Wilhelm Mild - IT Architect
10 / 2013

IBM

Networking in zEnterprise between z/OS and Linux on System z

Enterprise2013

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 ® 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.

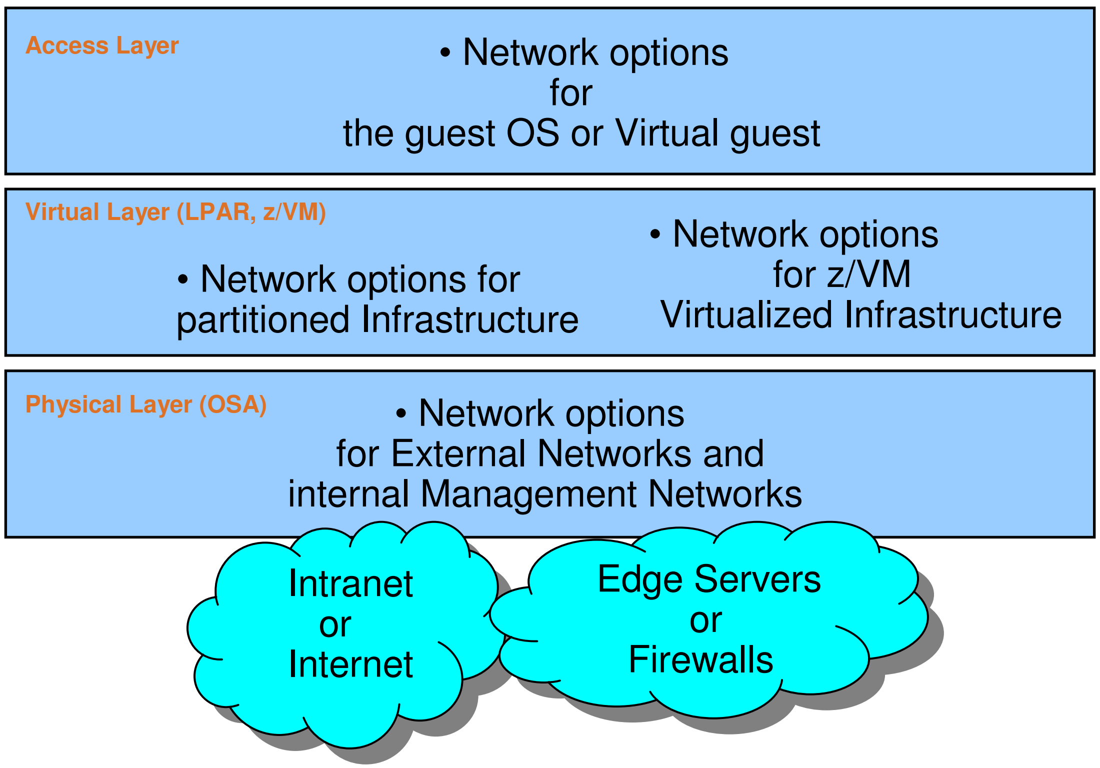
This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available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

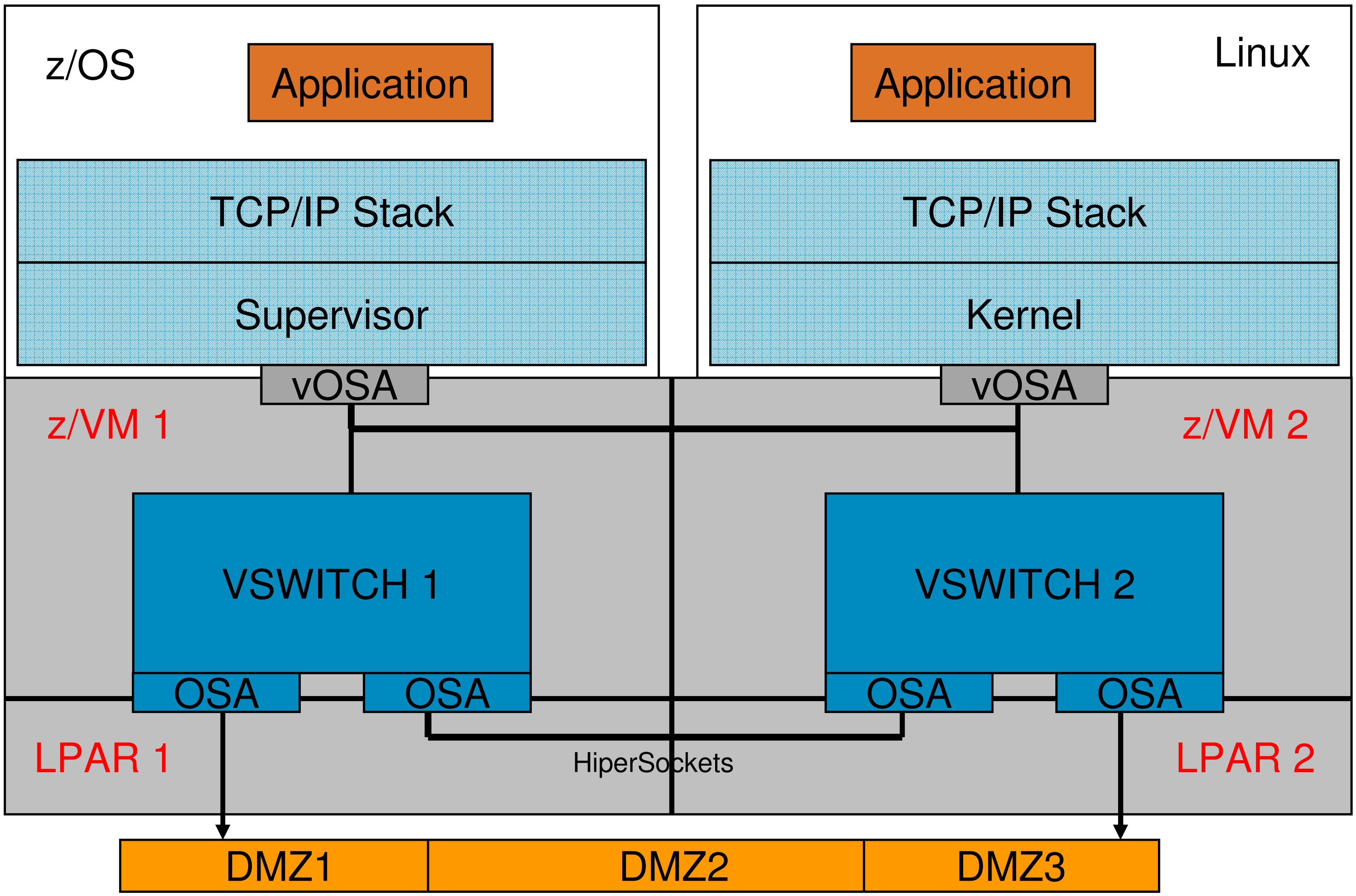
Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products 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.

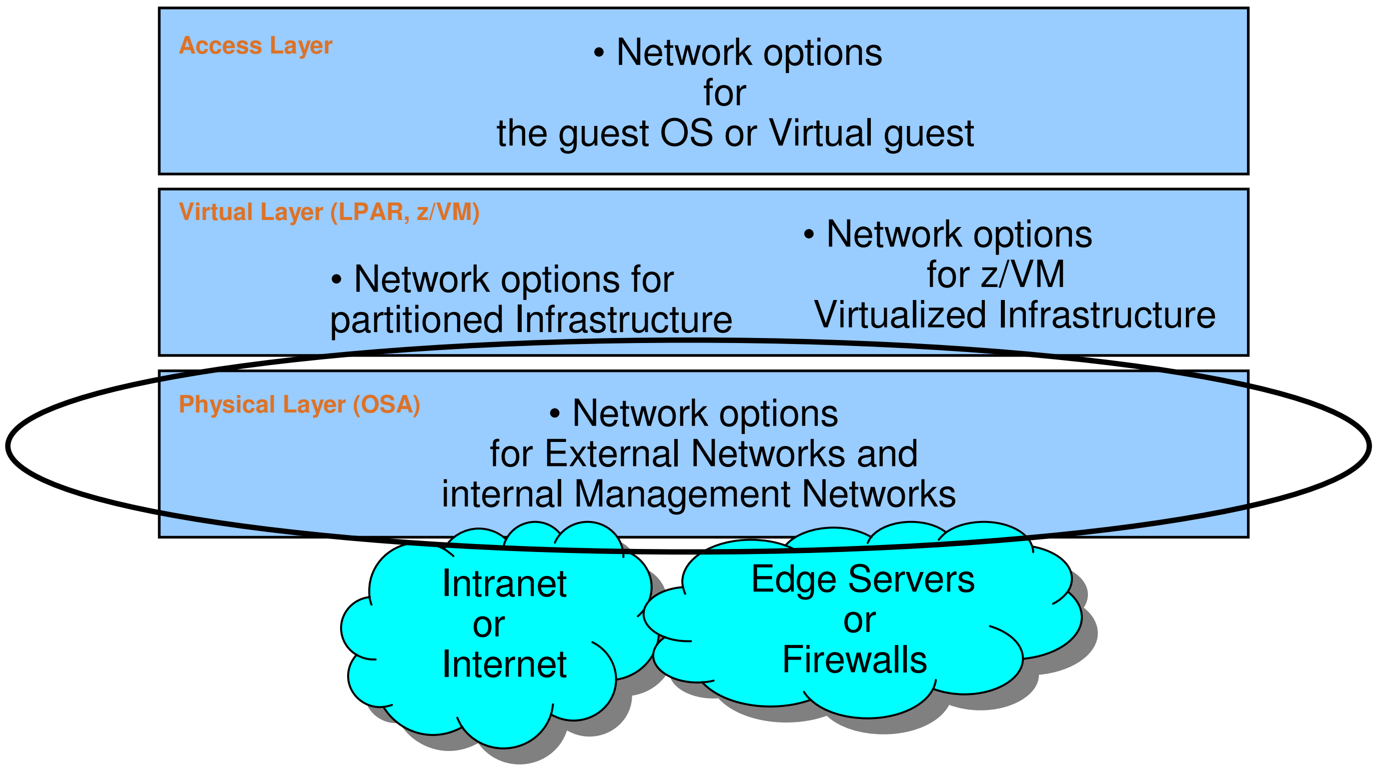
Reference Architecture for Networks with System z



System z Networking - Operational Diagram



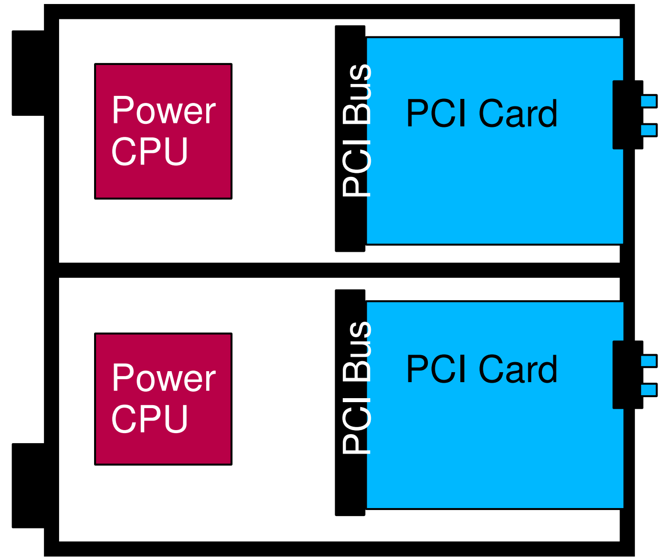
Reference Architecture for Networks with System z



Note:
- All network Connections in System z and zEnterprise are realized via **OSA (Open System Adapter**) cards
- Different types of network possible with the same OSA card
- **Hipersockets** - the network in the box, has no setup requirements from a HW perspective
- New: **RoCE** cards for ‚remote Hipersockets‘ network communications

OSA Express communication characteristics

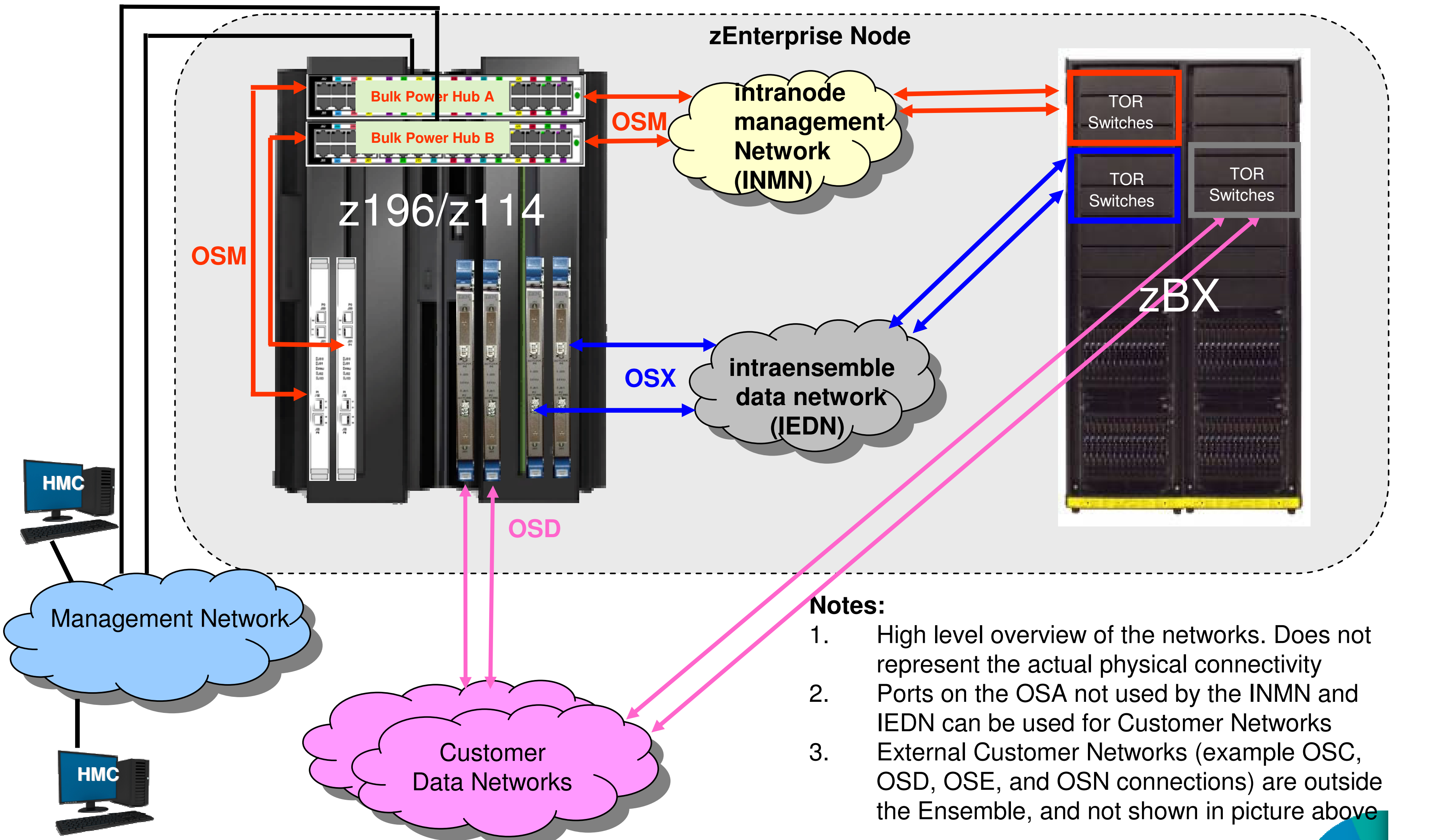
- 'Integrated Power computer' with network card
- Shared between up to 640 OSA devices
- Three device numbers (ccw devices) per OSA device:
  - Read device (control data ← OSA)
  - Write device (control data → OSA)
  - Data device (network traffic)
- OSA Address Table: which OS image has which IP address
- 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:

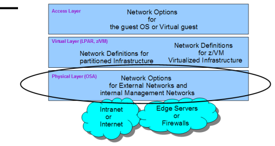
- **Communication is asynchronous** – from an application perspective
- **Communication is at OSA card clock speed** (typically lower than Hipersockets)

zEnterprise – What is INMN, IEDN and Customer network



Notes:

1. High level overview of the networks. Does not represent the actual physical connectivity
2. Ports on the OSA not used by the INMN and IEDN can be used for Customer Networks
3. External Customer Networks (example OSC, OSD, OSE, and OSN connections) are outside the Ensemble, and not shown in picture above



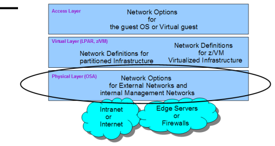
OSA Express – Network types

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

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

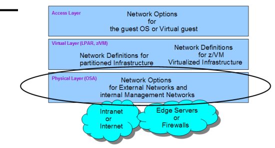
- **CHPID types**
  - **OSC** OSA-ICC (for emulation of TN3270E and non-SNA DFT 3270, IPL CPCs, and LPARs, OS system console operations)
  - **OSD** Queue Direct Input/Output (QDIO) architecture
  - **OSE** non-QDIO Mode (OSA-2, for SNA/APPN connections)
  - **OSN** OSA-Express for NCP: Appears to z/OS 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 its 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 zEnterprise Ensemble management. OSM ports connect to the Intranode Management Network (INMN)
  - **IQDX** Hipersockets Bridge to zBX

Open Systems Adapter (OSA) performance

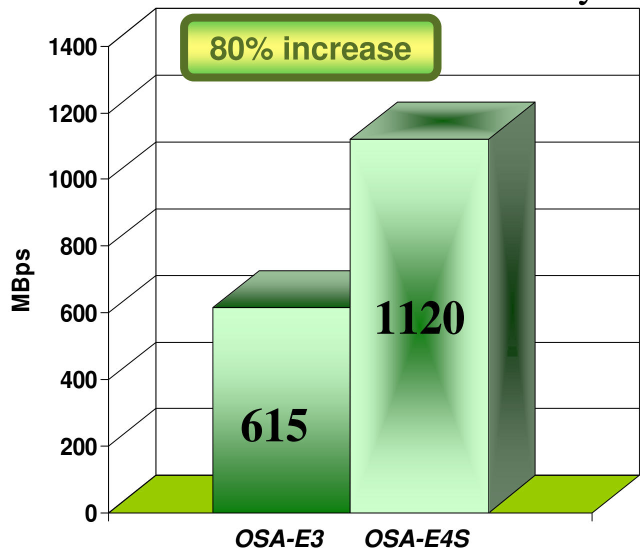


- OSA processor becomes more efficient as throughput increases
- **Window size**
  - TCP window determines amount of data that the sender can transmit to receiver without needing an acknowledgment from the receiver
  - Faster and longer networks require larger windows to keep data flowing smoothly
- **Blocking**
  - Performance is affected by the amount of data blocked together for transfer between OSA and TCP
- **Frame size**
  - Larger frames perform better
  - Larger frames reduce host and OSA processing costs
  - Size of frame depends on LAN type, MTU setting, size of data sent
- Measure: throughput, transaction response time, server utilization
  - Bulk data transfer and interactive transactions
- QDIO and jumbo frames (8992 byte MTUs) yield the highest streams
  - Ethernet II (DIX) – MTU 1500 and jumbo frame MTU 9000 (most common)
  - IEEE802 – MTU 1492 and jumbo frame MTU 8992

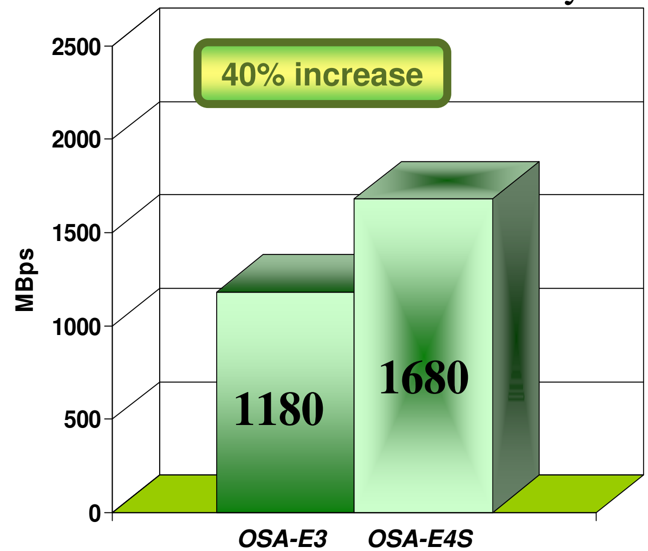
OSA-Express4S 10 GbE Performance (Lab. Measurements)



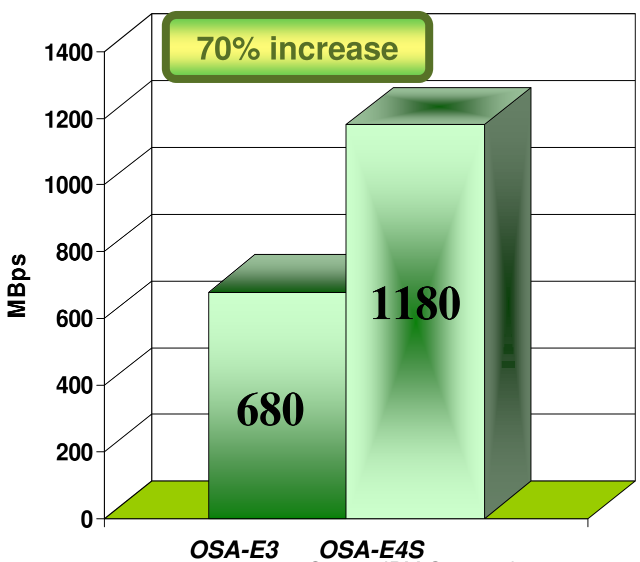
Inbound Streams – 1492 Byte MTUs



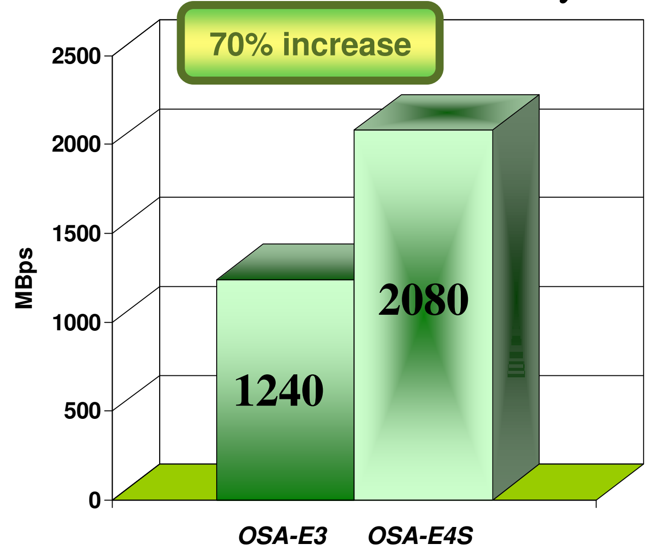
Mixed Streams – 1492 Byte MTUs



Inbound Streams – 8000 Byte MTUs



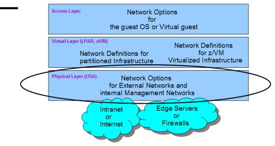
Mixed Streams – 8000 Byte MTUs



Notes:
- AWM on z/OS
- z/OS is doing checksum
- 1 megabyte per second (MBps) is 1,048,576 bytes per second

MBps represents payload throughput (does not count packet and frame headers)

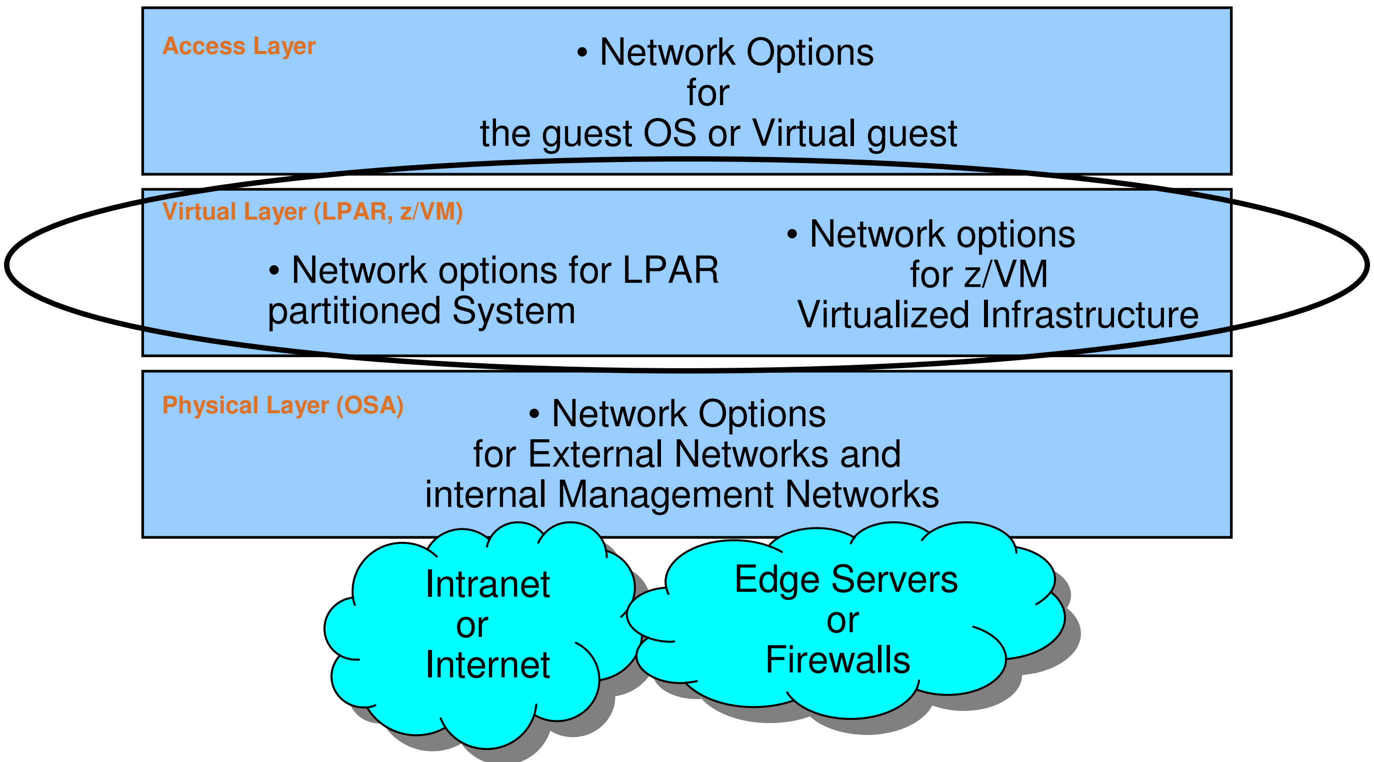
Enterprise2013



Summary: OSA-Express CHPID types to control operation

CHPID type	Purpose / Traffic	Operating Systems
OSC **1000BASE-T** **EC12,z196, z114, z10, z9** **z990, z890**	**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** **zEC12, z196, z114, z10, z9, zSeries**	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 System z
OSE **1000BASE-T** **zEC12, z196, z114, z10, z9, zSeries**	**Non-QDIO; for SNA/APPN/HPR** traffic and TCP/IP "passthru" traffic	z/OS, z/VM z/VSE
OSM **1000BASE-T** **zEC12, z196, z114**	**OSA-Express for Unified Resource Manager** Connectivity to Intra Node Management Network (INMN) from zEC12, z196, or z114 to Unified Resource Manager functions	z/OS, z/VM Linux on System z
OSN **GbE, 1000BASE-T** **zEC12, z196, z114, z10, z9** **No OSN support for** **OSA-Express4S 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 System z
OSX **10 GbE** **zEC12, z196, z114**	**OSA-Express for zBX** Connectivity and access control to intraensemble data network (IEDN) from zEC12, z196, or z114 to zBX	z/OS, z/VM, z/VSE 5.1, Linux on System z

Reference Architecture for Networks with System z

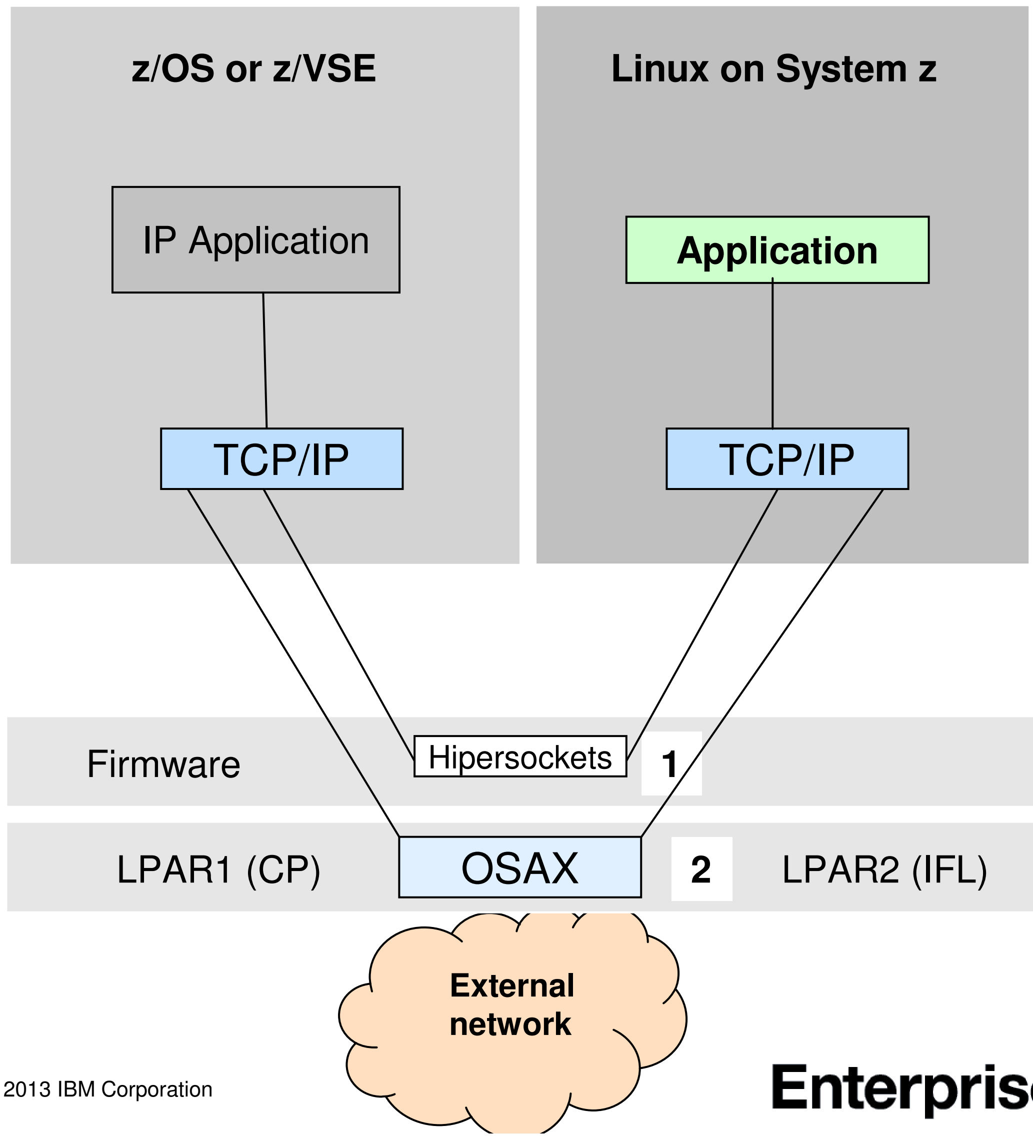


Recommendations:

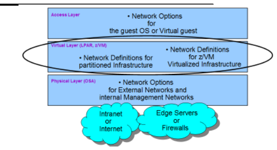
- LPAR network Connections in System z are realized
  - via **Hipersockets** between LPARs
  - via (shared) **OSA (Open System Adapter**) cards to external network
- z/VM network Connections are realized
  - via **Virtual Hipersockets**
  - via **VSWITCH** between guest Systems
  - via **IUCV** for special network Connections

Enterprise2013

System z Network alternatives between LPARs with OSA

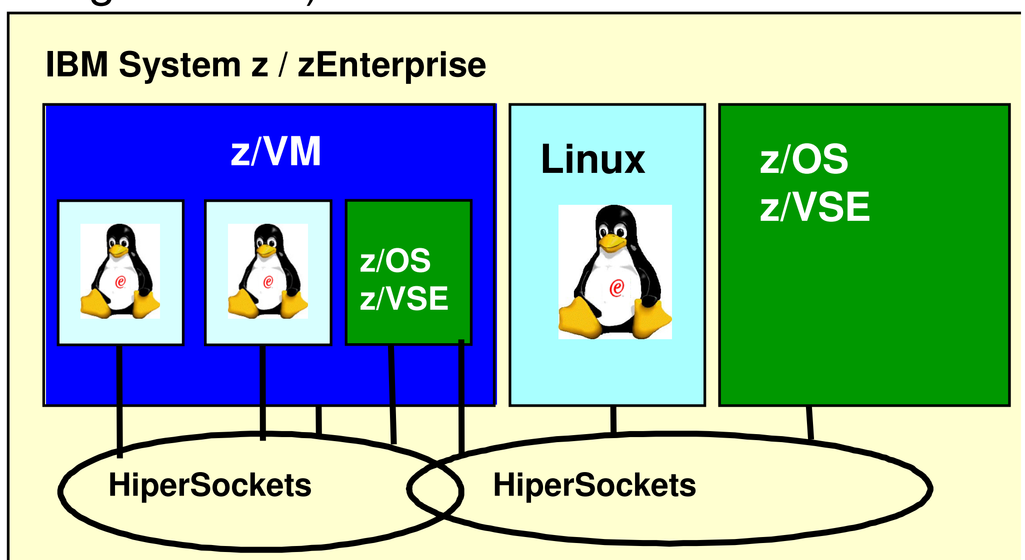


Network between LPARs on System z - Hipersockets network



- 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



Enterprise2013

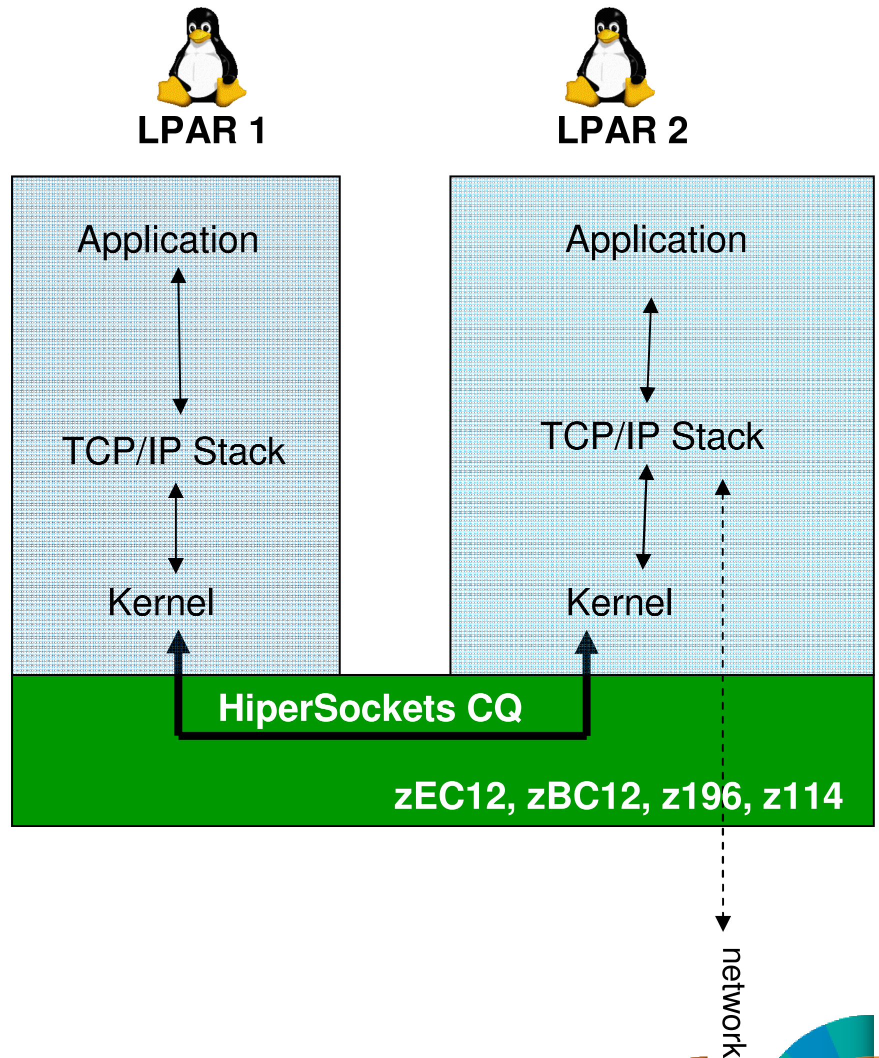
New: Hipersockets Completion Queue between LPARs

Avail since: 06/2012: Function for zEnterprise between LPARs

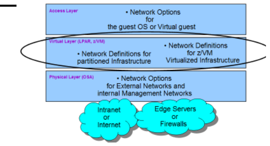
- When the remote side can receive data volume then data is sent synchronously.
- When the remote side cannot receive data volume then data is sent asynchronously.

HiperSockets transfers data synchronously if possible and asynchronously if necessary.

- Ultra-low latency with more tolerance for traffic peeks.
- Requires zEnterprise z196 or later hardware.
- Requires z/OS V1.13 or later software.



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

Enterprise2013

Optimize server to server networking – transparently – zEC12+zBC12

"HiperSockets™-like" capability across physical systems

Up to 50% CPU savings for FTP file transfers across z/OS systems versus standard TCP/IP **

Up to 48% reduction in response time and 10% CPU savings for a sample CICS workload exploiting IPIC using SMC-R versus TCP/IP ***

Up to 40% reduction in overall transaction response time for WAS workload accessing z/OS DB2 ****

Up to 3X increase in WebSphere MQ messages delivered across z/OS systems *****

Shared Memory Communications (SMC-R):

Exploit RDMA (Remote Direct Memory Access) over Converged Ethernet (RoCE) with qualities of service support for dynamic failover to redundant hardware

Typical Client Use Cases:

Help to reduce both latency and CPU resource consumption over traditional TCP/IP for communications across z/OS systems

Any z/OS TCP sockets based workload can **seamlessly** use SMC-R without requiring any application changes

NEW **z/OS V2.1 SMC-R**

NEW **z/VM V6.3 support for guests**

NEW **10GbE RoCE Express**

** Based on internal IBM benchmarks in a controlled environment using z/OS V2R1 Communications Server FTP client and FTP server, transferring a 1.2GB binary file using SMC-R (10GbE RoCE Express feature) vs standard TCP/IP (10GbE OSA Express4 feature). The actual CPU savings any user will experience may vary.

*** Based on internal IBM benchmarks using a modeled CICS workload driving a CICS transaction that performs 5 DPL (Distributed Program Link) calls to a CICS region on a remote z/OS system via CICS IP interconnectivity (IPIC), using 32K input/output containers. Response times and CPU savings measured on z/OS system initiating the DPL calls. The actual response times and CPU savings any user will experience will vary.

**** Based on projections and measurements completed in a controlled environment. Results may vary by customer based on individual workload, configuration and software levels.

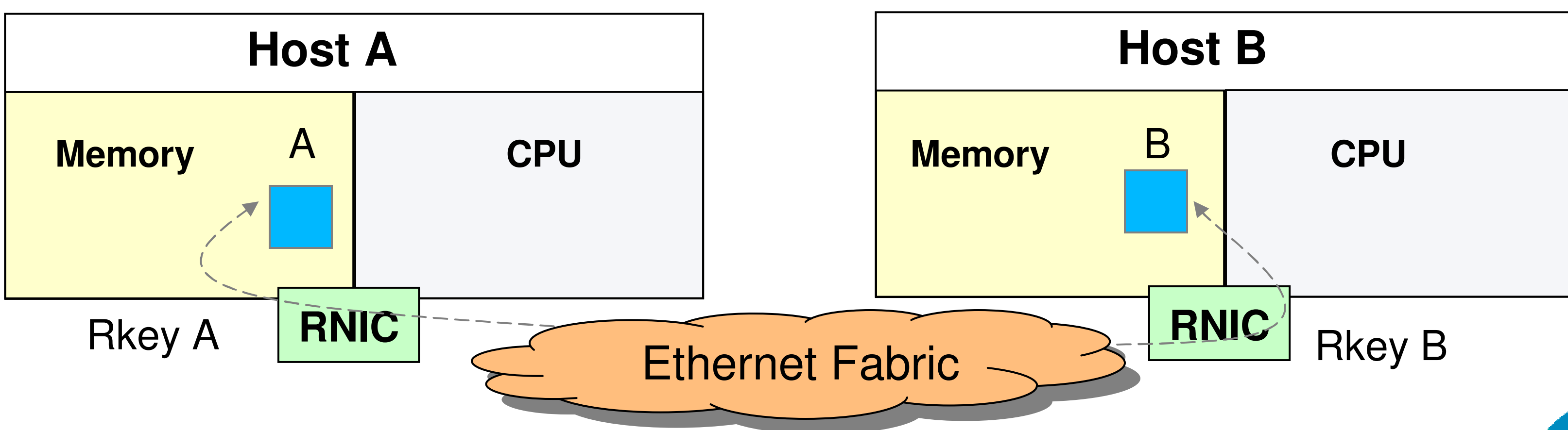
***** Based on internal IBM benchmarks using a modeled WebSphere MQ for z/OS workload driving non-persistent messages across z/OS systems in a request/response pattern. The benchmarks included various data sizes and number of channel pairs The actual throughput and CPU savings users will experience may vary based on the user workload and configuration.

Shared Memory Communications – Remote Direct Memory Access (SMC-R) Definition

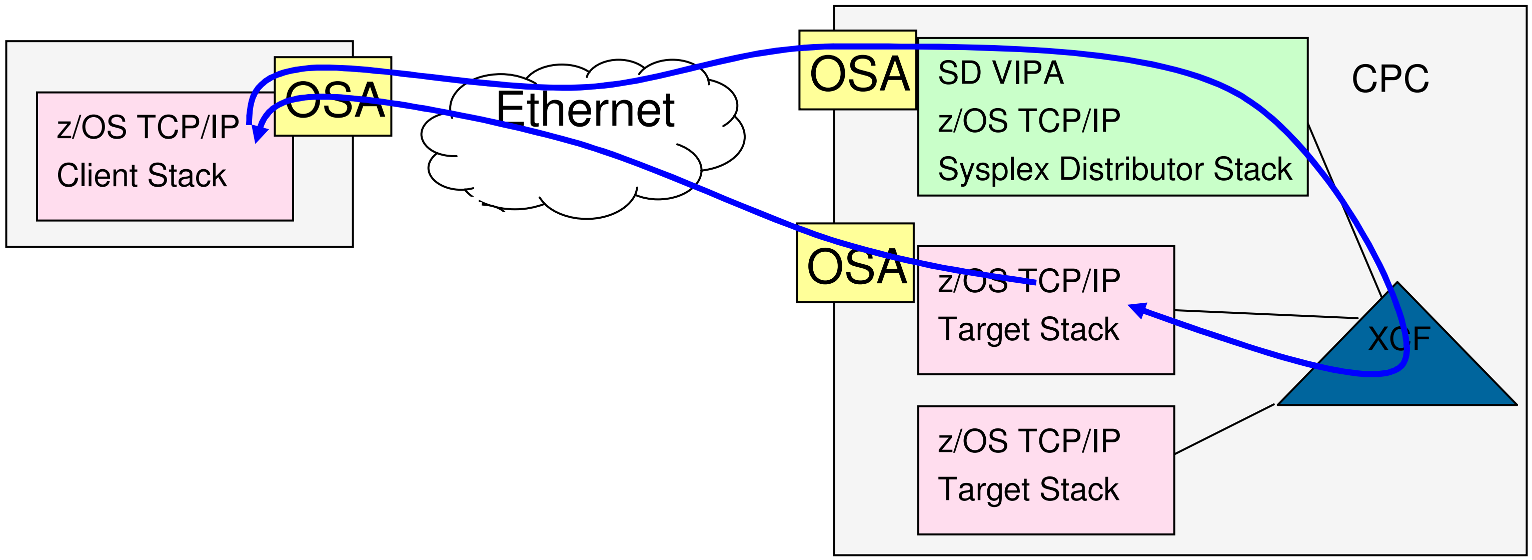
- Shared Memory Communications – Remote Direct Memory Access is a new communication protocol aimed at providing transparent acceleration for sockets-based TCP/IP applications and middleware
  - Remote Direct Memory Access (RDMA) technology provides low latency, high bandwidth, high throughput, low processor utilization attachment between hosts
  - SMC-R utilizes RDMA over Converged Ethernet (RoCE) as the physical transport layer
- SMC-R is built on the following concepts:
  - RDMA enablement of the communications fabric
  - Partitioning a part of OS host real memory into buffers and using RDMA technology to access this memory
  - Establishing an ‘out of band’ connection over which data is passed to the partner peer using RMDA writes and signaling

RDMA (Remote Direct Memory Access) Technology Overview

- Key attributes of RDMA
  - Enables a host to read or write directly from/to a remote host's memory ***without*** involving the remote host's CPU
    - By registering specific memory for RDMA partner use
    - Interrupts ***still required*** for notification (i.e. CPU cycles are not completely eliminated)
  - Reduced networking stack overhead by using streamlined, low level, RMDA interfaces
    - No requirement for TCP/IP protocols/stack, sockets, etc.
      - Low level APIs such as uDAPL, MPI or RDMA verbs allow optimized exploitation
        - > *For applications/middleware willing to exploit these interfaces*
  - Key requirements:
    - A reliable "lossless" network fabric (LAN for layer 2 data center network distance)
    - An RDMA capable NIC (RNIC) and ethernet fabric – switches (recommended) or point to point

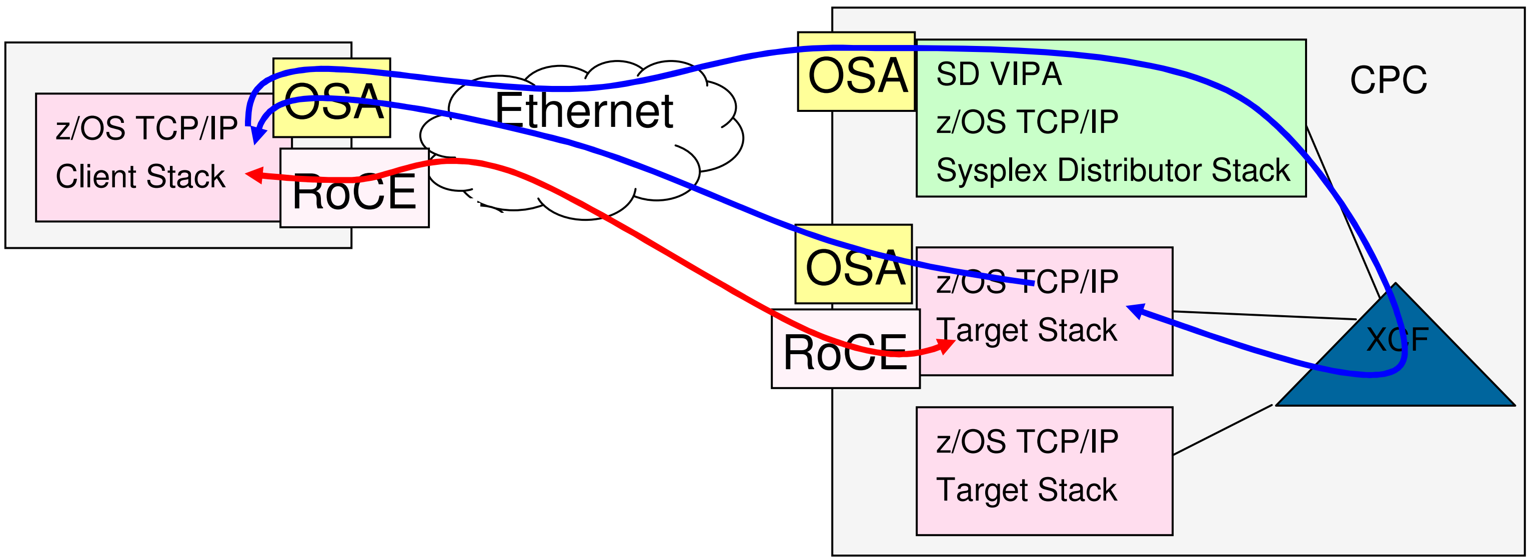


Sysplex Distributor before RoCE



- Traditional Sysplex Distributor
  - All traffic from the client to the target application goes through the Sysplex Distributor TCP/IP stack
  - All traffic from the target application goes directly back to the client using the TCP/IP routing table on the target TCP/IP stack.

Sysplex Distributor after RoCE

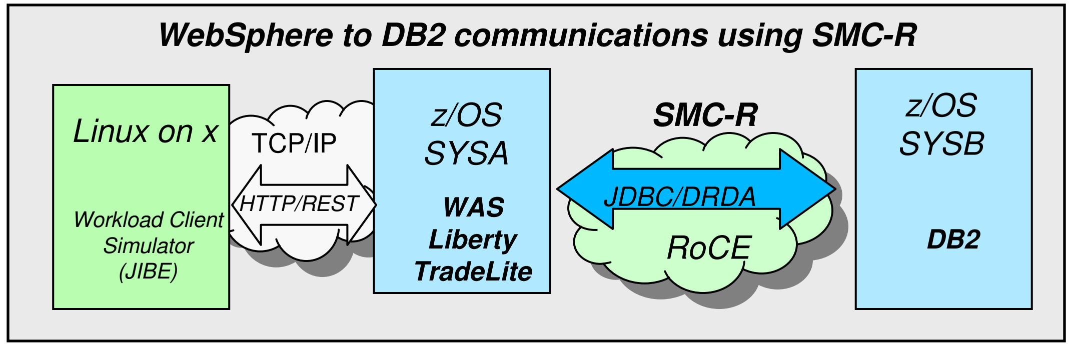


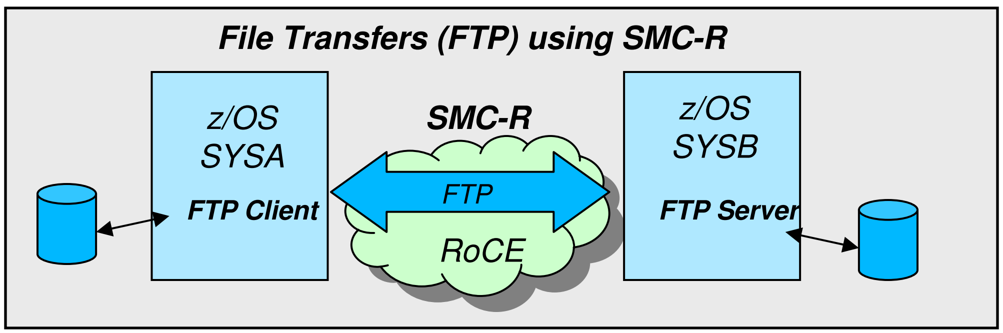
- RoCE Sysplex Distributor
  - The initial connection request goes through the Sysplex Distributor stack.
  - The session then flows directly between the client and the target over the RoCE features.

Note: As with all RoCE communication the session end also flows over the OSAs.

Impact of SMC-R on real z/OS workloads – early benchmark results

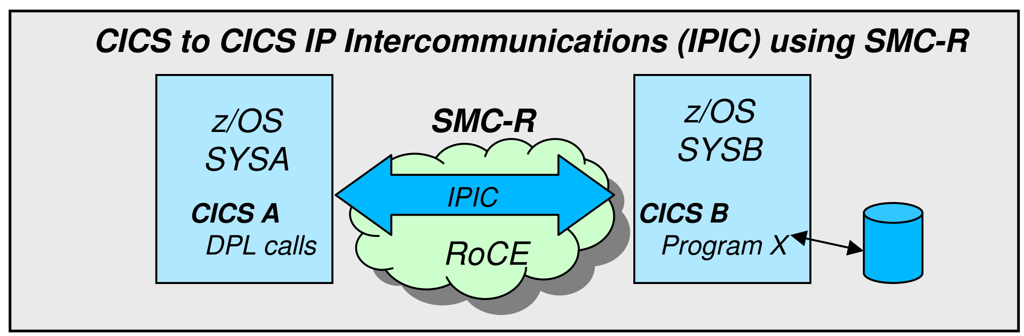
40% reduction in overall transaction response time for WebSphere 8.0 Liberty profile TradeLite workload accessing z/OS DB2 in another system measured in internal benchmarks *





Up to **50% CPU savings** for FTP binary file transfers across z/OS systems when using SMC-R vs standard TCP/IP **

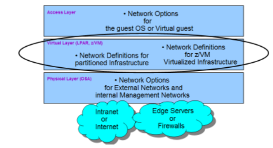
Up to 48% reduction in response time and up to 10% CPU savings for CICS transactions using DPL (Distributed Program Link) to invoke programs in remote CICS regions in another z/OS system via CICS IP interconnectivity (IPIC) when using SMC-R vs standard TCP/IP ***



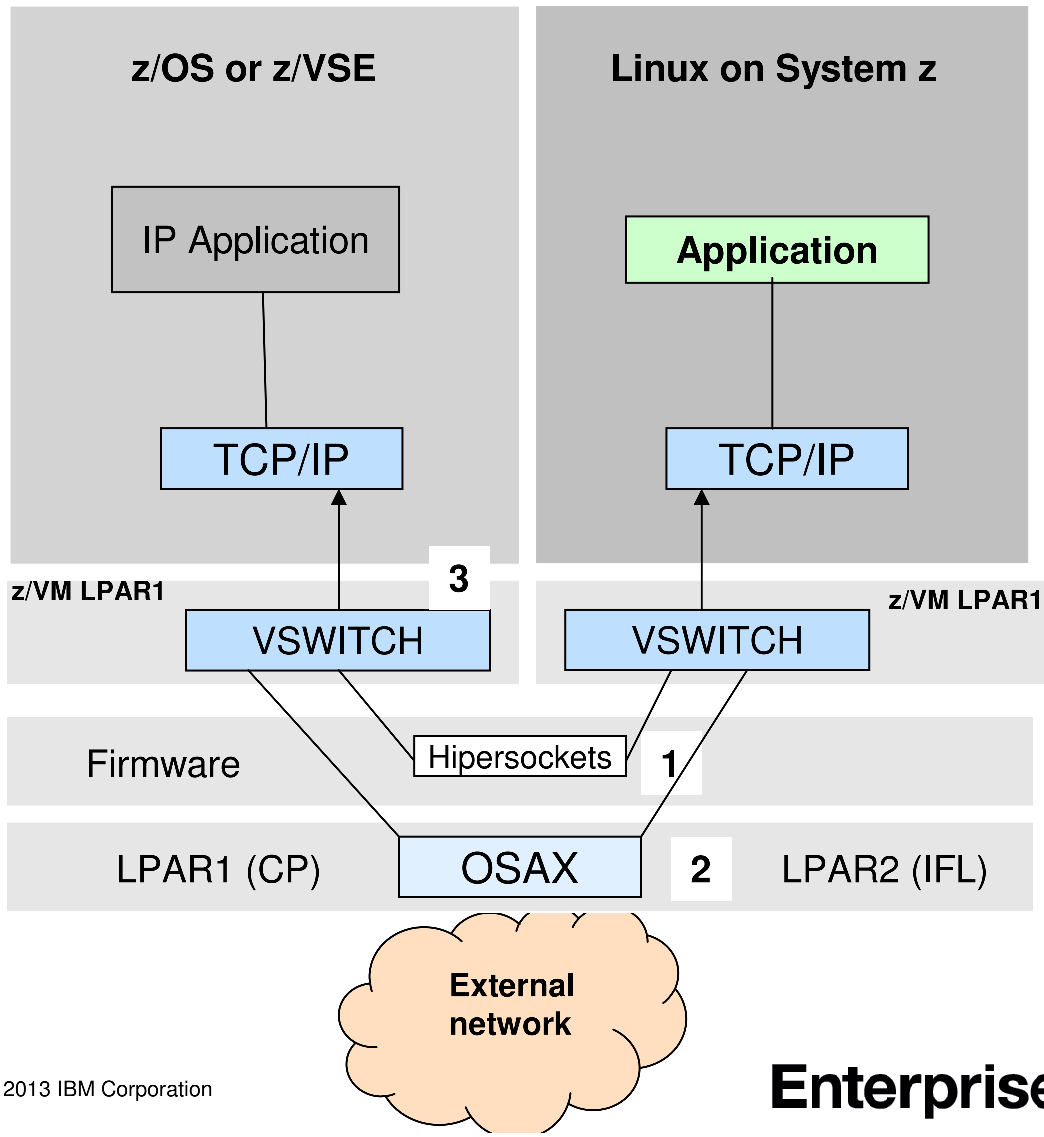
* Based on projections and measurements completed in a controlled environment. Results may vary by customer based on individual workload, configuration and software levels.

** Based on internal IBM benchmarks in a controlled environment using z/OS V2R1 Communications Server FTP client and FTP server, transferring a 1.2GB binary file using SMC-R (10GbE RoCE Express feature) vs standard TCP/IP (10GbE OSA Express4 feature). The actual CPU savings any user will experience may vary.

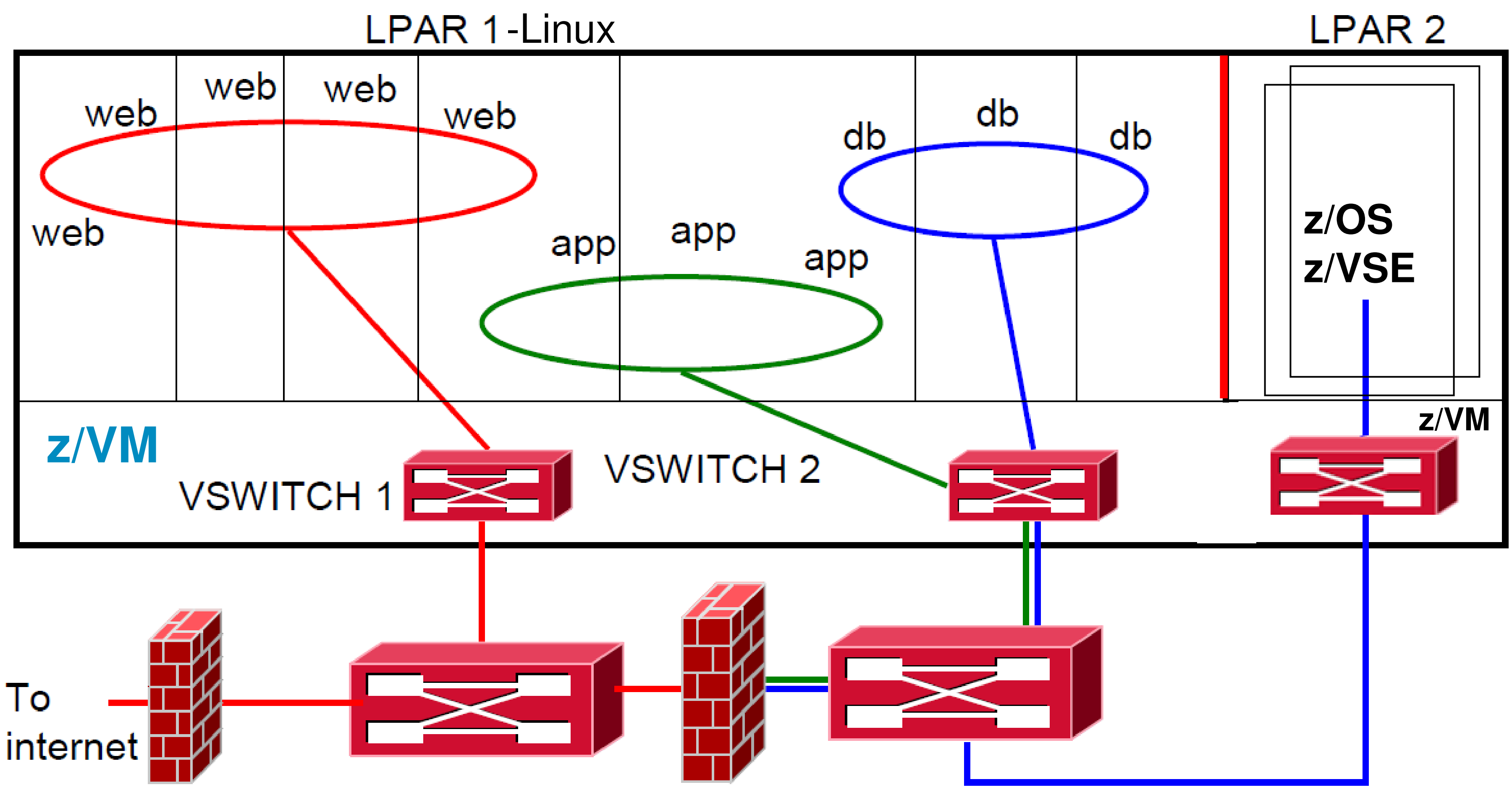
*** Based on internal IBM benchmarks using a modeled CICS workload driving a CICS transaction that performs 5 DPL calls to a CICS region on a remote z/OS system, using 32K input/output containers. Response times and CPU savings measured on z/OS system initiating the DPL calls. The actual response times and CPU savings any user will experience will vary.



System z Network alternatives between z/VM LPARs

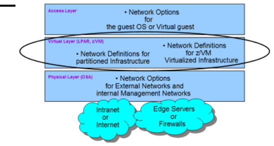


z/VM Multi-zone Network VSWITCH (red - physical isolation

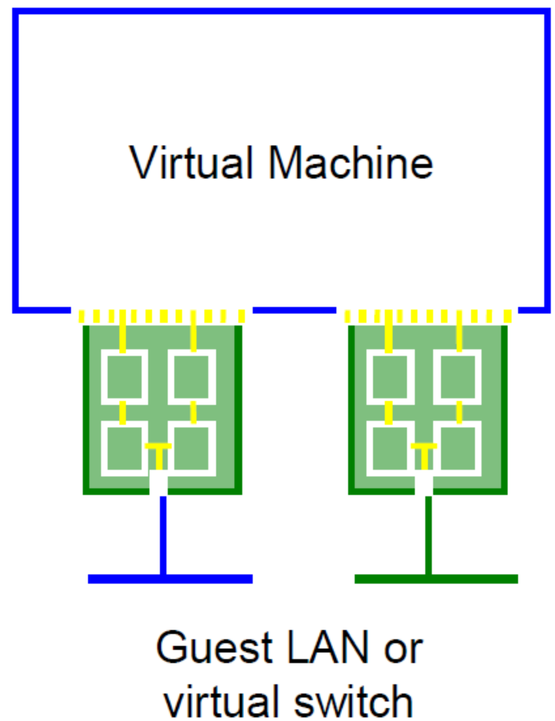


With 2 VSWITCHes, 3 VLANs, and a multi-domain firewall

Virtual Network Interface Card (vNIC)



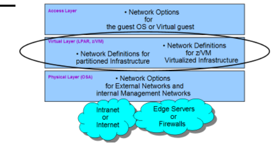
- A simulated network adapter
  - OSA-Express QDIO
  - HiperSockets
  - Must match LAN type
- Usually 3 devices per NIC
- Provides access to Guest LAN or VSWITCH
- Created by directory or *CP DEFINE NIC*



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

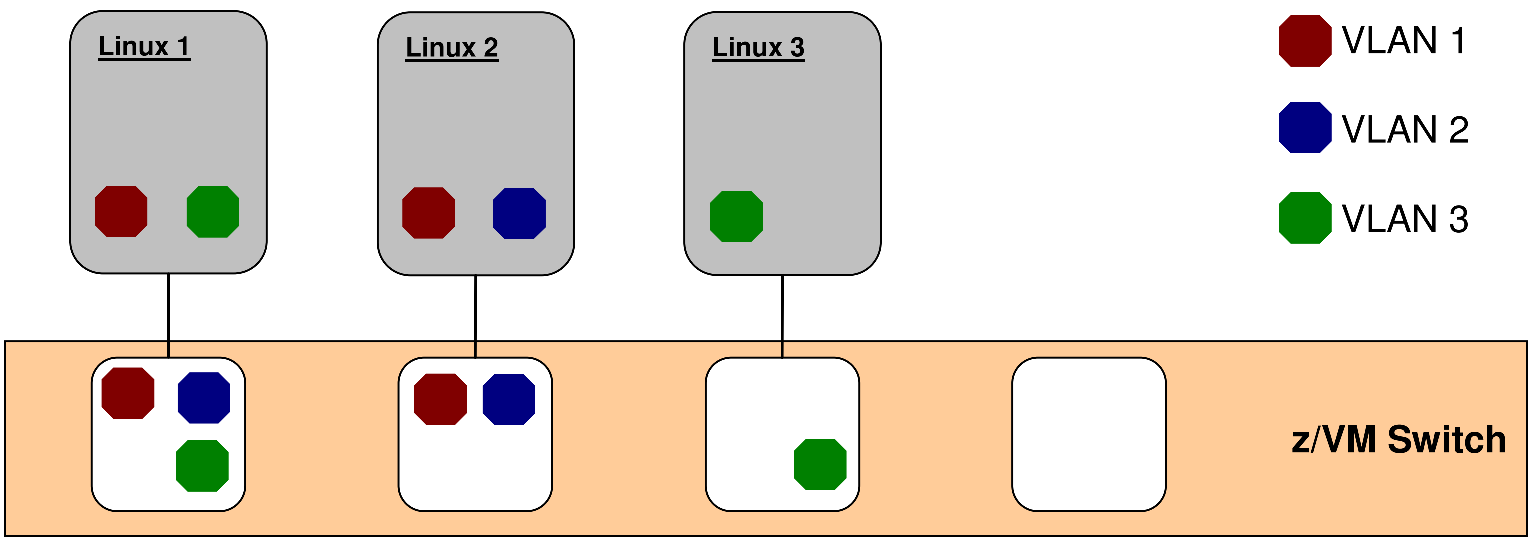
```
DEF NIC 600 TYPE QDIO
COUPLE 600 SYSTEM VSWITCH1
```

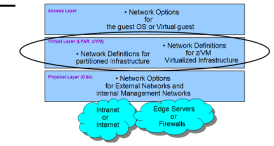

Enterprise2013



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





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.

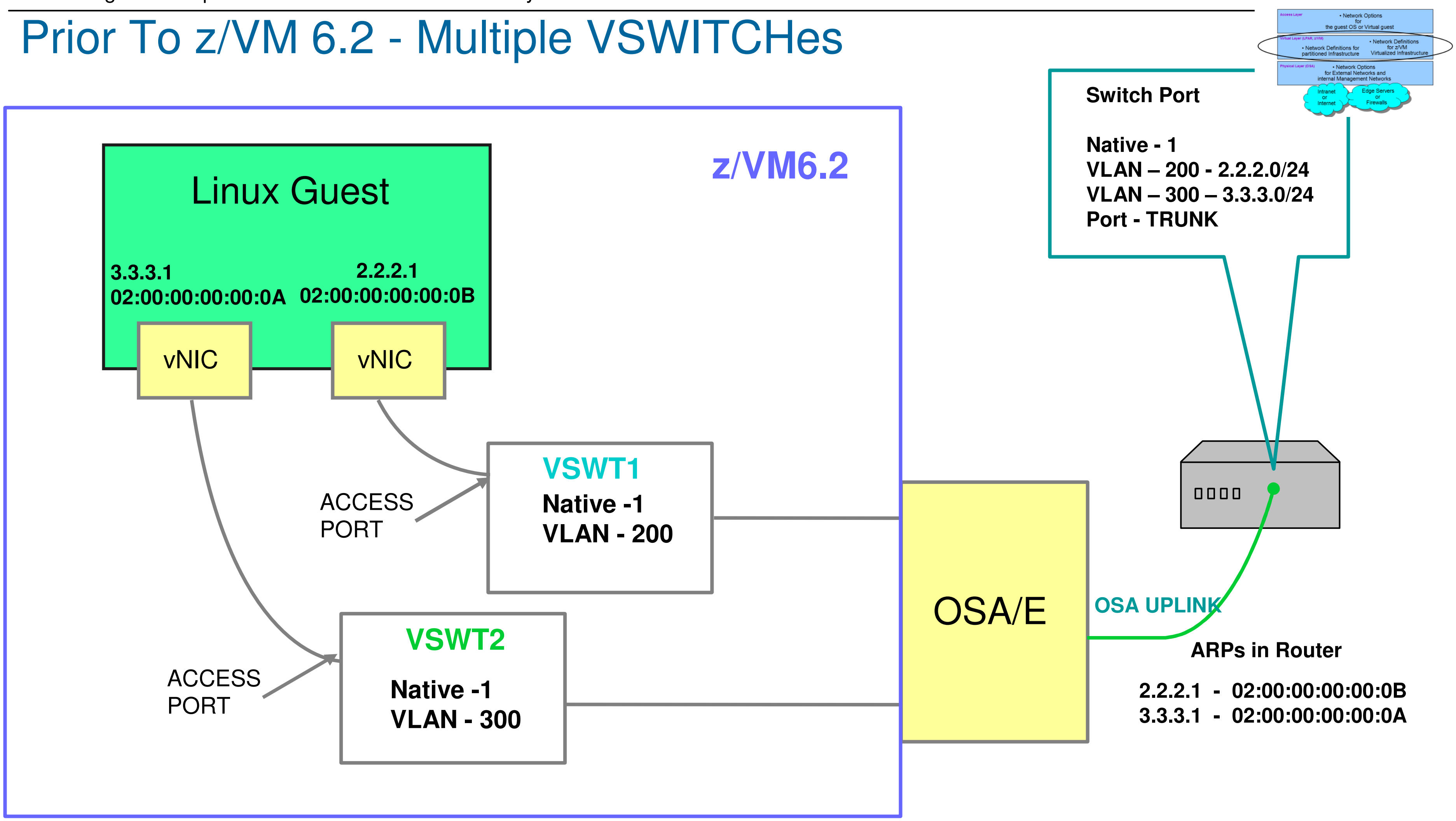
1. Have the Guest connect to multiple VSWITCHes.
   - Each VSWITCH would provide the Guest an ACCESS Port
   - Each vNIC would have a unique vMAC.

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

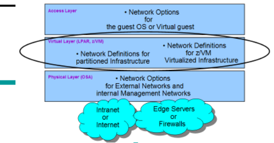
- Each vNIC would have a unique vMAC.

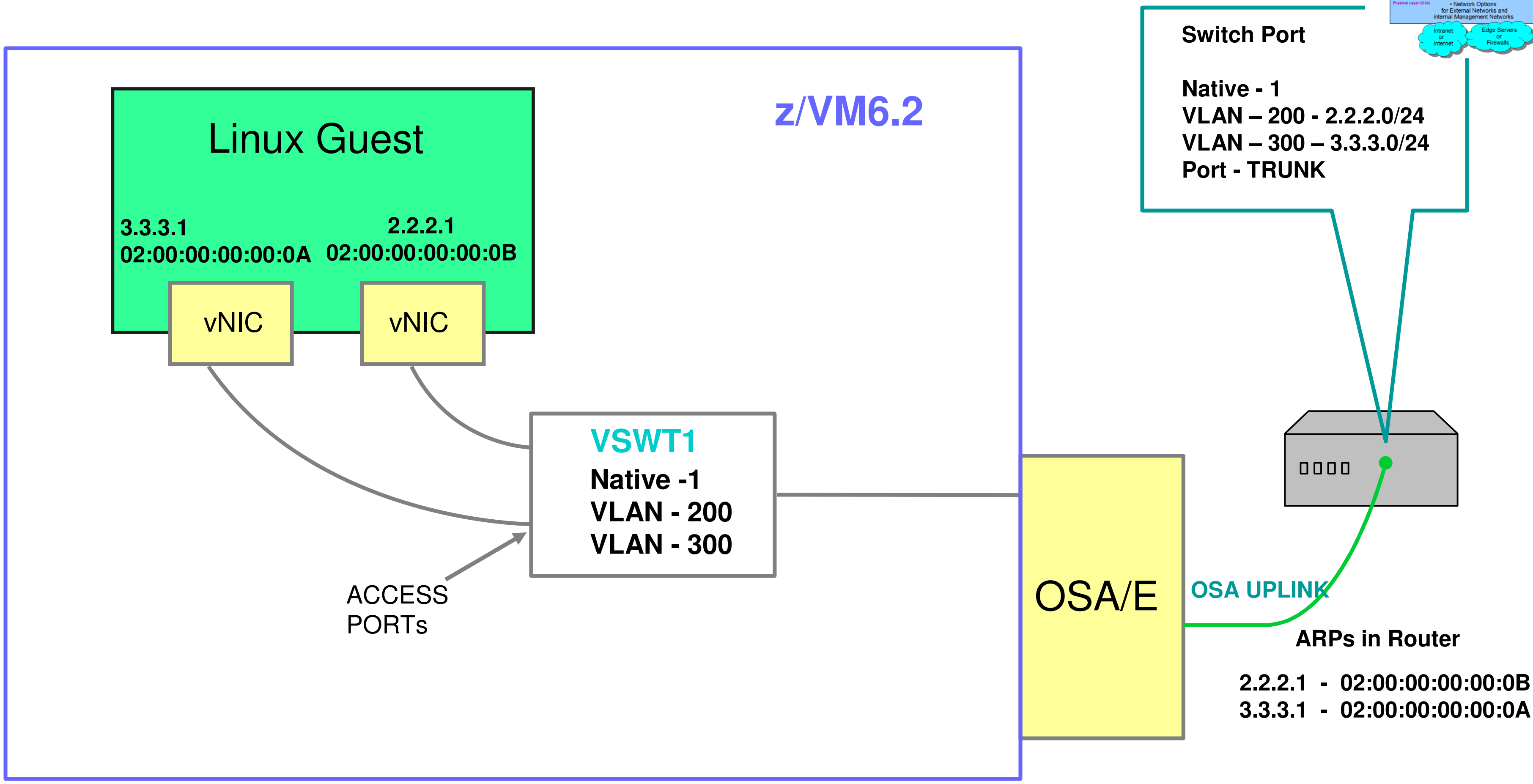
Prior To z/VM 6.2 - Multiple VSWITCHes



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

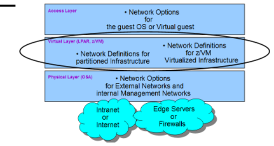
z/VM 6.2 Single VSWITCH, PORTBASED



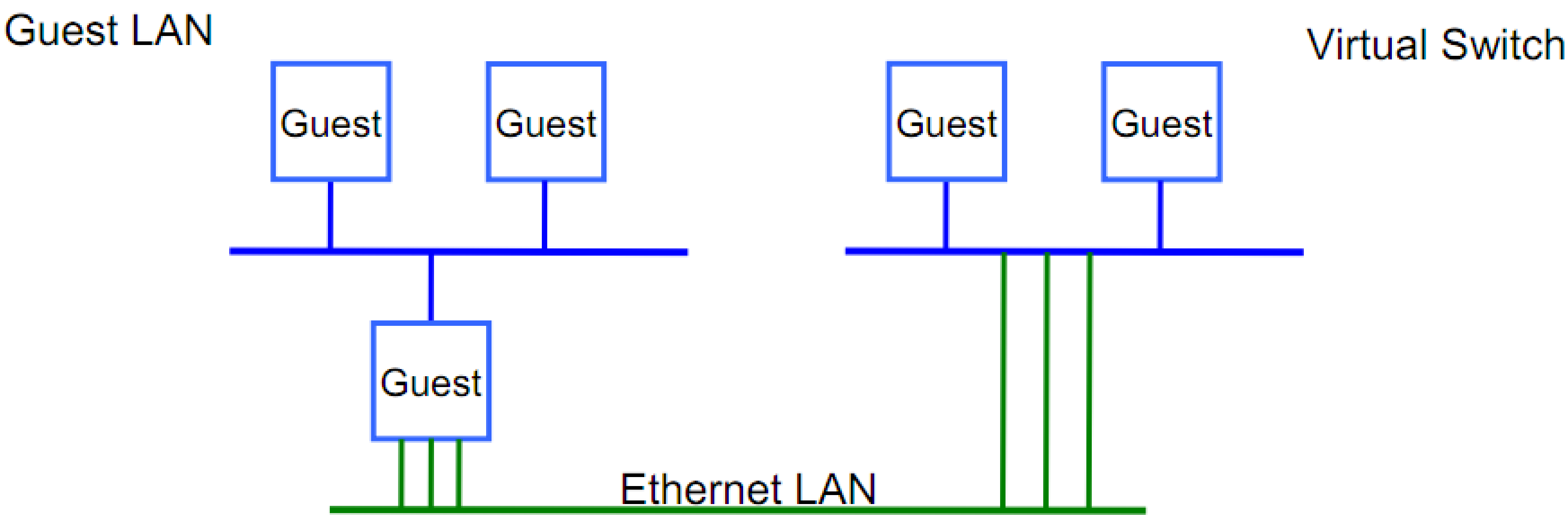


DEFINE VSWITCH VSWT1 RDEV 2A00 ETHERNET CONTROLLER * VLAN AWARE PORTBASED NAT 1
SET VSWITCH VSWT1 PORTNUMBER 20 GRANT GUEST VLAN 200
SET VSWITCH VSWT1 PORTNUMBER 21 GRANT GUEST VLAN 300

Enterprise2013

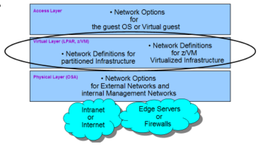


Guest LAN vs. Virtual Switch



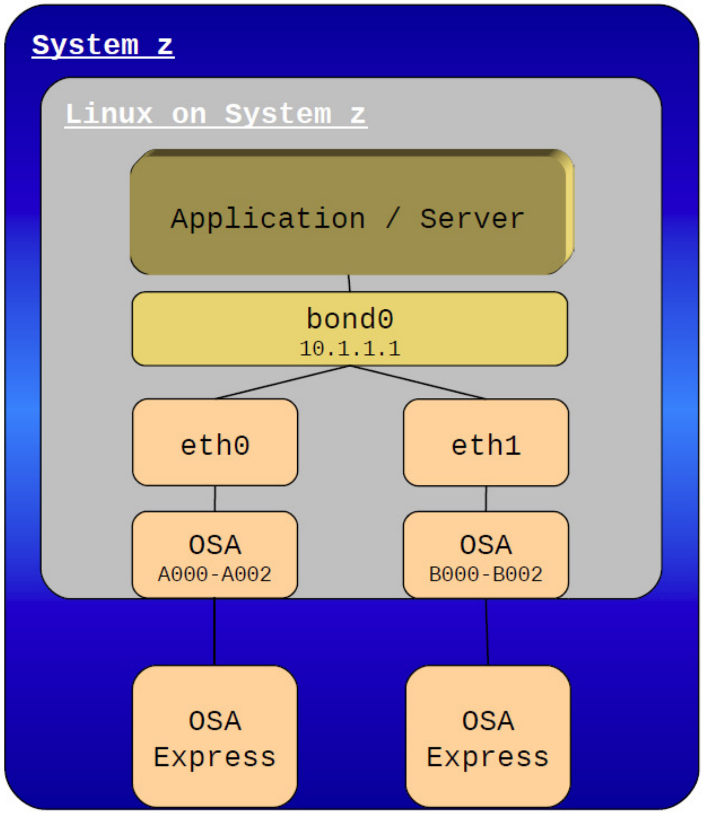
Guest LAN	Virtual Switch
Virtual router is required	No virtual router
Different subnets	Same subnets
External router awareness	Transparent bridge
Guest-managed failover	CP-managed failover

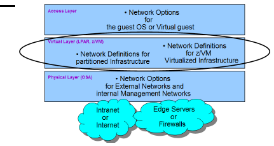
Enterprise2013



Channel Bonding - for higher bandwidth

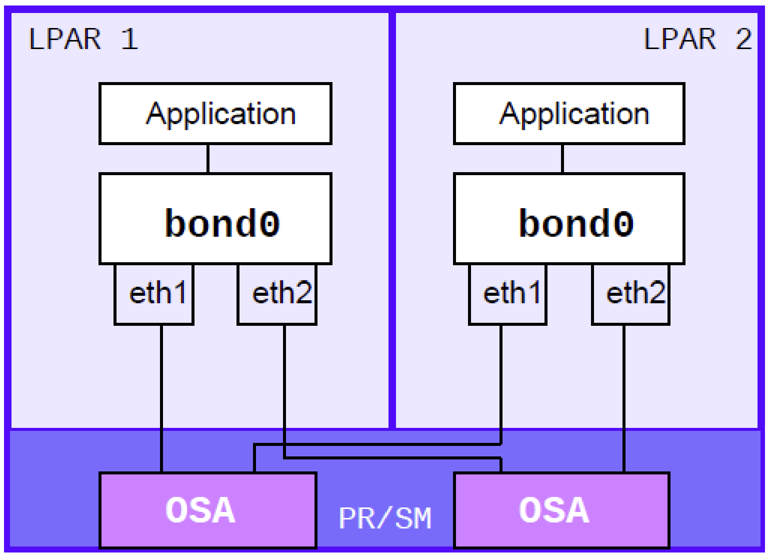
- The Linux bonding driver provides a
  - method for aggregating multiple
  - network interfaces into a single,
  - logical “bonded” interface
- Provides failover and/or load balancing functionality
- Better performance depending on bonding mode
- Requires layer2 devices
- Further information
  - http://sourceforge.net/projects/bonding





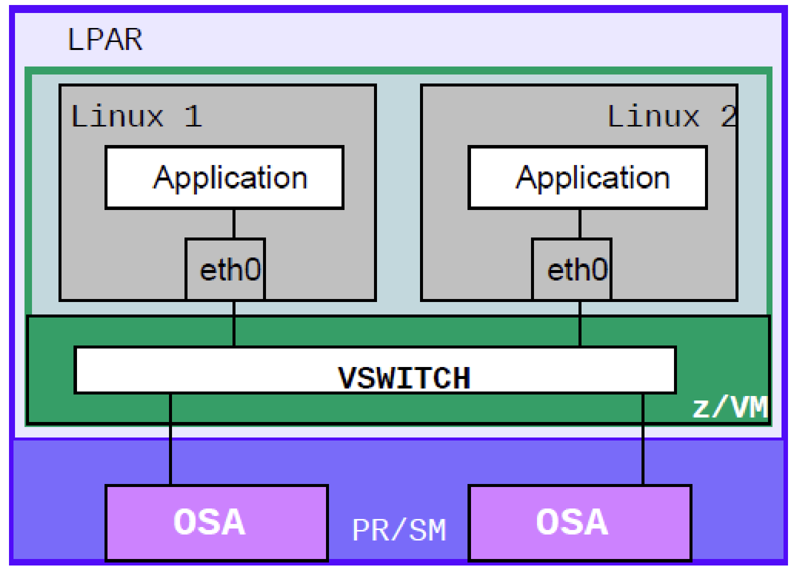
Network bandwidth enhancement and Automated Failover

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**

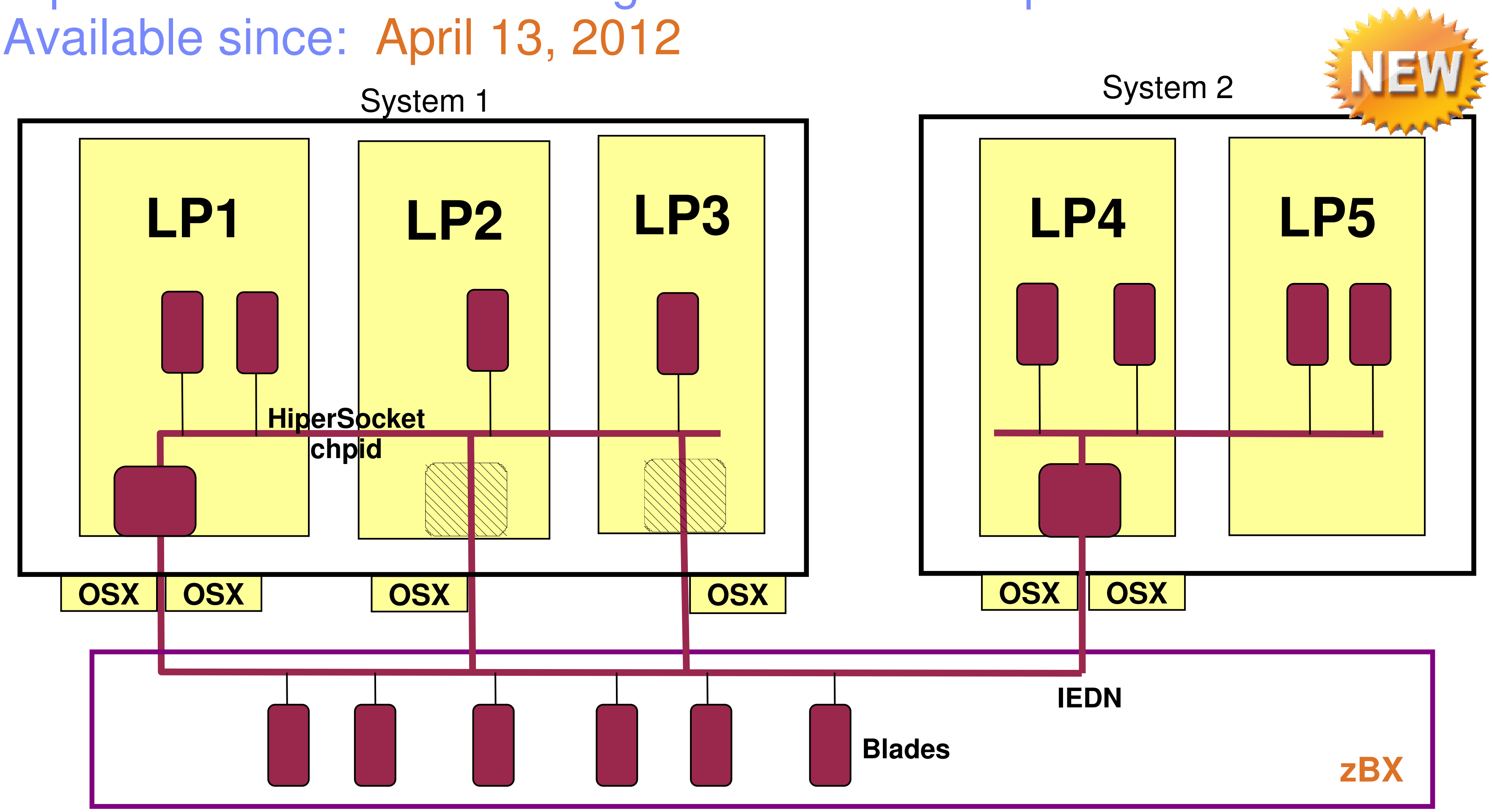
Network Virtualization: **z/VM Port aggregation**



- 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**

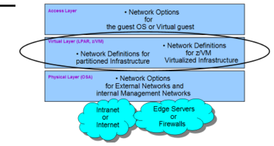
Enterprise2013

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



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

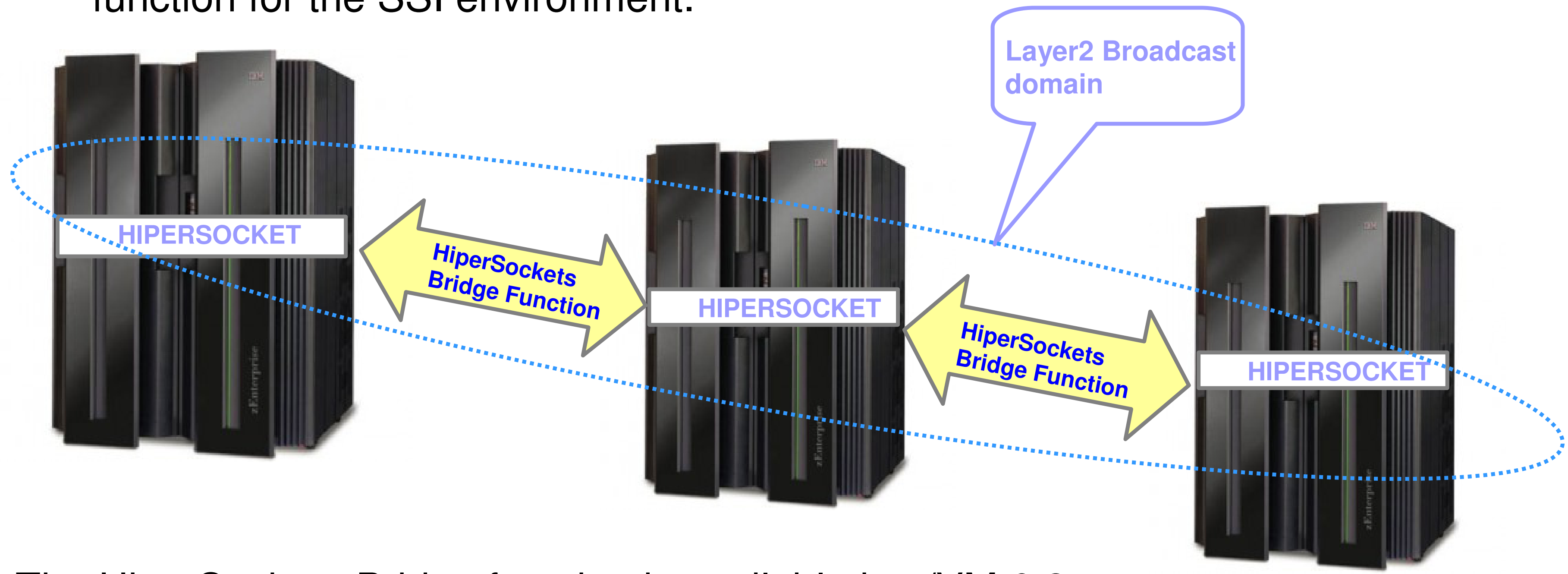
- Same or different LPAR
- One active bridge per CEC
- PMTU simulation



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.

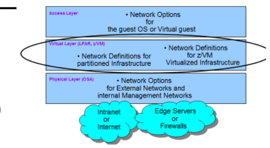


The HiperSockets Bridge function is available in z/VM 6.2 with a z114 or z196 or zEC12 processor. Must be sure to have z/VM and processor maintenance levels..

z/OS support for Hipersockets Bridge, CHPID = IQD

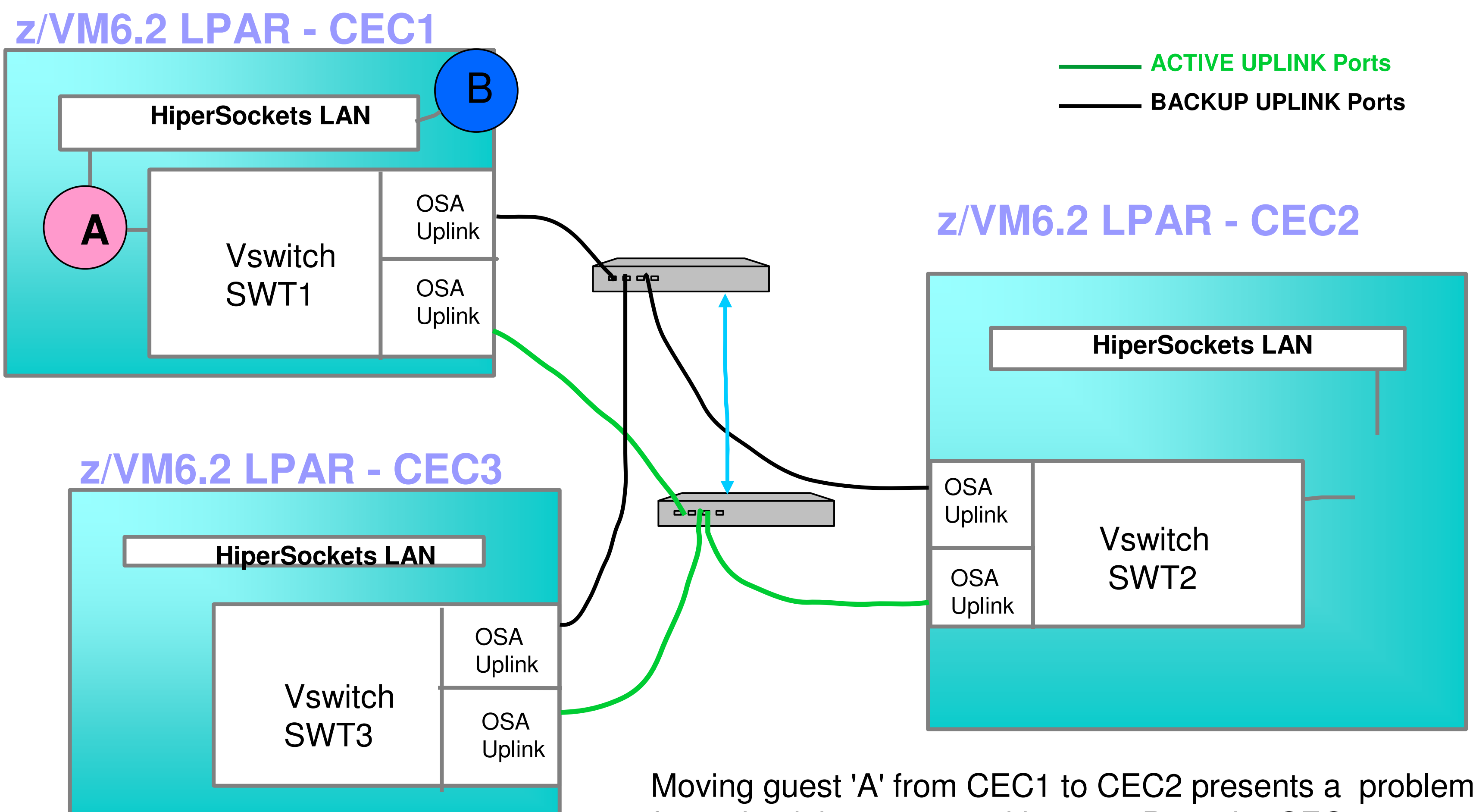
- z/OS does not support the bridging of non managed (zManager) networks, that is IQD configured as EXTERNAL_BRIDGED. For a Bridge HiperSockets (IQD) channel an explicit OSD and or IQD interface must be configured.
- For the IEDN z/OS provides a "Converged IQDX Link". This support provides transparent IQDX connectivity through the AUTOIQDX (GLOBAL CONFIG) which dynamically and transparently configures an IQDX interface under its existing OSX interface.
  - Once configured, Communications Server transparently splits and converges network traffic to this interface.
  - For outbound traffic on this single OSX interface, Communication Server will decide based on the destination IP address to either send the packet on the IQDX link or the OSX link. |
- Virtual switches (QDIO, OSX) deployed as HiperSockets bridges are Layer 2 transports only. z/OS does not support Layer 2 mode for OSD and OSX vNICs. Connectivity through a virtual switch to the bridged HiperSockets LAN is not possible.

Attention: Do not configure z/OS or any other guest that has both HiperSockets and OSA-Express interfaces to forward traffic from one link to the other, this may cause duplicate packets to be generated and in some cases initiate a broadcast storm.

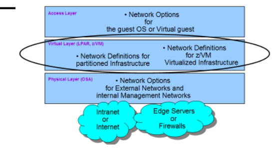


VSWITCH Topology

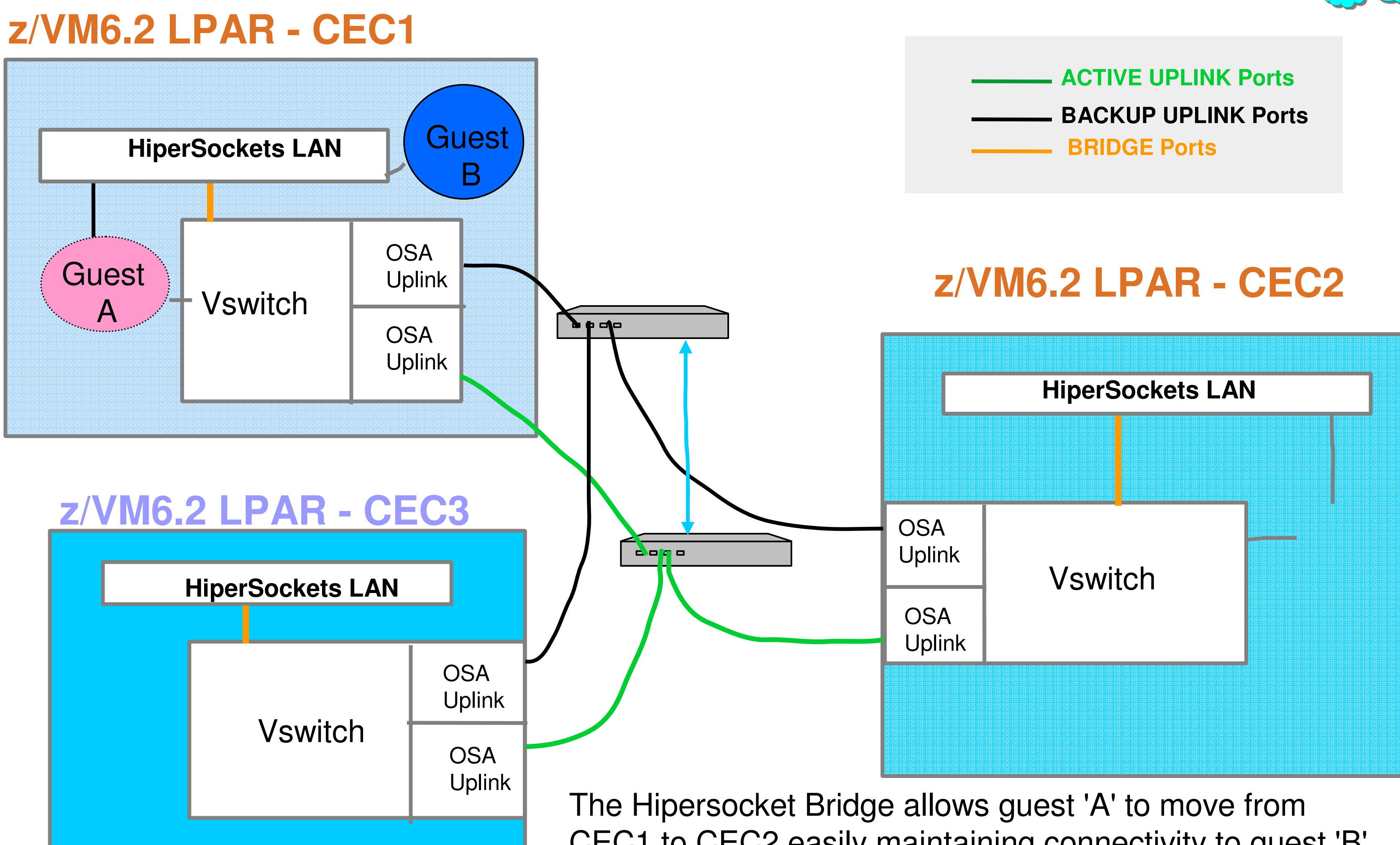
A typical Vswitch topology for multiple CECs. Active and Backup Uplink Ports to redundant Ethernet switches.



Moving guest 'A' from CEC1 to CEC2 presents a problem for maintaining contact with guest 'B' on the CEC1 hipersockets LAN segment.

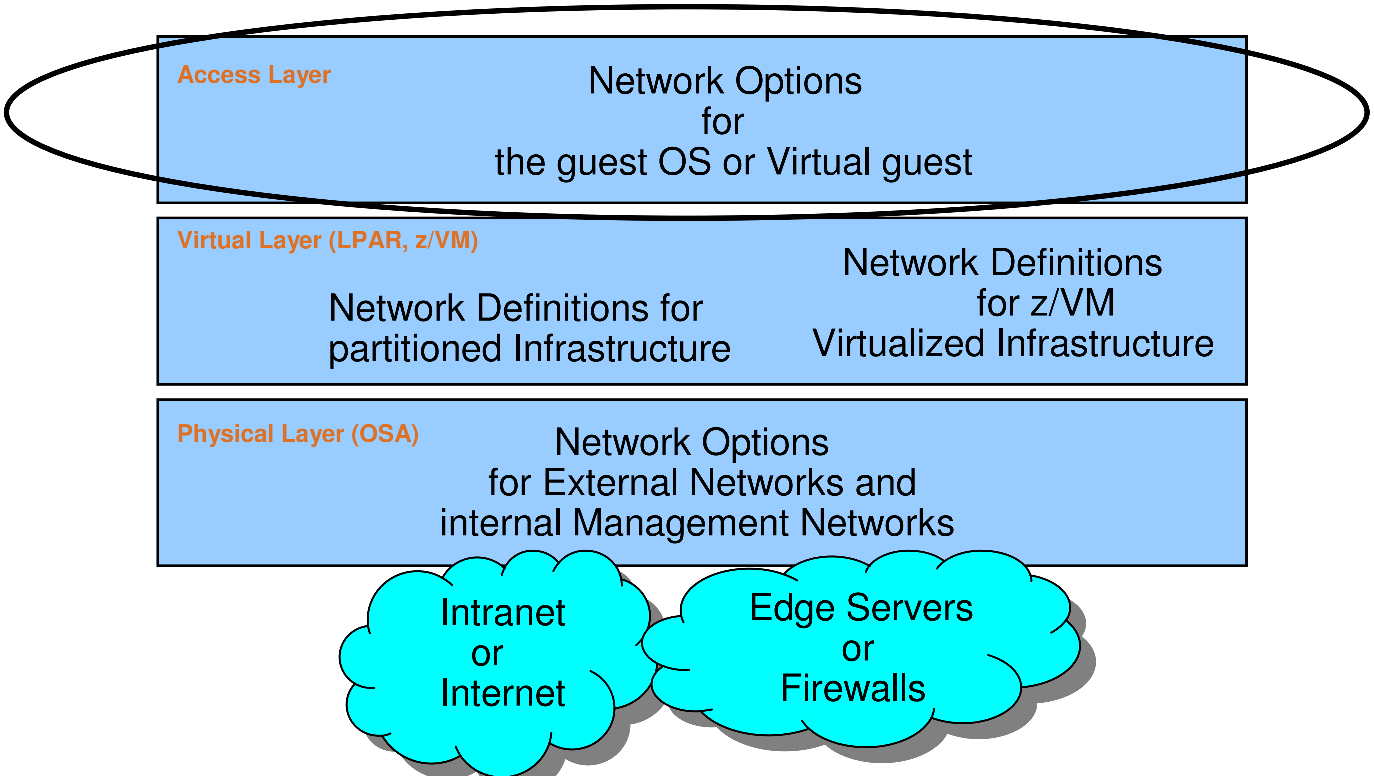


VSwitch – With Hipersocket Bridge



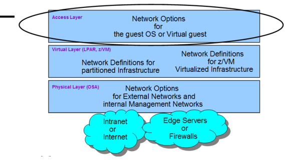
The Hipersocket Bridge allows guest 'A' to move from CEC1 to CEC2 easily maintaining connectivity to guest 'B'

Reference Architecture for Networks with System z

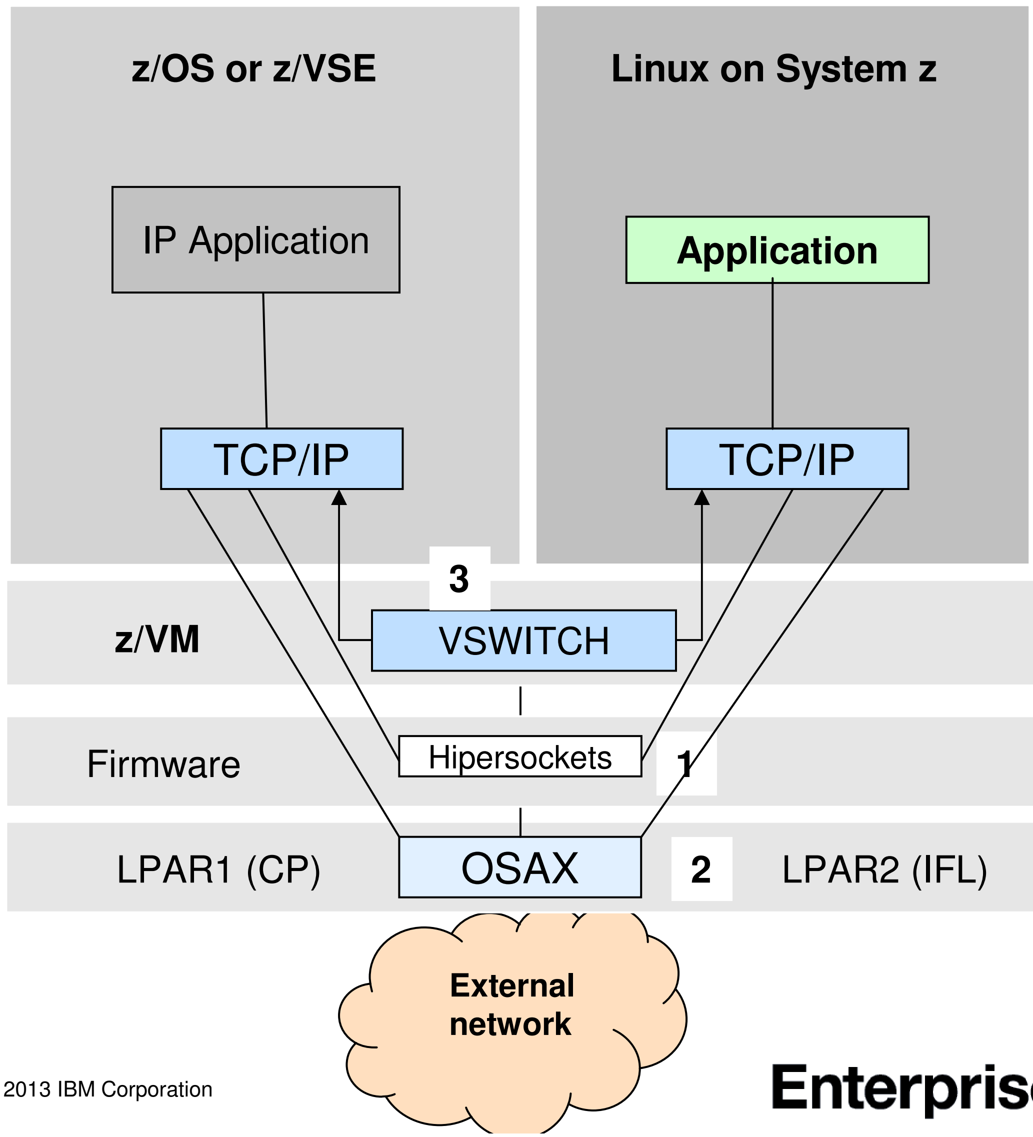


Recommendations for OS and guests running in:

- LPAR:
  - use **Hipersockets** between OS or guests
  - use direct attached (shared) **OSA (Open System Adapter**)
- z/VM :
  - use **Virtual Hipersockets** – great for HA and DR solutions
  - use **Hipersockets Bridge** for DR solutions
  - use **VSWITCH** between guest Systems – advantage in DR solutions
  - use **IUCV** for special network Connections (terminal server)

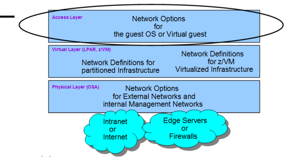


System z Network alternatives from z/VM Guests



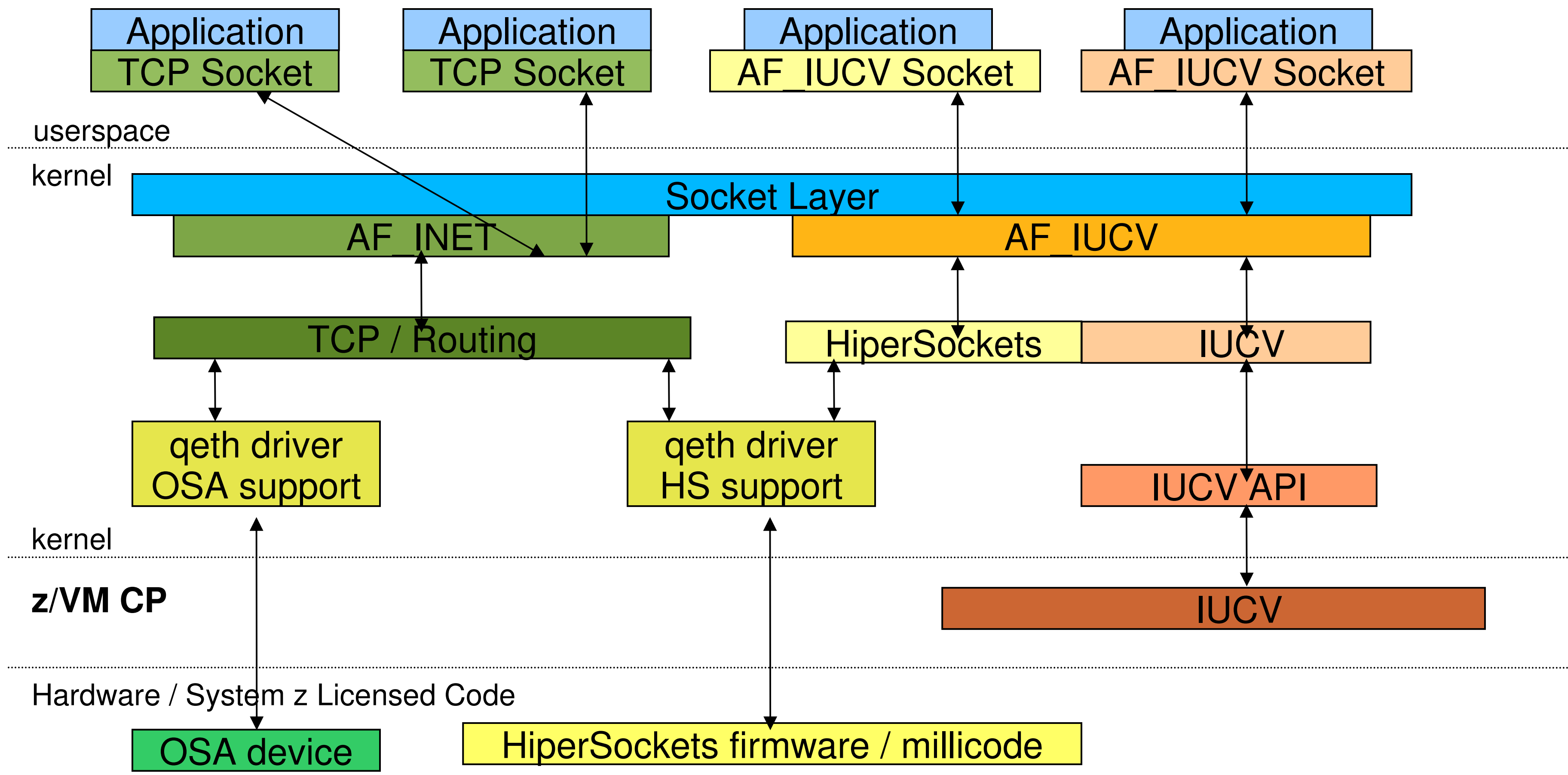
Enterprise2013

z/OS Network Exploitation Support (z/OS 1.13 and higher)

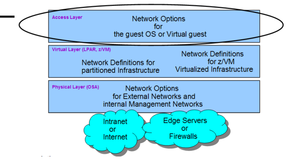


- Provides same functionality as that on the IBM zEnterprise 196
- IBM zEnterprise Unified Resource Manager
- Network and Performance Management
- Intranode Management Network (INMN)
- Intra ensemble data network (IEDN)
- OSA-Express3 and OSA-Express4S Inbound Workload queuing (IWQ)
  – Large Send for IPv6,
  – Inbound Workload queuing (IWQ) for Enterprise Extender traffic
- OSA-Express4S checksum offload for IPv6
- OSA-Express4S checksum offload for LPAR to LPAR traffic (both IPv4 and IPv6)
- HiperSockets optimization for intraensemble data networks (IEDN)

Linux on z network options and interfaces



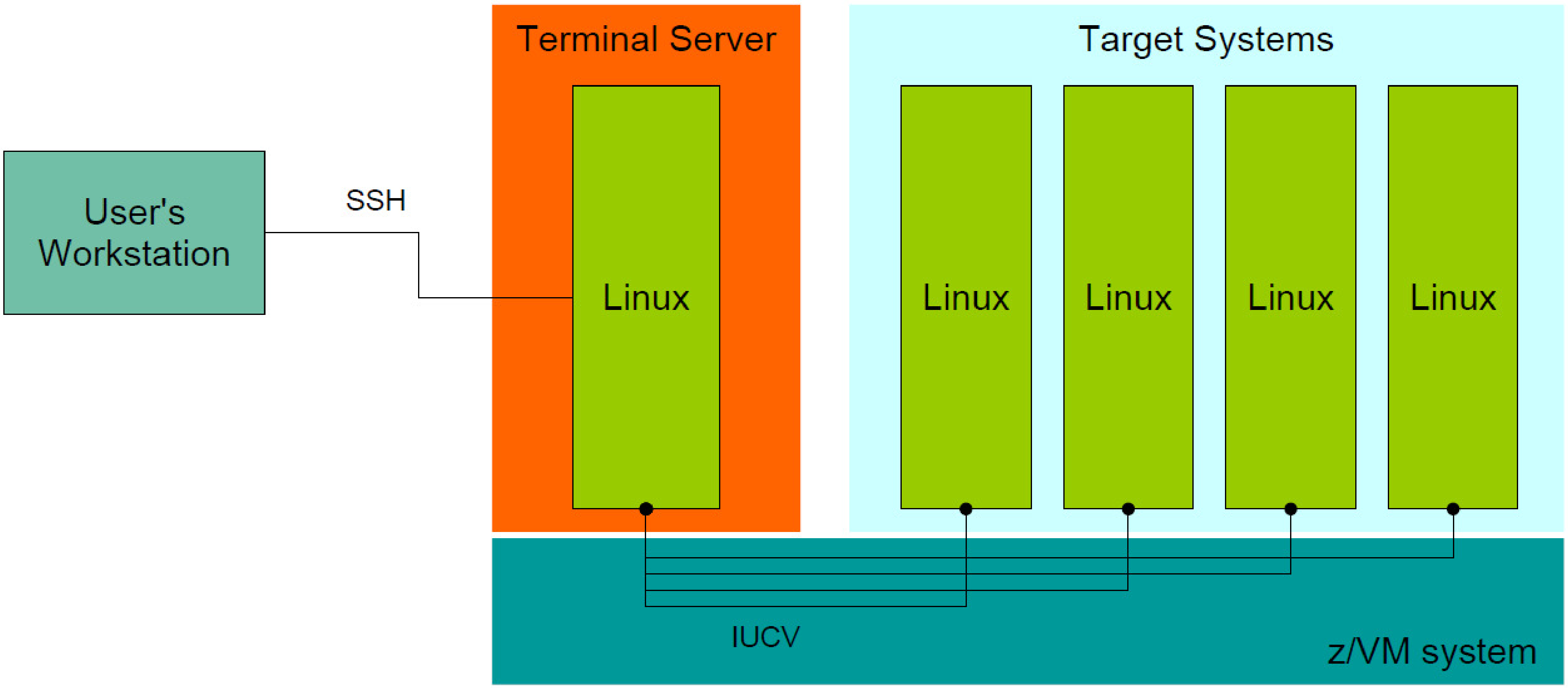
Enterprise2013



z/VM IUCV – Inter-User Communication Vehicle

- 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 (for instance VSE)
  - 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 without TCP



Enterprise2013

Summary Network Recommendations

- Which connectivity to use:
  - 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) – 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.

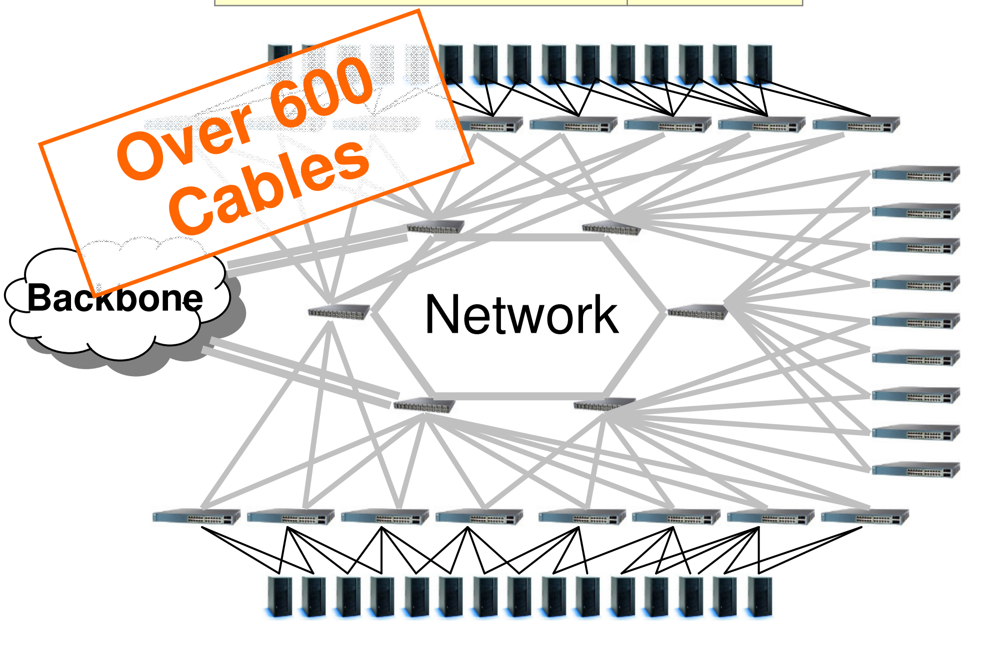
System z and zEnterprise Network value points

- **Network Simplification ("Network in a Box")**
- **Central point of Management (Unified Resource Manager via the HMC/SE)**
- **Single physical network and zBX "package" (physical network integration)**
- **Reduced network path length; reduced number of hops**
- **Secured internal 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 with integrated HA**
- **Special purpose dedicated**
  - **data network & OSA-Express**
  - **potential for reduced network encryption and HW encryption support**

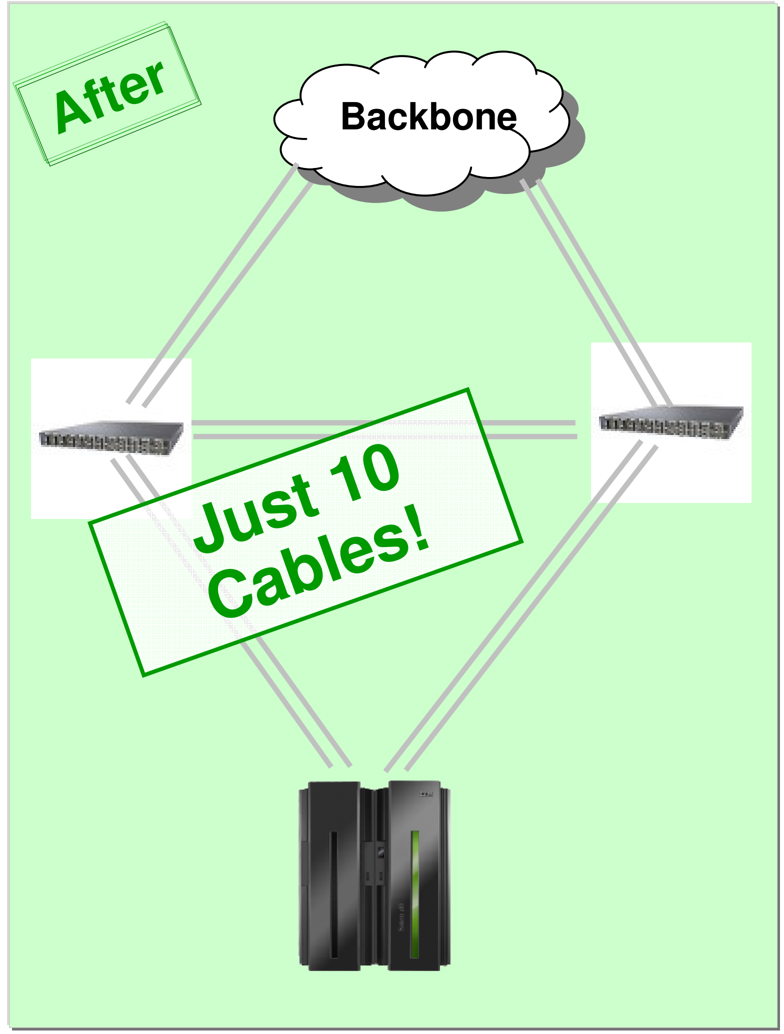
Insurance Company Consolidated 292 Servers to a z10

Before

3560E-24TD	8
3560E-24TD	17
3560E-12D	6
50 Ft UTP Cable	584
10GB Eth Fiber Cable	60



The diagram only shows **30** of **292** servers

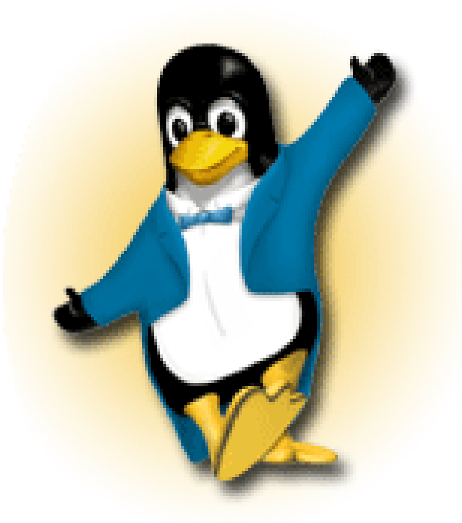


Data is based on real client opportunity and on internal standardized costing tools and methodologies.
Client results will vary by types of workloads, technology level of consolidated servers, utilization factor, and other implementation requirements. Savings will vary by client.

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/
- **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

THE Open GROUP
Master
Certified IT Architect

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

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