

Decide for the best network between z/VSE, z/VM and Linux on z Systems

Wilhelm Mild Executive IT Architect IBM Germany



© Copyright IBM Corporation 2015

Technical University/Symposia materials may not be reproduced in whole or in part without the prior written permission of IBM.



Networking with z13 and z Systems





IBM z Systems Networking - Operational Diagram



IBM z Systems Networking - Operational Diagram





z Systems networking – Operational Diagram

All components are configured, managed, and serviced the same way. The Unified Resource Manager enables end-to-end management capabilities for zBX and CPC. An ensemble consists of these networks:

- Customer provided management network
- Customer provided data network (OSD)
- Intranode management network (OSM)
- Intraensemble ensemble data network (OSX)

- zEC12-BPH Bulk Power hub
- z13 SCH System Control Hub





Patterns in Network Architecture for z13 & IBM z Systems



Note:

- Network Connections in z13 and IBM z Systems are mainly realized via OSA (Open System Adapter) cards
- Different types of network possible with the same OSA card
- New: RoCE cards for ,remote Hipersockets' network communications

OSA Express communication characteristics

- 'Integrated Power computer' with network card
- Shared between up to 640 OSA devices
- Three device numbers (ccw devices) per OSA device:
 - Read device (control data ← OSA)
 - Write device (control data \rightarrow OSA)
 - Data device (network traffic)
- OSA Address Table: which OS image has which IP address
- Network traffic OS (i.e. Linux) ↔ OSA
 - IP (layer3 mode)
 - Ethernet / data link layer level (layer2 mode)
 - OSA handles ARP- (Address Resolution Protocol)
 - One MAC address for all stacks

Note:

- Communication is asynchronous from an application perspective
- Communication is at OSA card clock speed (typically lower than Hipersockets)



Network types OSA Express 5s, OSA Express 4s, Hipersockets

March Charter Bag part of a roome guar March Charter March Char

its

- OSA Express supports various features such as:
 - 10 Gigabit Ethernet
 - Gigabit Ethernet
 - 1000BASE-T Ethernet

CHPID types (OSA and internal Hipersockets)

- OSCOSA-ICC (for emulation of TN3270E and non-SNA DFT 3270, IPL CPCs, and LPARs, OS system console operations)
- OSD Queue Direct Input/Output (QDIO) architecture
- OSE non-QDIO Mode (OSA-2, for SNA/APPN connections)
- OSNOSA-Express for NCP: Appears to z/OS and z/VSE as a device-

supporting channel data link control (CDLC) protocol.

- IQD The HiperSockets hardware device is represented by the IQD CHPID and associated subchannel devices. All LPARs that use the same IQD CHPID have internal connectivity and communicate using HiperSockets
- OSX OSA-Express for zBX. Provides connectivity and access control to the Intra-Ensemble Data Network (IEDN) from z196 and z114 to Unified Resource Manager functions.
- OSM OSA-Express for z Systems Ensemble management. OSM ports connect to the Intranode Management Network (INMN)
- IQDX Hipersockets Bridge to zBX

OSA-Express4S 1000BASE-T for the PCIe I/O Drawer

- OSA-Express4S 1000BASE-T Ethernet
 - Exclusive to zEC12
 - Auto-negotiation to 10, 100, 1000 Mbps
 - CHPID types
 - OSC, OSD, OSE, OSM, OSN
- Two ports
 - •1 PCHID/CHPID
- Operates at "line speed"
- Category 5 or Category 6 copper cable
- RJ-45 connector



Connector = RJ-45

Mode	CHPID	Description
OSA-ICC	OSC	TN3270E, non-SNA DFT, IPL CPCs, and LPARs, OS system console operations
QDIO	OSD	TCP/IP traffic when Layer 3, Protocol-independent when Layer 2
Non-QDIO	OSE	TCP/IP and/or SNA/APPN/HPR traffic
Unified Resource Manager	OSM	Connectivity to intranode management network (INMN) from zEC12 to Unified Resource Manager functions
OSA for NCP (LP-to-LP)	OSN	NCPs running under IBM Communication Controller for Linux (CDLC)

OSA-Express5S 1000BASE-T Ethernet Feature - PCIe I/O Drawer

- PCI-e form factor feature supported by PCIe I/O drawer
 - One two-port CHPID per feature
 Half the density of the OSA-Express3 version
- Small form factor pluggable (SFP+) transceivers
 - Concurrent repair/replace action for each SFP
- Exclusively Supports: Auto-negotiation to 100 or 1000 Mbps and full duplex only on Category 5 or better copper



- RJ-45 connector
- Operates at "line speed"
- CHPID TYPE Support:

Mode	TYPE	Description
OSA-ICC	osc	TN3270E, non-SNA DFT, OS system console operations
QDIO	OSD	TCP/IP traffic when Layer 3, Protocol-independent when Layer 2
Non-QDIO	OSE	TCP/IP and/or SNA/APPN/HPR traffic
Unified Resource Manager	OSM	Connectivity to intranode management network (INMN)
OSA for NCP (LP-to-LP)	OSN	NCPs running under IBM Communication Controller for Linux (CCL)

Note: OSA-Express5S feature are designed to have the same performance and to require the same software support as equivalent OSA-Express4S features.

© 2015 IBM Corporation



OSA-Express5S Fiber Optic Features

• 10 Gigabit Ethernet (10 GbE)

- -CHPID types: OSD, OSX
- -Single mode (LR) or multimode (SR) fiber
- -One port of LR or one port of SR
 - 1 PCHID/CHPID
- -Small form factor pluggable (SFP+) optics
 - Concurrent repair/replace action for each SFP
- -LC duplex

Gigabit Ethernet (1 GbE)

- -CHPID types: OSD (OSN not supported)
- -Single mode (LX) or multimode (SX) fiber
- -Two ports of LX or two ports of SX
 - 1 PCHID/CHPID
- -Small form factor pluggable (SFP+) optics
 - Concurrent repair/replace action for each SFP
- -LC Duplex

Note: OSA-Express5S features are designed to have the same performance and to require the same software support as equivalent OSA-Express4S features.





Summary: OSA-Express CHPID types to control operation

CHPID type	Purpose / Traffic	Operating Systems
OSC 1000BASE-T z13, zEC12, zBC12, z196, z114, z10, z9	OSA-Integrated Console Controller (OSA-ICC) Supports TN3270E, non-SNA DFT to IPL CPCs & LPs	z/OS, z/VM z/VSE
OSD All OSA features z13, zEC12, zBC12, z196, z114, z10, z9	Supports Queue Direct Input/Output (QDIO) architecture TCP/IP traffic when Layer 3 (uses IP address) Protocol-independent when Layer 2 (uses MAC address)	z/OS, z/VM z/VSE, z/TPF Linux on z Systems
OSE 1000BASE-T z13, zEC12, zBC12, z196, z114, z10, z9	Non-QDIO; for SNA/APPN/HPR traffic and TCP/IP "passthru" traffic	z/OS, z/VM z/VSE
OSM 1000BASE-T z13, zEC12, zBC12, z196, z114	OSA-Express for Unified Resource Manager Connectivity to intranode management network (INMN) from z13, zEC12, z196, or z114 to Unified Resource Manager functions	z/OS, z/VM* Linux on z Systems
OSN GbE, 1000BASE-T z13, zEC12, zBC12, z196, z114, z10, z9 No OSN support for OSA-Express4S and 5S GbE	OSA-Express for NCP Appears to OS as a device supporting CDLC protocol Enables Network Control Program (NCP) channel-related functions Provides LP-to-LP connectivity OS to IBM Communication Controller for Linux (CCL)	z/OS, z/VM z/VSE, z/TPF Linux on z Systems
OSX 10 GbE z13, zEC12, zBC12, z196, z114	OSA-Express for zBX Connectivity and access control to intraensemble data network (IEDN) from z13, zEC12, z196, or z114 to zBX	z/OS, z/VM*, z/VSE 5.1, Linux on z Systems

*CHPIDs OSX and OSM supported by z/VM V6.2 and z/VM V6.3 to define, modify, and delete OSX CHPID types when z/VM is the controlling LPAR for dynamic I/O



Patterns in Network Architecture for z13 & IBM z Systems



• via **VSWITCH** between guest Systems





Network between LPARs on z Systems - Hipersockets network

- Connectivity within a central processor complex without physical cabling
- Licensed Internal Code (LIC) function
 - -emulating Data Link Layer of an OSA-device (internal LAN)
- Internal Queued Input/Output (IQDIO) at memory speed
- 4 different MTU sizes supported (Max. Transmission Unit):
 - -8KB, 16KB, 32KB, 56KB
- Support of
 - -Broadcast, VLAN, IPv6, Layer2 (starting with z10)
- UP to 32 different, isolated networks
- Synchronous communication
- Bi-directional CPU speed communication
- Note: Hipersockets needs n/2 defined

buffers which should be at 32-128 for a good performance



New: Hipersockets Completion Queue between LPARs

Function for z Systems between LPARs

- •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.

- **HiperSockets** transfers data synchronously if possible and asynchronously if necessary.
- Ultra-low latency with more tolerance for traffic peeks.
- Requires z/VSE 5.1 (+PTFs) and newer
- Requires z/OS V1.13 or later software.





- → network

18

For z/VSE 5.1 + PTFs: Linux Fast Path in an LPAR environment

Faster communication between z/VSE and Linux applications

→ Exploits the HiperSockets Completion-Queue support of IBM zEnterprise (z196, zEC12, z13)





For z/VSE 5.1 + z/VM 6.1: z/VM IP Assist (VIA)

With z/VM IP Assist (VIA), no Linux is needed to utilize the LFP advantage



Recommendations



• Hipersockets and Hipersockets Completion Queue between LPARs

- for similar CPU / LPAR power, Hipersockets is recommended
- to increase stability and network speed use between 32- 64
 Hipersockets buffers and adjust MTU size

Shared OSA network

 for LPARs with limited CPU speed (below 1/2 of max CPU speed), the shared OSA network might be faster

IBM z Systems Network additional alternative with RoCE





10GbE RoCE Express Feature

Native PCIe networking card

- 10 Gigabit remote direct memory access (RDMA) capable network card
- Uses Infiniband RDMA over Converged Ethernet (RoCE) specification
- Up to 16 10GbE RoCE Express adapters per machine
- Reduced latency and lower CPU overhead
- Supports point-to-point connections and switch connection with an enterprise-class 10 GbE switch

Hardware & Software requirements

- IBM zEC12 (w/ appropriate updates), zBC12 (w/ appropriate updates), or z13
- z/VM 6.3 with APAR VM65417 Available
 - System Config option disabled by default
 - Required millicode fixes must be applied prior to enabling in system config
- z/OS 1.12, z/OS 1.13, z/OS 2.1 with APAR OA43256
- Linux support is available upstream and as tech preview in SLES12 / RHEL7
- Fulfills 2013 Statement of Direction









RoCE Express with the SMC-R Protocol RDMA Support for z/OS utilizing 10GbE



- SMC-R is a new protocol that allows existing TCP applications to transparently benefit from RDMA for transferring data
 - Initial deployment limited to z/OS<->z/OS communications, but goal to expand exploitation to additional operating systems and possibly appliances/accelerators
- InfiniBand Trade Association (IBTA) standardized RDMA over Converged Ethernet (RoCE) in April 2010
 - Management of RoCE fabrics can now readily be integrated with existing datacenter Ethernet fabrics (prior RDMA networks used InfiniBand)

Dynamic Transition from TCP to SMC-R



TCP connection transitions to SMC-R allowing application data to be exchanged using RDMA

IBM z Systems Network alternatives between z/VM LPARs <



z/VM Multi-zone Network VSWITCH (red - physical isolation



With 2 VSWITCHes, 3 VLANs, and a multi-domain firewall

HiperSockets

Must match LAN type

Virtual Network Interface Card (vNIC)

A simulated network adapter in z/VM

OSA-Express QDIO

- Usually 3 devices per NIC
- Provides access to Guest LAN or VSWITCH
- Created by directory or CP DEFINE NIC

z/VM guests (Linux, z/OS, z/VSE,...)

DEF NIC 600 TYPE QDIO COUPLE 600 SYSTEM VSWITCH1



Guest LAN or virtual switch





Virtual LAN (VLAN)

- VLAN allows a physical network to be divided administratively into separate logical network
- These logical networks operate as if they are physically independent of each other
- A VLAN tag is inserted into the Link Layer Header
 - 3 bit priority: an be used to prioritize different classes of traffic (voice, video, data)
 - 12 bit VLAN ID: specifies the VLAN to which the frame belongs





Source: Wikipedia: http://en.wikipedia.org/wiki/File:TCPIP_802.1Q.jpg



Virtual LAN (VLAN) – Trunc Port / Access Port

Switches have different types of ports

- Access Port
 - Not VLAN-aware
 - Un-tagged frames
- Trunc Port
 - VLAN-aware
 - Tagged frames







VSWITCH Definition for multiple VLANs

Prior to z/VM 6.2 if a Guest required access to multiple VLANs there Were two ways to define this connectivity.

- 1. Have the Guest connect to multiple VSWITCHes.
 - Each VSWITCH would provide the Guest an ACCESS Port
 - Each vNIC would have a unique vMAC.
- 2. Have the Guest GRANT to a Vswitch with a TRUNK Port.
 - The Guest would load the 8021Q module
 - Configure a VLAN with VCONFIG there is one vMAC.

For ease-of-use it would be desirable to have this connectivity with one VSWITCH where the Guest vNICs are GRANTed as ACCESS Ports

• Each vNIC would have a unique vMAC.



DEFINE VSWITCH VSWT1 RDEV 2A00 ETHERNET CONTROLLER * VLAN 200 NAT 1 DEFINE VSWITCH VSWT2 REDV 2A03 ETHERNET CONTROLLER * VLAN 300 NAT 1 SET VSWITCH VSWT1 GRANT GUEST SET VSWITCH VSWT2 GRANT GUEST



DEFINE VSWITCH VSWT1 RDEV 2A00 ETHERNET CONTROLLER * VLAN AWARE PORTBASED NAT 1 SET VSWITCH VSWT1 PORTNUMBER 20 GRANT GUEST VLAN 200 SET VSWITCH VSWT1 PORTNUMBER 21 GRANT GUEST VLAN 300



Guest LAN vs. Virtual Switch Guest LAN Virtual Switch Guest Guest Guest Guest Guest Ethernet LAN

- Virtual router is required
- Different subnets
- External router awareness
- Guest-managed failover

- No virtual router
- Same subnets
- Transparent bridge
- CP-managed failover

Channel Bonding in Linux on z - for higher bandwidth

- The Linux bonding driver provides a
 - -method for aggregating multiple
 - -network interfaces into a single,
 - -logical "bonded" interface
- Provides failover and/or load balancing functionality
- Better performance depending on bonding mode
- Requires layer2 devices
- Further information
 - -http://sourceforge.net/projects/bonding





Network bandwidth enhancement and Automated Failover



Resource Virtualization: OSA Channel Bonding in Linux



- Linux bonding driver enslaves multiple OSA connections to create a single logical network interface card (NIC)
- Detects loss of NIC connectivity and automatically fails over to surviving NIC
- Active/backup & aggregation modes
- Separately configured for each Linux

Network Virtualization:

z/VM Port aggregation



- z/VM VSWITCH enslaves multiple OSA connections. Creates virtual NICs for each Linux guest
- Detects loss of physical NIC connectivity and automatically fails over to surviving NIC
- Active/backup & aggregation modes
- Centralized configuration benefits all guests



z/VM – Multi-VSWITCH Link Aggregation

- Makes it possible to do Link Aggregation with VSwitches without the requirement for dedicated OSAs
- Allows a port group of OSA-Express features to span VSwitches within a single or multiple z/VM systems.
 - Cannot be shared with non-z/VM logical partitions or z/VM systems without support
- APARs VM65583 and PI21053 for z/VM 6.3 only
 PTFs planned to be available June 26, 2015
- Only available on z13
 - Requires OSA enhancements introduced with the z13
- Allows better consolidation and availability while improving TCO





The Global virtual switch

Collection of virtual switches that share the same networking characteristics

- Same name
- Reside on systems that are part of the same IVL domain
 - In order to create a global virtual switch "member"
- The system's IVL virtual switch must be defined and its UPLINK port connected
- The virtual switch's attributes must match any existing virtual switches with the same name in the IVL domain
 - Created when first member is defined, deleted when last member is detached

DEFINE VSWITCH RICK TYPE QDIO GLOBAL ETHERNET UPLINK GROUP MARY



Shared port group

Set of OSA-Express features configured to be shared by one or more global virtual switches and global virtual switch members

- Can only be used with global virtual switches and members
 - Provides connectivity across multiple virtual switches and multiple z/VM systems
 - Propagated to all systems in the same IVL domain
 - Single point of control for all virtual switches using the shared port group

SET PORT GROUP MARY LACP ACTIVE SHARED SET PORT GROUP MARY JOIN 400 500

HIPERSOCKET Bridge – Network HA Overview

A HiperSockets channel is an intra-CEC communications path.

The HiperSockets Bridge provides inter-CEC Hipersockets LAN connection to combine many HiperSockets LANs into a Layer 2 broadcast domain. An ideal function for the SSI environment.

The HiperSockets Bridge function is available in z/VM 6.2 with a z114 or z196 or zEC12 processor. Must be sure to have z/VM and processor maintenance levels..







VSWITCH Topology

A typical Vswitch topology for multiple CECs. Active and Backup Uplink Ports to redundant Ethernet switches.





VSwitch – With Hipersocket Bridge

z/VM6.2 LPAR - CEC1



Patterns in Network Architecture for z13 & IBM z Systems



- use VSWITCH between guest Systems advantage in DR solutions
- use IUCV for special network Connections (terminal server)

IBM z Systems Network alternatives from z/VM Guests





Linux on z network options and interfaces



z/VM IUCV – Inter-User Communication Vehicle

IBM Very Construction Very Con

- A communications facility inside of z/VM
- A program running in a z/VM guest communicating
 - With another virtual machine within same z/VM
 - Running Linux on z/VM
 - Running other Operating System
 - With a CP system service
 - With itself
- IUCV interrupt control functions to
 - -establish and remove communication paths
 - -transfer messages

A Terminal Server environment using z/VM IUCV



TCP/IP is enabled only in the Linux running the Terminal Server.

HS versus IUCV communication in a z/VM-mode LPAR - Supported by z/VM, Linux and z/VSE (Linux Fast Path)

Faster communication between applications





Summary network Recommendations

- Which connectivity to use with z13 or z Systems:
 - External connectivity:
 - LPAR: 10 GbE cards
 - z/VM: VSWITCH with 10GbE card(s) attached
 - For maximum throughput and minimal CPU utilization attach OSA directly to Linux guest

- Internal connectivity:

- LPAR:
 - HiperSockets for LPAR-LPAR communication (if fullsize CPUs used)
 - Hipersockest Completion Queue (CQ) helpfull for asynchronous communication
 - Hipersocket Bridge for SSI and multi CEC environments
 - Shared OSA for capped CPUs in one of the LPARs
- z/VM:
 - VSWITCH for guest-guest communication
 - VLANs for network isolation
 - IUCV for inter guest communication and Linux access without TCP/IP
- For high network workload and **cloud use**:
 - z/VM VSWITCH with link aggregation
 - OSA channel bonding

Both include high availability and automatic failover.

z13 and z Systems network value points

- Network Simplification ("Network in a Box")
- Central point of Management
- Single physical network and high virtualization with isolation
- Reduced network overhead across CPC with RoCE
- Reduced network path length and reduced number of hops
- Secured internal communications
- Physical security (internal / dedicated network equipment)
- Logical security (controlled access)
- High Availability network
- Redundant network hardware
- Logical failover
- Unique z Systems QoS
- Isolated / dedicated equipment with integrated HA
- Special purpose dedicated
 - -data network & OSA-Express
 - -potential for reduced network encryption and HW encryption support



Insurance Company Consolidated 292 Servers to z Systems





Data is based on real client opportunity and on internal standardized costing tools and methodologies.

Client results will vary by types of workloads, technology level of consolidated servers, utilization factor, and other implementation requirements. Savings will vary by client.



Additional Documentation

IBM System z Networking

http://www.ibm.com/systems/z/hardware/networking/

IBM System z Connectivity Handbook

http://www.redbooks.ibm.com/redpieces/abstracts/sg245444.html

VM Networking

http://www.vm.ibm.com/virtualnetwork/

z/VSE Networking

Redbook SG24-8091: http://www.redbooks.ibm.com/abstracts/sg248091.html?Open

Linux on System z documentation

http://www.ibm.com/developerworks/linux/linux390/documentation_dev.html

Linux on System z - Tuning Hints & Tips

http://www.ibm.com/developerworks/linux/linux390/perf/index.html

Linux on System z on developerWorks

http://www.ibm.com/developerworks/linux/linux390

Linux on System z – Downloads

http://www.ibm.com/developerworks/linux/linux390/development_recommended.html



Questions?



