

Networking options on zEnterprise

Wilhelm Mild IT Architect mildw@de.ibm.com





Trademarks

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

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market.

Those trademarks followed by (B) are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

*, AS/400®, e business(logo)®, DBE, ESCO, eServer, FICON, IBM®, IBM (logo)®, iSeries®, MVS, OS/390®, pSeries®, RS/6000®, S/30, VM/ESA®, VSE/ESA, WebSphere®, xSeries®, z/OS®, zSeries®, z/VM®, System i, System i5, System p, System p5, System x, System z, System z9®, BladeCenter®

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. 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.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. 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.

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.

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

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

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.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective 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



zEnterprise Positioning

- Networking Options on zEnterprise
- Wrap-up



The Data Center Challenge - Controlling IT complexity and cost while maintaining daily operations

- An Integrated system of multiple architectures for optimizing the deployment of multi-tier workloads
- Creating a single point of control for management and administration to reduce operational overhead by up to 80%, including:
 - Power and Facilities
 - Labor
 - Software License

zEnterprise

- Lowers cost of acquisition by up to 56%
- Reduces cost of ownership by up to 55%*



A strategic systems platform....

Helping to free up resources for critical projects and establish a base for the future

[•] Based on IBM analysis of a large Financial Services company Datacenter. See details on ibm.com/systems/zenterprise/ Deployment configurations based on IBM studies and will vary based on workload characteristics. Price calculations based on publicly available US list prices, prices will vary by country.



IBM zEnterprise System – one for everything !

Re-write the rulebook and set new standards for business-centric IT with IBM System z, to be the world's premier workload-optimized platform for enterprise applications.



Our Vision:

An IT environment driven with one centralized System
- IBM zEnterprise System -

Deliver the best of all worlds - Mainframe, UNIX, x86 and single function processors - integrated in a single system for ultimate flexibility and simplicity to optimize service, risk, and cost across multiple heterogeneous workloads.



z/VSE 5 Support for IBM zEnterprise - IEDN to zBX





zEnterprise Network characteristics

- Network Simplification ("Network in a Box")
- Single physical network and zBX "package" (physical network integration)
- Central point of Management (Unified Resource Manager via the HMC/SE)
- Reduced network path length; reduced number of hops
- Secure communications
- Physical security (internal / dedicated network equipment)
- Logical security (controlled access)
- Network Virtualization and Isolation
- High Availability network
- Redundant network hardware
- Logical failover
- Unique System z QoS
- Isolated / dedicated
 - equipment
- Special purpose dedicated
 - data network & OSA-Express
 - potential for reduced network encryption



Agenda

zEnterprise Positioning

Networking options on zEnterprise

Wrap-up





zEnterprise – What are the INMN, IEDN and Customer networks



Ensemble private Networks in zEnterprise





Intranode management network (INMN)

- 1000Base-T OSA-Express3 (copper) --- QDIO (CHPID Type OSM)
 - Cables are 3.2 meters long from OSM to BPH in CEC and 26 meters from BPH to TOR
- HMC security is implemented with standard practices PLUS additional security mechanisms:
 - Isolated IPv6 network with "link-local" addresses only; authentication and authorization and access control, etc.

Intraensemble data network (IEDN)

- 10 Gigabit OSA-Express3 --- QDIO (CHPID Type OSX) Cables are maximum of 26 meters long to TOR & 10km long-range
- Security is implemented with standard practices PLUS additional security mechanisms: VLAN ID enforcement, access control, authentication, authorization, application security, routing table restrictions, IP Filtering, etc.
- Networks can be further isolated using VLAN and VMAC segmentation of the network connections



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 \leftarrow OSA)
 - Write device (control data \rightarrow OSA)
 - Data device (network traffic)
- OSA Address Table: which OS image has which IP address
- Network traffic Linux ↔ OSA, either
 - IP (layer3 mode)
 - One MAC address for all stacks
 - OSA handles ARP- (Address Resolution Protocol)
 - Ethernet / data link layer level (layer2 mode)
- Communication is asynchronous –from an application perspective
- Communication is at OSA card clock speed (lower than Hipersockets)





OSA Express – Network types

OSA Express 4s, OSA Express 3, OSA Express 2

• OSA Express supports various features such as:

- 10 Gigabit Ethernet
- Gigabit Ethernet
- 1000BASE-T Ethernet

CHPID types

- **OSC** OSA-ICC (for emulation of TN3270E and non-SNA DFT 3270)
- **OSD** Queue Direct Input/Output (QDIO) architecture
- **OSE** non-QDIO Mode (OSA-2, for SNA/APPN connections)
- **OSN** OSA-Express for NCP: Appears to z/VSE as a device-supporting channel data link control (CDLC) protocol.
- 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.
- OSA-Express for zEnterprise Ensemble management. OSM ports connect to the Intranode Management Network (INMN) over which the Unified Resource Manager defines, accesses, and manages the members of the ensemble.





System z Hipersockets

- 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: -8KB, 16KB, 32KB, 56KB
- Support of
 - -Broadcast, VLAN, IPv6, Layer2 (starting with z10)
- UP to 32 different, isolated networks
- Synchronous communication
- CPU speed communication



© 2012 IBM Internal



HiperSockets – internal network in zEnterprise

• "Network within the box" functionality

- allows high speed any-to-any connectivity among operating systems
- without requiring any physical cabling

CHPID type IQD

- Uses the QDIO (Queue Direct I/O) architecture
- For an HiperSockets adapter, you need 3 devices
 - A read device
 - A write device
 - A datapath device
- Add the devices in the IPL procedure as device type OSAX with mode 01:

• ADD cuu1-cuu3, OSAX, 01

- Frame size is defined via OS parameter:
 - OS=00 (default): 16K (MTU=8K)
 - OS=40 24K (MTU=16K)
 - OS=80 40K (MTU=32K)
 - OS=C0 64K (MTU=56K)





Hipersockets and OSA Express Features

Feature	z196, z114	System z10	System z9	zSeries
HiperSockets	Yes	Yes	Yes (layer3)	Yes (layer3)
OSA-Express4	Gigabit Ethernet 10 Gigabit Eth.	Not supported	Not supported	Not supported
OSA-Express3	Gigabit Ethernet 10 Gigabit Eth. 1000Base-T Eth.	Gigabit Ethernet 10 Gigabit Eth. 1000Base-T Eth.	Not supported	Not supported
OSA-Express2	Gigabit Ethernet 1000Base-T Eth.	Gigabit Ethernet 10 Gigabit Eth. 1000Base-T Eth.	Gigabit Ethernet 10 Gigabit Eth. 1000Base-T Eth.	Not supported
OSA-Express	Not supported	Not supported	Fast Ethernet Gigabit Ethernet 1000Base-T Eth.	Fast Ethernet Gigabit Ethernet 1000Base-T Eth. Token Ring ATM



Network Layer 2 vs. Layer 3

OSI Model	TCP/IF	TCP/IP Model	
	Layer	Protocol	
7 Application			
6 Presentation	Application	FTP, HTTP	
5 Session			
4 Transport	Transport	TCP, UDP	
3 Network	Internet	IP, ICMP, IPV6	•
2 Data Link	Network Access	Ethernet	ARP
1 Physical			

ARP – Address Resolution Protocol





z/VSE support Layer 2 and Layer 3 Mode

• Layer 2:

Supported by IPv6/VSE product (BSI) with IPv6
 OSA Express adapter (OSD, OSX) only, no HiperSockets

• Layer 3:

- Supported by IPv6/VSE product (BSI) with IPv4 and IPv6
- Supported by TCP/IP for VSE product (CSI) with IPv4

VLAN

- z/VSE provides VLAN support for OSA Express (CHPID type OSD and OSX) and HiperSockets devices
- In a Layer 3 configuration, VLANs can be transparently used by IPv6/VSE and TCP/IP for VSE/ESA
- VLANs for OSA-Express (CHPID type OSD and OSX) devices in a Layer 2 configuration can be defined using the IPv6/VSE product and IPv6 traffic



• VSWITCH:

- z/VM allows to define VSWITCH in Layer 2 or layer 3 mode
- z/VSE V4.2 and 4.3:
 - Supports Layer 3 VSWITCH (IPv4 only)
- z/VSE V5.1:
 - Supports Layer 2 VSWITCH (IPv4 and IPv6)
 - Supports Layer 3 VSWITCH (IPv4 only)



Intra-Ensemble Data Network (IEDN) support

OSA-Express for zBX (CHPID type OSX)

 Provides connectivity and access control to the Intra-Ensemble Data Network (IEDN) from zEnterprise 196 and 114 to Unified Resource Manager functions

• An Intra-Ensemble Data Network (IEDN) provides connectivity between:

- A zEnterprise CEC (Central Electrical Complex) and System z Blade Center Extensions (zBXs)
- Two or more zEnterprise CECs
- z/VSE supports the IEDN network of a zEnterprise 196 or 114
 - z/VSE V4.2, V4.3 and V5.1:
 - z/VM VSWITCH and OSDSIM mode in a z/VM 6.1 guest environment
 z/VSE V5.1:
 - OSA Express for zBX devices either in an LPAR or z/VM guest environment with dedicated OSAX devices
 - This requires VLAN support





Virtual Network Interface Card (NIC)

- A simulated network adapter
 - OSA-Express QDIO
 - HiperSockets
 - Must match LAN type
- Usually 3 devices per NIC
- Provides access to Guest LAN or VSWITCH
- Created by directory or *CP DEFINE NIC*



DEF NIC 600 TYPE QDIO COUPLE 600 SYSTEM VSWITCH1



Guest LAN or virtual switch



Virtual LAN (VLAN) Support

- IEEE Standard 802.1Q
- Reduce broadcast traffic
- Divide LANs logically into subnets to optimize bandwidth utilization
- Network devices supporting VLAN:
 - real OSA card, HiperSockets, z/VM GuestLAN, z/VM VSWITCH





Virtual Switch

- Create simulated Layer 2 or Layer 3 switch device
- VM access control and VLAN authorization
- Create ports
- Connect NIC to Virtual Switch (LAN Segment)
- Full MAC address management
 - generation and assignment
- 1 n VSWITCHs per z/VM Image

Create VSWITCH from PRIVCLASS B User ID

DEF VSWITCH VSWITCH1 ETHERNET SET VSWITCH VSWITCH1 GRANT {user ID} From Linux Virtual Machines

DEF NIC 600 TYPE QDIO COUPLE 600 SYSTEM VSWITCH1



Guest LAN vs. Virtual Switch



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

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

Multi-zone Network VSWITCH (red zone physical isolation)



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



z/VM Guest LANs vs. VSWITCH

z/VM Guest LAN

- A simulated LAN segment
- Types:
 - QDIO:
 - IP: layer3, IPv4 only
 - Ethernet: layer2, IPv4 or IPv6
 - HiperSockets: IPv4 and IPv6 (layer3)
- No physical connection
- Unrestricted / restricted
- Persistent / transient
- As many as you want

z/VM VSWITCH

- Special purpose GuestLAN
- Type QDIO only
 - IP: layer3, IPv4 only
 - Ethernet: layer2, IPv4 or IPv6
- Built-in IEEE 802.1q bridge to outside network
- 1-8 associated OSA-connections
- Restricted
- Persistent
- Failover and Link Aggregation
- Port Isolation
- Virtual Network Devices NICs (Virtual Network Interface Cards)
- Defined by directory or CP DEFINE NIC command
 - Type QDIO or HIPERS (must match LAN type)
- The only thing visible to Linux



Channel Bonding

- 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 Interface Redundancy 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 VSWITCH



- 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



Linux Network alternatives





Agenda

- zEnterprise Positioning
- Networking options on zEnterprise
 - Special network option Linux with z/VSE
- Wrap-up





Network alternatives





Linux Fast Path in a z/VM-mode LPAR

- Supported by z/VM, Linux and z/VSE 4.3 + 5.1 Faster communication between z/VSE and Linux applications





z/VSE z/VM IP Assist (VIA) - Supported by z/VSE V5 + z/VM V6.2

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





HiperSocket VSWITCH Integration with zEnterprise IEDN Available: April 13, 2012



- Built-in failover and failback
- Bridge new IQDX chpid to OSX chpid
- Also works for IQD to OSD

- Same or different LPAR
- One active bridge per CEC
- PMTU simulation



Linux Fast Path in an LPAR





Linux for System z Network Device Drivers

- LCS LAN Channel Station
- CTC(M) (stabilized)
- NETIUCV (stabilized)
- QETH



LAN Channel Station (LCS) Device Driver

- Supports:
 - OSA Express (in non-QDIO mode)
 - (HighSpeed TokenRing)
 - (ATM (running Ethernet LAN Emulation))
- May be preferred instead of QETH for security reasons
 - Administrator defines OSA Address Table \rightarrow restricted access, whereas with QETH each Linux registers its own IP address
- But: performance is inferior to QETH's performance!!!



Change Required - for CTC and IUCV users

- CTC = Channel-to-Channel connection
- IUCV = Inter User Communication Vehicle
- CTC(M) and NETIUCV device drivers are deprecated (Linux 2.6+)
- Device drivers are still available for backward compatibility
- Please consider migration
 - Virtual CTC and IUCV (under z/VM) ==> guest LAN HiperSocket or guest LAN type QDIO
 - CTC inside a CEC
 ==> Hipersockets
 - CTC

==> OSA-Express (QDIO)



QETH Device Driver

- Supports
 - OSA Express / OSA Express2 / OSA Express3 OSD type (=QDIO)
 - Fast/Giga/10GBit Ethernet (fiber infrastructure)
 - 1000Base-T Ethernet (copper infrastructure)
 - System z HiperSockets
 - z/VM
 - GuestLAN Type QDIO (layer2 / layer3), Type Hiper
 - z/VM VSWITCH (layer2 / layer3)
 - IPv4, IPv6, VLAN, VIPA, Proxy ARP, IP Address Takeover, Channel Bonding
- Primary network driver for Linux on System z
- Main focus in current and future development



Hipersockets Network verification

ifconfig output

ifconfig	ifconfig hsi0				
hsi0	Link encap:Ethernet HWaddr 06:00:F2:01:00:1B				
	inet addr:10.10.32.5 Bcast:10.10.63.255 Mask:255.255.224.0				
	inet6 addr: fe80::400:f2ff:fe01:1b/64 Scope:Link				
	UP BROADCAST RUNNING NOARP MULTICAST MTU:32768 Metric:1				
	RX packets:32285 errers:0 dropped:0 overruns:0 frame:0				
	TX packets:4453 (errors:382 oropped:0 overruns:0 carrier:0				

s390 debug feature

- Check for geth errors:



dbginfo file

- Check for buffer count of receiving partner:

lsqeth hsi0 | grep buffer_count
buffer_count : 16



Hipersockets network connection verification

netstat –s

netstat -s			
•••			
Tcp:			
•••			
39409 segments received			
51294 segments send out			
70 segments retransmited			
0 bad segments received.			

- Original Problem: Input buffers of receiving partner are full:
 - Too few input buffers
 - increase buffer_count to 128
 - CPUs busy with something else
 - LPAR share / VM guest share too low
 - Increase sysctl values
 - net.ipv4.tcp_rmem, sample: 4096 131072 4194240
 - net.core.netdev_max_backlog, sample: 3000
- Imbalance between sender and receiver
 - If CPU of Linux and z/VSE has high difference use Shared OSA or z/VM VSWITCH



Network Recommendations

• Which connectivity to use:

- -External connectivity:
 - LPAR: 10 GbE cards
 - z/VM: VSWITCH with 10GbE card(s) attached
 - z/VM: For maximum throughput and minimal CPU utilization attach OSA directly to Linux guest
- -Internal connectivity:
 - LPAR: HiperSockets for LPAR-LPAR communication
 - z/VM: VSWITCH for guest-guest communication
- For highly utilized network devices consider
 - -to use z/VM VSWITCH with link aggregation

-to use channel bonding

-that channel bonding for high availability has low overhead



z/VSE V5 Strategy with zEnterprise - More options, highly integrated

Traditional + Linux on z

and zBX

Reduce

- Routers
- Switches
- Firewalls
- Centralize
 - DNS Server
 - Network filtering
 - Work balancer
 - Edge Server
- LDAP security integration

Uses the internal IEDN network.

No need for additional DMZ security zones

➤use standard Intel based software for non-critical applications to Linux on z databases





Summary

The demands placed on the data center have never been greater.

IBM System zEnterprise:

- 1. Enables mixed workload Business Processes to be deployed, and centrally managed
- 2. Allows **optimized integration** of data, applications, and web serving
- 3. Delivers dynamically responsive IT with lower acquisition and operating costs
- 4. Meets the need of heterogeneous data centers



A strategic systems platform....

Helping to free up resources for critical projects and establish a base for the future



The Future runs on System z, the largest scalable server



... System z delivers extreme business value by helping you reduce cost, manage risk, and improve service.



Additional Documentation

- IBM System z Networking <u>http://www.ibm.com/systems/z/hardware/networking/</u>
- IBM System z Connectivity Handbook http://www.redbooks.ibm.com/redpieces/abstracts/sg245444.html
- z/VSE documentation

http://www-03.ibm.com/systems/z/os/zvse/documentation/#tcpipz/

- VM Networking <u>http://www.vm.ibm.com/virtualnetwork/</u>
- 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?





Wilhelm Mild IBM IT Architect



▙▟▛▟▝▙

IBM Deutschland Research & Development GmbH Schönaicher Strasse 220 71032 Böblingen, Germany

Office: +49 (0)7031-16-3796 mildw@de.ibm.com