

HiperSockets for System z

Newest Functions

Alan Altmark – Senior Managing z/VM and Linux Consultant
IBM Systems Lab Services and Training

Alexandra Winter – HiperSockets Architect
IBM System z Firmware Development



Session 13206



Trademarks

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

* 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.

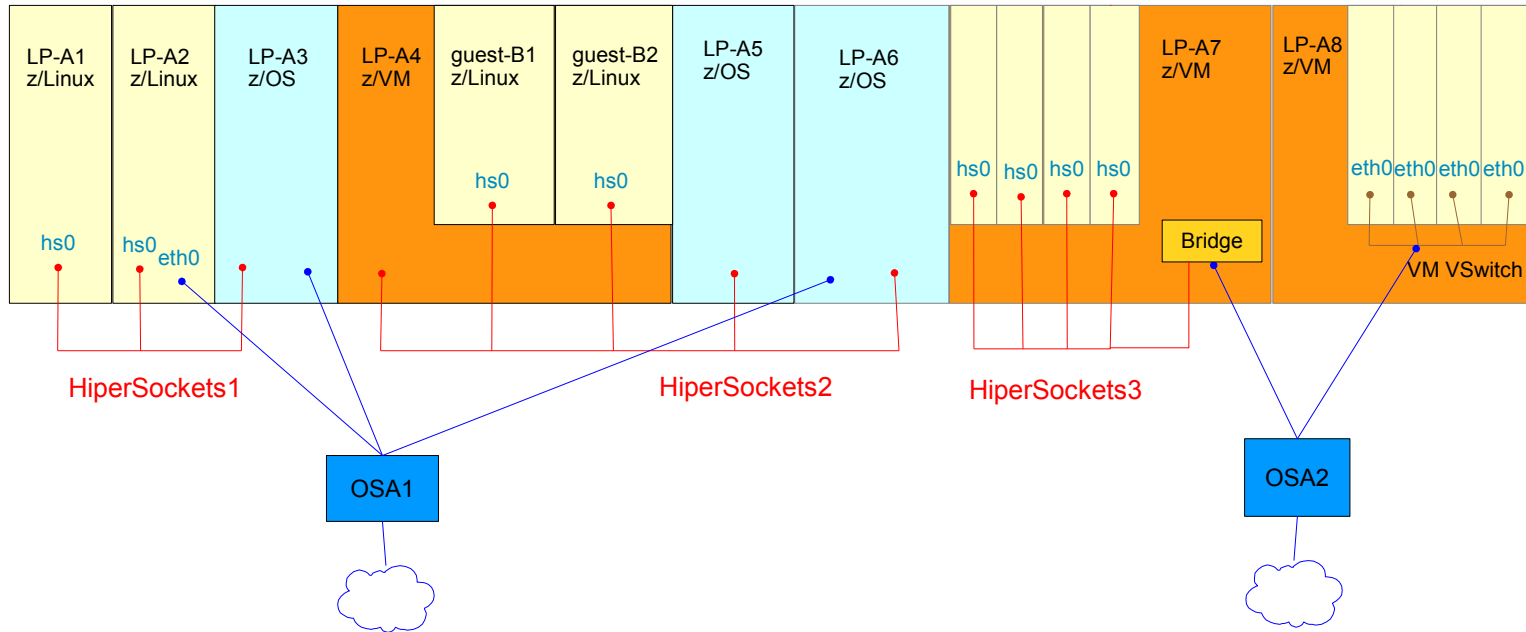
This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g. zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the "Authorized Use Table for IBM Machines" provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html ("AUT"). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

Agenda

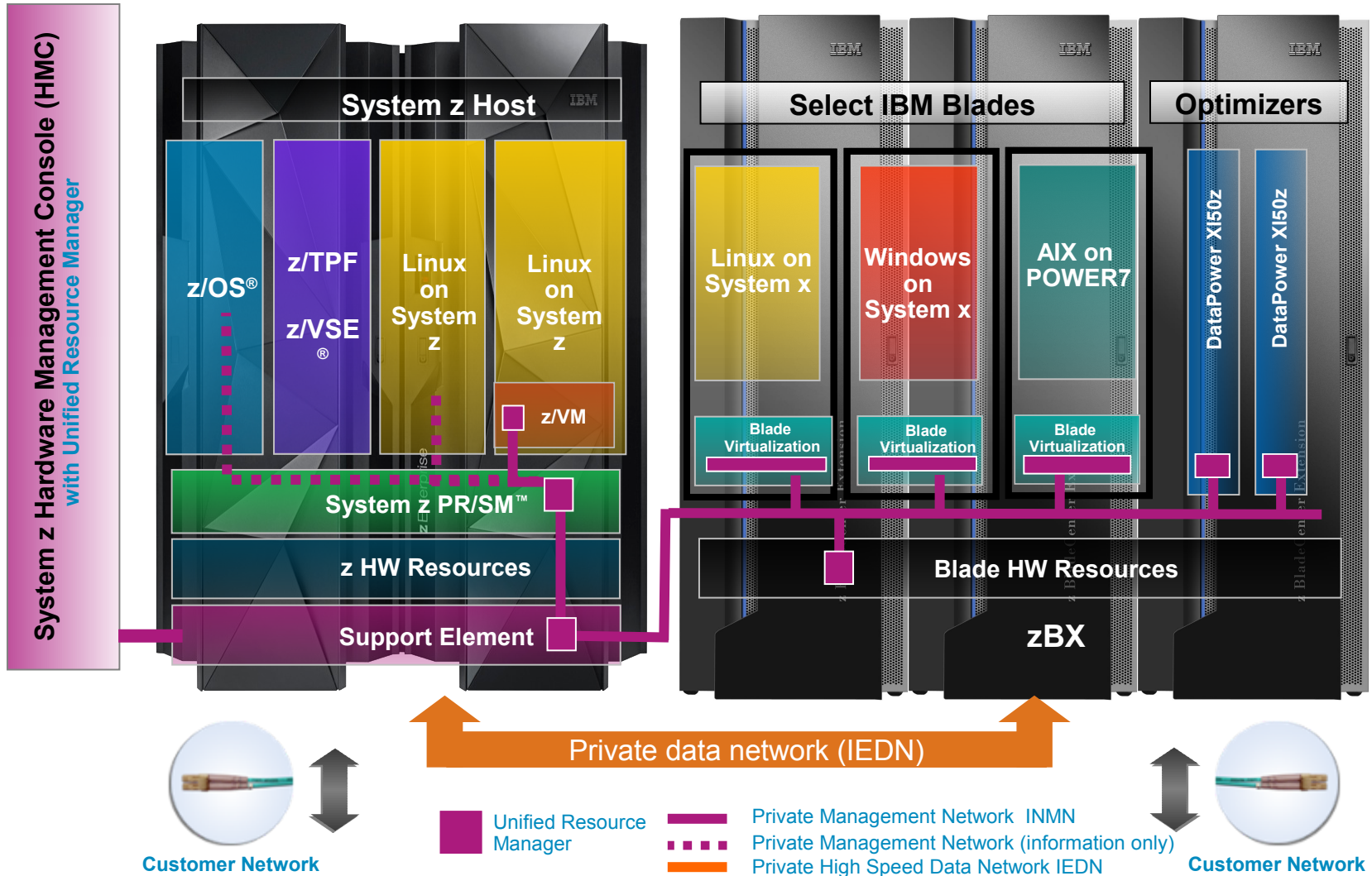
- ▶ Review
- ▶ Features and functions
 - ▶ QDIO ASSIST
 - ▶ Layer 2 vs Layer 3
 - ▶ **VM vSwitch HiperSockets Bridge**
 - ▶ System z Network Virtualization Manager
 - ▶ IEDN
 - ▶ **Completion Queues**
 - ▶ ...
- ▶ Where to find more information



System z Networking Review



zBX zEnterprise BladeCenter Extension

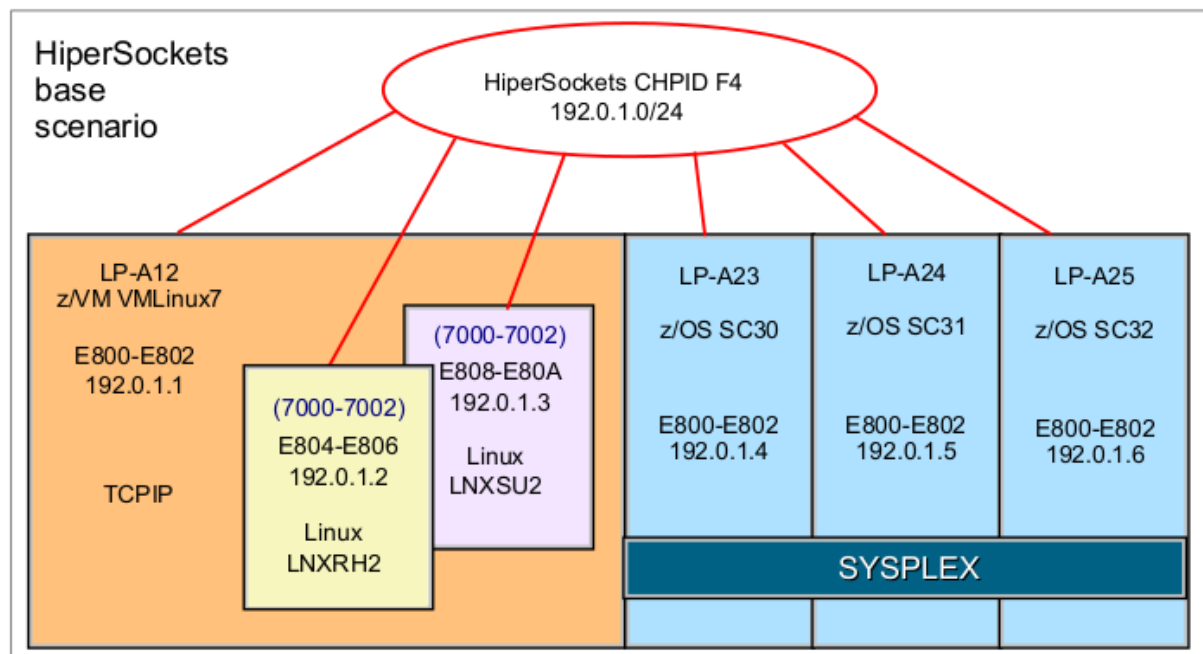


HiperSockets Features and functions

- ▶ What is available to you?
- ▶ What is new?



Dedicated QDIO devices for VM guests



Source: HiperSockets Implementation Guide www.redbooks.ibm.com

◆ QDIO ASSIST / QEBSM (also for OSA, FCP)

- ◆ interface definition with VM Hypervisor
 - ◆ (1:1 mapping of virtual devices to real devices)
- ◆ support in guest OS required
 - ◆ available in zLinux and zVSE
- ➔ direct pass-through for data transfer, without interception to the VM Hypervisor
- ➔ delivery of interrupts to the VM guest without interception to the VM Hypervisor

Layer 3 versus Layer 2

- ◆ A HiperSockets VNIC can be defined by the device driver either as
 - ◆ **Layer 2** device (MAC addressing, ethernet frames) or as
 - ◆ **Layer 3** device (IPv4 or IPv6)
- ◆ L2 and L3 devices can be defined on the same channel, but cannot communicate with each other!
- ◆ Only L2 devices can be activated on IQDX / IEDN and External Bridge Channels

Miscellaneous features

◆ **Multiple Write**

- ◆ exploited by z/OS, send multiple output buffers at one time

◆ **Network Traffic Analyzer**

- ◆ set one IQD VNIC in 'promiscuous mode' and get a copy of all traffic on this channel
- ◆ Authorization and 'filtering' on SE required
 - ◆ Which LPAR is authorized to run a NTA?
 - ◆ Traffic between which LPARs will be sniffed?
- ◆ Linux exploitation for `tcpdump` is available
(see ZSQ03039USEN white paper)

◆ **VLAN**

- ◆ VLAN support available
- ◆ device driver defines which VLAN this device is allowed to use
- ◆ out-of-band VLAN management only for IQDX (zManager)

◆ **Network concentrator**

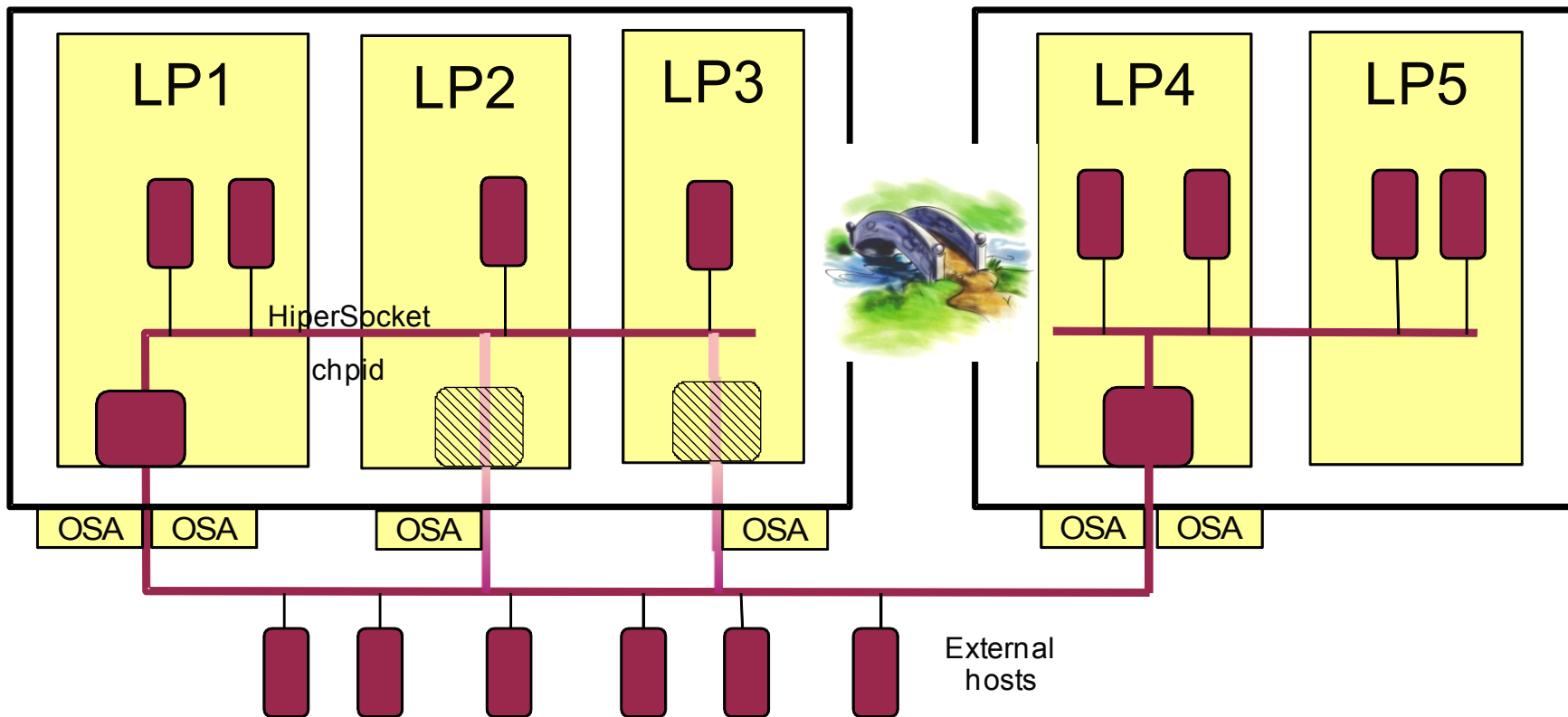
- ◆ Linux tool to connect L3 IPv4 HiperSockets to external network
 - ◆ see “Linux on System z, Device Drivers, Features, and Commands”
www.ibm.com/developerworks
- ◆ see also VM Bridge

z/VM VSWITCH HiperSockets Bridge

- Connect HiperSocket LAN to ethernet LAN without a router
 - Same subnet as ethernet LAN

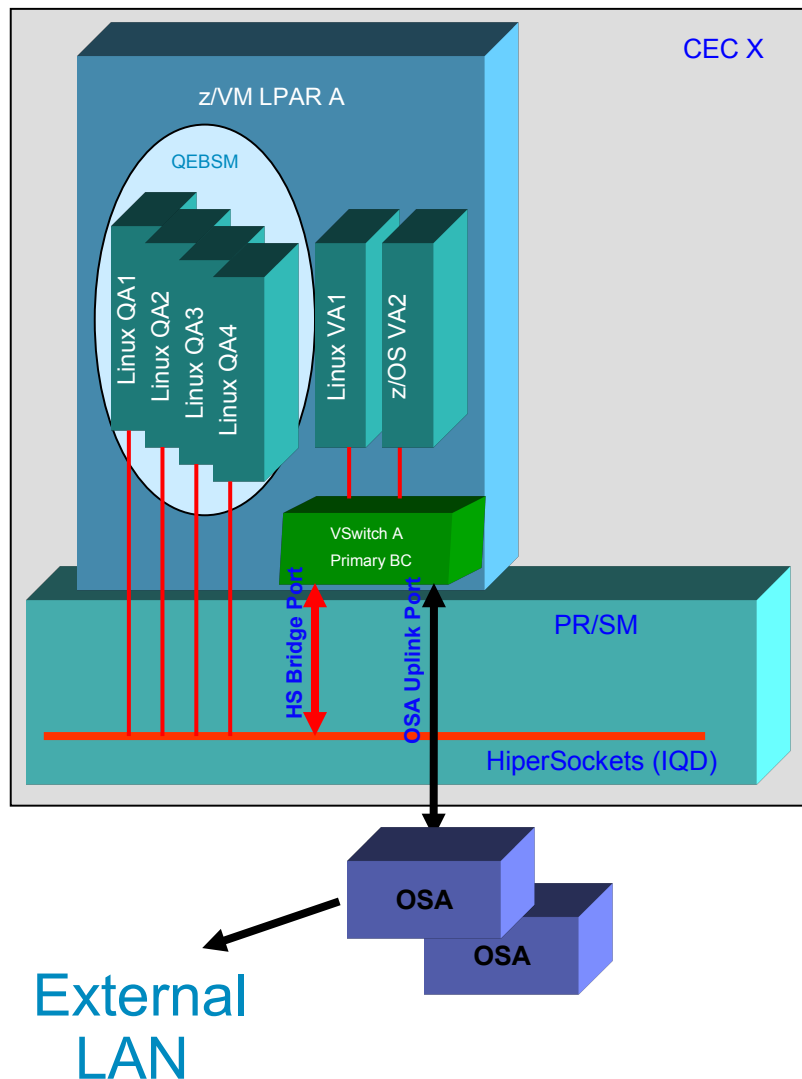
- Full redundancy
 - Up to 5 bridges per CPC (CEC)
 - Automatic failover with optional failback
 - Each bridge can have more than one OSA uplink (typical)

z/VM VSWITCH HiperSockets Bridge



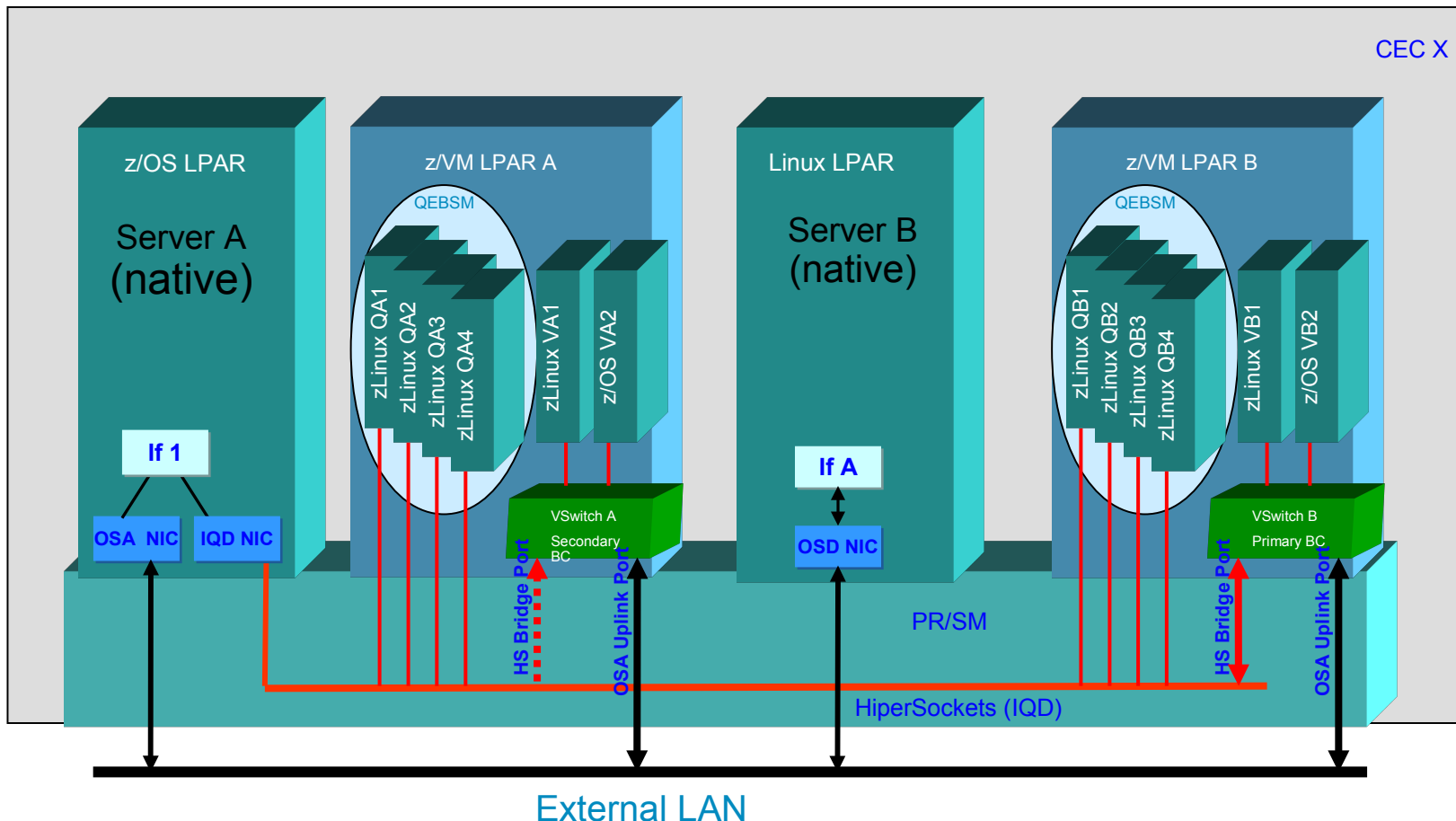
- One active bridge per HiperSocket CHPID
- Path MTU discovery support
 - Large frames inside
 - Small frames outside

z/VM VSWITCH HiperSockets Bridge



- Layer 2 only
 - No transport mode conversions
- Bridges both **IEDN** and **Customer** networks
- Only traffic to/from QEBSM NICs will flow over the bridge
- Guests QA1,QA2, QA3 and QA4 have real (*dedicated*) QEBSM connections to HS CHPID.
 - Requires almost no z/VM involvement
 - Bridged by default (if bridge is defined)
- Guests VA1 and VA2 have virtual NIC connections through VSWITCH A
 - Optimum performance for guests that are not deployed with QEBSM on z/VM. Eliminates “shadow queue” overhead
 - Connectivity to HS and external LAN segments
- OSA uplink port BAU
 - No changes in current support

z/VM VSWITCH HiperSockets Bridge



- ◆ One active bridge port per IQD channel; max of 1 primary and 4 secondary
- ◆ native LPARs are not bridged
- ◆ z/OS uses concept of converged devices (IEDN only)

z/VM VSWITCH HiperSockets Bridge

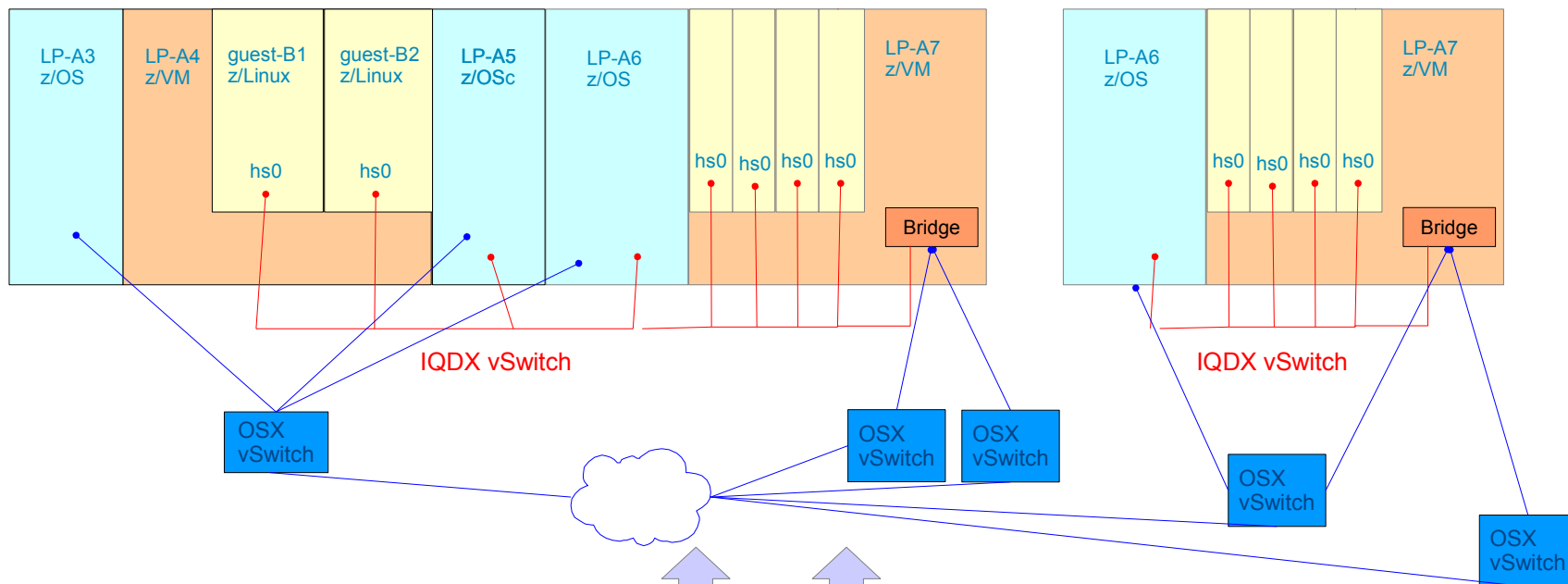
```
DEFINE VSWITCH switch
```

```
(all the traditional keywords)
```

```
ETHERNET BRIDGEPORT RDEV hipersocket_rdev [PRIMARY]
```

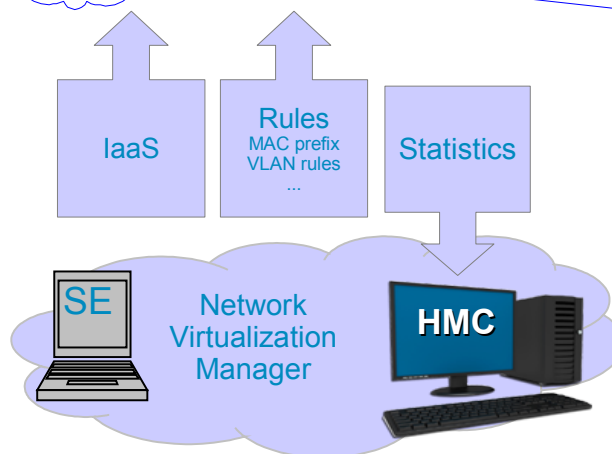
- The HiperSocket device must be on a CHPID defined in the IOCP with CHPARM=x4
- CP DEFINE CHPID **EXTERNAL_BRIDGED** is available for dynamic I/O

System z Network Virtualization Manager (z196+)



IEDN

- ◆ single flat L2 network
- ◆ connects CECs and zBXs in an ensemble
- ◆ separation via VLANs
- ◆ z/VM 6.1 and 6.2



IEDN / IQDX

- ◆ Only one IQDX channel per CEC
- ◆ Layer 2 only
- ◆ VLAN mandatory
- ◆ Bridged via z/VM bridges to OSX (Linux as z/VM guest) or merged interface with OSX vNIC (z/OS)
- ◆ Managed by Network Virtualization Manager (NVM) component of zManager / URM
 - ◆ MAC address management (prefix)
 - ◆ VLAN management
 - ◆ Monitoring
 - ◆ definition of z/VM bridges to OSX / IEDN

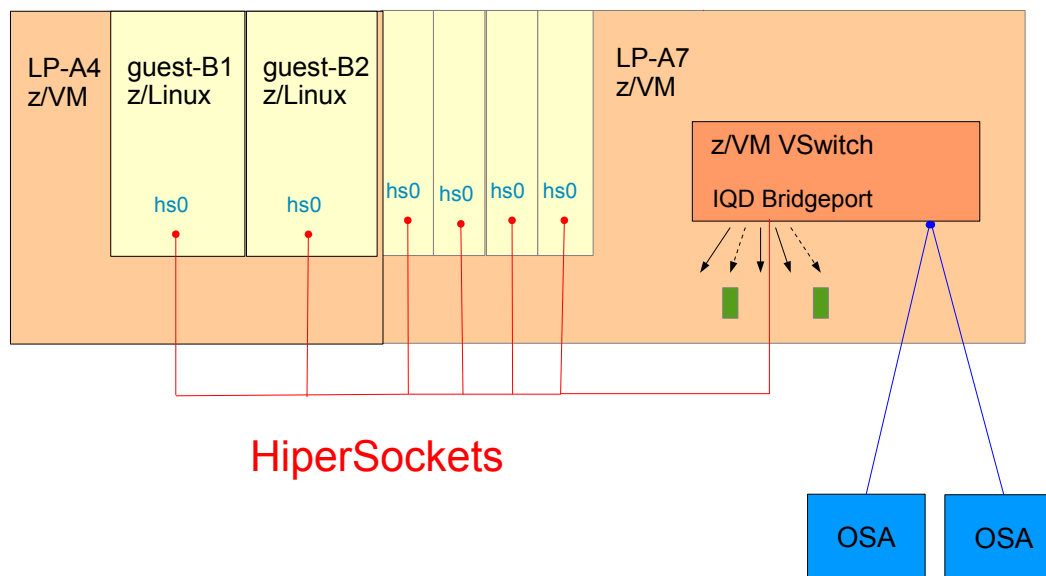
Completion Queues

- ◆ HiperSockets messages are sent **synchronously, in-order and reliably**
- ◆ If the target has no free input buffers, error is delivered to the sender
 - Sender can retry, but does not know when new target buffers are available
 - Performance impact!
- ◆ OSA has the capability to buffer 512 packets. In a high sharing environment OSA may perform better than Hipersockets, packet buffering may be a reason.

- ◆ **Completion queues:**
 - ◆ **Deliver synchronously if possible, asynchronously if necessary**
 - ◆ Messages remain at sender
 - ◆ When target provides free input buffers, messages are delivered and completion messages are reported to sender
 - ◆ IBM zEnterprise System 196 (z196) and later

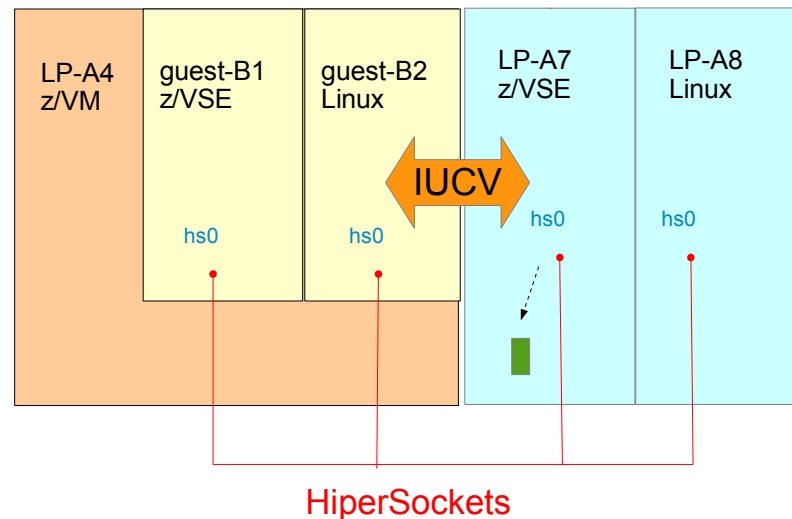
Completion Queue exploitation

- ◆ Exploitation possible per server only sender needs support
- ◆ Amount of buffered messages counted by **Resource Measurement Facility (RMF)** and **NVM Monitoring** as 'unavailable receive buffers'
- ◆ Exploited today by **z/VM bridge ports**



Completion Queue exploitation

- ◆ Exploited today by **IUCV Sockets over HiperSockets** (Linux, z/VSE)
 - ◆ Inter-User Communication Vehicle (IUCV) is traditionally provided by z/VM for communication between two z/VM guests in same z/VM LPAR
 - ◆ point-to-point connection
 - ◆ Used to provide z/VSE Fast Path to Linux (LFP)
- ◆ **IUCV over HiperSockets**
 - ◆ flow control by completion messages
 - ◆ Available for z/VM guests and native LPARs
 - ◆ Available for communication between z/VM guests in different z/VM LPARs



Functional Matrix

HiperSockets Features	z/OS	z/VM	Linux	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
▪ 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)	No	Yes	Yes	No
▪ Merged IEDN interfaces (OSX / IQDX)	Yes	No	No	No
▪ Virtual Switch Bridge Support	No	Yes	No	No
▪ IUCV over HiperSockets	No	No	Yes	Yes
▪ Completion Queue	No	Yes	No	Yes

HiperSockets CHPARM

Maximum Frame Size / Maximum Transfer Unit:

CHPID Parameter	MFS	max. MTU
CHPARM=0x (default)	16kByte	8kByte
CHPARM=4x	24kByte	16kByte
CHPARM=8x	40kByte	32kByte
CHPARM=Cx	64kByte	56kByte

- ◆ Allows optimization per HiperSockets LAN for small packets versus large streams
- ◆ MFS == size of 1 input buffer
- ◆ MTU defined for device driver \leq max. MTU in CHPARM;
device driver may put multiple frames in a HiperSockets message

Channel flavor:

CHPID Parameter	Usage
CHPARM=x0 (default)	Traditional HiperSockets
CHPARM=x2	HiperSockets for IEDN (IQDX)
CHPARM=x4	HiperSockets for External Bridge

More information

- ◆ www.ibm.com/developerworks
 - ◆ “Linux on System z, Device Drivers, Features, and Commands”

- ◆ IBM Redbooks
 - ◆ <http://www.redbooks.ibm.com>
 - HiperSockets Implementation Guide, SG24-6816
 - IBM System z Connectivity Handbook, SG24-5444
 - I/O Configuration Using z/OS HCD and HCM, SG24-7804
 - Building an Ensemble Using Unified Resource Manager, SG24-7921

- ◆ System z HiperSockets web page:
 - ◆ <http://www.ibm.com/systems/z/hardware/networking/products.html>

- ◆ IBM ATS Technical Documents:
 - ◆ <http://www.ibm.com/support/techdocs>

- ◆ IBM Information Center
 - ◆ <http://www.ibm.com/support/documentation/us/en>



THANK YOU



Session 13206