HiperSockets and Shared Memory Communications

Linda Harrison
lharriso@us.ibm.com
IBM Washington System Center

Trademarks

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX*	IBM i*	POWER7*	PureSystems	Tivoli*
BladeCenter*	IBM logo*	Power Systems	Storwize*	WebSphere*
DB2*	Informix*	PowerVM	System Storage*	zEnterprise*
Easy TIER*	PartnerWorld *	PureApplication	System x*	
IBM*	Power*	PureFlex	System z*	

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Windows Server and the Windows logo are trademarks of the Microsoft group of countries.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

* Other product and service names might be trademarks of IBM or other companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

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

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the 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.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Agenda

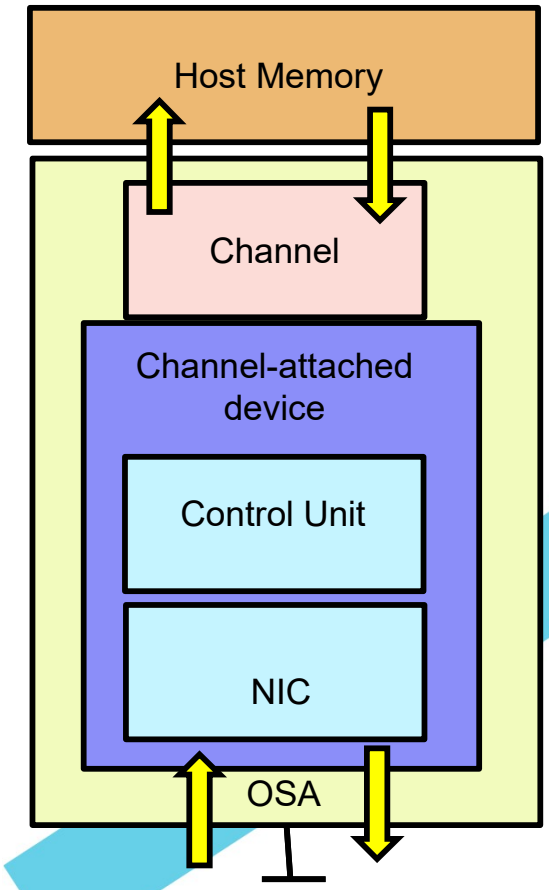
- OSA
- HiperSockets
- HiperSockets Features
- HiperSockets Converged Interface (HSCI)
- Shared Memory Communications
- Comparison
- Configuration
- SMC Connection Eligibility
- More Information

OSA

QDIO

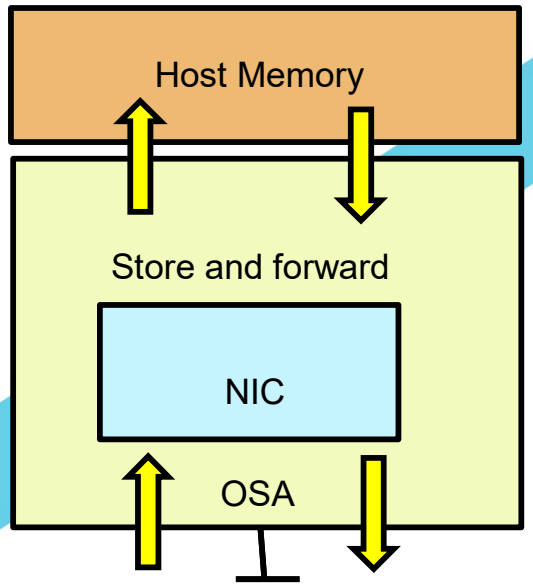
- Queued Direct Input Output – Provides Ethernet LAN Access – IP Layer 3 TCP/IP only

OSE mode (non-QDIO)
LAN Channel Station
(TCP/IP uses LCS protocol)
Link Station Architecture
(SNA uses LSA protocol)



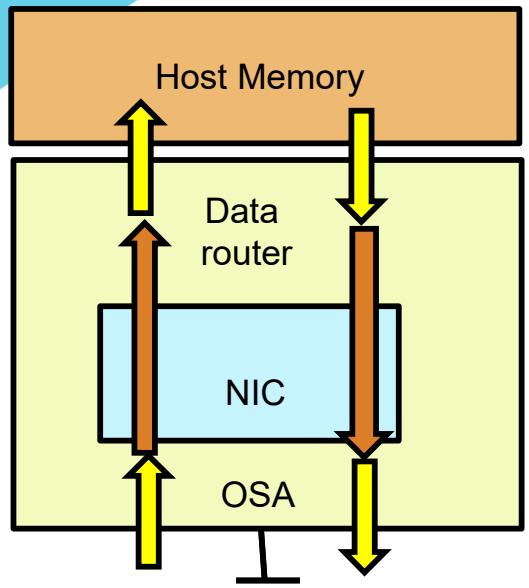
Same as old external 3172

OSD mode (QDIO)
OSA-Express2



OSD and System Z can access shared memory. OSA interrupts the system using Peripheral Component Interconnect (PCI). The system interrupts OSA via Signal Adapter (SIGA) Instruction.

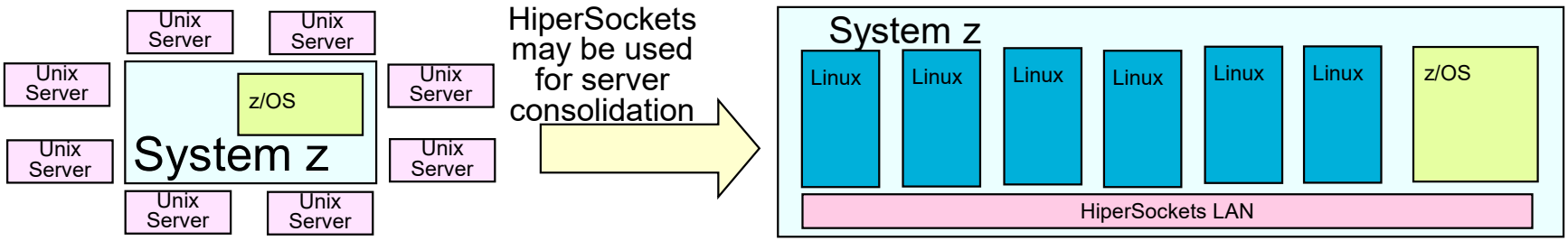
OSD mode (QDIO)
OSA-Express3+



Latest OSAs use Direct Memory Access (DMA) and hardware data router, allowing data to flow through OSA without slower store and forward process.

HiperSockets

HiperSockets (iQDIO)

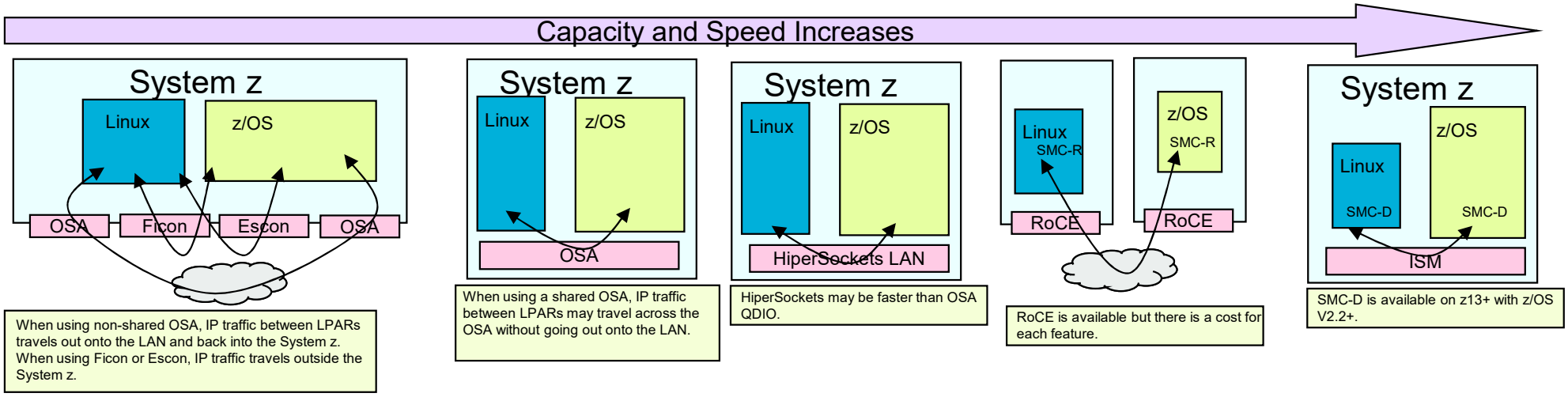


- HiperSockets = Internal Queued Direct Input Output (iQDIO)
  - Was developed from the OSA QDIO architecture
    - Is limited to TCP/IP protocol only to/from z/OS (z/VM and Linux on Z support Layer 2)
  - Also known as HiperSockets device or System z internal virtual LAN or HiperSockets LAN
  - LPAR to LPAR communication via shared memory
    - High speed, low latency, similar to cross-address-space memory move using memory bus
    - Provides better performance than channel protocols for network access.
  - Multiple HiperSockets may be configured as internal LANs on the System z box.
  - A HiperSockets LAN may be configured to be part of TCP/IP DynamicXCF.
    - A TCP/IP stack may only define a single HiperSockets LAN for DynamicXCF.
      - Some TCP/IP stacks may use one HiperSockets LAN for DynamicXCF connectivity while other TCP/IP stacks use a different HiperSockets LAN for DynamicXCF connectivity.
    - Not recommended for some LPARs to define a HiperSockets LAN for DynamicXCF and other LPARs to manually define the same HiperSockets LAN.
  - Common Lookup Table is stored in the Hardware System Area, the same as QDIO.
    - HiperSockets LAN, IP address, TCP/IP stack

Hardware HiperSockets

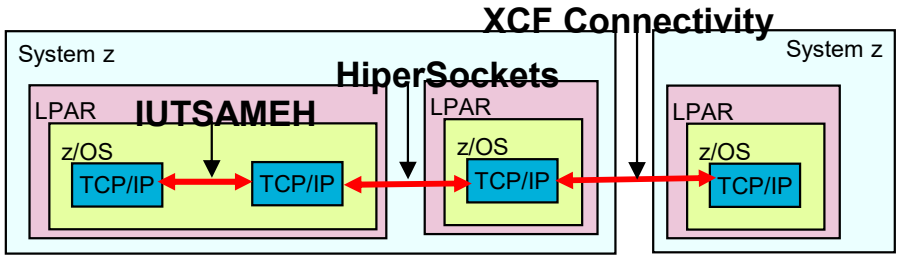
- HiperSockets are virtual LANs provided by the System Z platform without any additional fee. Because they are internal virtual LANs on System Z there is no exposed cable or wire and therefore provide a secure connection between LPARs in the same System Z machine.
- This document details HiperSockets in z/OS running standalone on a System Z processor LPAR. For the additional tasks required for configuring HiperSockets in z/OS running as a guest under z/VM, or HiperSockets in z/OS running as a guest under z/VM using Guest LAN(s) support, or for the tasks required for HiperSockets in z/Linux or in z/VM, please consult the “HiperSockets Implementation Guide” Redbook, SG24-6816

TCP/IP LPAR to LPAR Communication Path



- The System Z does a direct move from one LPAR's memory to another LPAR's memory.
- HiperSockets is usually faster than OSA between LPARs
  - RoCE and SMC-D offer an additional option (RoCE has a cost associated with it).
  - SMC-D is usually preferred to HiperSockets.
- HiperSockets Maximums
  - Current limits are documented in Redbook "IBM System z Connectivity Handbook", SG24-5444
- HiperSockets CHPIDs may span across multiple Logical Channel SubSystems (LCSS)

TCP/IP DynamicXCF Transport Choices

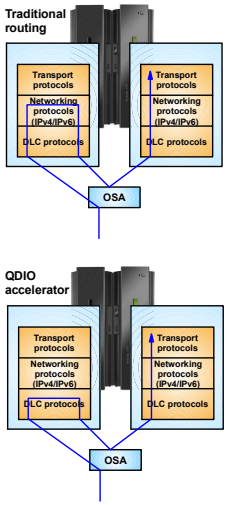


- TCP/IP DynamicXCF is capable of dynamically creating multiple device, link, and interfaces all with the same IPv4 or IPv6 address:
  - Same host: device IUSAMEH, link EZASAMEMVS (IPv4), interface EZ6SAMEMVS (IPv6)
  - Hipersockets: device IUTIQDIO, link IQDLNK*nnnnnnnn* (IPv4), interface IQDIOINTF6 (IPv6)
    - where *nnnnnnnn* is the hexadecimal representation of the IP address specified on the IPCONFIG DYNAMICXCF statement
  - XCF connectivity: device CPName, link EZAXCF*nn* (IPv4), interface EZ6XCF*nn* (IPv6)
    - where *nn* is the 2-character &SYSCLONE value
- TCP/IP DynamicXCF automatically chooses the fastest path:
  - Same host is used between TCP/IP stacks inside the same LPAR
  - HiperSockets is used between TCP/IP stacks inside same System z CEC (when configured)
  - XCF connectivity is used between TCP/IP stacks outside the CEC
- VTAM start-options required for TCP/IP DynamicXCF use of HiperSockets:
  - IQDCHPID=nn (where nn is the HiperSockets LAN CHPID, ie. FA)
  - XCFINIT=YES (required for TCP/IP DynamicXCF with or without HiperSockets)
    - Generates XCF device ISTLSXCF in VTAM
    - Requires prerequisite Start Options like HPR=RTP, which requires minimal APPN enablement.
- TCP/IP Profile options required to define DynamicXCF:
  - IPCONFIG DYNAMICXCF 10.1.2.101 ...
  - IPCONFIG6 DYNAMICXCF 2001:0DB8:1:0:50C9:C2D4:0:1 ...
- When HiperSockets DynamicXCF is configured:
  - HiperSockets DEVICE/LINK/INTERFACE/HOME and VTAM TRLE are all dynamically built

HiperSockets Features

QDIO Accelerator

- Accelerator support includes all combinations of QDIO and iQDIO traffic
  - When traffic is routed through z/OS.
  - Inbound over OSA or HiperSockets and Outbound over OSA or HiperSockets
- The first packet will travel up thru QDIO to the Accelerator stack and down thru iQDIO device drivers to reach the backend LPAR IP address. After that first packet, all the rest of the packets flow via the accelerated path through the DLC layer, thus bypassing the IP layer in z/OS and reducing path length and improving performance.



	Outbound QDIO	Outbound iQDIO
Inbound QDIO	Yes	Yes
Inbound iQDIO	Yes	Yes

Accelerator requires IP Forwarding to be enabled for non-Sysplex Distributor acceleration.
No Acceleration for:
•IPv6
•Traffic which requires fragmentation in order to be forwarded
•Incoming fragments for a Sysplex Distributor connection
•OSA port interfaces using Optimized Latency Mode (OLM)

- Supports Sysplex Distributor (SD)
  - When traffic to target stack is sent over HiperSockets Dynamic XCF or QDIO as a result of VIPAROUTE definition.

IPCONFIG QDIOACCELERATOR

Multiple Write Facility and zIIP Offload

- HiperSockets can move multiple output data buffers in one write operation
  - Reduces CPU utilization

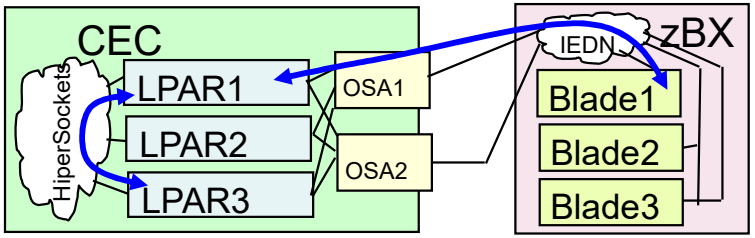
GLOBALCONFIG IQDMULTIWRITE ZIIP IQDIOMULTIWRITE

- Multiwrite operation can be offloaded to a zIIP
  - Only for TCP traffic that originates in this host
    - **Only large TCP outbound messages (32KB+)**

Write operation (System z9)

Write operation (System z10)

HiperSockets Integration with OSA for IEDN

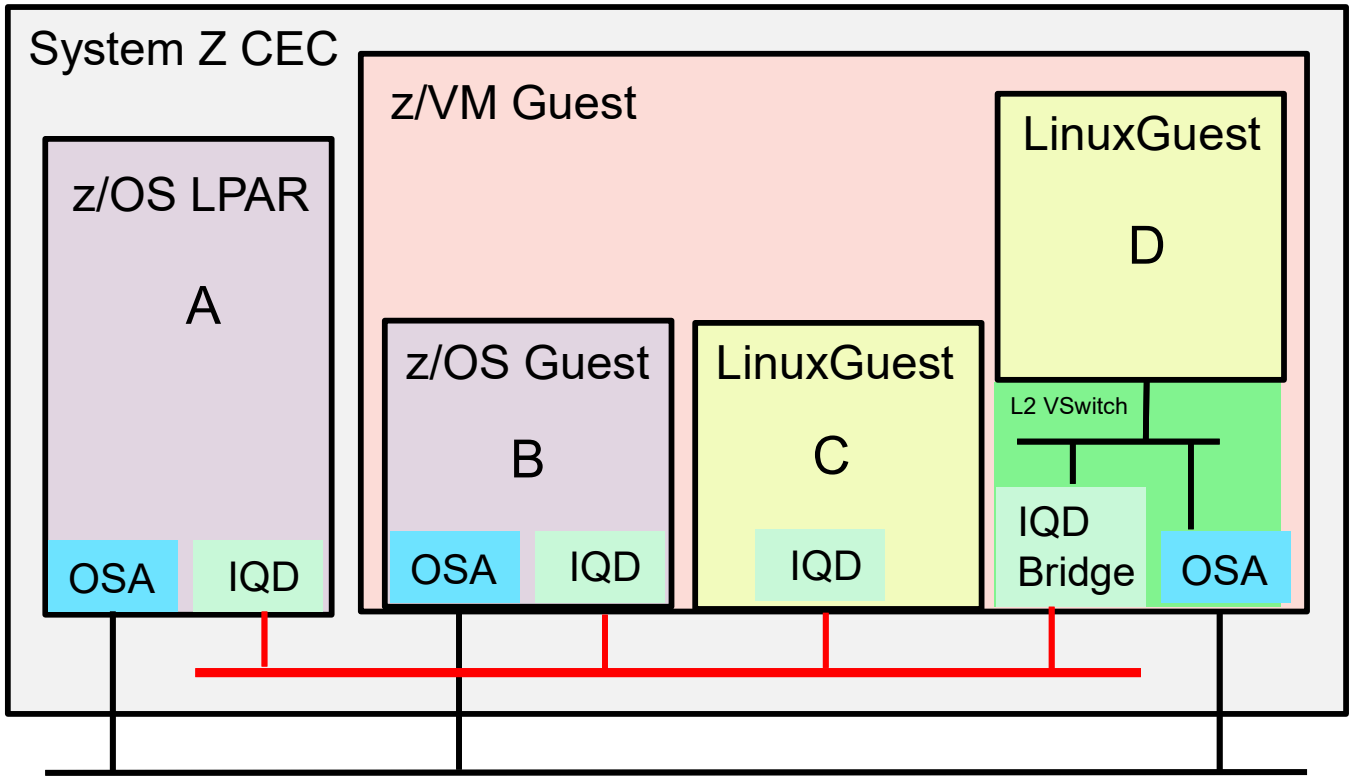


CHPARM=x2 See previous discussion in this presentation.
GlobalConfig AUTOIQDX ALLTRAFFIC
or GlobalConfig AUTOIQDX NOLARGEDATA
or GlobalConfig NOAUTOIQDX
Dynamically created TRLE is IUTIQXxx or IUTIQ6xx and dynamically created interface is EZAIQXxx or EZ6IQXxx where xx is the OSX CHPID.

- A single HiperSockets LAN may be defined such that it is automatically used when the destination is an LPAR on the same CEC belonging to the same OSX/HiperSockets LAN.
  - The OSA OSX devices are assigned IP Addresses.
  - The HiperSockets LAN (IQDX) is not assigned an IP Address.
- Background
  - With VIPA and Dynamic Routing
    - The application may bind to the VIPA.
    - Dynamic routing causes traffic between LPARs to be routed over HiperSockets.
    - Dynamic routing causes traffic to remote partners (outside the CEC) to be routed over OSA.
  - Without VIPA and Dynamic Routing
    - It is a challenge to cause same CEC traffic to flow over HiperSockets at the same time that remote traffic flows over OSA.
    - Static Host routes may be used.
      - The application binds to the OSA IP address.
      - A static route is used on each LPAR such that when the other LPAR OSA address is the destination then the HiperSockets LAN is used to route the traffic.
      - With a large number of LPARs the administration of these static host routes is onerous.

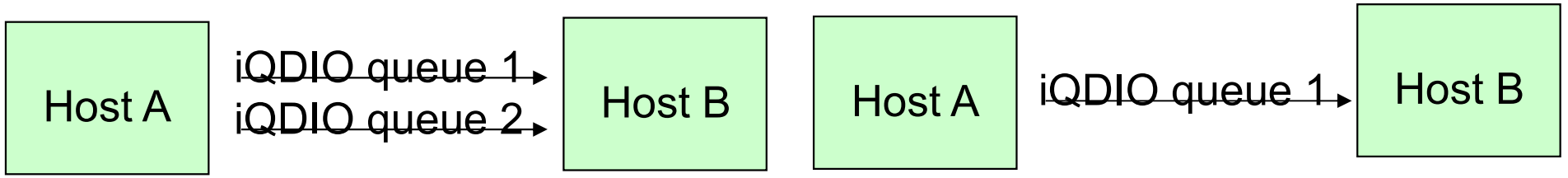
HCD (IOCP)
Define 10 subchannel addresses for each IQDX CHPID that is in use for IPv4.
Define 10 subchannel addresses for each IQDX CHPID that is in use for IPv6.
Multiple VLAN does not affect the required number of subchannel addresses.

HiperSockets Converged Interface



- HiperSockets LAN may be defined such that it is automatically used when the destination is an LPAR on the same CEC belonging to the same OSD/HiperSockets LAN.
  - The OSA OSD devices are assigned IP Addresses.
  - The HiperSockets LAN (IQDX) is not assigned an IP Address.

HiperSockets Completion Queue



When the remote side can receive data volume then data is sent synchronously.

When the remote side cannot receive data volume then data is sent asynchronously.

- Automatic capability
- HiperSockets transfers data synchronously if possible and asynchronously if necessary.
- Ultra-low latency with more tolerance for traffic peeks.

Non-z/OS Support

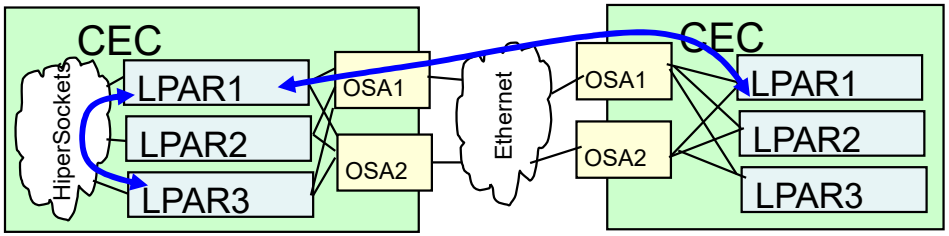
- z/VM Virtual HiperSockets Support
  - In addition to CEC HiperSockets that are available to all LPARs, z/VM is able to support virtual HiperSockets available to all guests running on that z/VM image.
- z/VM HiperSockets Virtual Switch Bridge Support (also referred to as External Bridge)
  - A single HiperSockets LAN may be defined such that it is automatically used when the destination is an LPAR on the same CEC belonging to the same OSD/HiperSockets or OSX/HiperSockets Virtual Switch.
  - The OSA OSD or OSX devices are assigned IP Addresses. Like z/OS HiperSockets Converged Interface (HSCI)
  - The HiperSockets LAN is not assigned an IP Address.
- zLinux HiperSockets Network Concentrator Like z/OS QDIO Accelerator
  - zLinux with HiperSockets and OSD devices is able to bridge traffic without routing overhead providing increased performance.
- zLinux HiperSockets Network Traffic Analyzer
  - Allows Linux on System z to control tracing of the internal virtual LAN.
  - Requires System z10 or later hardware.
  - Requires Linux for System z software.
- z/VM and zLinux Layer 2 Support
  - Both z/VM and zLinux support HiperSockets Layer 2 as well as Layer 3
  - z/OS only supports Layer 3
- z/VSE Fast Path to Linux (LFP) Support
  - Allows communications of z/VSE TCP/IP applications to Linux without a TCP/IP stack on z/VSE.
  - Requires:
    - z196 or later
    - z/VSE V5.1.1
    - LFP in an LPAR
    - HiperSockets Completion Queue

HiperSockets Features

HiperSockets Supported Features	z/OS	z/VM	Linux on System z	z/VSE
IPv4 Support	Yes	Yes	Yes	Yes
IPv6 Support	Yes	Yes	Yes	Yes
VLAN Support	Yes	Yes	Yes	Yes
Network Concentrator	No	No	Yes	No
QDIO Accelerator	Yes	No	No	No
Layer 2 Support	No	Yes	Yes	No
Multiple Write Facility	Yes	No	No	No
zIIP Assisted Multiple Write Facility	Yes	No	No	No
HiperSockets NTA (Network Traffic Analyzer)	No	No	Yes	No
Integration with IEDN (IQDX)	Yes	No*	Yes	No
Virtual Switch Bridge Support	No	Yes	No	No
HiperSockets Converged Interface (HSCI)	Yes	No	No	No
Fast Path to Linux (LFP) Support / IUCV over HiperSockets	No	No	Yes	Yes
Completion Queue	Yes	No	Yes	Yes
* Depends upon the z/VM release				

HiperSockets Converged Interface (HSCI)

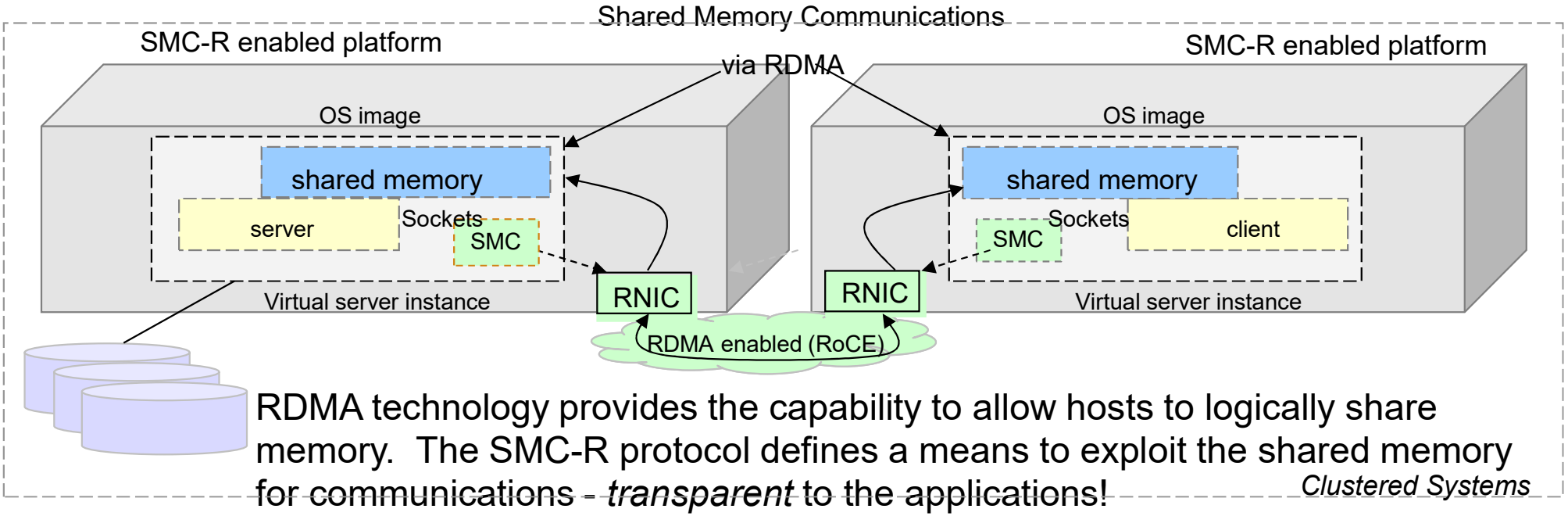
HiperSockets Converged Interface



- HiperSockets LAN may be defined such that it is automatically used when the destination is an LPAR on the same CEC belonging to the same OSD/HiperSockets LAN.
  - The OSA OSD devices are assigned IP Addresses.
  - The HiperSockets LAN (IQDX) is not assigned an IP Address.
- Background
  - With VIPA and Dynamic Routing
    - The application may bind to the VIPA.
    - Dynamic routing causes traffic between LPARs to be routed over HiperSockets.
    - Dynamic routing causes traffic to remote partners (outside the CEC) to be routed over OSA.
  - Without VIPA and Dynamic Routing
    - It is a challenge to cause same CEC traffic to flow over HiperSockets at the same time that remote traffic flows over OSA.
    - Static Host routes may be used.
      - The application binds to the OSA IP address.
      - A static route is used on each LPAR such that when the other LPAR OSA address is the destination then the HiperSockets LAN is used to route the traffic.
      - With a large number of LPARs the administration of these static host routes is onerous.

Shared Memory Communications

SMC-R and RoCE Alternative to TCP/IP

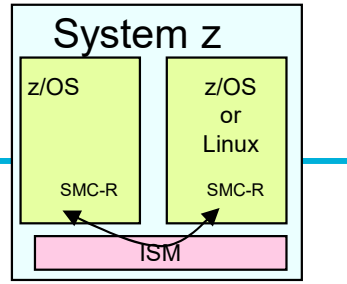


RDMA technology provides the capability to allow hosts to logically share memory. The SMC-R protocol defines a means to exploit the shared memory for communications - *transparent* to the applications!

Clustered Systems

Shared Memory Communications – Remote (SMC-R) is an *open* sockets over Remote Direct Memory Access (RDMA) protocol that provides transparent exploitation of RDMA (for TCP based applications) while preserving key functions and qualities of service from the TCP/IP ecosystem that enterprise level servers/network depend on! SMC-R uses RDMA over Converged Ethernet (RoCE).

SMC-D



SMC-R is available on z13 and z13s with z/OS V2.2.

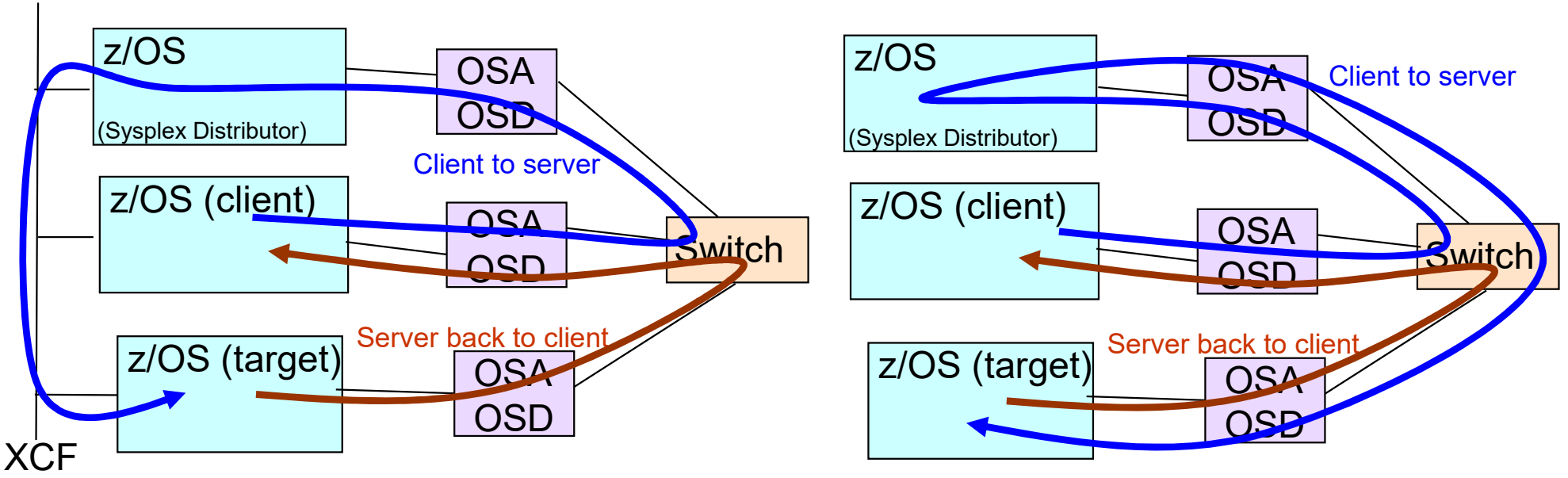
- Shared Memory Communications – Direct Memory Access (SMC-D) (z13s and z13 GA2)
  - High bandwidth, low latency LPAR-to-LPAR TCP/IP traffic using the direct memory access software protocols over virtual Internal Shared Memory (ISM) devices. Designed to provide application transparent RDMA communications to TCP endpoints for sockets-based connections with reduced latency, improved throughput, and reduced CPU cost compared to HiperSockets, OSA, or SMC-R (RoCE).
  - Streaming workload test case results*:
    - Up to 89% reduction in latency, 9 times the throughput, and 87% reduction in CPU cost compared to HiperSockets*
    - Up to 95% reduction in latency, 20 times the throughput, and 83% reduction in CPU cost compared to OSA*
    - Up to 94% reduction in latency, 16 times the throughput, and 58% reduction in CPU cost compared to SMC-R (RoCE)*
- SMC-D performance test results:
  - https://www.ibm.com/support/pages/node/317829

*This performance data was measured in a controlled environment under z/OS. The actual latency, throughput, and CPU cost that any **client will experience will vary depending upon** considerations such the I/O configuration, the storage configuration, and the characteristics of the communications workload.

SMC-R and SMC-D

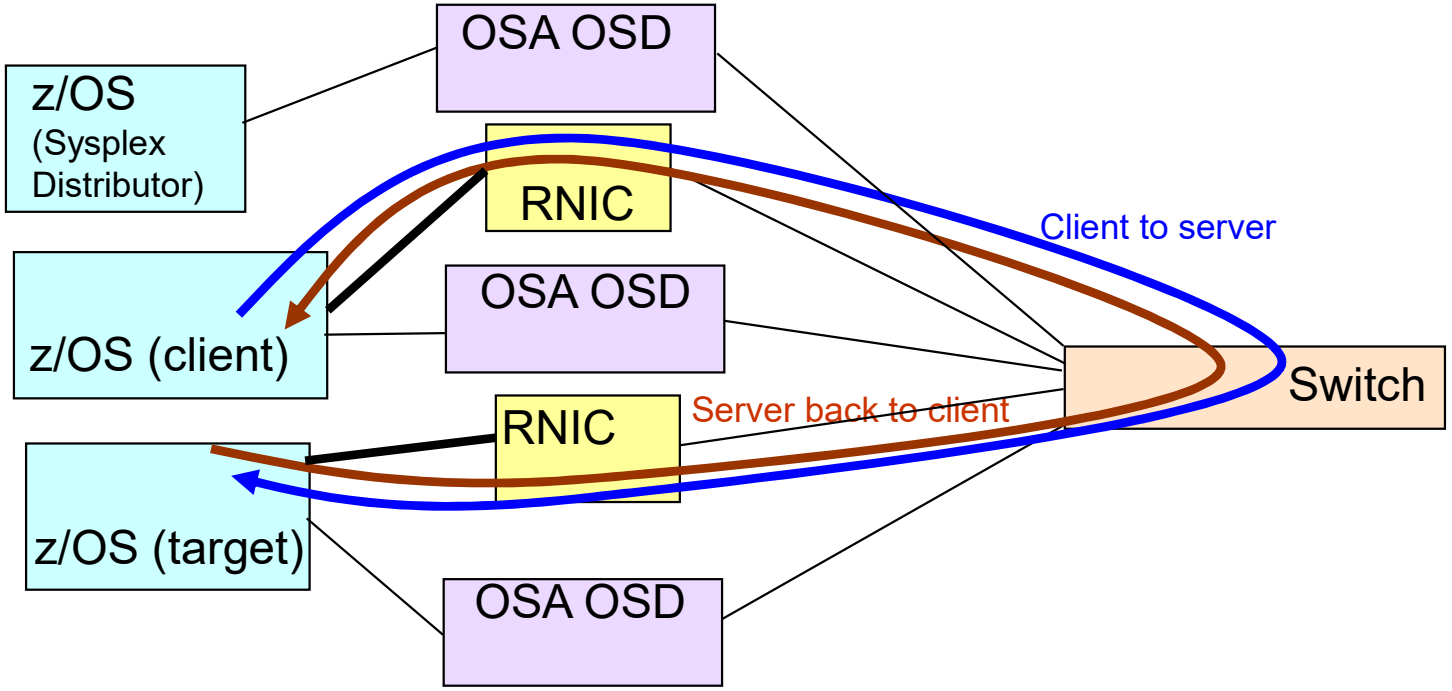
- TCP traffic only. No UDP traffic support. No Enterprise Extender support.
- The applications still send data to the TCP/IP stack.
  - No application change required.
  - The application has no idea that RDMA is being used.
- In the initial TCP session setup handshake SMC-R and SMC-D eligibility is determined.
  - If SMC-R or SMC-D eligibility is determined the traffic is “switched” to RDMA.
  - The TCP/IP session remains active for the life of the session for Keep Alive, etc.
- If the RDMA path fails, the session automatically will failover to another RDMA path, but it will not “fall back” to TCP/IP. New sessions remain on the TCP/IP path and not “switch” to the RDMA path if the RDMA path is not available.
- By using RDMA instead of TCP/IP a performance improvement is achieved.
- Original SMC (SMCv1) requires that partners are both in the same subnet (Layer 2).
- SMCv2 supports partners in different subnets (Layer 3 IP Routing).
- When communicating with SMC-R between an SMC-Rv1 host and an SMC-Rv2 Ethernet Switch Port Trunk mode is recommended.

Sysplex Distributor without Shared Memory



- Client to Sysplex Distributor to target.
- With VIPAROUTE parameter the traffic can be sent over the OSA network rather than the XCF network.
  - VIPADYNAMIC
    - VIPADEFINE 255.255.255.0 10.15.63.111
    - VIPADISTRIBUTE 10.15.63.111
    - VIPAROUTE 10.201.22.89 10.173.44.89
  - ENDVIPADYNAMIC
- Return traffic back does not have to go through Sysplex Distributor system.

Syplex Distributor with Shared Memory



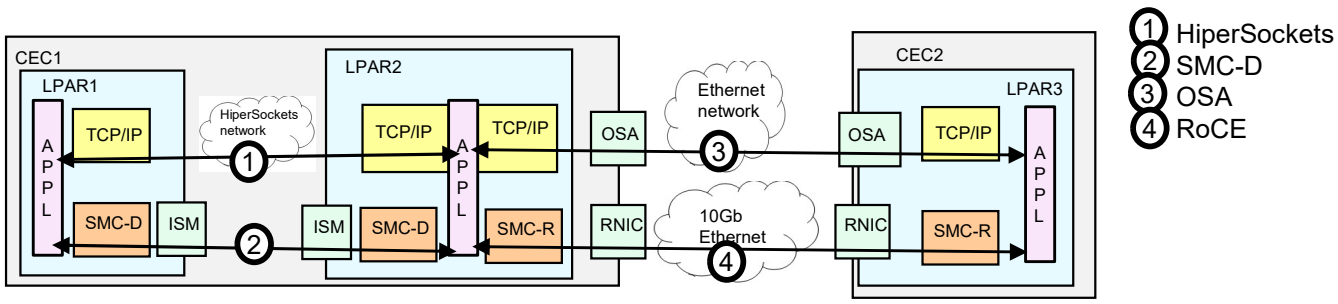
- When VIPAROUTE sends traffic over OSA rather than XCF, traffic is SMC-R RoCE eligible.
  - No need to go through Sysplex Distributor.
  - The initial TCP session setup is still done from client to Sysplex Distributor to target server.

SMC-AT and SMC Preference

- SMC - Applicability Tool (SMC-AT)
  - Reports percentage of traffic that is eligible for and well suited for SMC.
- Performance Improvement and Lower Overhead
  - All things being equal, traffic between partners in the same subnet (Layer 2) perform better than partners in different subnets that have to traverse a Layer 3 router.
  - HiperSockets provides the biggest performance improvement when messages are 32K and larger.
  - SMC provides the biggest performance and overhead improvement when messages are 32K and larger as well.
- Routing Issues
  - z/OS partners using dynamic routing may prefer HiperSockets over OSA.
  - When Linux on Z System implements HiperSockets static Host Routes must be used for routing.
    - HiperSockets Converged Interface solves this problem because the routing is automatically handled by the TCP/IP stack.
  - SMC doesn't have a routing problem because it doesn't use TCP/IP.

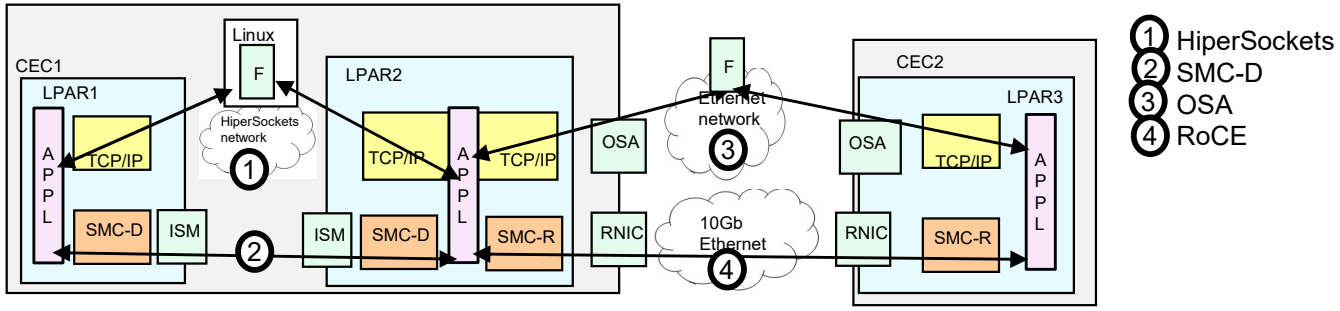
Comparison

Cross CEC and OS



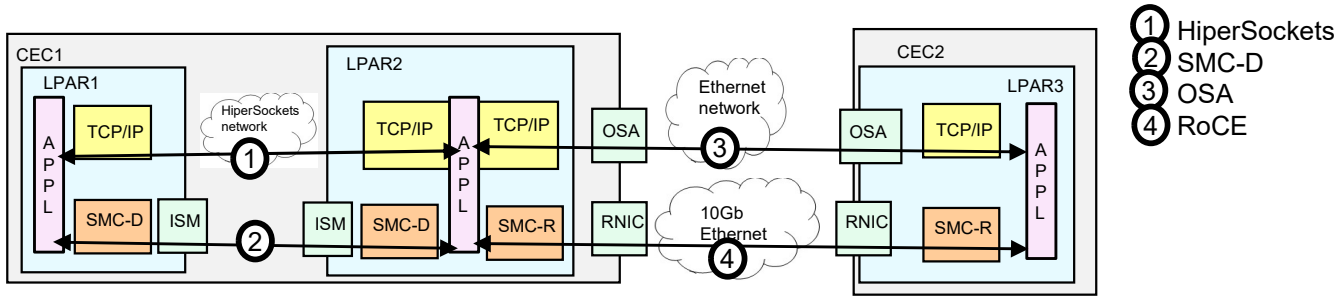
- All used for data transfer between hosts.
- Cross CEC traffic
  - HiperSockets and SMC-D only supports traffic between LPARs on a single CEC.
  - OSA and RoCE can be used to send traffic between CECs.
- Different Operating Systems
  - HiperSockets and OSA are supported by multiple Operating Systems (ie. z/OS, z/VM, Linux on System z, etc.)
  - SMC-D and RoCE are supported by z/OS and Linux on Z
  - There is a different non-Z System RoCE that supports SMC-R on AIX.

Firewall and Protocol



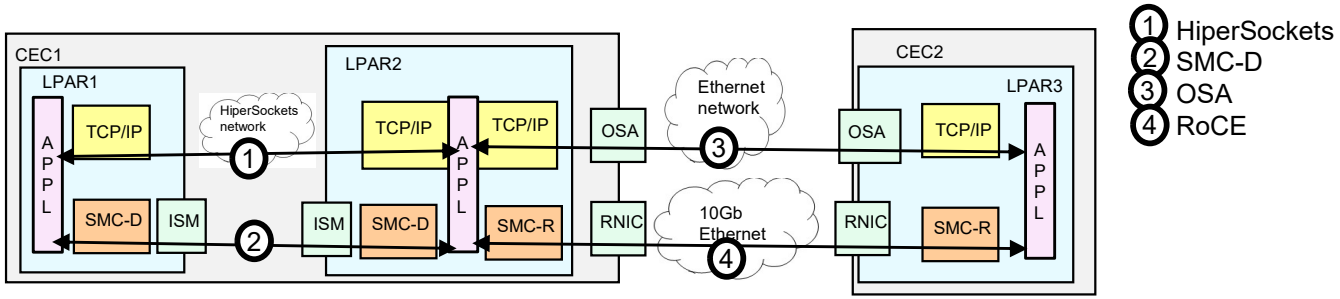
- Firewall with stateful packet inspection (a PCI (Payment Card Industry) requirement for some traffic)
  - HiperSockets traffic cannot be sent over a Firewall with stateful packet inspection support (unless routed traffic).
  - OSA traffic can be sent over a LAN to a Firewall with stateful packet inspection support.
  - SMCv1 traffic cannot be sent over a Firewall with stateful packet inspection support (does not support routed traffic).
  - SMCv2 traffic can be sent over a Firewall with stateful packet inspection support (does support routed traffic).
- Protocol Support
  - HiperSockets only supports IP traffic, it does not support native SNA (so EE must be used to send SNA).
  - OSA supports all TCP/IP protocols and even supports SNA protocol (natively in OSE mode, or when sent with UDP (Enterprise Extender (EE))).
  - SMC-D and RoCE only support TCP traffic, it does not support native SNA or UDP (EE).

Hardware



- Required hardware feature
  - HiperSockets is part of z System Firmware so it does not require any additional hardware / adapter card purchase.
  - SMC-D requires either OSA or HiperSockets but only for minimal traffic flow.
  - OSA requires OSA card(s).
  - SMC-R requires RoCE feature(s).

Overhead



- CP Overhead
  - HiperSockets supports zIIP offload to reduce associated cost.
  - OSA provides many different types of offload to the adapter that reduces CP overhead (ARP, Check Sum, Segmentation, etc.)
  - RoCE and SMC-D reduce overhead by using RDMA protocol instead of TCP/IP protocol.
- Storage Usage
  - HiperSockets and OSA use CSM fixed storage backed by 64-bit real for data buffers and use Hardware System Area (HSA) memory for routing table.
  - SMC-D and SMC-R use pinned Fixed Memory (mostly 64-bit Common), not CSM managed memory. The maximum amount of memory available for SMC-R or SMC-D is definable. When the maximum is reached, new connections will not be SMC-R or SMC-D eligible.

Routing and VLAN

- Dynamic Backup
  - HiperSockets with dynamic routing protocol may be “backed up” by OSA.
    - When HiperSockets is defined as part of a DynamicXCF network routing is handled automatically.
    - When HiperSockets is defined with Integration with IEDN (IQDX) routing is handled automatically.
    - When HiperSockets Converged Interface is defined routing is handled automatically.
    - Backup might not be a requirement because if HiperSockets fails there is probably major CEC problems occurring.
  - OSA with dynamic routing protocol may be “backed up” by another OSA or HiperSockets. OSA with stack routing may be “backed up” to another OSA in the same subnet.
  - SMC-D traffic starts over OSA/HiperSockets but once it “switches” to SMC-D there should be no need for “backup” since SMC-D, like HiperSockets, has no hardware piece to fail.
  - SMC-R traffic starts over OSA but once it “swiches” to SMC-R it may be “backed up” by other RoCE pairs.
    - Active RoCE or SMC-D sessions are dropped if the connection fails and there is no alternate path, but new sessions will flow using OSA (or HiperSockets for SMC-D).
- IP Routed Traffic
  - HiperSockets supports routed traffic (ie. traffic may come in over OSA and then be routed over HiperSockets). Routed traffic is optimized on z/OS with QDIO Accelerator.
  - SMC-D and RoCE support routed traffic with SMCv2. All connections are over a single subnet (Layer 2) network for the original SMC (SMCv1).
  - OSA supports routed traffic. Routed traffic is optimized on z/OS with QDIO Accelerator.
- VLAN Support
  - HiperSockets supports VLAN tagging.
  - SMC-D and RoCE inherit VLAN IDs from all associated OSAs (or HiperSockets for SMC-D) and supports VLAN tagging.
  - OSA supports VLAN tagging.

Configuration

OSA and HiperSockets Definitions

- OSA
  - HCD OSD CHPID for QDIO
    - PNET – must match for OSA and RoCE/ISM
  - VTAM Transport Resource List (TRL)
  - TCP/IP Profile Interface IPAQENET
    - IPADDR with subnet mask
    - SMCR is enabled by default on the OSA
      - PFID **required for SMC-Rv2**
      - SMCRIPADDR **required for SMC-Rv2**
        - must be IPv4 within the same IP subnet as the OSA
        - is not visible to or usable by z/OS applications (not added to the homelist).(IP address duplication is enforced)
- HiperSockets
  - HCD VCHID type IQD
    - PNET – must match for HiperSockets and ISM
  - VTAM IQDCHIP defines CHPID for DynamicXCF
  - TCP/IP Profile
    - Interface IPAQIDIO
    - GLOBALCONFIG
      - AUTOIQDC – enables HiperSockets Converged Interface, PNET must match for OSA and HiperSockets
      - AUTOIQDX – enables HiperSockets Integration with IEDN
      - IQDMULTIWRITE – enables HiperSockets Multi-Write
      - IQDVLANID – to define a DynamicXCF HiperSockets VLAN ID for subplexing
      - ZIIP IQDIOMULTIWRITE – offloads HiperSockets processing to zIIP engine(s) (requires IQDMULTIWRITE)
    - IPCONFIG
      - QDIOACCELERATOR – enables QDIO/iQDIO Accelerator

HiperSockets Dynamically Created Items

- TRLEs in ISTTRL Major Node (where xx = CHPID)
  - Manual HiperSockets
    - Device/Link MPCIPA/IPAQIDIO or Interface IPAQIDIO6 TRLE = IUTIQDxx
    - Interface IPAQIDIO TRLE = **IUTIQ4xx**
  - IPv4 DynamicXCF HiperSockets TRLE = **IUTIQDIO**
  - IPv6 DynamicXCF HiperSockrets TRLE = **IQDIOINTF6**
  - IPv4 HiperSockets Integration with IEDN (IQDX) TRLE = **IUTIQXxx**
  - IPv6 HiperSockets Integration with IEDN (IQDX) TRLE = **IUTIQ6xx**
- Device, Link, and Home (where where nnnnnnnn is hexadecimal representation of the IP address)
  - IPv4 DynamicXCF
    - DEVICE = **IUTIQDIO**
    - LINK = **IQDIOLNKnnnnnnnn**
    - HOME for IQDIOLNKnnnnnnnn
- Interface (where xx = CHPID)
  - IPv6 DynamicXCF INTERFACE = **IQDIOINTF6**
  - HiperSockets Integration with IEDN (IQDX)
    - IPv4 Interface IPAQIQDX = **EZAIQX*xx***
    - IPv6 Interface IPAQIQDX6 = **EZ6IQXxx**

HiperSockets CHPID

CHPID Parameter	MFS	MTU
CHPARM=00	16K	8K
CHPARM=40	24K	16K
CHPARM=80	40K	32K
CHPARM=C0	64K	56K

Default

"CHPARM" parameter was originally "OS" parameter.

On z196 and later processor the CHPID is also used to identify usage.
First character still indicates frame size:
0x, 4x, 8x, and Cx (as documented on the left)
The second character indicates usage:
x0 indicates Normal HiperSockets
x2 indicates HiperSockets for IEDN
x4 indicates HiperSockets for z/VM External Bridge
Where all "x" characters are wild cards.

- Each CHPID has configurable frame size (16K, 24K, 40K, 64K)
  - Allows optimization per HiperSockets LAN for small packets versus large streams
  - Affects MTU size of 8K, 16K, 32K, 56K
- HiperSockets LANs
  - Each HiperSockets LAN has its own CHPID, type IQD
    - IBM recommends starting from x"FF" and working your way backwards through the CHPID numbers, picking addresses from the high range to avoid addressing conflicts
    - May be shared by all defined LPARs
    - Delivered as object code only (OCO)
      - HiperSockets CHPIDs do not reside physically in the hardware but these CHPIDs cannot be used by other devices.
      - No physical media constraint, so no priority queuing or cabling required .
    - Each Operating System image configures its own usage of available HiperSockets CHPIDs.

SMC-R and SMC-D Definitions

- SMC-R with RoCE
  - HCD (Hardware Configuration Definition)
    - PCHID (Physical Channel ID) – 3 digit hex value of physical slot location
    - PFID (PCIe Function ID) – unique 3 digit hex value per PCHID
    - VF (Virtual Function) – unique 2 digit decimal value per PCHID
    - Different PFID and VF per TCP/IP stack/port
    - PNET (Physical Network)(PNET on RoCE and OSA must match)
    - Port Number - **SMCv2 port is defined in HCD** and SMCv1 port is defined in PROFILE.TCPIP
  - PROFILE.TCPIP PORT 4791 UDP
    - Required for **SMCv2** traffic to be routed. Open port on Routers.
  - PROFILE.TCPIP GLOBALCONFIG SMCR
    - PFID MTU (PCIe Function ID)(Maximum Transmission Unit) – **SMCv1**
  - PROFILE.TCPIP GLOBALCONFIG SMCEID/ENDSMCEID
    - UEIDs (User-defined Enterprise IDs) – **SMCv2**
- SMC-D with ISM
  - HCD (Hardware Configuration Definition)
    - VCHID (Virtual Channel ID)
    - PNET – **required for SMCv1 but not recommended for SMCv2**
  - PROFILE.TCPIP GLOBALCONFIG SMCD
    - SYSTEMEID – causes System EID (SEID) to be generated – **SMC-Dv2 requires SEID or UEID**
  - PROFILE.TCPIP GLOBALCONFIG SMCEID/ENDSMCEID
    - UEIDs (User-defined Enterprise IDs) – **SMC-Dv2 requires SEID or UEID**

Other Optional PROFILE.TCPIP Definitions

- SMC-R with RoCE and SMC-D with ISM
  - PROFILE.TCPIP GLOBALCONFIG SMCD and SMCR
    - FIXEDMEMORY - Specifies the maximum amount of 64-bit storage that the stack can use for the send and receive buffers that are required for SMC-D and SMC-R communications. 256 megabytes default.
    - TCPKEEPMININTERVAL - This interval specifies the minimum interval that TCP keepalive packets are sent on the TCP path of an SMC-D or SMC-R link. 300 seconds default.
  - PROFILE.TCPIP GLOBALCONFIG
    - AUTOCACHE - Specifies whether this stack caches unsuccessful attempts to use SMC communication. AUTOCACHE is the default.
    - AUTOSMC - Specifies whether this stack monitors inbound TCP connections to dynamically determine whether SMC is beneficial for a local TCP server application. AUTOSMC is the default.
    - SMCPERMIT/ENDSMCPERMIT - Specifies the SMC filter that allows SMC negotiation with peers within the listed TCP/IP address(es)/subnet(s).
    - SMCEXCLUDE/ENDSMCEXCLUDE - Specifies the SMC filter that prevents SMC negotiation with peers within the listed TCP/IP address(es)/subnet(s).

SMC Connection Eligibility

SMC-R and RoCE

- Both SMC-Rv1 and RoCEv1 Client and Server must have:
  - Profile GLOBALCONFIG SMCR with PFID
  - Direct network access to the same physical Layer-2 network (same physical LAN)
  - Direct network interface configured with the same IP Subnet (for IPv4 or IP mask for IPv6)
  - and VLAN ID (if VLAN is applicable)
  - Both hosts must also have access to an RNIC and be enabled for SMC-R.
- Both SMC-Rv2 and RoCEv2 Client and Server must have:
  - Profile GLOBALCONFIG SMCR
  - Profile GLOBALCONFIG SMCGLOBAL SMCEID/ENDSMCEID
  - Network interface configured with PFID and SMCIPADDR
  - Both hosts must also have access to an RNIC and be enabled for SMC-R.

SMC-D and ISM

- Both SMCv1 and ISMv1 Client and Server must:
  - Must have access to a common ISM VCHID and PNET
  - Must be be enabled for SMCD in the Profile
  - Must not be associated with a TCP/IP connection that requires IPsec encryption
- Both SMCv2 and ISMv2 Client and Server must:
  - Be updated to support SMC-Dv2 (z/OS 2.4 with PTFs)
  - Execute on the same IBM Z System CPC that supports ISMv2 (z15+)
  - Must have access to a common ISM VCHID
  - Must be SMCv2 enabled, defined with the same EID (user defined (UEID) or system defined(SEID))
  - Must not be associated with a TCP/IP connection that requires IPsec encryption

More Information

Links

- SMC-R and RoCE FAQ (Frequently Asked Questions)
  - https://www.ibm.com/support/pages/node/6414391
- SMC-D and ISM FAQ (Frequently Asked Questions)
  - https://www.ibm.com/support/pages/node/6414395
- SMC-AT Overview
  - http://ftpmirror.your.org/pub/misc/ftp.software.ibm.com/software/os/systemz/pdf/SMC_Applicability_Tool_Overview_03-10-16.pdf
- SMC-Rv2 Configuration
  - https://www.ibm.com/support/pages/configuring-shared-memory-communications-version-over-rdma-version-2-smc-rv2-your-zos-environment
- SMC-Rv2 Profile Configuration using the z/OSMF Network Configuration Assistant Tool
  - https://www.youtube.com/watch?v=ypzt1EttRCo
- SMC-Dv2 Profile Configuration using the z/OSMF Network Configuration Assistant Tool
  - https://www.ibm.com/support/pages/node/6985517

The End