700	125 m	1.95
	MM 50 µm	32 Gbps	500	35 m	1.57
			2000	100 m	1.75
			4700	125 m	1.86
Gigabit Ethernet LX	SM 9 µm	1 Gbps	N/A	5 km	4.6
	MM 50 µm[b]	1 Gbps	500	550 m	2.4
Gigabit Ethernet SX	MM 62.5 µm	1 Gbps	200	275 m	2.6
	MM 50 µm		500	550 m	3.6
	MM 50 µm		2000	550 m	3.6
10-gigabit Ethernet LR	SM 9 µm	10 Gbps	N/A	10 km	6

Feature type	Fiber type	Link data rate	Fiber bandwidth (MHz-km)	Maximum distance[a]	Link budget (dB)
10-gigabit Ethernet SR	MM 62.5 µm	10 Gbps	200	33 m	1.6
	MM 50 µm	10 Gbps	500	82 m	1.8
			2000	300 m	2.6
			4700	400 m	2.9
25-gigabit Ethernet SR	MM 50 µm	25 Gbps	2000	70 m	1.8
			4700	100 m	1.9
25-gigabit Ethernet LR	SM 9 µm	25 Gbps	N/A	10 km	5.5

a. Some features have extended distance with an RPQ.
b. Requires fiber optic mode-conditioning patch (MCP) cables.

The following notes apply to Table 10-1 on page 156:

- Single-Byte Command Code Sets Connection (SBCON) is the American National Standards Institute (ANSI) standard for the command set that is used by FICON over a Fibre Channel physical interface. It is also known as FC-SB.
- All industry-standard links (FICON, gigabit Ethernet) follow published industry standards. The minimum fiber bandwidth requirement to achieve the distances that are listed is applicable for multimode (MM) fiber only. There is no minimum bandwidth requirement for single mode (SM) fiber.
- The bit rates that are given might not correspond to the effective channel data rate in a particular application because of protocol overhead and other factors.
- LC duplex and SC duplex connectors are keyed per the ANSI Fibre Channel Standard specifications.
- *Mode-conditioning patch* (MCP) cable is required to operate certain links over multimode fiber.
- The FICON Express16S features, which are available on IBM z15, IBM z14, and IBM z14 ZR1 (carry forward only), allow an auto-negotiated link speed of either 4 Gbps, 8 Gbps, or 16 Gbps.
- The FICON Express16S+ features, which are available on IBM z16 A01, IBM z16 A02 and IBM z16 AGZ (carry forward only), IBM z15 T01 (carry forward only), IBM z15 T02, and IBM z14, allow an auto-negotiated link speed of 4 Gbps, 8 Gbps, or 16 Gbps.
- The FICON Express16SA features, which are available on IBM z15 T01 and IBM z16 A01 (carry forward only), allow an auto-negotiated link speed of 8 Gbps or 16 Gbps.
- The FICON Express32S features, which are available on IBM z16, allow an auto-negotiated link speed of either 8 Gbps, 16 Gbps, or 32 Gbps.
- As light signals traverse a fiber optic cable, the signal loses some of its strength. Decibels (dB) is the metric that is used to measure light power loss. The significant factors that contribute to light power loss are the length of the fiber, the number of splices, and the number of connections. The amount of light power loss (dB) across a link is known as the link budget.

  All links are rated for a maximum link budget (the sum of the applicable light power loss factors must be less than the link budget) and a maximum distance (exceeding the

maximum distance causes undetectable data integrity exposures). Another factor that limits distance is jitter, but that is typically not a problem at these distances.

- Measure link budget and fiber bandwidth at the appropriate wavelength:
  - Long wavelength (1300 nm)
  - Short wavelength (850 nm)

  For planning purposes, the following worst case values can be used to estimate the link budget. See the references that are listed and contact the fiber vendor for specific values, which might be different for your configuration:
  - Link loss at 1300 nm = 0.50 db/km
  - Link loss per splice = 0.15 db/splice (not dependent on wavelength)
  - Link loss per connection = 0.50 db/connection (not dependent on wavelength)
- Deviations from these specifications (longer distance or link budget) might be possible. These deviations are evaluated on an individual basis by submitting a request for price quotation (RPQ) to IBM.

Note: For more information about extended distances, see 10.4, “Wavelength-division multiplexing” on page 161.

10.2 Fibre Channel connection

This section describes architectural requirements and implementation solutions for Fibre Channel connection (FICON) channel connectivity over unrepeated and repeated distances. The term FICON represents the architecture as defined by the International Committee of Information Technology Standards (INCITS) and published as ANSI standards. FICON also represents the names of the IBM zSystems platform feature types:

- FICON Express32S
- FICON Express16SA
- FICON Express16S+
- FICON Express16S
- FICON Express8S
- FICON Express8

All feature types support a long wavelength (LX) laser version and a short wavelength (SX) laser version. They support native FICON (Fibre Channel (FC) FCTC) and Fibre Channel Protocol (FCP) channel modes.

For more information, see 3.3, “Connectivity” on page 60.

10.2.1 FICON unrepeated distance

The unrepeated distance that is supported by IBM zSystems FICON features depends on these factors:

- The feature port transceiver type (LX or SX)
- The fiber type being used:
  - 9 µm single mode
  - 50 µm or 62.5 µm multimode

  Also, for multimode, the fiber bandwidth (MHz-km) of the fiber.

- The speed at which the feature port is operating
- Whether there are MCP cables in the fiber optic link

For more information, see Table 10-1 on page 156, and *Planning for Fiber Optic Links,* GA23-1409.

10.2.2 FICON repeated distance solutions

This section describes several extended distance connectivity solutions for FICON channel-attached I/O control units and devices.

The repeated distance for a FICON channel is IBM zSystems qualified to a maximum of 100 km. For all FICON features that use repeaters, the end-to-end distance between the FICON channel and the control unit is IBM zSystems qualified for up to 100 km (62 miles) only. RPQ 8P2981 is available for customers who require distances over 100 km.

FICON channel to control unit (CU) distance can be increased by placing channel repeaters such FICON Directors between the host channel port and the CU. Common practice is to place a FICON Director at each site between the IBM zSystems host channel port at one site and the CU at the other site (usually two Directors at each site for redundancy). The supported distance between the FICON Directors is vendor-specific.

The links between the two directors are called Inter-Switch Links (ISLs). FICON Multihop allows support for cascading up to four switches with three hops.

FICON Directors with ISLs add flexibility to the system configuration because they allow one to many and many to many links to be defined.

Another way of extending the distance between a FICON channel and CU is with Qualified Wavelength Division Multiplexer (QWDM) infrastructure between the two sites (see 10.4, “Wavelength-division multiplexing” on page 161).

FICON channel to CU end-to-end distance can be increased up to 100 km without data rate performance droop occurring if the FICON Director buffer credits are set. The number of buffer credits that are required depends on the link data rate and the maximum number of buffer credits that are supported by the FICON Director or control unit, and application and workload characteristics.

Although it is possible for FICON to maintain high bandwidth at distances greater than 100 km, these distances have not been qualified for use with IBM zSystems. They are achievable only if enough buffer credits exist to support the link speed. Support for distances over 100kms can be requested with RPQ 8P2981.

FICON extended distance example

This example describes a single hop between two FICON Directors by using ISLs that can extend the end-to-end link distance.

The maximum supported distance of the FICON channel path between two FICON Directors is FICON Director vendor specific. Each ISL requires one fiber trunk (two fibers) between the FICON Directors.

Figure 10-1 shows an example of such a configuration. Assuming that the distance between the two FICON Directors is 10 km (6.21 miles), the maximum supported distance is 30 km (18.64 miles) with FICON Express8S 10KM LX features. The example is also valid for the FICON Express16S, FICON Express16S+, and FICON Express16SA.

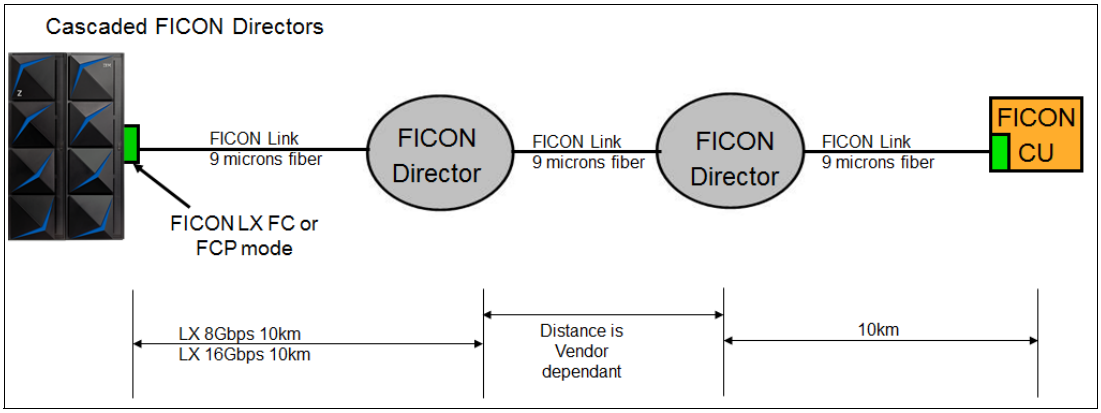


Figure 10-1 FICON LX path with cascaded FICON Directors

FICON channel path: Transceiver intermix

A FICON channel path through one or two FICON Directors consists of multiple optical fiber links. Each link in the channel path can be either LX or SX, allowing the channel path to be made up of a mixture of link types. This configuration is possible because the FICON Director converts optical to electrical and back to optical (known as an OEO conversion, for optical-electrical-optical) of the channel path as it passes through the director.

Note: The transceiver type (LX or SX) at each end of a particular link must match.

WDM technologies

Other extended distance connectivity technologies are available to extend FICON and other link types, for example, WDM. WDM technology also provides increased flexibility in that multiple links and protocol types can be transported over a single dark fiber trunk. For more information, see 10.4, “Wavelength-division multiplexing” on page 161.

10.3 Coupling links

This section describes architectural requirements and implementation solutions for coupling link connectivity over unrepeated and repeated distances.

Coupling link unrepeated distance

Table 10-2 on page 161 lists the maximum unrepeated distances and link data rates that are supported for coupling links on IBM zSystems platforms. For more information, see Table 10-2 on page 161 and its notes and *Planning for Fiber Optic Links*, GA23-1409.

Table 10-2 Coupling link unrepeated distance and link data rate support

Coupling link type	Maximum unrepeated distance	Link data rate
IC (internal)	N/A	Memory-to-memory (the highest bandwidth)
ICA SR	150 meters[a]	8 Gbps
Coupling Express LR	10 km	10 Gbps
12x InfiniBand[b]	150 meters	6 Gbps
1x InfiniBand[b]	10 km	5 Gbps

a. 150 meters distance is achieved by using OM4 fiber types only; with OM3 fiber, the distance is 100 meters maximum.
b. Cannot be configured if an IBM z16 is part of the Parallel Sysplex or Coordinated Time Network configuration.

10.4 Wavelength-division multiplexing

WDM is a technique that is used to transmit several independent bit streams over a single fiber optic link (see Figure 10-2). It is an approach to opening up the conventional optical fiber bandwidth by breaking it up into many channels, each at a different optical wavelength (a different color of light). Each wavelength can carry a signal at any bit rate less than an upper limit that is defined by the electronics, typically up to several gigabits per second.

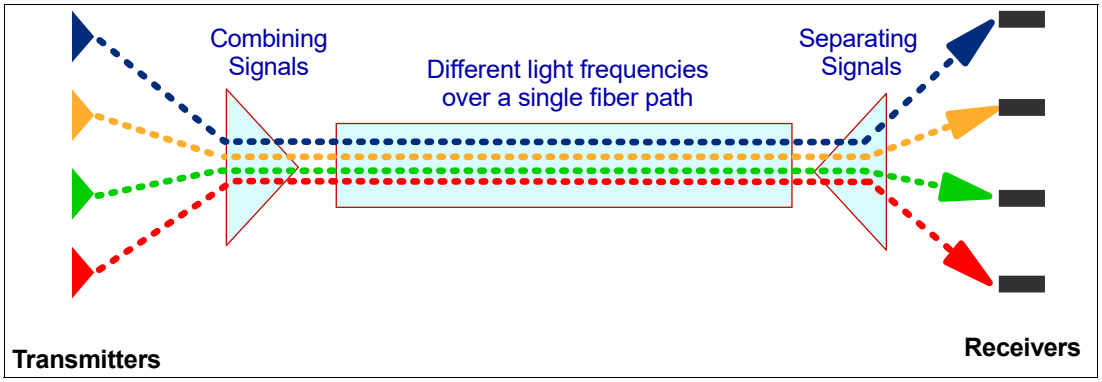


Figure 10-2 WDM transmission technique

The channels are protocol-independent, so a wide variety of protocols is supported, including FICON, FCP, coupling links, Server Time Protocol (STP), and gigabit Ethernet.

The actual signal bandwidth that the electronics can handle over one wavelength is a small fraction of the inter-channel spacing. So, the signals do not interfere with one another and can therefore be multiplexed into a single fiber by using a passive grating multiplexer.

There are several extended distance uses of WDM technology:

- FICON/FCP channel connections to remote control units and devices
- LAN and network protocol connections to remote sites
- IBM System Storage Metro Mirror (synchronous Peer-to-Peer Remote Copy (PPRC))
- IBM System Storage Global Mirror
- IBM System Storage z/OS Global Mirror (asynchronous Extended Remote Copy (XRC))

- Peer-to-Peer Virtual Tape Server (PtP VTS), a form of remote copying tape data
- Coupling Express LR connections between CPCs

10.4.1 GDPS qualification

GDPS is an enterprise-wide continuous availability and disaster recovery automation solution that can manage recovery from planned and unplanned outages across distributed servers and IBM zSystems platforms. GDPS can be configured in either a single site or in a multi-site configuration. It is designed to manage remote copy services between storage subsystems, automate Parallel Sysplex operational tasks, and perform failure recovery from a single point of control, improving application availability.

Historically, this solution was known as a *Geographically Dispersed Parallel Sysplex*. Today, GDPS continues to be applied as a general term for a suite of business continuity solutions, including the ones that do not require a dispersed or multi-site sysplex environment.

GDPS supports the following forms of remote copy in multi-site solutions:

- IBM Metro Mirror, synchronous Peer-to-Peer Remote Copy
- IBM Global Mirror, asynchronous Peer-to-Peer Remote Copy
- IBM z/OS Global Mirror, asynchronous XRC

The GDPS solution is also independent of disk vendors provided the vendor meets the specific levels of IBM Metro Mirror, IBM Global Mirror, and IBM z/OS Global Mirror architectures. For more information, see the GDPS web page.

IBM supports only WDM products that are IBM zSystems qualified for use in GDPS solutions. To obtain this qualification, WDM vendors obtain licensed IBM patents, intellectual property, and know-how that are related to the GDPS architecture. This access allows vendors to use proprietary IBM protocols and applications that are used in a GDPS environment, including coupling and STP links, Metro Mirror, Global Mirror, and z/OS Global Mirror.

Licensing of IBM patents also provides the WDM vendor with technical information about future IBM releases. Qualified vendors typically license this information for an extended period, which allows them to subscribe to the latest GDPS architecture changes and to be among the first to market with offerings that support these features.

Note: Check with your WDM vendor for current licensing status.

In addition, these vendor products are tested and qualified by IBM technicians with the same test environment and procedures that are used to test the protocols that provide connectivity for a GDPS configuration. This testing includes functions, recovery, and, in certain cases, performance measurements.

Having access to these test facilities allows IBM to configure a fully functional sysplex and simulate failure and recovery actions that cannot be tested as part of an operational customer environment.

IBM has facilities to test and qualify these products with both current and previous generation equipment within the IBM Vendor Solutions Connectivity Lab in Poughkeepsie, New York, in the United States. This qualification testing allows IBM specialists to reproduce any concerns that might arise when using this equipment in a client's application.

Components

The following GDPS components are used during the qualification process:

- IBM Parallel Sysplex
- IBM System Storage
- Optical Wavelength Division Multiplexer (WDM)
- IBM System Storage Metro Mirror (PPRC), a synchronous form of remote copy
- IBM System Storage Global Mirror
- IBM System Storage z/OS Global Mirror (XRC), an asynchronous form of remote copy

Protocols

The following GDPS connectivity protocols are tested during the qualification process:

- Fibre Channel connection (FICON)
- Fibre Channel protocol (FCP)
- Fibre Channel Inter-Switch Links (ISL)
- Server Time Protocol (STP)
- Coupling Express LR links
- 1x InfiniBand coupling links
- 10-gigabit Ethernet and RDMA over Converged Enhanced Ethernet (RoCE and RoCE Express2) using Shared Memory Communications - RDMA (SMC-R)

Often, these tested protocols are used in non-GDPS environments as well. The robust testing that is performed during the qualification process provides a high level of confidence when you use these IBM zSystems qualified optical WDM vendor products in non-GDPS environments.

10.4.2 IBM zSystems qualified WDM vendor products

The latest list of qualified WDM vendor products can be found through IBM Resource Link.

Select **Hardware products for servers** in the resource link, and then, select the page titled **IBM System z® Qualified Wavelength Division Multiplexer (WDM) products for GDPS solutions**.

Note: It is important to select the particular WDM vendor link in Resource Link and download the qualification letter to verify the details about the WDM product, model, firmware level, and the IBM zSystems server models for which it is qualified.

10.5 References

For more information about fiber optic link distance, see the following publications:

- *Coupling Facility Channel I/O Interface Physical Layer,* SA23-0395
- *Planning for Fiber Optic Links,* GA23-1409
- *Fiber Transport Services Direct Attach Planning,* GA22-7234

For more information about IBM zSystems connectivity, see this web page.

For more information about GDPS solutions, see the following resources:

- GDPS home page
- Parallel Sysplex home page
- *IBM GDPS Family: An Introduction to Concepts and Capabilities*, SG24-6374

For more information about IBM zSystems qualified WDM vendor products, use this IBM Redbooks publications search result.

A

Cryptographic solutions

This appendix briefly describes the optional Peripheral Component Interconnect Express (PCIe) cryptographic features of IBM zSystems platform.

This appendix includes the following topics:

- “Overview” on page 166
- “Crypto Express8S features (1 HSM and 2 HSM)” on page 167
- “Crypto Express7S (1 port or 2 port)” on page 168
- “Crypto Express6S” on page 169

Overview

Public Key Cryptography Standards (PKCS) #11[1] and the IBM Common Cryptographic Architecture (CCA) define various cryptographic functions, external interfaces, and a set of key cryptographic algorithms. These specifications provide a consistent, end to end cryptographic architecture across IBM z/OS, IBM AIX, and IBM i operating systems and other platforms, including Linux and Microsoft Windows.

The following cryptographic features and functions are part of the IBM zSystems environment:

- Central Processor Assist for Cryptographic Function (CPACF-CP Assist). CPACF-CP Assist offers a set of symmetric cryptographic functions for high encrypting and decrypting performance of clear key operations. This interface is for Secure Sockets Layer/Transport Layer Security (SSL/TLS), VPN, and data-storing applications that do not require US Federal Information Processing Standard (FIPS[2]) 140-2 Level 4 security. The CP Assist for Cryptographic Function is implemented as a coprocessor in the processor core of the IBM zSystems platform.

  The on-core design consists of one Compression Coprocessor (CMPCS), one CPACF, and one IBM Integrated Accelerator for IBM Z Sort. The CPACF is embedded in each processing unit (PU core) of the IBM zSystems PU chip.

  The coprocessor (COP) supports SMT. For increased throughput, the IBM z14 COP was further developed so that the coprocessor results are now stored directly in the L1 cache (on-core), which was carried over to later IBM zSystems processor generations.

  Figure A-1 illustrates the on-core COP. The highlighted (red rectangle) area shows the functional blocks that belong to the CPACF. Exploitation of this unit requires ordering the CPACF feature (Feature Code (FC) 3863).

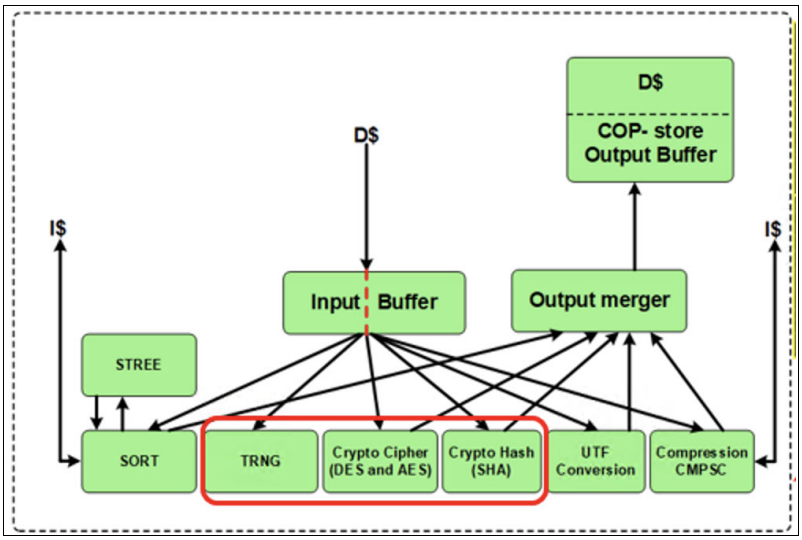


Figure A-1 Compression and cryptography accelerators on a core in the chip

- Crypto Express features are optional and available in different generations. All Crypto Express features can be configured during installation as either a secure coprocessor or as an accelerator. Changing the operational mode (coprocessor, EP11, or accelerator) can be changed, but the process is disruptive to the feature operation. This disruption can be avoided by using redundant features and planning and configuration.

[1] One of the industry-accepted Public Key Cryptography Standards (PKCS) provided by RSA Laboratories from RSA, the security division of Dell EMC Corporation.
[2] Federal Information Processing Standards (FIPS)140-2 Security Requirements for Cryptographic Modules.

The support of the different generations of Crypto Express features depends on the IBM zSystems generation.

Crypto Express features provide a secure hardware and programming environment for cryptographic processes. Each cryptographic coprocessor includes a general-purpose processor, non-volatile storage, and specialized cryptographic electronics.

The cryptographic coprocessor functions are supported by the Integrated Cryptographic Service Facility for z/OS and the IBM Common Cryptographic Architecture Support Program for Linux.

Crypto Express8S features (1 HSM and 2 HSM)

The Crypto Express8S features are supported on the IBM z16 only. They include the following characteristics:

- The Crypto Express8S (2 HSM) (FC 0908) has two IBM PCIe Cryptographic Coprocessors (PCIeCCs[3]). The feature occupies one I/O slot in the PCIe+ I/O drawer and has two PCHIDs.
- Crypto Express8S (1 HSM) (FC 0909) has one IBM PCIe Cryptographic Coprocessor (PCIeCC). It occupies one I/O slot in the PCIe+ I/O drawer and has one PCHID.
- Each Crypto Express8S PCHID can be configured in one of the following modes:
  - The Secure IBM CCA coprocessor (CEX8C) includes secure key functions. The Crypto Express8S supports user-defined extensions (UDX) through which you can define and load customized cryptographic functions.
  - The Secure IBM Enterprise PKCS #11 (EP11) coprocessor (CEX8P) implements an industry-standardized set of services that adhere to the PKCS #11 specification V2.20.

    This cryptographic coprocessor mode introduced the PKCS #11 secure key function.

    A Trusted Key Entry (TKE) workstation is required to support the administration of the Crypto Express8S 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.

  These modes can be configured by using the Support Element. The PCIe adapter must be configured offline to change the mode.

  Note: When the Crypto Express8S PCIe adapter is configured as a secure IBM CCA coprocessor, it still provides accelerator functions. However, you can achieve up to three times better performance for those functions if the Crypto Express8S PCIe adapter is configured as an accelerator.

- IBM z16 A01 supports up to 30 Crypto Express8S (2 HSM), or up to 16 Crypto Express8S (1 HSM) features are supported.

  IBM z16 A01 supports up to 60 HSMs in any combination (Crypto Express8S (2 HSM), Crypto Express8S (1 HSM) and carry forward features, Crypto Express7S, and Crypto Express6S) are supported.

[3] The IBM PCIeCC is a Hardware Security Module that acts as a PCIe card.

- IBM z16 A02 and IBM z16 AGZ support up to 20 Crypto Express8S (2 HSM), or up to 16 Crypto Express8S (1 HSM) features are supported.

  IBM z16 A02 and IBM z16 AGZ support up to 40 HSMs in any combination (Crypto Express8S (2 HSM), Crypto Express8S (1 HSM)) and carry forward features, Crypto Express7S, and Crypto Express6S) are supported.
- The Crypto Express8S HSM supports up to 85 domains on IBM z16 A01 and 40 domains on IBM z16 A02 and IBM z16 AGZ. This enhancement is based on the new Adjunct Processor Extended Addressing (APXA), which enables the z/Architecture to support up to 256 domains in an Adjunct Processor (AP).

Crypto Express7S (1 port or 2 port)

The Crypto Express7S features are supported on IBM z15 and IBM z16 (carry forward only) systems. They include the following characteristics:

- The Crypto Express7S (2 port), FC 0898, has two IBM PCIe Cryptographic Coprocessors (PCIeCC[4]). It occupies one I/O slot in the PCIe+ I/O drawer and has two PCHIDs.
- Crypto Express7S (1 port), FC 0899, has one IBM PCIe Cryptographic Coprocessor (PCIeCC). It occupies one I/O slot in the PCIe+ I/O drawer and has one PCHID.
- Each Crypto Express7S PCHID can be configured in one of the following modes:
  - Secure IBM CCA coprocessor (CEX7C) for FIPS 140-2 Level 4 certification. This mode includes secure key functions. The Crypto Express7S supports user-defined extensions (UDX), which you can use to define and load customized cryptographic functions.
  - Secure IBM Enterprise PKCS #11 (EP11) coprocessor (CEX7P) implements an industry-standardized set of services that adhere to the PKCS #11 specification v2.20.

    This cryptographic coprocessor mode introduced the PKCS #11 secure key function.

    A Trusted Key Entry (TKE) workstation is required to support the administration of the Crypto Express7S when it is configured in EP11 mode.
  - Accelerator (CEX7A) for acceleration of public key and private key cryptographic operations that are used with SSL/TLS processing.

  These modes can be configured by using the Support Element. The PCIe adapter must be configured offline to change the mode.

  Note: When the Crypto Express7S PCIe adapter is configured as a secure IBM CCA coprocessor, it still provides accelerator functions. However, you can achieve up to three times better performance for those functions if the Crypto Express7S PCIe adapter is configured as an accelerator.

- IBM z15 T01 and IBM z16 A01 support up to 30 Crypto Express7S (2 port) or up to 16 Crypto Express7S (1 port) features for up to 60 HSMs in supported combinations.

  Note: Crypto Express7S and Crypto Express6S features can be carried forward to IBM z16. Crypto Express5S is *not* supported on IBM z16.

[4] The IBM PCIeCC is a Hardware Security Module that acts as a PCIe card.

- IBM z16 A02, IBM z16 AGZ, and IBM z15 T02 support up to 20 Crypto Express7S (2-port), or up to 16 Crypto Express7S (1-port) features are supported. Up to 40 HSMs in any combination (Crypto Express7S (2 port), Crypto Express7S (1 port) and carry forward feature, Crypto Express6S is supported.
- The Crypto Express7S adapters support up to 85 domains on IBM z16 A01 and z15 T01, and up to 40 domains on IBM z16 A02, IBM z16 AGZ, and z15 T02 for logical partitions (LPARs). This enhancement is based on the new APXA, which enables the z/Architecture to support up to 256 domains in an AP.

Crypto Express6S

The Crypto Express6S feature is supported on IBM z16 and IBM z15 (carry forward only), and IBM z14. It has the following characteristics:

- It occupies one I/O slot in the PCIe I/O drawer and has one PCIe adapter (PCIeCC or HSM), with one PCHID assigned to it according to its physical location.
- Each Crypto Express6S PCIe adapter can be configured in one of the following modes:
  - Secure IBM CCA coprocessor (CEX6C) for FIPS 140-2 Level 4 certification. This mode includes secure key functions. The Crypto Express6s supports user-defined extensions (UDX), which you can use to define and load customized cryptographic functions.
  - Secure IBM Enterprise PKCS #11 (EP11) coprocessor (CEX6P) implements an industry-standardized set of services that adhere to the PKCS #11 specification V2.20.

    This cryptographic coprocessor mode introduced the PKCS #11 secure key function.

    A Trusted Key Entry (TKE) workstation is required to support the administration of the Crypto Express4S when it is configured in EP11 mode.
  - Accelerator (CEX6A) for acceleration of public key and private key cryptographic operations that are used with SSL/TLS processing.

  These modes can be configured by using the Support Element. The PCIe adapter must be configured offline to change the mode.

  Note: When the Crypto Express6S PCIe adapter is configured as a secure IBM CCA coprocessor, it still provides accelerator functions. However, you can achieve up to three times better performance for those functions if the Crypto Express6S PCIe adapter is configured as an accelerator.

- Up to 16 Crypto Express6S features are supported (16 PCIe adapters per supported IBM zSystems platform).
- Up to 85 domains on IBM z16 A01, IBM z15 T01, and IBM z14 M0x, and up to 40 domains on IBM z16 A02, IBM z16 AGZ, IBM z15 T02 and IBM z14 ZR1 for LPARs or IBM z/VM guests, are supported. This enhancement is based on the new APXA, which enables the z/Architecture to support up to 256 domains in an AP.

References

For more information, see the following publications:

- *IBM z16 (3931) Technical Guide*, SG24-8951
- *IBM z16 Configuration Setup*, SG24-8960
- *IBM z15 (8561) Technical Guide,* SG24-8851
- *IBM z15 Configuration Setup*, SG24-8860
- *IBM z14 Technical Guide,* SG24-8451
- *IBM z14 Configuration Setup,* SG24-8460

B

Channel conversion options

This appendix describes the possibilities of conversion from FICON channel connectivity to ESCON or connectivity to Bus and Tag (B/T) devices (parallel channel). An ESCON channel feature is not supported on any marketable IBM mainframes.

This appendix includes the following topic:

► “Conversion solutions” on page 172

Conversion solutions

In this section, the available solutions for channel conversion are described.

FICON to ESCON conversion

Prizm Protocol Convert from Optica Technologies Inc. provides a FICON-to-ESCON conversion function that is qualified for use with IBM zSystems. To view qualification letters, see the IBM zSystems I/O Connectivity web page.

Select the **Products** tab; then, **FICON / FCP Connectivity**. Scroll down to the “other supported devices” area on the web page.

The Optica Prizm converter can be implemented in point-to-point, switched, and cascaded FICON configurations, as shown in Figure B-1.

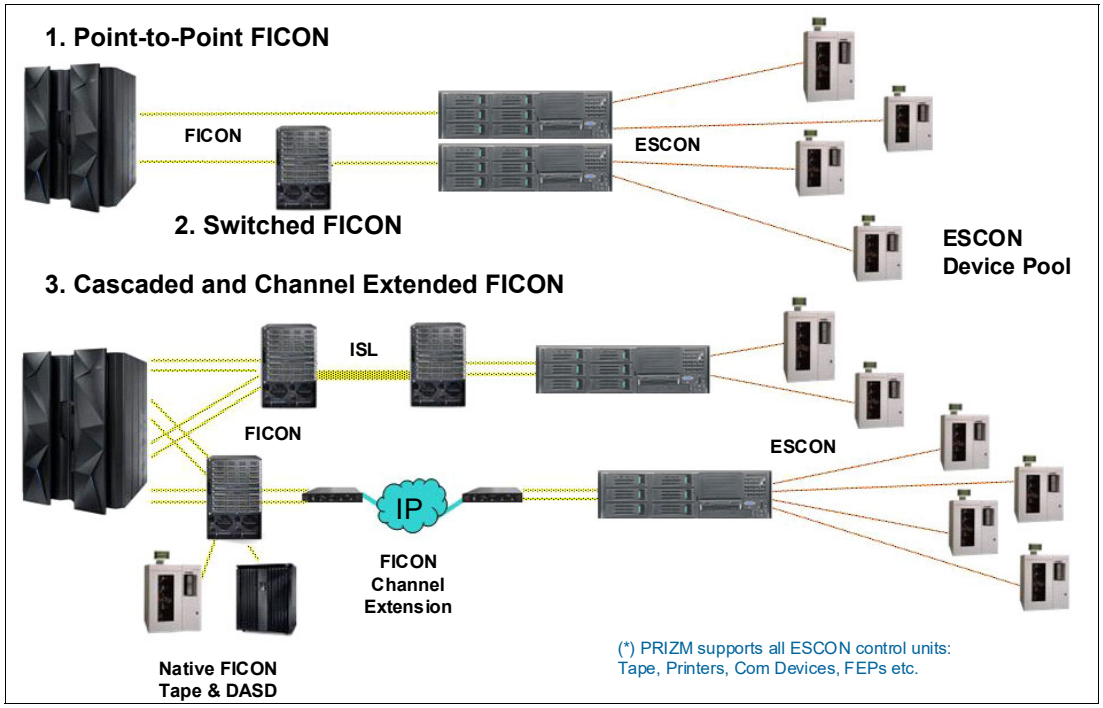


Figure B-1 Optica Prizm possible configurations

FICON to Bus and Tag conversion

For IBM zSystems platform that still require connectivity to B/T devices, a combination of two Optica converters can be used: The Optica Prizm to convert FICON to ESCON and the Optica ESBT to convert ESCON to B/T that is operating for devices in block mode only, as shown in Figure B-2.

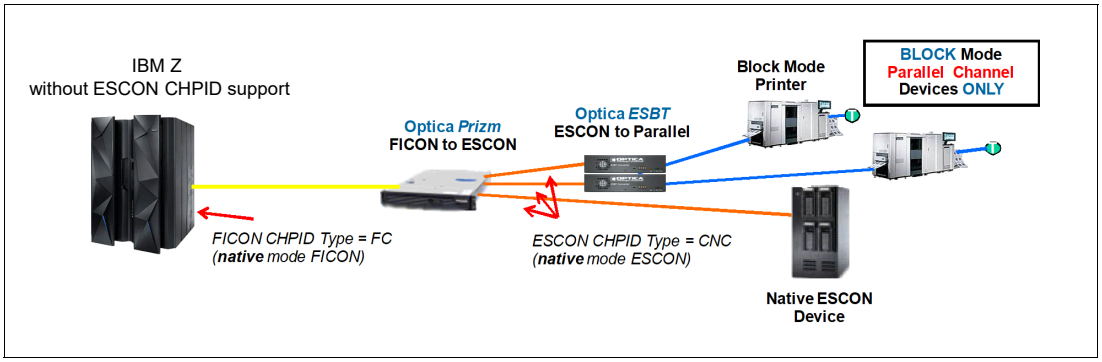


Figure B-2 IBM zSystems connections to ESCON and B/T devices

For more information about Prizm Protocol Convert, see the Optica website.

Note: IBM cannot confirm the accuracy of compatibility, performance, or any other claims by vendors for products that have not been qualified for use with IBM zSystems platforms. Address questions regarding these capabilities and device support to the suppliers of those products.

IBM Facilities Cabling Services ESCON-to-FICON migration services can help you take advantage of high-speed FICON to support an upgraded IBM zSystems environment. Also, it is possible to keep using ESCON-attached devices to reduce migration costs.

C

Channel feature attributes

This appendix lists the cable types and attributes of channel features that are supported on the IBM zSystems platform. Not all channel features can be ordered for all server families. Certain features are available only when carried forward on a server upgrade.

This appendix includes the following topic:

- “Cable types and attributes” on page 176

Cable types and attributes

Table C-1 lists the cable types and attributes of the channel types that are supported on the IBM zSystems platform and the maximum number of channels or ports that are supported per system.

The connector type for most fiber optic cable types is LC duplex, with the following exceptions:

zHyperLink Express MTP connector
12x InfiniBand MPO connector
ICA SR MTP connector
1000BASE-T Ethernet RJ-45 connector (UTP copper Ethernet cable)

For more information about fiber optic cables and connector types, see Appendix D, “Fiber optic cables” on page 183.

The footnotes in Table C-1 reflect special considerations for certain features. Check the referenced chapters for in-depth information.

The special conditions for each feature are as follows:

- NB stands for new build, that is, the features can be added for a new build system.
- Carry forward (CF) means that the feature can be transferred to the new platform only by using the miscellaneous equipment specification (MES) process. The feature cannot be ordered for a new build system.
- WDFM means that the feature is supported on the platform. However, because the feature at this platform is withdrawn from marketing, no ordering is possible.

The entry in this column belongs always to the supported platform information on the left.

For more information about support of extended distances, see Chapter 10, “Extended distance solutions” on page 155.

Table C-1 IBM zSystems channel feature support

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Platform	New build (NB)[c], carry forward (CF), or WDFM
Storage Connectivity		Chapter 3, “Fibre Channel connectivity” on page 33				
zHyperLink Express1.1	0451	8 GBps	MM 50 µm OM3 OM4	100 m (2000) 150 m (4700)	IBM z16 and IBM z15	NB or CF
zHyperLink Express	0431	8 GBps	MM 50 µm OM3 OM4	100 m (2000) 150 m (4700)	IBM z16, IBM z15, and IBM z14	CF
FICON Express32S LX	0461	32 GBps	SM 9 µm	5[b] km	IBM z16	NB
		8, or 16 GBps		10 km		

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Platform	New build (NB)[c], carry forward (CF), or WDFM
FICON Express32S SX	0462	32 GBps	MM 50 µm OM2 OM3 OM4	20 m (500) 70 m (2000) 100 m (4700)	IBM z16	NB
		16 GBps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	15 m (200) 35 m (500) 100 m (2000) 125 m (4700)		
		8 GBps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)		
FICON Express16SA LX	0437	8, or 16 Gbps	SM 9 µm	10 km	IBM z15 T01 IBM z16 A01	NB CF
FICON Express16SA SX	0438	16 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	15 m (200) 35 m (500) 100 m (2000) 125 m (4700)	IBM z15 T01 IBM z16 A01	NB CF
		8 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)		
FICON Express16S+ LX	0427	4, 8, or 16 Gbps	SM 9 µm	10 km	IBM z14, IBM z15 T01 IBM z15 T02 IBM z16	NB CF NB or CF CF
FICON Express16S+ SX	0428	16 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	15 m (200) 35 m (500) 100 m (2000) 125 m (4700)	IBM z14, IBM z15 T01 IBM z15 T02 IBM z16	NB CF NB or CF CF
		8 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)		
		4 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)		
FICON Express16S LX	0418	4, 8, or 16 Gbps	SM 9 µm	10 km	IBM z15 and IBM z14	CF or WDFM

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Platform	New build (NB)[c], carry forward (CF), or WDFM
FICON Express16S SX	0419	16 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	15 m (200) 35 m (500) 100 m (2000) 125 m (4700)	IBM z15 and IBM z14	CF or WDFM
		8 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)		
		4 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)		
FICON Express8S LX	0409	2, 4, or 8 Gbps	SM 9 µm	10 km	IBM z15 and IBM z14	CF or WDFM
FICON Express8S SX	0410	8 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)	IBM z15, IBM z14, and IBM z14 ZR1	CF or WDFM
		4 Gbps	MM 62.5 µm MM 50 µm OM2 OM3 OM4	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)		
		2 Gbps	MM 62.5 µm MM 50 µm OM2 OM3	150 m (200) 300 m (500) 500 m (2000)		
OSA-Express		Chapter 5, “IBM Open Systems Adapter Express” on page 73				
OSA-Express7S 1.2 25 GbE LR	0459	25 Gbps	SM 9 µm	10 km	IBM z16	NB
OSA-Express7S 1.2 25 GbE SR	0460	25 Gbps	MM 50 µm OM3 OM4	70 m (2000) 100 m (4700)	IBM z16	NB
OSA-Express7S 1.2 10 GbE LR	0456	10 Gbps	SM 9 µm	10 km	IBM z16	NB
OSA-Express7S 1.2 10 GbE SR	0457	10 Gbps	MM 62.5 µm	33 m	IBM z16	NB
			MM 50 µm OM2 OM3	82 m (500) 300 m (2000)		
OSA-Express7S 1.2 1 GbE LX	0454	1 Gbps	SM 9 µm	10 km	IBM z16	NB
OSA-Express7S 1.2 1 GbE SX	0455	1 Gbps	MM 62.5 µm	275 m (200)	IBM z16	NB
			MM 50 µm OM3	550 m (500)		
OSA-Express7S 1.2 1000BASE-T Ethernet	0458	1000 Mbps	UTP Cat5 or Cat6	100 m	IBM z16	NB

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Platform	New build (NB)[c], carry forward (CF), or WDFM
OSA-Express7S 25 GbE SR1.1	0449	25 Gbps	MM 50 µm	70 m (2000) 100 m (4700)	IBM z15 T01 IBM z16 A01	NB CF
OSA-Express7S 25 GbE SR	0429	25 Gbps	MM 50 µm	70 m (2000) 100 m (4700)	IBM z14, IBM z15 T01 IBM z15 T02 IBM z16	NB CF NB or CF CF
OSA-Express7S 10 GbE LR	0444	10 Gbps	SM 9 µm	10 km	IBM z15 T01 IBM z16 A01	NB CF
OSA-Express7S 10 GbE SR	0445	10 Gbps	MM 62.5 µm	33 m	IBM z15 T01 IBM z16 A01	NB CF
			MM 50 µm OM2 OM3	82 m (500) 300 m (2000)		
OSA-Express7S GbE LX	0442	1 Gbps	SM 9 µm	5 km	IBM z15 T01 IBM z16 A01	NB CF
OSA-Express7S GbE SX	0443	1 Gbps	MM 62.5 µm	275 m (200)	IBM z15 T01 IBM z16 A01	NB CF
			MM 50 µm OM3	550 m (500)		
OSA-Express7S 1000BASE-T Ethernet	0446	1000 Mbps	UTP Cat5 or Cat6	100 m	IBM z15 T01 IBM z16 A01	NB CF
OSA-Express6S 10 GbE LR	0424	10 Gbps	SM 9 µm	10 km	IBM z15 T01 IBM z15 T02 IBM z14 IBM z16	CF NB or CF NB CF
OSA-Express6S 10 GbE SR	0425	10 Gbps	MM 62.5 µm	33 m	IBM z15 T01 IBM z15 T02 IBM z14 IBM z16	CF NB or CF NB CF
			MM 50 µm OM2 OM3	82 m (500) 300 m (2000)		
OSA-Express6S GbE LX	0422	1 Gbps	SM 9 µm	5 km	IBM z15 T01 IBM z15 T02 IBM z14 IBM z16	CF NB or CF NB CF
			MM 50 µm	550 m (500)		
OSA-Express6S GbE SX	0423	1 Gbps	MM 62.5 µm	275 m (200)	IBM z15 T01 IBM z15 T02 IBM z14 IBM z16	CF NB or CF NB CF
			MM 50 µm OM3	550 m (500)		
OSA-Express6S 1000BASE-T Ethernet	0426	100/1000 Mbps	UTP Cat5 or 6	100 m	IBM z15 T01 IBM z15 T02 IBM z14 IBM z16	CF NB or CF NB CF
RoCE and SMC-Direct		Chapter 7, “Shared Memory Communications” on page 113				
25 GbE RoCE Express3 SR	0440	25 Gbps	MM 50 µm OM3 OM4	70 m (2000) 100 m (4700)	IBM z16	NB
25 GbE RoCE Express3 LR	0441	25 Gbps	SM 9µm	10 km	IBM z16	NB

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Platform	New build (NB)[c], carry forward (CF), or WDFM
10 GbE RoCE Express3 SR	0442	10 Gbps	MM 62.5 µm	33 m	IBM z16	NB
			MM 50 µm OM2 OM3	82 m (500) 300 m (2000)		
10 GbE RoCE Express3 LR	0443	10 Gbps	SM 9µm	10 km	IBM z16	NB
25GbE RoCE Express2.1	0450	25 Gbps	MM 50 µm	70 m (2000) 100 m (4700)	IBM z15 IBM z16	NB CF
25GbE RoCE Express2	0430	25 Gbps	MM 50 µm	70 m (2000) 100 m (4700)	IBM z15 IBM z16 IBM z14	CF CF NB
10GbE RoCE Express2.1	0432	10 Gbps	MM 62.5 µm	33 m (200)	IBM z15 IBM z16	NB CF
			MM 50 µm	82 m (500) 300 m (2000)		
10GbE RoCE Express2	0412	10 Gbps	MM 62.5 µm	33 m	IBM z15 IBM z16 IBM z14	CF CF NB
			MM 50 µm OM2 OM3	82 m (500) 300 m (2000)		
SMC-Direct	N/A		N/A	N/A	IBM z16, IBM z15, and IBM z14	N/A
HiperSockets		Chapter 8, “HiperSockets” on page 127				
HiperSockets	N/A		N/A	N/A	IBM z16, IBM z15, and IBM z14	N/A
Coupling links		Chapter 9, “Coupling links and common time” on page 141				
IC	N/A		N/A	N/A	IBM z16, IBM z15, and IBM z14	N/A
CE2 LR	0434	10 Gbps	SM 9 µm	10 km	IBM z16	NB
CE LR	0433	10 Gbps	SM 9 µm	10 km	IBM z15 and IBM z14	NB[c] and CF
ICA SR1.1	0176	8 GBps	MM 50 µm OM3 OM4	100 m (2000) 150 m (4700)	IBM z16 and IBM z15	NB and CF
ICA SR	0172	8 GBps	MM 50 µm OM3 OM4	100 m (2000) 150 m (4700)	IBM z16 IBM z15 IBM z14	CF NB and CF NB and CF
HCA3-O LR (1x InfiniBand)	0170	5 Gbps or 2.5 Gbps	SM 9 µm	10 km	IBM z14 M0x[d]	NB[c], CF, or WdFM
HCA3-O (12x InfiniBand)	0171	6 GBps	MM 50 µm OM3	150 m (2000)	IBM z14 M0x[d]	NB[c], CF, or WdFM

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Platform	New build (NB)[c], carry forward (CF), or WDFM
Crypto		Appendix A, “Cryptographic solutions” on page 165				
Crypto Express8S (2 HSM)	0908	N/A	N/A	N/A	IBM z16	NB
Crypto Express8S (1 HSM)	0909	N/A	N/A	N/A	IBM z16	NB
Crypto Express7S (2 port)	0898	N/A	N/A	N/A	IBM z16 IBM z15	CF NB
Crypto Express7S (1 port)	0899	N/A	N/A	N/A	IBM z16 IBM z15	CF NB
Crypto Express6S	0893	N/A	N/A	N/A	IBM z16 IBM z15, IBM z14	CF CF NB
zEDC Express						
IBM zEDC Express[e]	0420	N/A	N/A	N/A	IBM z14	WDFM

a. Minimum fiber bandwidths in MHz/km for multimode fiber optic links are included in parentheses where applicable.
b. 5 km for a point-to-point link running at 32 GBps (direct connection to a switch or another device).
c. NB is a new build. Features can be added for a new build system or as carry forward during an MES or upgrade.
d. InfiniBand coupling and timing links cannot be used in a sysplex or CTN where IBM z16 is a member.
e. The IBM zEDC Express PCIe feature was replaced in IBM z15 and newer IBM zSystems generations by the on-chip
IBM Integrated Accelerator for zEnterprise Data Compression (zEDC).

D

Fiber optic cables

This appendix describes the physical attributes of fiber optic technologies that are supported on IBM zSystems platform.

This appendix includes the following topics:

- “Description” on page 184
- “Connector types for fiber cables” on page 185
- “Mode-conditioning patch cables” on page 185
- “zHyperLink Express and ICA SR cables” on page 187
- “Conversion kits” on page 188
- “References” on page 189

Description

Fiber optic cables use light for data transmission, rather than electrical current on copper cables. Fiber optic cables have many advantages:

- Many times lighter and have substantially less bulk
- No pins
- A smaller and more reliable connector
- Reduced loss and distortion
- Free from signal skew and the effects of electro-magnetic interference

Figure D-1 shows the following types of optical fiber that are used in a data center environment with IBM zSystems platform:

- Multimode
- Single mode

The difference between these modes the way that light travels along the fiber. Multimode features multiple light paths; single mode features only one light path.

Each fiber type consists of three parts:

- The core can be 50 or 62.5 µm in diameter for multimode or 9 µm in diameter for single mode.
- The cladding that surrounds the core is 125 µm in diameter.
- The outer coating is 250 µm in diameter.

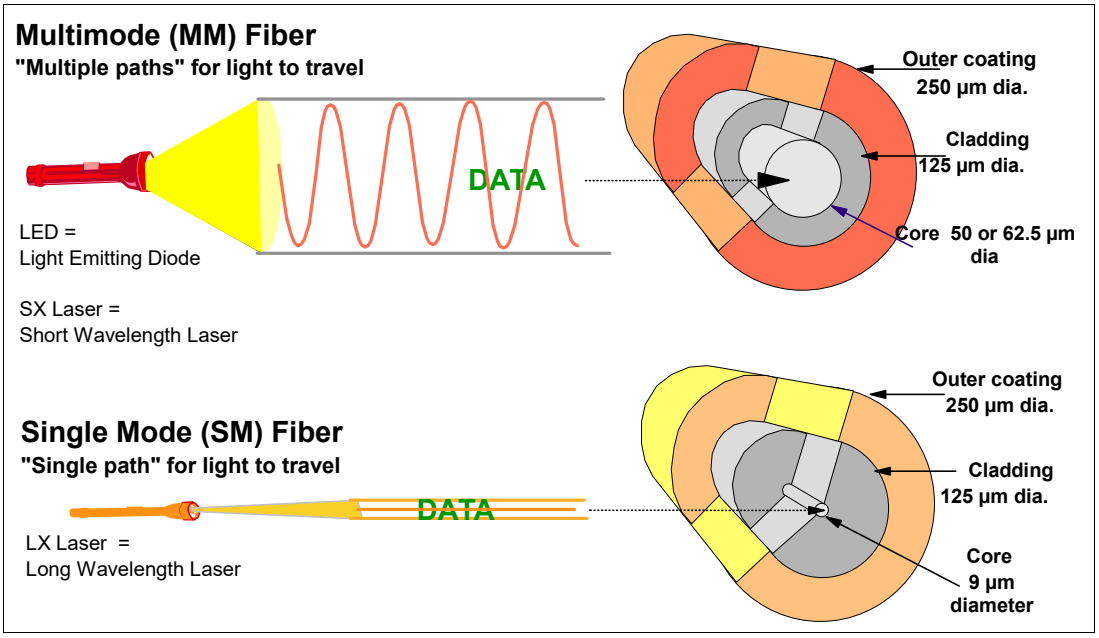


Figure D-1 Fiber optic cable types

Note: To keep the data flowing, thorough cleaning of fiber optic connectors is critical. Make sure that you have the necessary fiber optic-cleaning procedures in place.

Connector types for fiber cables

For all optical links, the connector type is LC duplex, except the ESCON connector, which has an MT-RJ type connector, and 12x InfiniBand, which has an MPO connector.

Figure D-2 shows the most common fiber cable connectors that are used in data center environments.

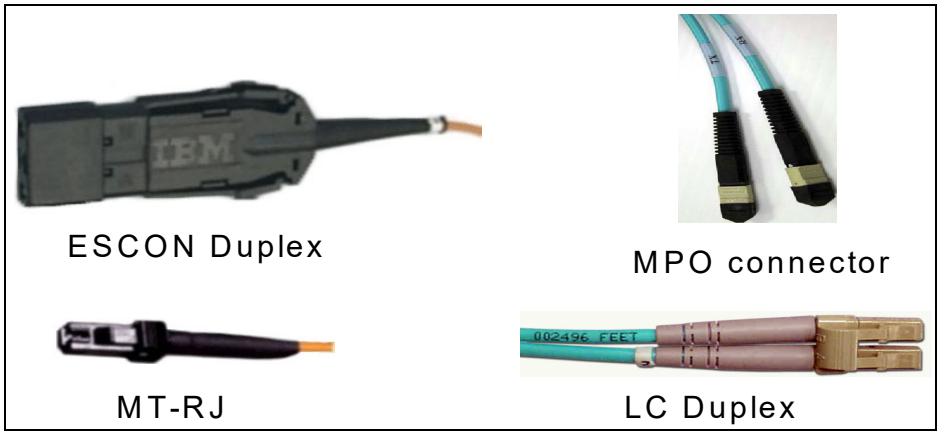


Figure D-2 The connectors that are commonly used for optical cables

Mode-conditioning patch cables

In certain situations, for reusing existing multimode fiber optic cabling infrastructure, it is possible to connect a long wavelength (1300 nm) single-mode transceiver with multimode fiber by placing a special device called a mode-conditioning patch (MCP) cable. The MCP cables are 2 meters long and have a link loss of up to 5.0 dB.

The MCP cable *must* be installed on both ends of a link and occupy the same space as a standard 2-meter jumper cable. Adapter kits containing the MCPs are available with either SC Duplex connectors (to support coupling links) or ESCON connectors (to support ESCON-to-FICON migration). MCP adapters differ for 50 or 62.5 µm fiber. MCPs reduce the maximum link distance to 550 meters for Gigabit links.

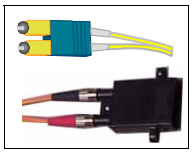
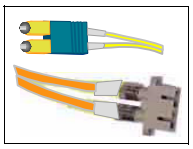
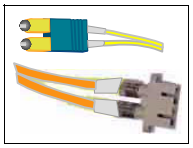
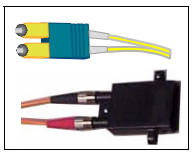
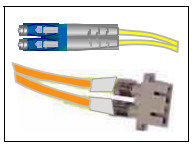
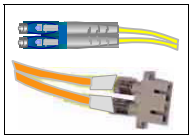
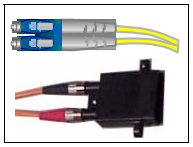
Optical mode conditioners are supported for FICON, coupling links, and OSA. For more information, see *Planning for Fiber Optic Links*, GA23-1409.

Important: One MCP cable must be plugged into the long wavelength transceiver at each end of the link.

Fiber optic MCP cables cannot be ordered as product feature codes for IBM zSystems.

Fiber optic-mode-conditioning patch cables (listed in Table D-1) can be ordered through the IBM Networking Services fiber cabling services options.

Table D-1 MCP cables

MCP cable description	MCP cable connector or receptacle description	MCP cable connector or receptacle illustration
9 µm single mode to 50 µm multimode	SC duplex connector to ESCON duplex receptacle	
9 µm single mode to 50 µm multimode	SC duplex connector to SC duplex receptacle	
9 µm single mode to 62.5 µm multimode	SC duplex connector to SC duplex receptacle	
9 µm single mode to 62.5 µm multimode	SC duplex connector to ESCON duplex receptacle	
ISC-3 compatibility 9 µm single mode to 50 µm multimode	LC duplex connector to SC duplex receptacle	
9 µm single mode to 62.5 µm multimode	LC duplex connector to SC duplex receptacle	
9 µm single mode to 62.5 µm multimode	LC duplex connector to ESCON duplex receptacle	

zHyperLink Express and ICA SR cables

The HyperLink Express and ICA SR features are designed to drive distances up to 150 meters and support a link data rate of 8 GBps by using customer-supplied OM4 (4.7 GHz-Km @ 850 nm) fiber optic cables. With OM3 (2.0 GHz-Km @ 850 nm) fiber optic cables, the zHyperLink Express and ICA SR distance drops to 100 meters. Figure D-3 shows the OM4 fiber cable with 24-fibers (12 transmit plus 12 receive fibers) and Multi-fiber Termination Push-on (MTP) connectors.

Figure D-3 OM4 50/125 µm multimode fiber cable with MTP connectors

Custom cable lengths or standard cable lengths that are shown in Table D-2 are available from IBM GTS or through other vendors, such as Anixter Inc.

Table D-2 ICA-SR cables: Standard lengths

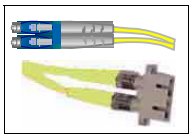
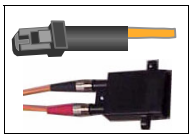
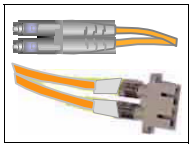
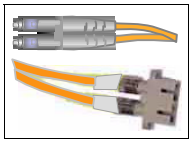
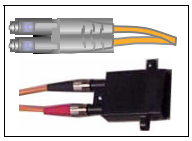
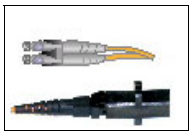
Part Number	Length meters (feet)
00JA683	1 m (3.28’)
00JA684	2 m (6.56’)
00JA685	3 m (9.84’)
00JA686	5 m (16.40’)
00JA687	8 m (26.24’)
00LU282	10 m (32.80’)
00LU283	13 m (42.65’)
00JA688	15 m (49.21’)
00LU669	20 m (65.61’)
00LU284	40 m (131.23’)
00LU285	80 m (262.36’)
00LU286	120 m (393.78’)
00LU287	150 m (492.12’)

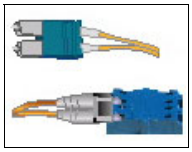
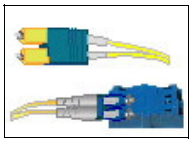
For more information, see *Planning for Fiber Optic Links,* GA23-1409. This publication is available at the Library section of Resource Link.

Conversion kits

Conversion kits allow for the reuse of already installed cables that are the same fiber optic mode but have different connectors than the ones that are required (see Table D-3).

Table D-3 Conversion kit cables

Conversion kit cable description	Conversion kit cable connector or receptacle description	Conversion kit cable connector or receptacle illustration
9 µm single mode	LC Duplex Connector to SC Duplex Receptacle	
62.5 µm multimode	MT-RJ Connector to ESCON Duplex Receptacle	
50 µm Multimode	LC Duplex Connector to SC Duplex Receptacle	
62.5 µm Multimode	LC Duplex Connector to SC Duplex Receptacle	
62.5 µm Multimode	LC Duplex Connector to ESCON Duplex Receptacle	
62.5 µm Multimode	LC Duplex Connector to MT-RJ Connector with Coupler	

Conversion kit cable description	Conversion kit cable connector or receptacle description	Conversion kit cable connector or receptacle illustration
62.5 µm Multimode	SC Duplex Connector to LC Duplex Connector with Coupler	
9 µm Single Mode	SC Duplex Connector to LC Duplex Connector with Coupler	

Note: Fiber optic conversion kits are not orderable as product feature codes for IBM zSystems.

Fiber optic conversion kits can be ordered by using the IBM Networking Services fiber cabling services options. Each conversion kit contains one cable.

References

For more information, see the following publications:

- *Planning for Fiber Optic Links*, GA23-1409
- *ESCON I/O Interface Physical Layer Document,* SA23-0394
- *Fiber Channel Connection for S/390 I/O Interface Physical Layer,* SA24-7172
- *Coupling Facility Channel I/O Interface Physical Layer,* SA23-0395
- *S/390 Fiber Optic Link (ESCON, FICON, Coupling Links and OSA) Maintenance Information,* SY27-2597
- *Fiber Transport Services Direct Attach Planning,* GA22-7234

Abbreviations and acronyms

AH	Authentication Header
AID	adapter ID
ANSI	American National Standards Institute
AP	Adjunct Processor
APXA	Adjunct Processor Extended Addressing
ARP	Address Resolution Protocol
BMS	Base Management Card
BTS	Backup Time Server
CCA	Common Cryptographic Architecture
CCW	channel command word
CE LR	Coupling Express Long Reach
CF	carry forward or coupling facility
CFCC	Coupling Facility Control Code
CHPID	channel path identifier
CMPCS	Compression Coprocessor
COP	coprocessor
CP	central processor
CPACF	Central Processor Assist for Cryptographic Function
CPC	central processor complex
CRC	cyclical redundancy check
CSS	channel subsystem
CTC	channel-to-channel
CTN	Coordinated Timing Network
CTS	Current Time Server
CU	control unit
DASD	direct access storage device
dB	decibels
DBR	device-base routing
DIF	Data Integrity Field
DIX	Data Integrity Extensions
DMA	Direct Memory Access
DR	disaster recovery
DX	Drawer
EADM	Extended Asynchronous Data Mover
ECAR	Enhanced Console Assisted Recovery
ECC	error-correcting code
ECKD	Extended Count Key Data
EE	Enterprise Extender
ELF	Express Logon Facility
ELS	Extended Link Services
EP11	Enterprise PKCS #11
ESP	Encapsulated Security Payload
ETS	External Time Source
FC	Fibre Channel or Feature Code
FCP	Fibre Channel Protocol
FCS	Fibre Channel Standard
FCTC	FICON channel-to-channel
FEC	Forward Error Correction
FICON	Fibre Channel connection
FID	Function ID
FIDR	FICON Dynamic Routing
FIPS	Federal Information Processing Standards
FLOGI	fabric login
GbE	gigabit Ethernet
Gbps	gigabits per second
GDPS	Geographically Dispersed Parallel Sysplex
GRE	generic routing encapsulation
GUI	graphical user interface
HA	high availability
HADR	high availability and disaster recovery
HBA	host bus adapter
HCD	Hardware Configuration Definition
HMC	Hardware Management Console
HPMA	Host Page-Management Assist
HSA	hardware system area
HSM	Hardware Security Module
IBM	International Business Machines Corporation
IC	Internal Coupling
ICA SR	Integrated Coupling Adapter Short Reach
IETF	Internet Engineering Task Force
IFP	integrated firmware processor
IMS	IBM Information Management System

INCITS International Committee of Information Technology Standards
IOCDS I/O configuration data set
IOCP Input/Output Configuration Program
IODF input/output definition file
IOS Input/Output Supervisor
IP Internet Protocol
IPA Internet Protocol Assist
iPDU intelligent Power Distribution Unit
IPIC Internet Protocol interconnectivity
IPv4 Internet Protocol version 4
IPv6 Internet Protocol version 6
iQDIO internal queued direct input/output
ISL Inter-Switch Link
ISM Internal Shared Memory
IU Information Unit
IWQ Inbound workload queuing
LCS LAN channel station
LCSS Logical Channel Subsystem
LGR Live Guest Relocation
LIC Licensed Internal Code
LPAR logical partition
LRC longitudinal redundancy check
LUN logical unit number
LX long wavelength
MAC Media Access Control
MCP mode-conditioning patch
MENA Middle East/North Africa
MES miscellaneous equipment specification
MIB Management Information Base
MIDAW Modified Indirect Data Address Word
MIF multiple image facility
MM multimode
MPO Multifiber Push-On
MSS multiple subchannel sets
MTP Multi-fiber Termination Push-on
NAT network address translation
NPIV N-Port ID Virtualization
NTP Network Time Protocol
OAT OSA Address Table
OCP OpenShift Container Platform
OLM optimized latency mode
OLS Offline Signal
ORB operation request block
OSA Open Systems Adapter
OSC OSA Integrated Console Controller
OSD OSA devices for QDIO
OxID Open Exchange ID Routing
PAV parallel access volume
PBR Port-based routing
PCHID physical channel ID
PCI Peripheral Component Interconnect
PCI program-controlled interrupt
PCIe Peripheral Component Interconnect Express
PCIeCC PCIe Cryptographic Coprocessor
PCU physical control unit
PDSE partitioned data set extended
PKCS Public Key Cryptography Standards
POR power-on-reset
PPRC Peer-to-Peer Remote Copy
PPS Pulse Per Second
PRLI Process Login
PSP Preventive Service Planning
PTF program temporary fix
PTP Precision Time Protocol
PTS Preferred Time Server
PU processing unit
QDIO queued direct input/output
QEBSM QDIO Enhanced Buffer-State Management
QoS quality of service
QP queue pair
QWDM Qualified Wavelength Division Multiplexer
RAS reliability, availability, and serviceability
RDMA Remote Direct Memory Access
RIP Routing Information Protocol
RMF Resource Measurement Facility
RNIC RDMA-capable network interface card (NIC)
RNID Request Node Identification
RoCE RDMA over Converged Ethernet
RPQ request for price quotation
SAN storage area network
SAP system assist processor

SBCON	Single-Byte Command Code Sets Connection
SCSI	Small Computer System Interface
SE	Support Element
SFP	Small Form-factor Pluggable
SIGA	Signal Adapter
SKE	Secure Key Exchange
SM	single mode
SMC	Shared Memory Communications
SMCv1	Shared Memory Communication Version 1
SMCv2	Shared Memory Communication Version 2
SMP	symmetric multiprocessor
SNA	System Network Architecture
SoD	Statement of Direction
SS	subchannel set
SSCH	start subchannel
SSID	subsystem identifier
SSL	Secure Sockets Layer
STP	Server Time Protocol
SX	short wavelength
TKE	Trusted Key Entry
TOD	time-of-day
UCB	unit control block
UDX	user-defined extension
UTP	unshielded twisted pair
VCHID	virtual channel identifier
VEPA	Virtual Ethernet Port Aggregator
VF	Virtual Function
VFM	Virtual Flash Memory
VIPA	virtual IP address
VLAN	virtual local area network
VMAC	virtual MAC
VPN	virtual private network
VSAM	Virtual Storage Access Method
WDFM	withdrawn from marketing
WDfM	withdrawn from marketing
WDM	wavelength-division multiplexing or Wavelength Division Multiplexer
WWN	worldwide names
WWNN	worldwide node name
WWPN	worldwide port name
zDAC	IBM z/OS Discovery and Automatic Configuration
zEDC	IBM zEnterprise Data Compression
zFS	IBM z/OS File System
zHPF	IBM High Performance FICON for IBM Z
zIIP	IBM Z Integrated Information Processor

Related publications

The publications that are listed in this section are considered suitable for a more detailed description of the topics that are covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide more information about the topics in this document. Some publications that are referenced in this list might be available in softcopy only.

- *Enterprise Extender Implementation Guide*, SG24-7359
- *FICON Planning and Implementation Guide*, SG24-6497
- *IBM Communication Controller for Linux on System z V1.2.1 Implementation Guide*, SG24-7223
- *IBM Communication Controller Migration Guide*, SG24-6298
- *IBM HiperSockets Implementation Guide*, SG24-6816
- *IBM z14 (3931) Technical Guide,* SG24-8951
- *IBM z16Technical Introduction*, SG24-8950
- *IBM z14 (3906) Technical Guide*, SG24-8451
- *IBM z14 Model ZR1 Technical Introduction*, SG24-8550
- *IBM z14 Technical Introduction*, SG24-8450
- *IBM z14 ZR1 Technical Guide*, SG24-8651
- *IBM z15 (8561) Technical Guide*, SG24-8851
- *IBM z15 Technical Introduction*, SG24-8850
- *OSA-Express Implementation Guide*, SG24-5948

You can search for, view, download, or order these documents and other Redbooks, Redpapers, web docs, drafts, and additional materials, at the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- *Communications Server: IP Configuration*, SC31-8513
- *Communications Server: SNA Network Implementation Guide*, SC31-8777
- *Communications Server: SNA Resource Definition Reference,* SC31-8565
- *Enterprise Systems Architecture/390 Principles of Operation*, SA22-7201
- *Fiber Optic Link Planning*, GA23-0367
- *Fiber Optic Links (ESCON, FICON, Coupling LInks and OSA) Maintenance Information*, SY27-2597

- *FICON I/O Interface Physical Layer,* SA24-7172
- *Hardware Configuration Definition (HCD) Planning*, GA32-0907
- *Hardware Configuration Definition: User's Guide,* SC28-1848
- *IBM 3931 Installation Manual for Physical Planning, GC28-7015*
- *IBM 8561 Installation Manual for Physical Planning,* GC28-7002
- *Processor Resource/Systems Manager Planning Guide*, SB10-7178
- *RMF Report Analysis*, SC28-1950
- *Stand-Alone Input/Output Configuration Program User's Guide*, SB10-7180
- *System z ESCON and FICON Channel-to-channel Reference*, SB10-7034
- *z/Architecture Principles of Operation,* SA22-7832

Online resources

These websites are also relevant as further information sources:

- Fibre Channel standards
  - http://www.t10.org
  - http://www.t11.org
- FICON Director vendors

  http://www.ibm.com/systems/storage/san/enterprise/index.html
- IBM Parallel Sysplex

  http://www.ibm.com/servers/eserver/zseries/pso
- IBM Resource Link for documentation and tools

  http://www.ibm.com/servers/resourcelink
- IBM Z I/O connectivity

  http://www.ibm.com/systems/z/hardware/connectivity/index.html
- IBM Z networking

  http://www.ibm.com/systems/z/hardware/networking/

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

To determine the spine width of a book, you divide the paper PPI into the number of pages in the book. An example is a 250 page book using Plainfield opaque 50# smooth which has a PPI of 526. Divided 250 by 526 which equals a spine width of .4752". In this case, you would use the .5” spine. Now select the Spine width for the book and hide the others: **Special>Conditional Text>Show/Hide>SpineSize(-->Hide:)>Set** . Move the changed Conditional text settings to all files in your book by opening the book file with the spine.fm still open and **File>Import>Formats** the Conditional Text Settings (ONLY!) to the book files.

(0.2”spine)
0.17”<->0.473”
90<->249 pages

To determine the spine width of a book, you divide the paper PPI into the number of pages in the book. An example is a 250 page book using Plainfield opaque 50# smooth which has a PPI of 526. Divided 250 by 526 which equals a spine width of .4752". In this case, you would use the .5” spine. Now select the Spine width for the book and hide the others: **Special>Conditional Text>Show/Hide>SpineSize(-->Hide:)>Set** . Move the changed Conditional text settings to all files in your book by opening the book file with the spine.fm still open and **File>Import>Formats** the Conditional Text Settings (ONLY!) to the book files.

SG24-5444-22

ISBN 0738460516

Printed in U.S.A.