#SHAREorg

SHARE
Technology · Connections · Results

13196: HiperSockets on System zEC12 – Overview

Alexandra Winter – HiperSockets Architect

IBM System z Firmware Development

(Presented by Linda Harrison)

16 August 2013

SHARE in Boston

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.

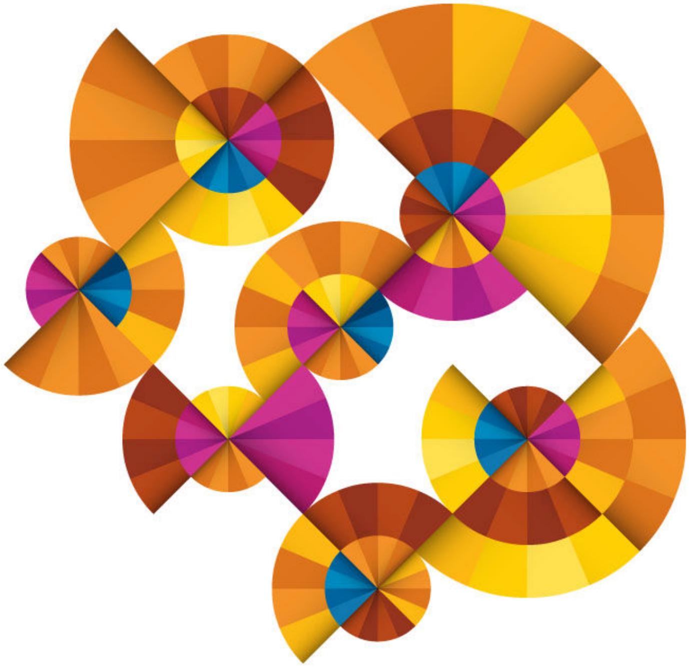
Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

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.

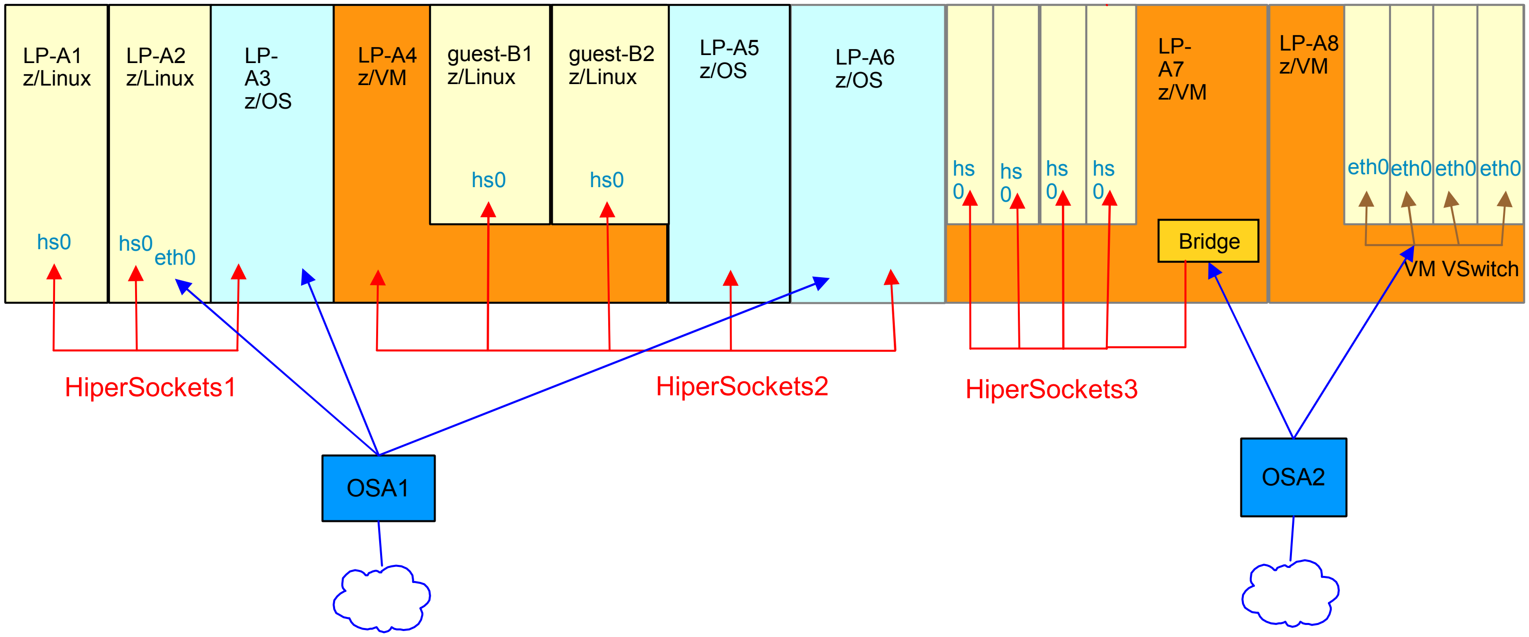
Agenda

- System z Networking

- HiperSockets Insights
  - Configuration
  - **How does it work?**
  - What are the implications?
  - **Performance considerations**

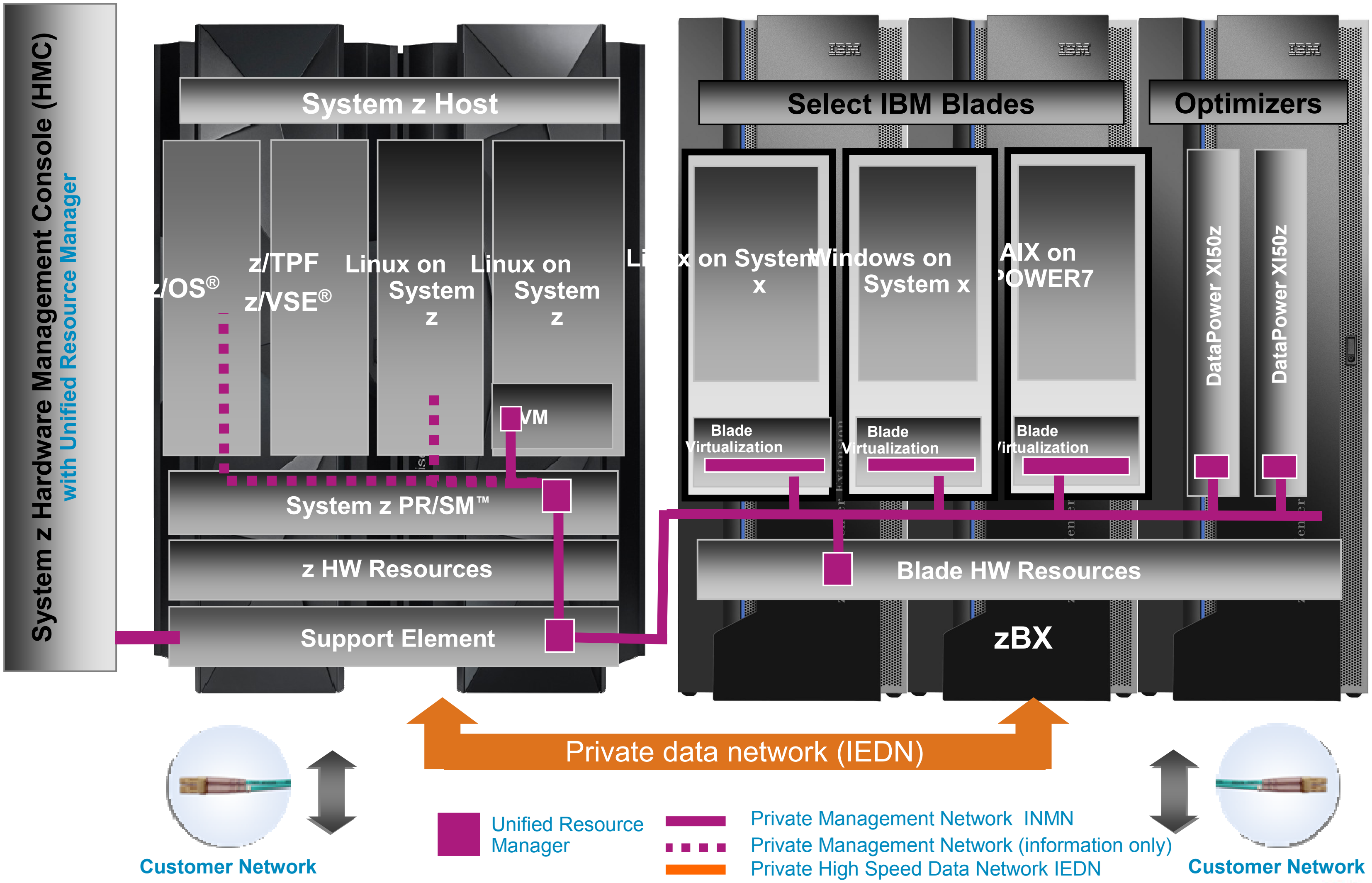


System z Networking



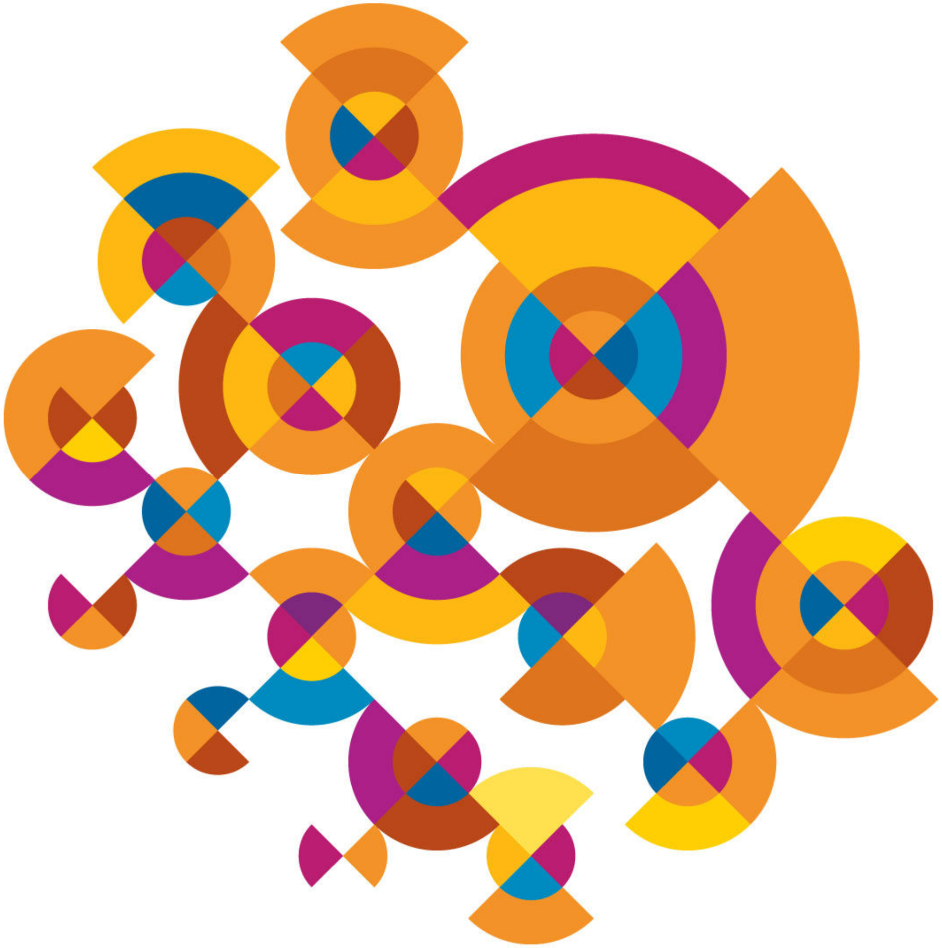
Complete your sessions evaluations online at SHARE.org/BostonEval

zBX zEnterprise BladeCenter Extension



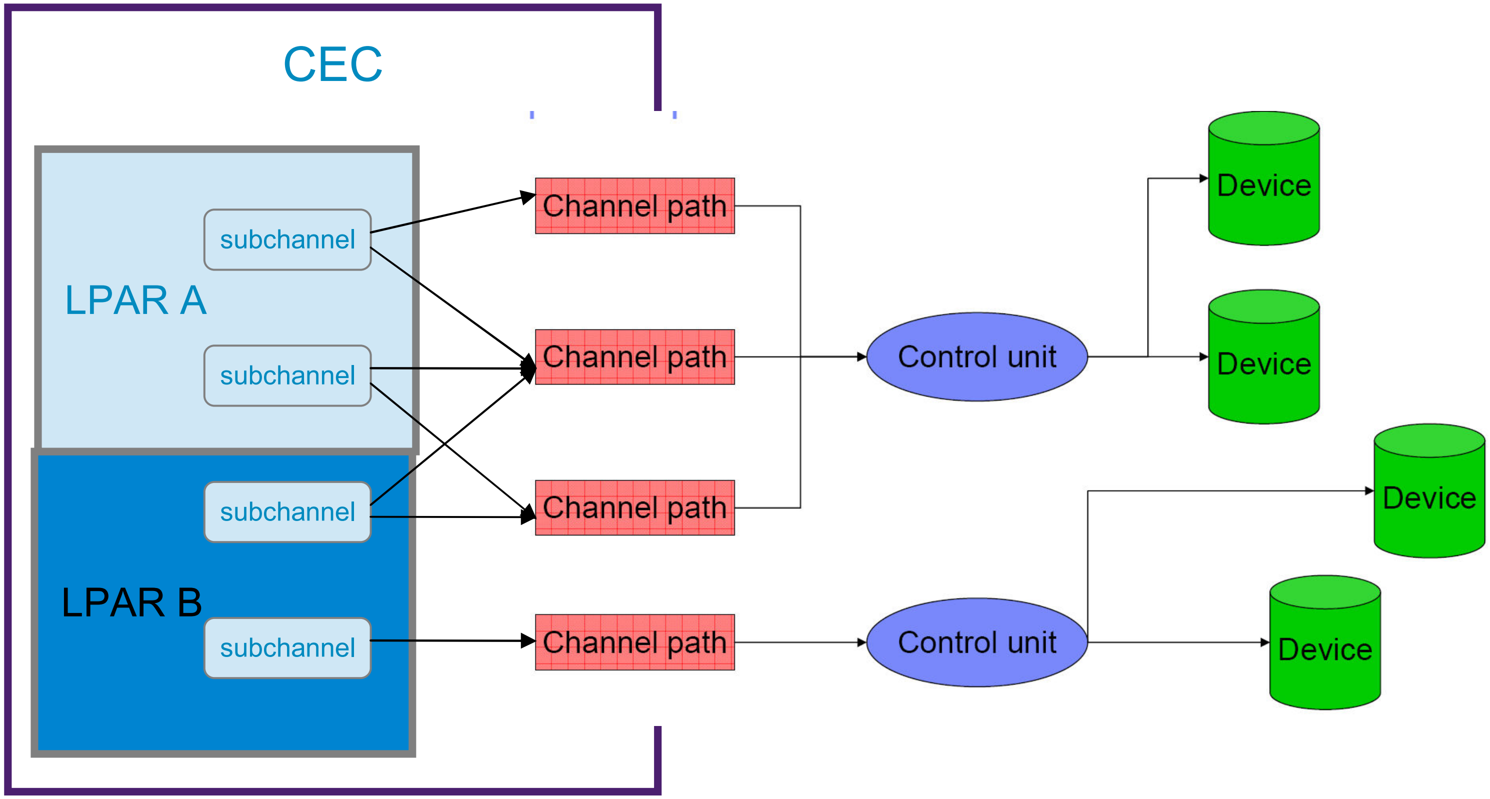
HiperSockets Insights

- How does it work?
- Why does it matter?



Complete your sessions evaluations online at SHARE.org/BostonEval

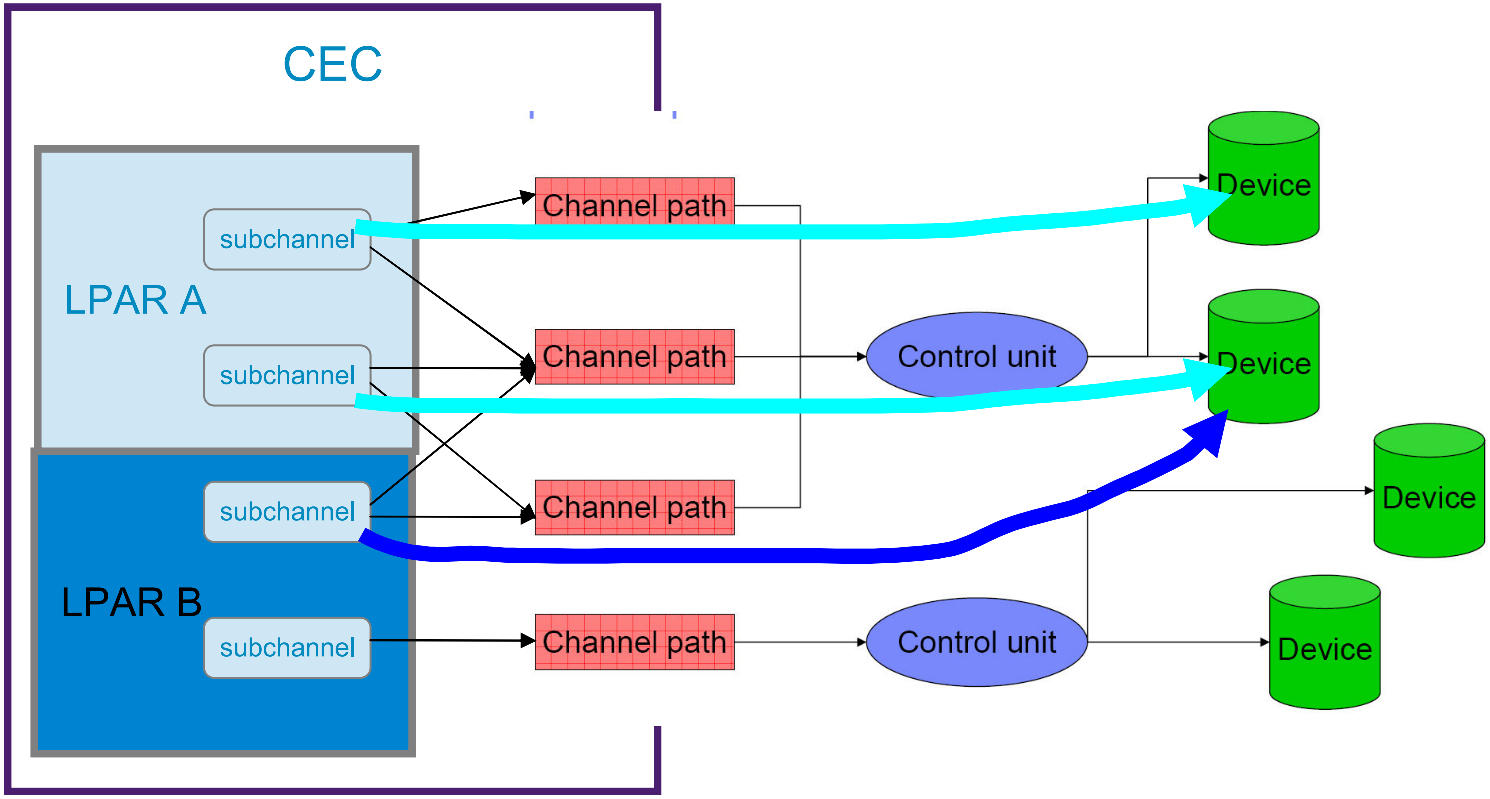
Traditional System z I/O model



IOCDS:

```
CHPID PCHID=240,PATH=(CSS(0),40),TYPE=FC,PART=(LP1, LP3)
CNTLUNIT CUNUMBR=240,UNITADD=((00,4)),UNIT=FC,PATH=((CSS(0),40,41))
IODEVICE ADDRESS=(2C00,4),CUNUMBR=240,UNIT=FC,UNITADD=00,              *
               STADET=Y,SCHSET=1
```

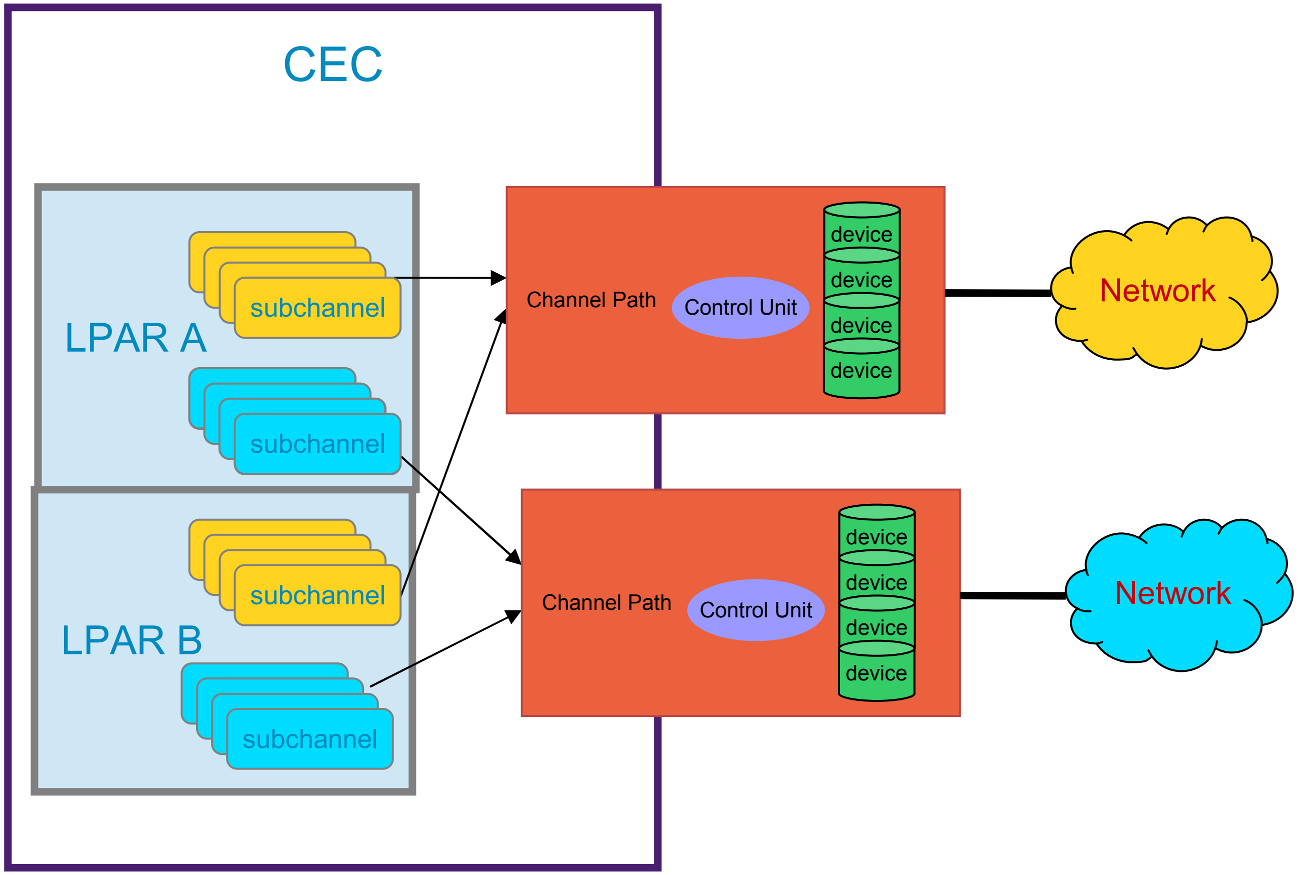
Traditional System z I/O model



IOCDS:

```
CHPID PCHID=240,PATH=(CSS(0),40),TYPE=FC,PART=(LP1, LP3)
CNTLUNIT CUNUMBR=240,UNITADD=((00,4)),UNIT=FC,PATH=((CSS(0),40,41))
IODEVICE ADDRESS=(2C00,4),CUNUMBR=240,UNIT=FC,UNITADD=00,          *
               STADET=Y,SCHSET=1
```

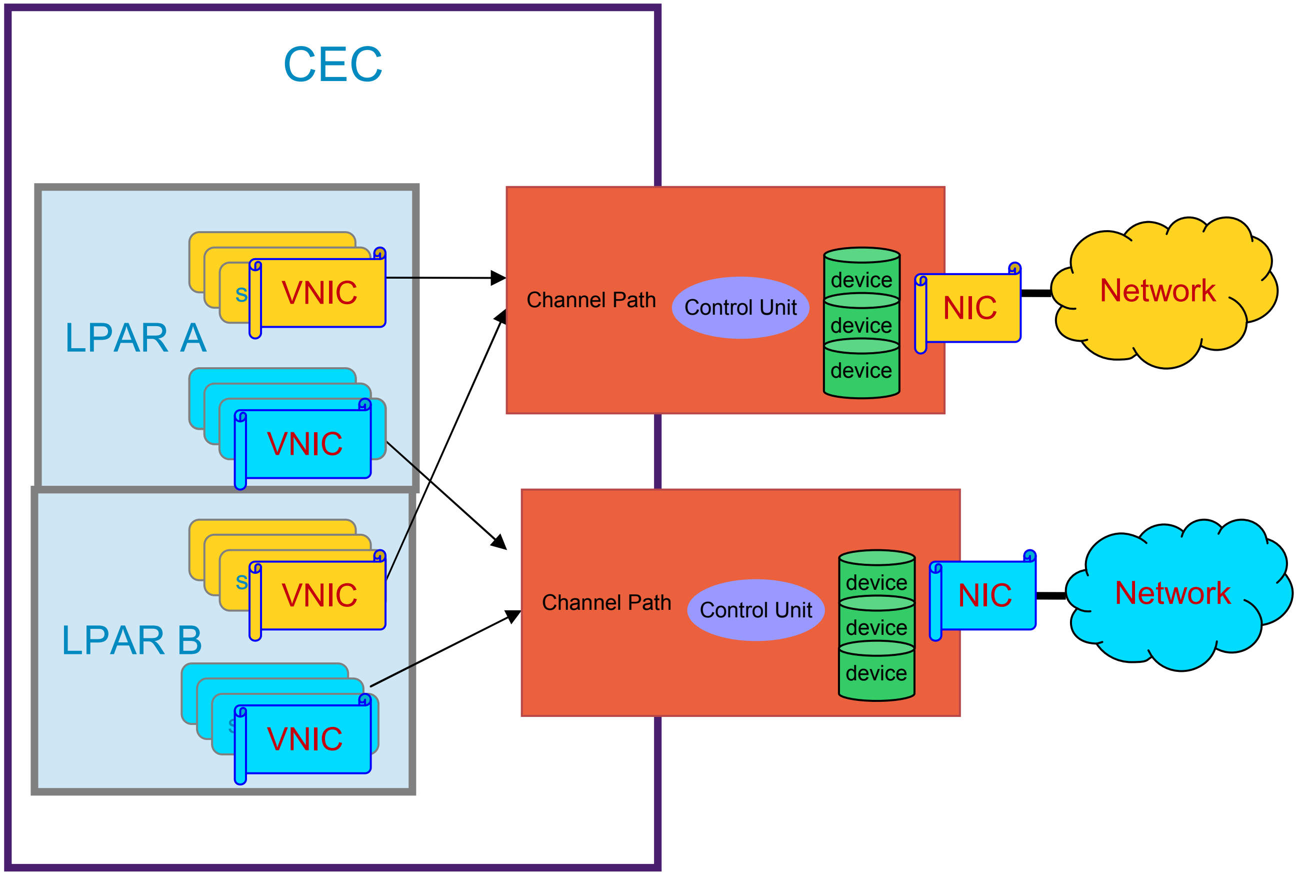
Network Channels



IOCDS:

```
CHPID PCHID=272,PATH=(CSS(0),72),TYPE=OSC,PART=(LP1, LP3)
CNTLUNIT CUNUMBR=272,PATH=(CSS(0),72),UNIT=OSA
IODEVICE ADDRESS=(2720,4),CUNUMBR=272,UNIT=OSA,UNITADD=00
```

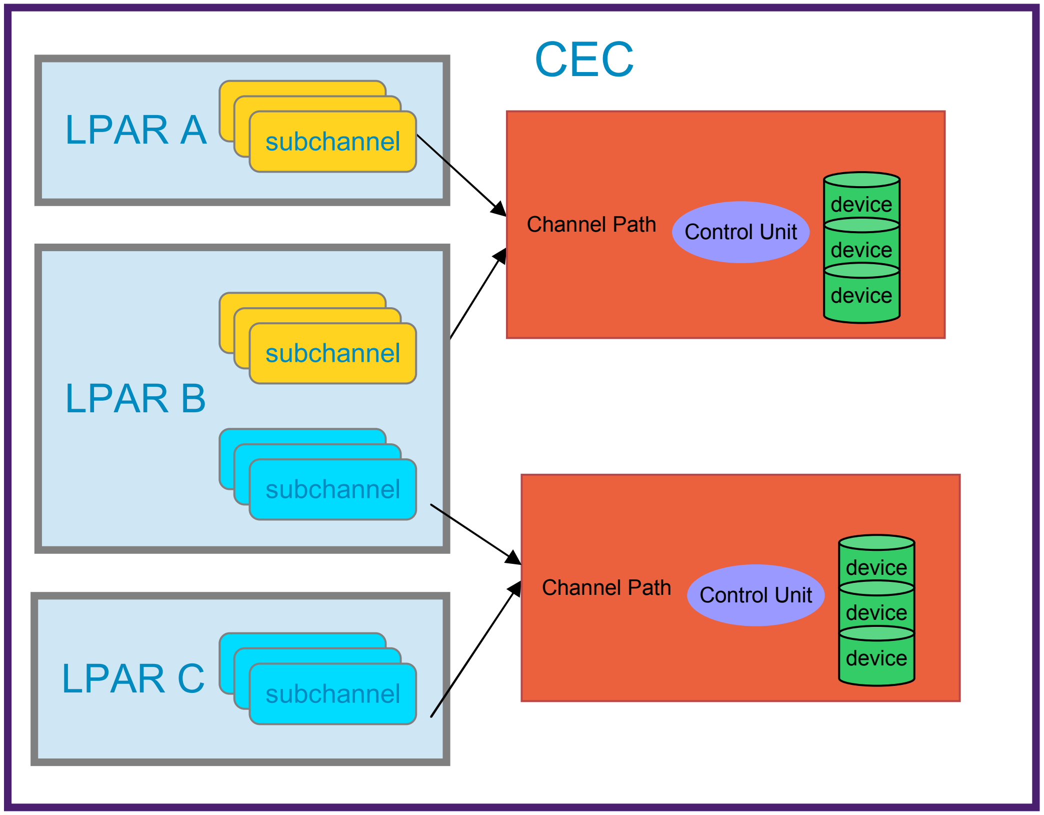
Network Channels



IOCDS:

```
CHPID PCHID=272,PATH=(CSS(0),72),TYPE=OSC,PART=(LP1, LP3)
CNTLUNIT CUNUMBR=272,PATH=(CSS(0),72),UNIT=OSA
IODEVICE ADDRESS=(2720,4),CUNUMBR=272,UNIT=OSA,UNITADD=00
```

HiperSockets Channels

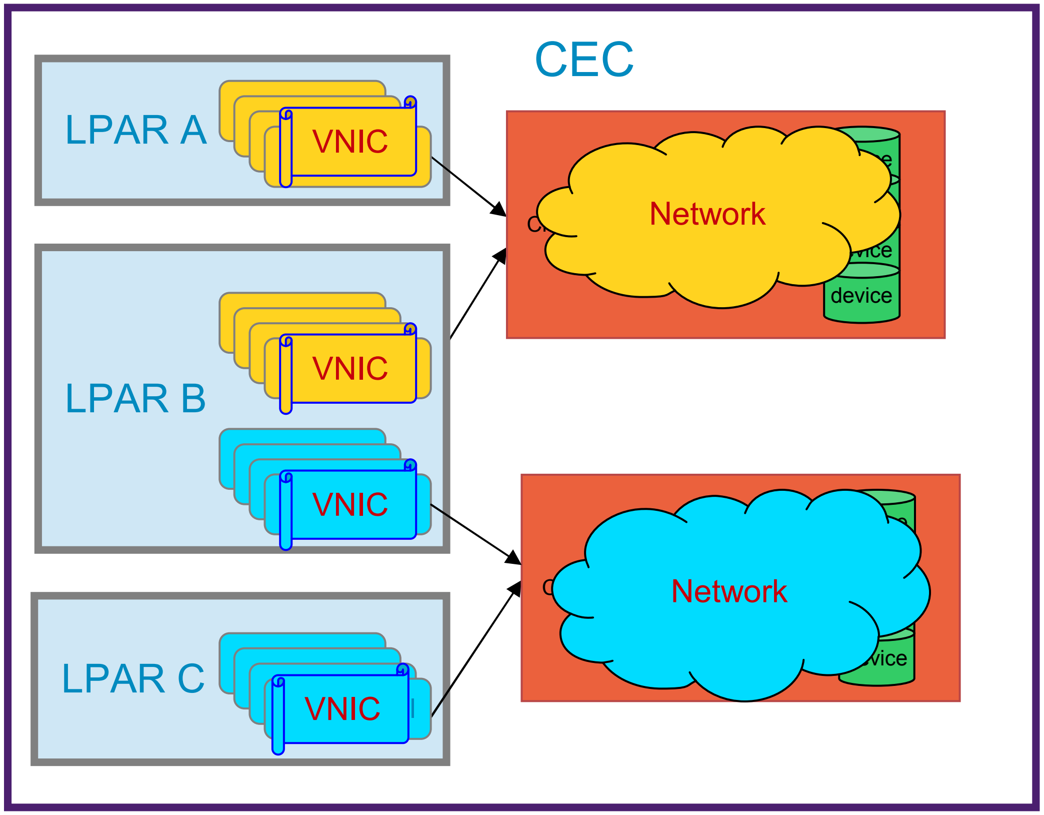


Maxima per CEC:
32 IQD CHPIDs
12 288 IQD subchannels

IOCDS:

```
CHPID PATH=(CSS(3),FA),TYPE=IQD,PART=(LP1, LP3),CHPARM=C0
CNTLUNIT CUNUMBR=FA,PATH=(CSS(3),FA),UNIT=IQD
IODEVICE ADDRESS=(FA00,16),CUNUMBR=FA,UNIT=IQD,UNITADD=00
```

HiperSockets Channels



Maxima per CEC:
32 IQD CHPIDs
12 288 IQD subchannels

IOCDS:

```
CHPID PATH=(CSS(3),FA),TYPE=IQD,PART=(LP1, LP3),CHPARM=C0
CNTLUNIT CUNUMBR=FA,PATH=(CSS(3),FA),UNIT=IQD
IODEVICE ADDRESS=(FA00,16),CUNUMBR=FA,UNIT=IQD,UNITADD=00
```

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 <= max. MTU in CHPARM;
  device driver may put multiple frames in a HiperSockets message

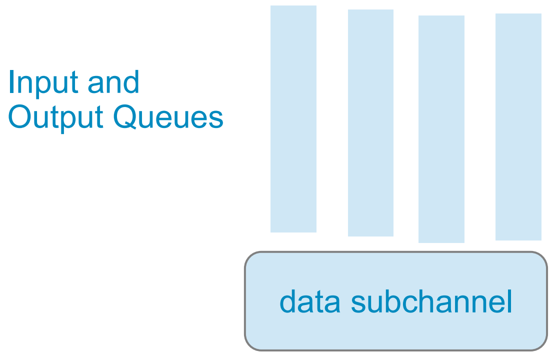
Channel flavor:

CHPID Parameter	usage
CHPARM=x0 (default)	traditional HiperSockets
CHPARM=x2	HiperSocktets for IEDN (IQDX)
CHPARM=x4	HiperSockets for External Bridge

QDIO architecture (history)

- **Queued Direct IO**
- Adapter accesses **Queues in customer memory** without using SSCH and CCW as in traditional IO
- Invented for OSA Ethernet adapters
- HiperSockets was modeled after OSA (**iQDIO: internal QDIO**)
- QDIO is also exploited by FCP
- VM GuestLAN and VM VSwitch emulate either OSA devices or HiperSockets devices

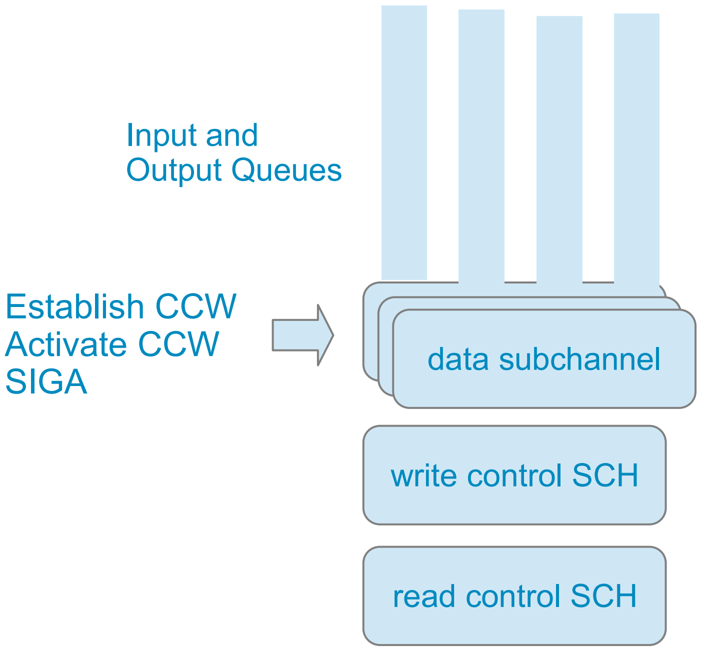
QDIO subchannels



- Establish CCW to data sch – links queues to subchannel
- Activate CCW to data sch – long running CCW
  - SCH status : active
  - HALT SCH / CLEAR SCH to de-activate subchannel
  - device end by Firmware in case of problems

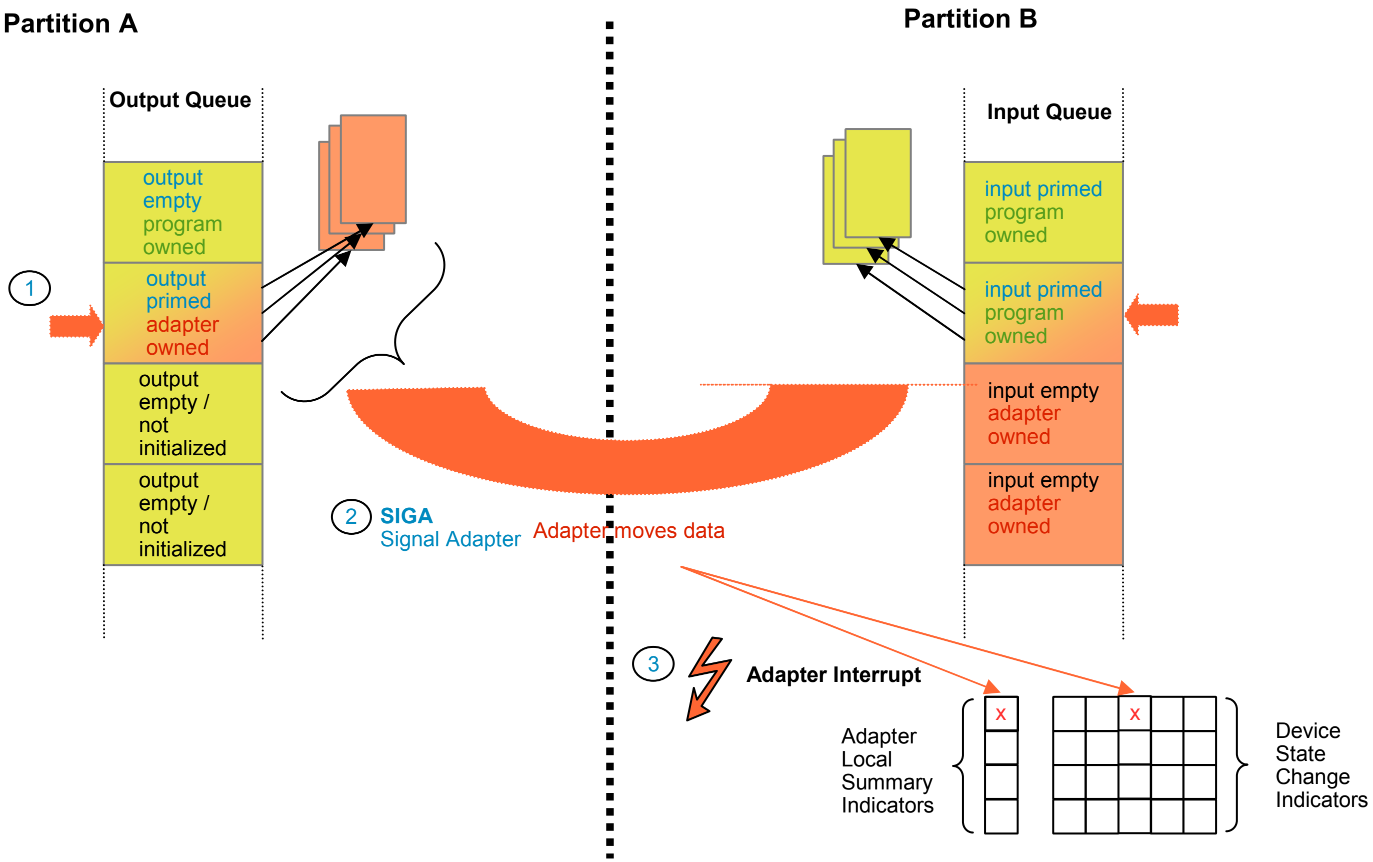
- Signal Adapter Opcode to trigger data transfer
- Adapter Interrupts to indicate reception of incoming data

QDIO subchannels - 2



- Write Control to send Control information to adapter
  - IP / MAC Address
  - VLAN information, etc...
- Read Control to receive responses / Control Information
- MPC group: 1 write control, 1 read control, 1-8 data
  - Linux: 1 data – MPC group: 3 subchannels => max 4k VNICs / CEC

QDIO architecture


Partition A
Partition B
Output Queue
output empty program owned
1
output primed adapter owned
output empty / not initialized
output empty / not initialized
Input Queue
input primed program owned
input primed program owned
input empty adapter owned
input empty adapter owned
2
SIGA
Signal Adapter
Adapter moves data
3
Adapter Interrupt
x
x
Adapter Local Summary Indicators
Device State Change Indicators

- HiperSockets SIGA is synchronous!!
  Sending CPU is moving the data inside SIGA operation !!

Performance example

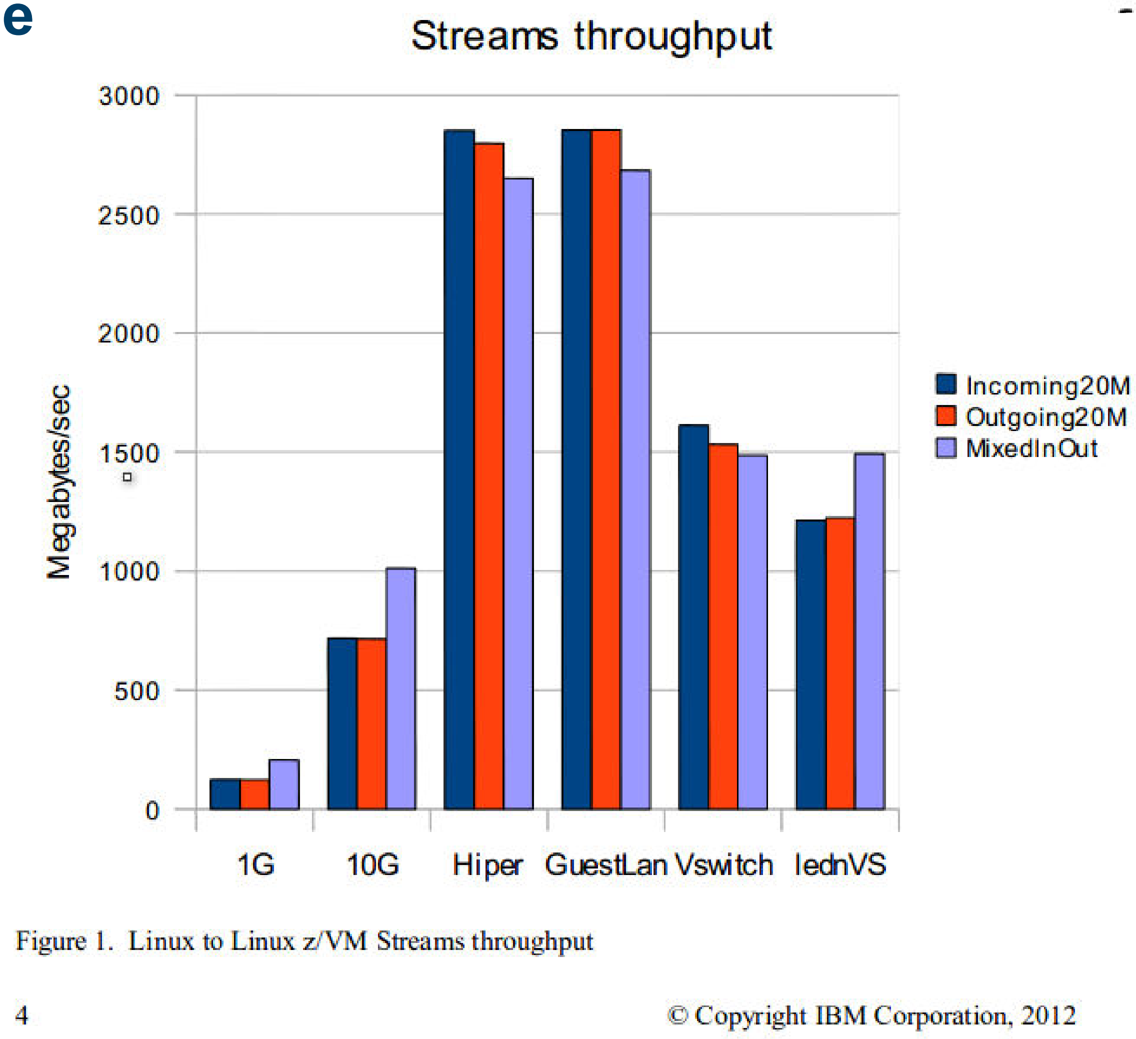


Figure 1. Linux to Linux z/VM Streams throughput

4

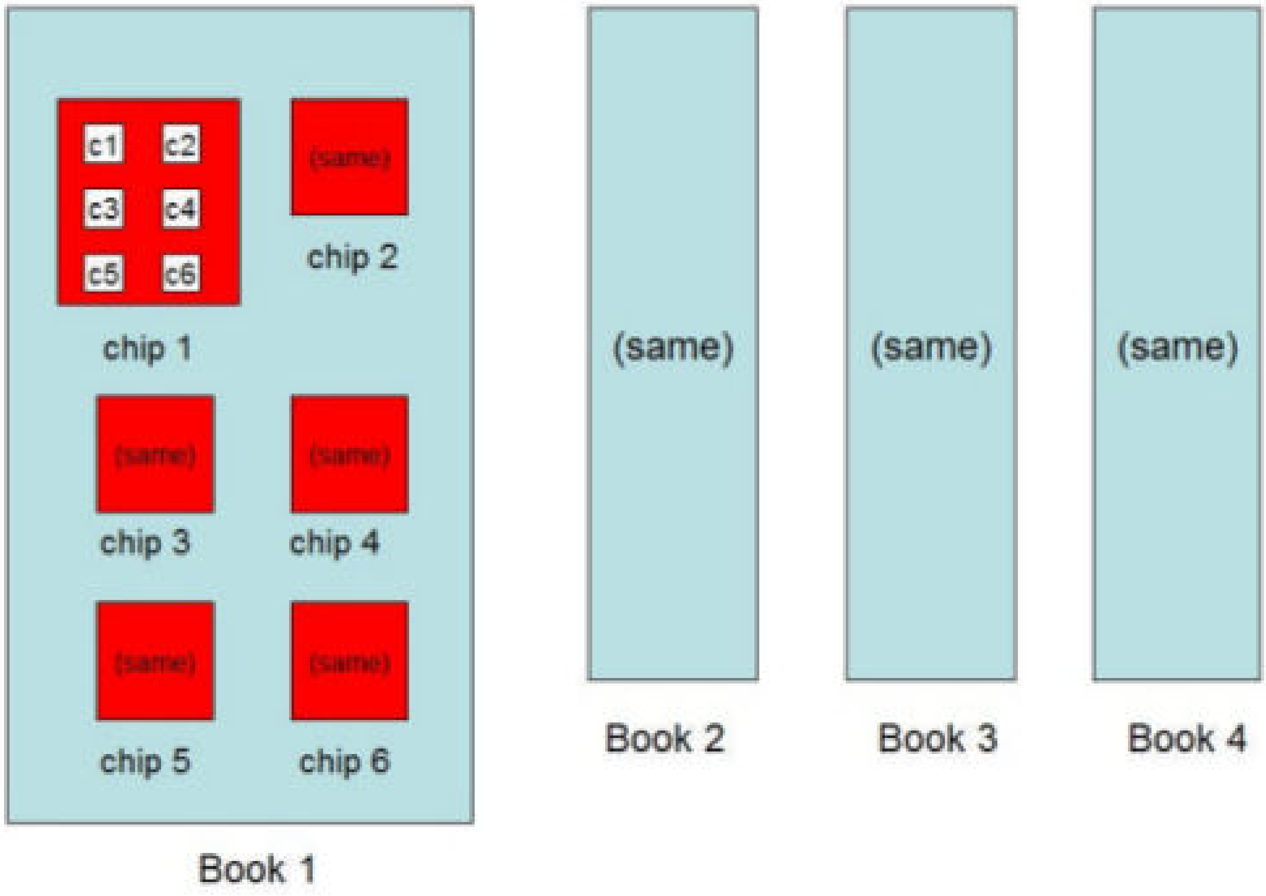
© Copyright IBM Corporation, 2012

- Source: IBM Techdocs - A Study of Network Performance on the IBM System z196 http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102175

Performance

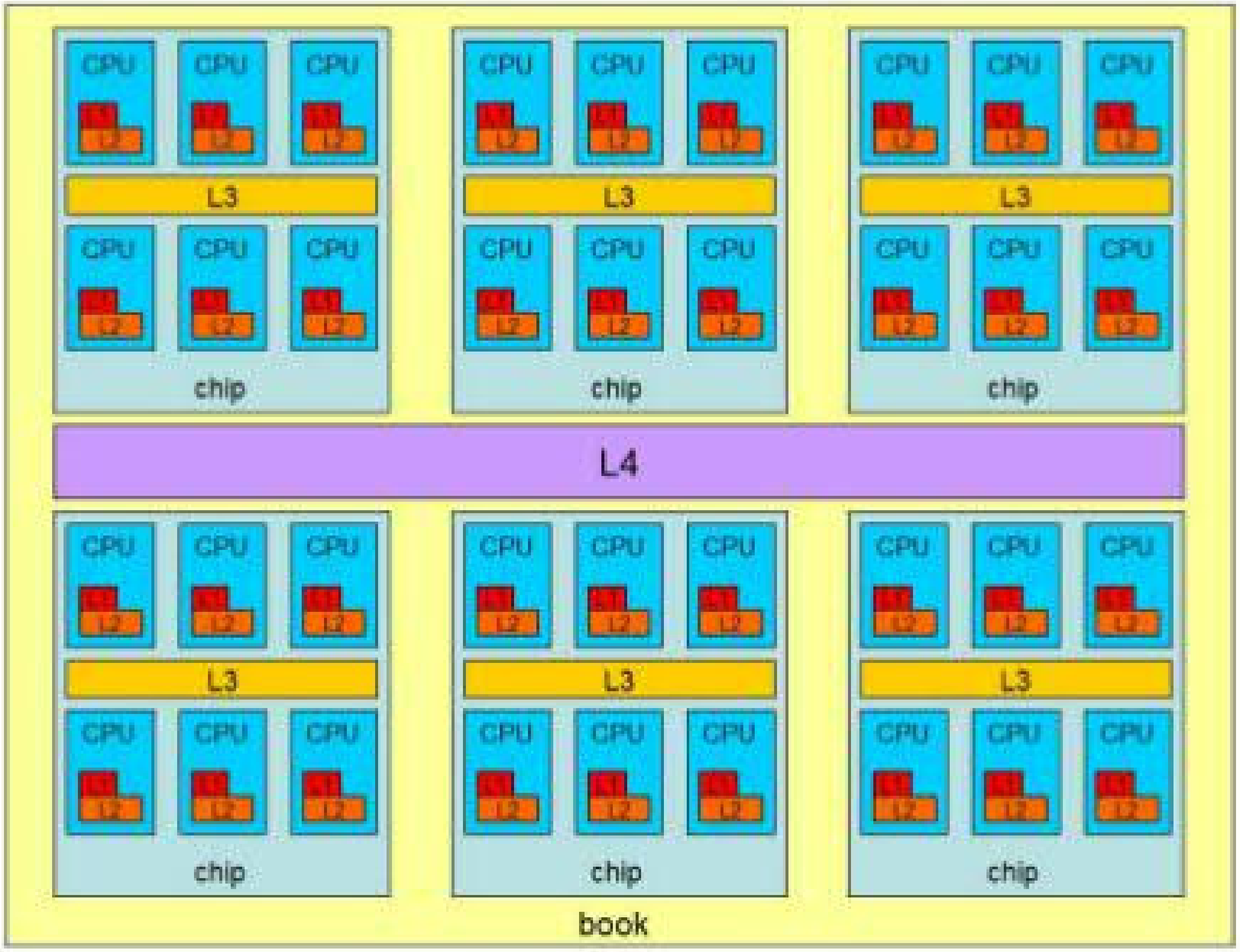
- **synchronous memory moves**
  - very **low latency**
  - **high throughput** for large data
  - throughput for small packets is more restricted by SW path length
- What influences HiperSockets performance?
  - **available CPU** processing power (number, sharing rate, ..)
  - HW model (sub-capacity CPs do not reduce Firmware performance)
  - physical structure of memory (1 book vs multiple books, cache effects)
  - MTU size for streaming (→ memory consumption for input queues)
  - **input buffer count** (→ memory consumption for input queues)
  - software settings (TCP/IP buffer size, number of sessions, scheduling, memory management, ..)

IBM System z CPU–Chip-Book Relationship



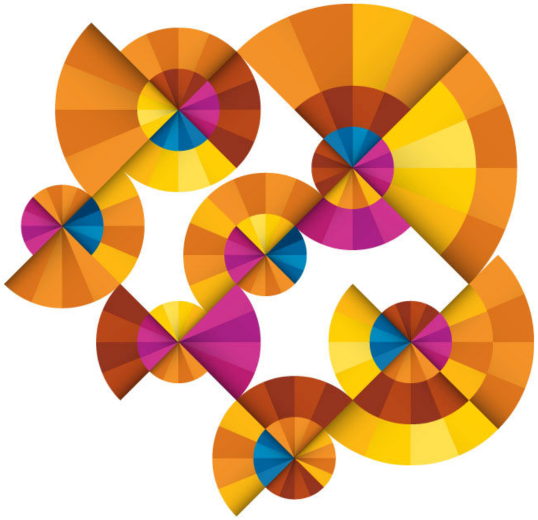
Drawing by B. Wade zVM development

IBM System z Cache Layering



Drawing by B. Wade zVM development

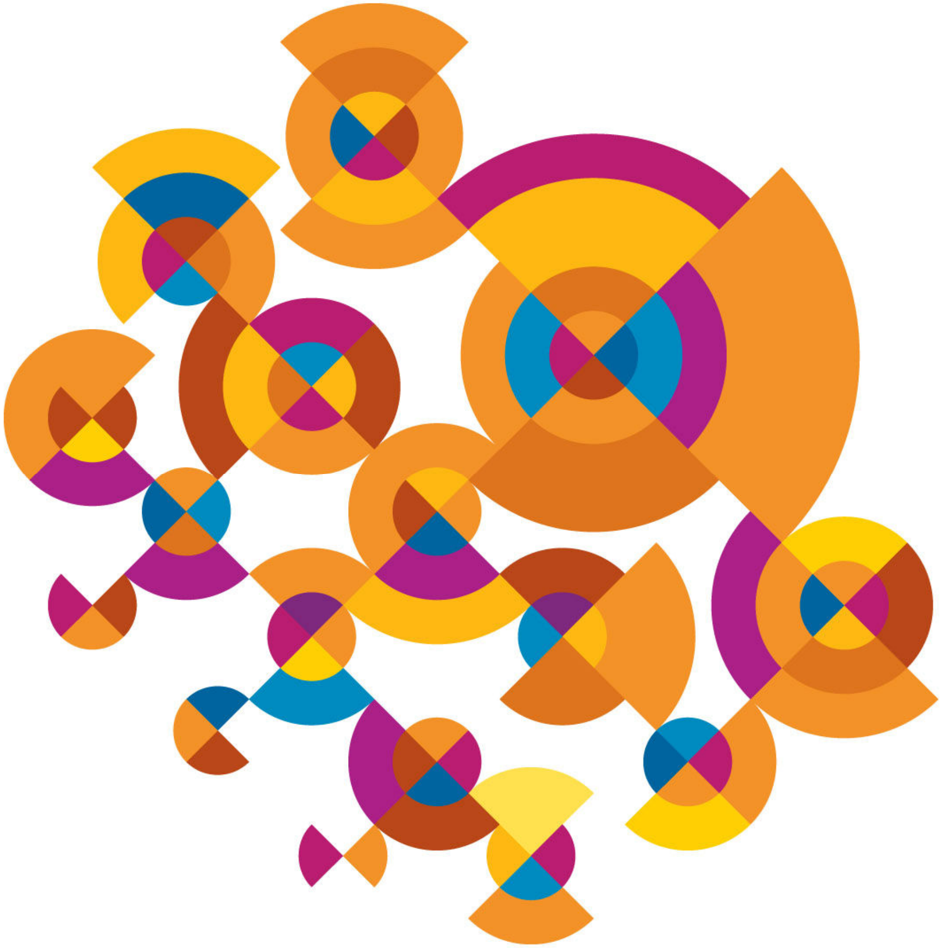
Why HiperSockets?



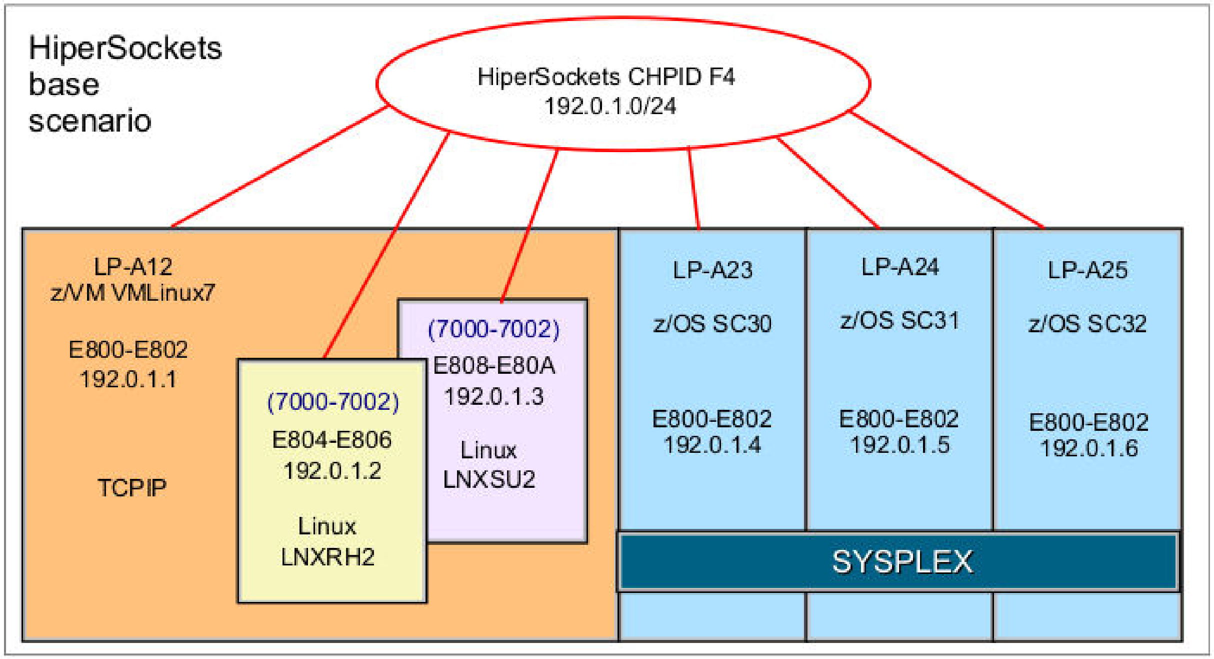
- **Performance**
  - low latency
  - high throughput
- **No hardware required**
  - **Flexibility**
    - no re-cabling for network configuration changes
  - OSA bandwidth fully available for external traffic
  - **Security**
    - no wires or external components
    - no encryption required (→ performance)
  - **Reliability**
    - no mechanical parts

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 Hipervisor
  - (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 Hipervisor
- delivery of interrupts to the VM guest without interception to the VM Hipervisor

Complete your sessions evaluations online at SHARE.org/BostonEval

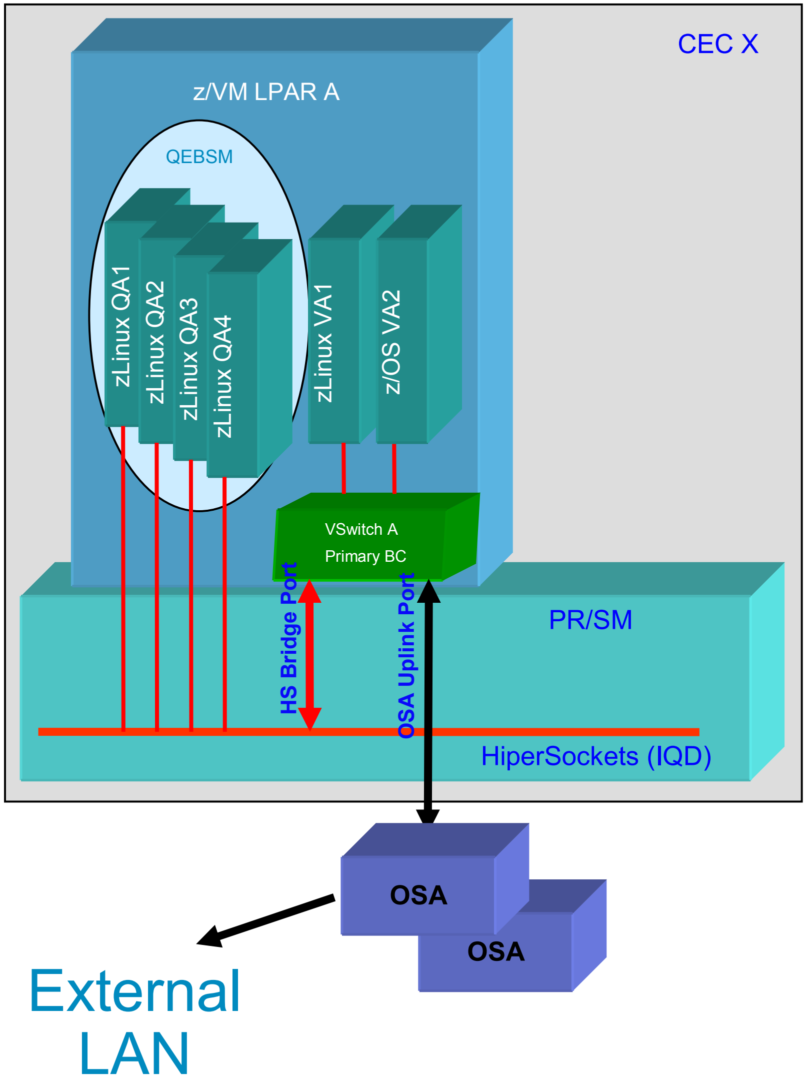
Layer 3 versus Layer 2

- 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

Miscalleaneous features

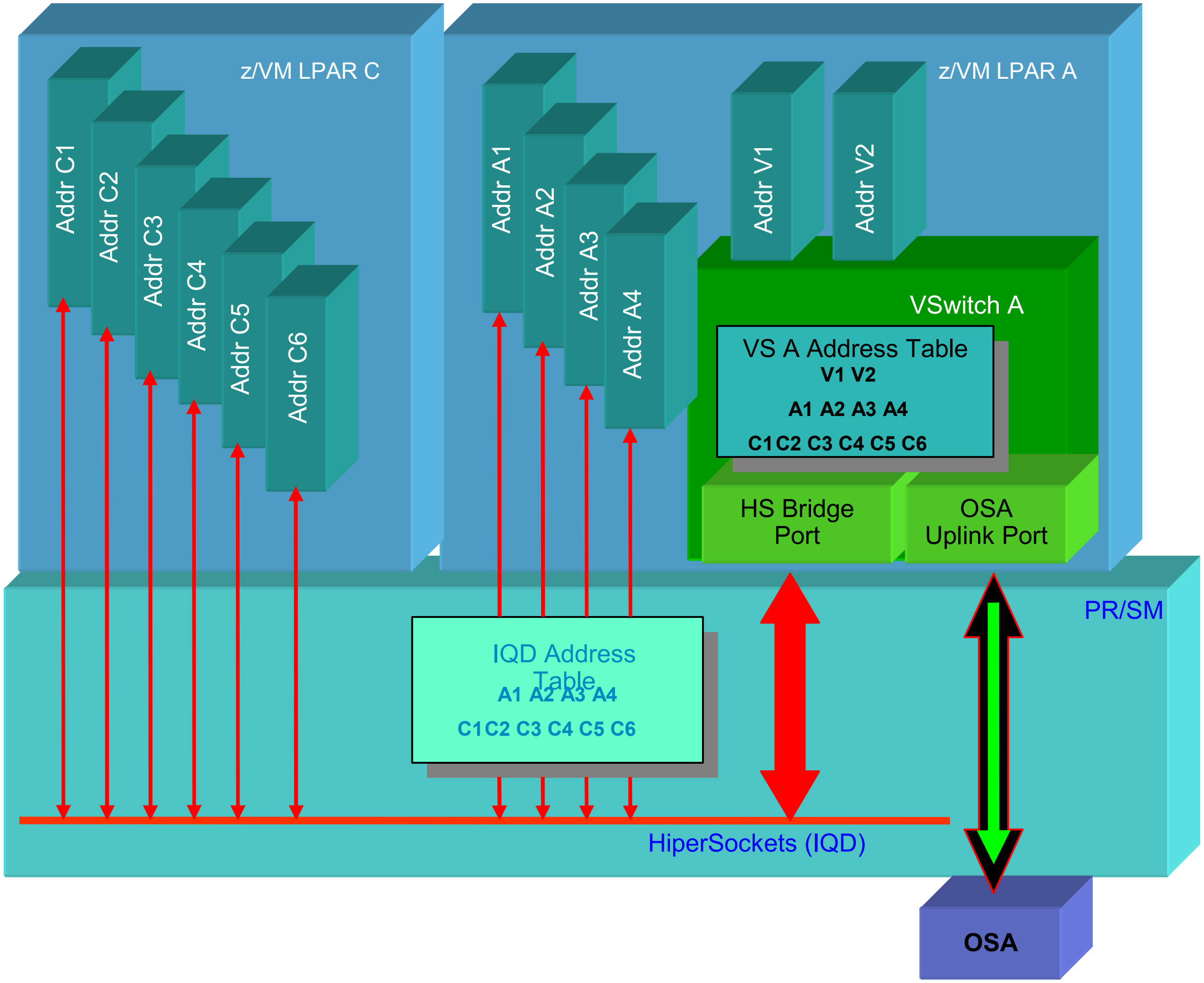
- **MULTIWRITE**
  - exploited by z/OS, send multiple output buffers in one SIGA
- **Network Traffic Analyzer**
  - set one IQD VNIC in 'promiscuos mode' and get a copy of all traffic on this channel
  - Authorization and 'filtering' on SE required
    - Which LPAR is authorized to run a NTA?
    - Traffic 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 HS Bridge (z196)



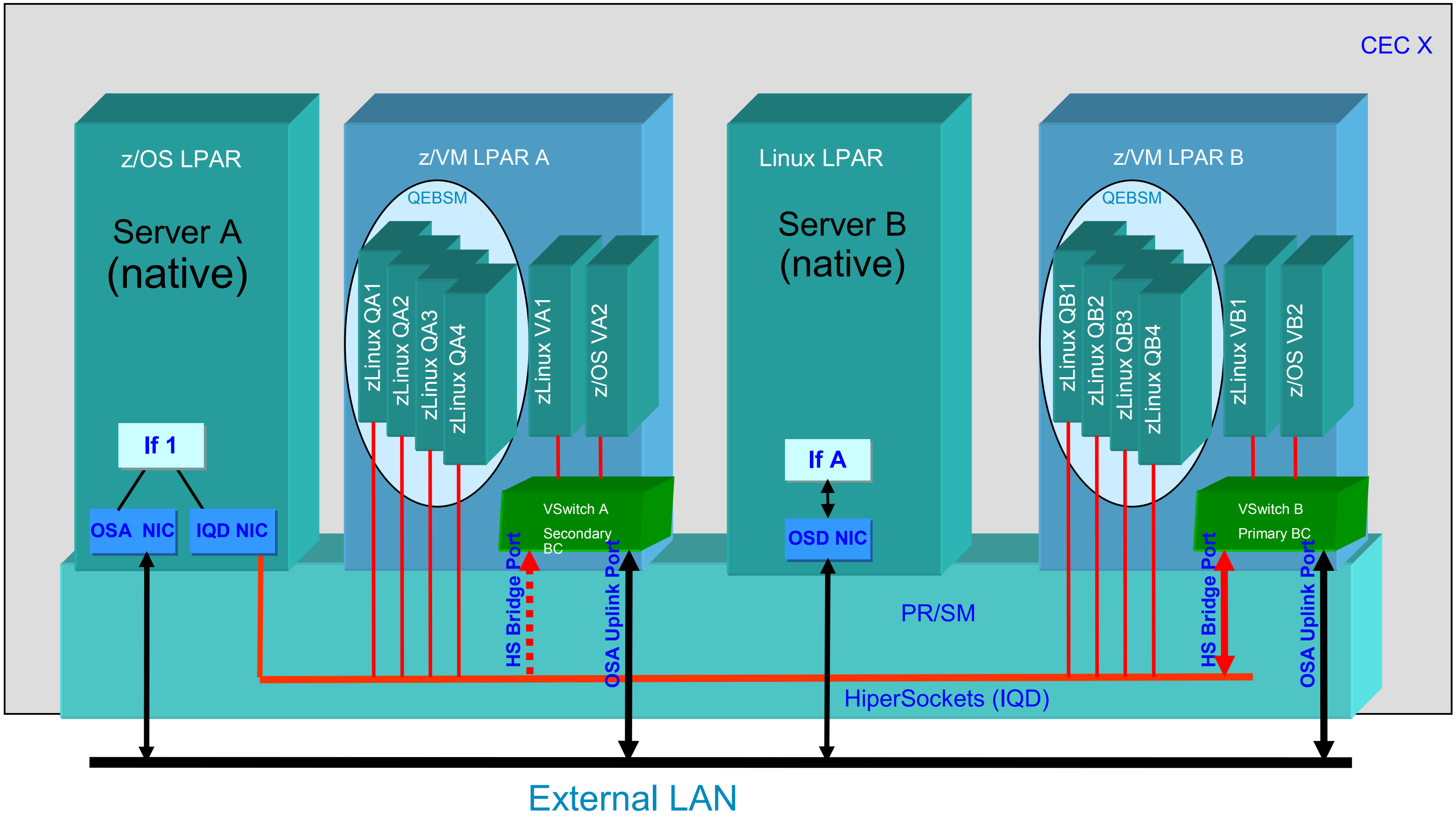
- No transport mode conversions
  - The z/VM VSwitch and HS must be operating in **layer 2**
- Supports both **IEDN** and **External Bridged** IQD channel customer networks
- Only traffic to/from QEBSM VNICs will flow over the bridge
- Guests QA1,QA2, QA3 and QA4 have real (*dedicated*) QEBSM connections to HS CHPID.
  - Optimum performance config requiring almost no z/VM involvement
  - Bridged by default (Connectivity to HS and external LAN segments0
- Guests VA1 and VA2 have virtual (NIC) connections through VSwitch A
  - Optimum performance config 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 is current support

z/VM VvSwitch HS Bridge Port - 2



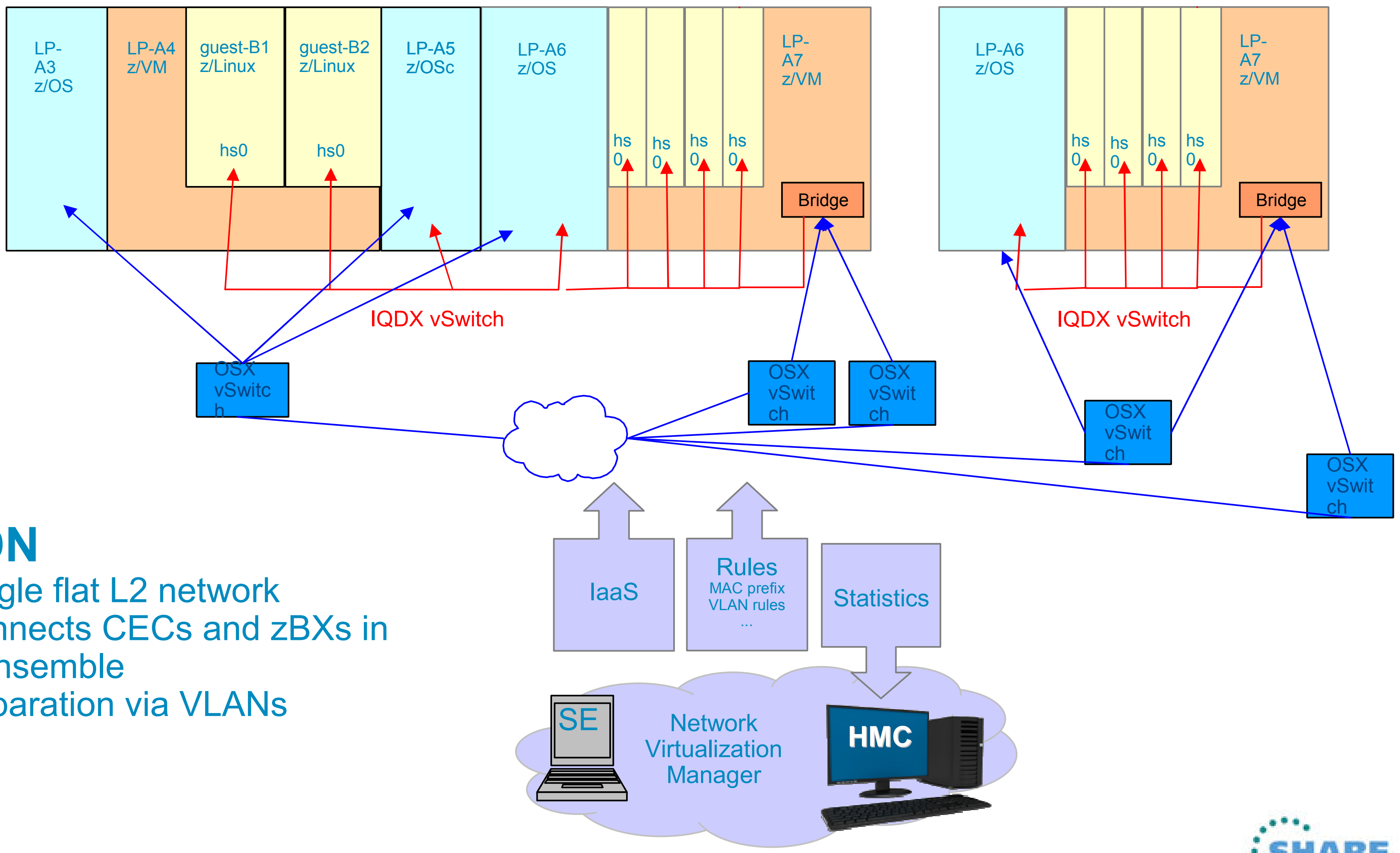
- VSwitch A provides bridging service for both LPARs
- Bridgeable servers may be added or removed dynamically
- HS keeps z/VM VSwitch A bridge in synch through asynchronous table entry updates
- z/VM VSwitch keeps OSA in synch with LAN residency

z/VM Vswitch HS Bridge Failover Configuration



- 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)

System z Network Virtualization Manager (z196)



IEDN

- single flat L2 network
- connects CECs and zBXs in an ensemble
- separation via VLANs

IEDN / IQDX

- Only 1 IQDX channel per CEC
- Layer 2 only
- VLAN mandatory
- Bridged via VM bridges to OSX (Linux as VM guest) or

merged interface with OSX VNIC (z/OS)

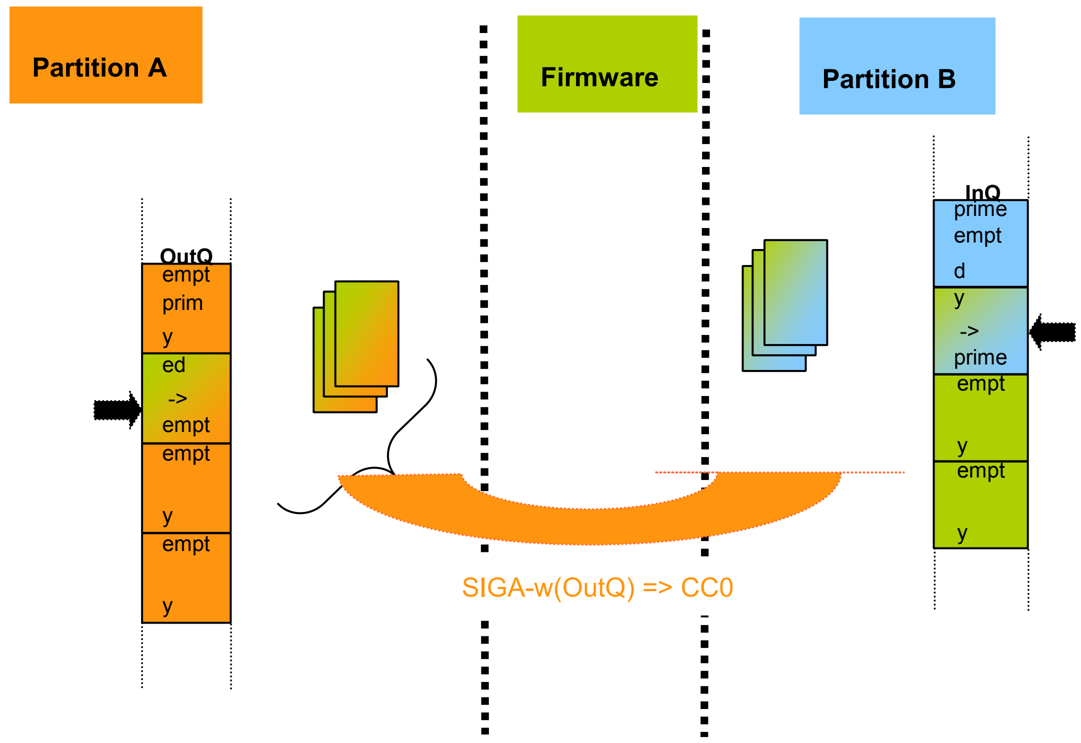
- Managed by Network Virtualization Manager (NVM) component of zManager / URM
  - MAC address management (prefix)
  - VLAN management
  - Monitoring
  - definition of VM bridges to OSX / IEDN

Completion Queues (z196)

- HiperSockets messages are sent **synchronously, in-order and reliably**
- if the target has no free input buffers, bad return code 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

SIGA-w with normal completion (CC0)



Partition A

Firmware

Partition B

OutQ

empt
prim
y

ed
->
empt

empt
y

empt
y

InQ

prime
empt
d

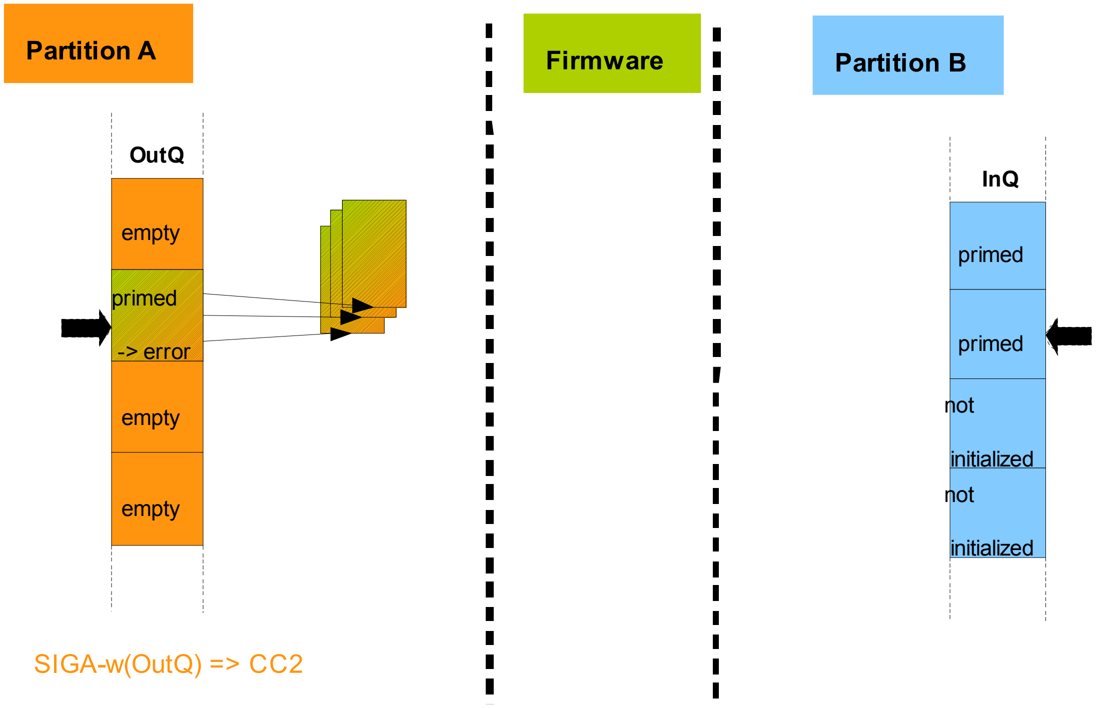
y
->
prime

empt
y

empt
y

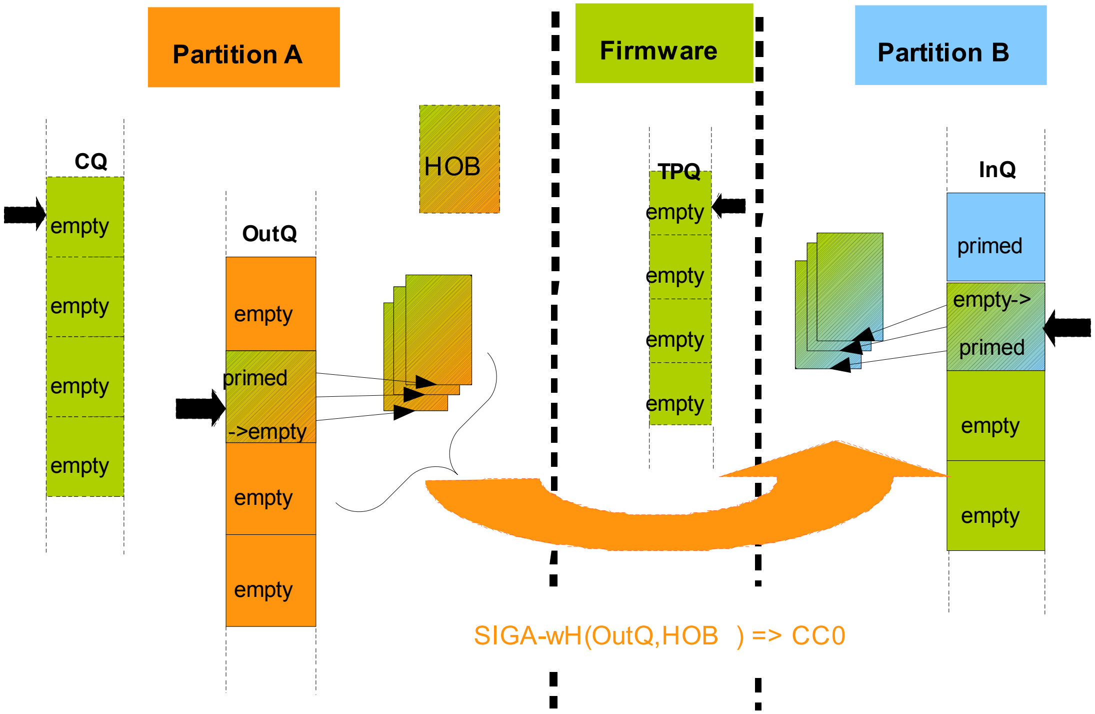
SIGA-w(OutQ) => CC0

SIGA-w with target Q full (CC2)



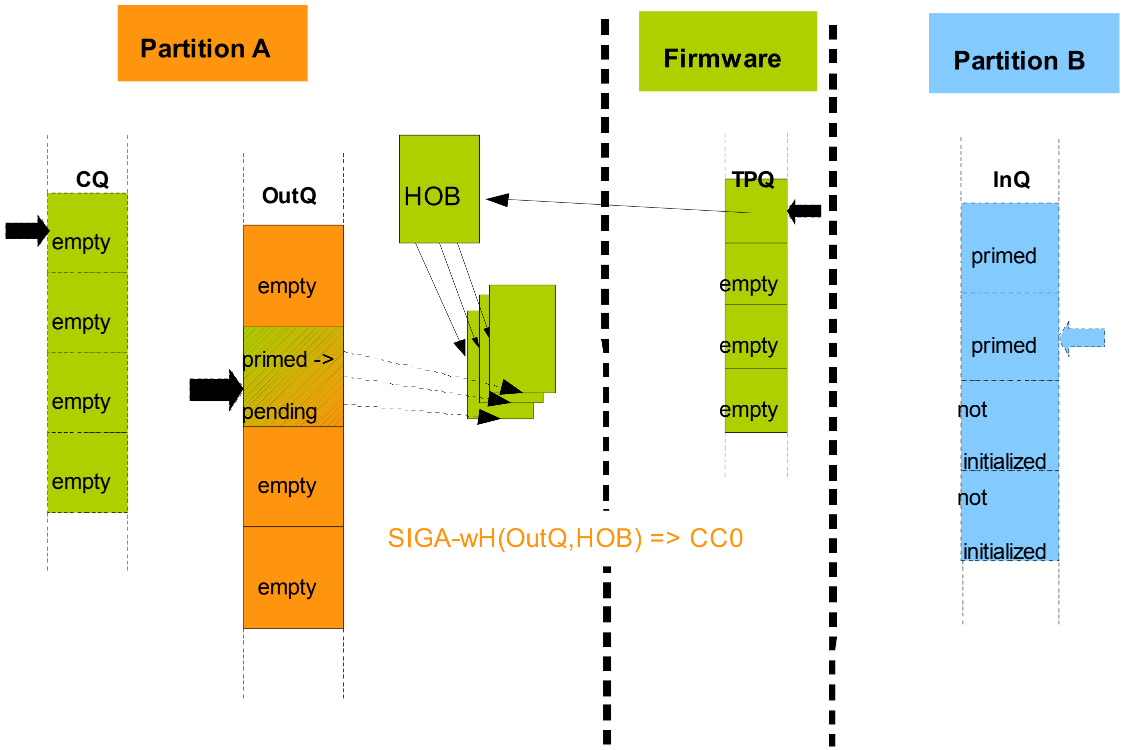
SIGA-w(OutQ) => CC2

SIGA-wq good case (CC0)

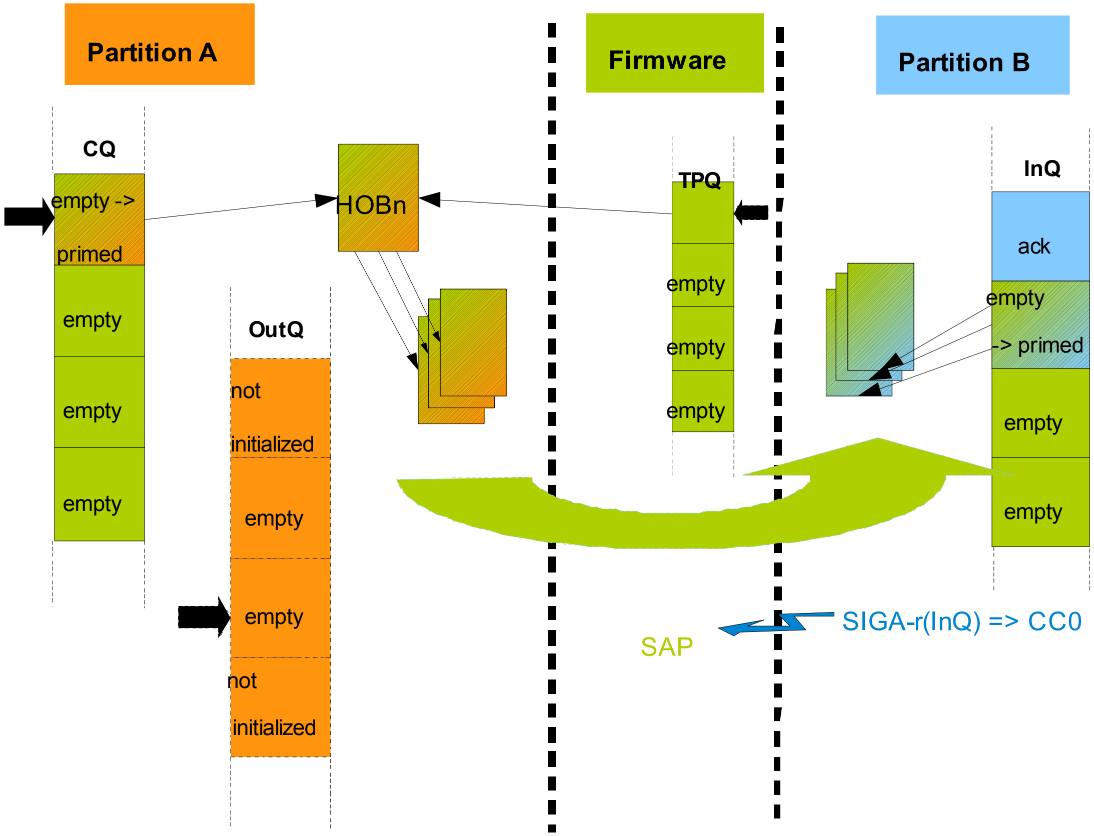


Complete your sessions evaluations online at SHARE.org/BostonEval

SIGA-wq with target queue full –> pending state

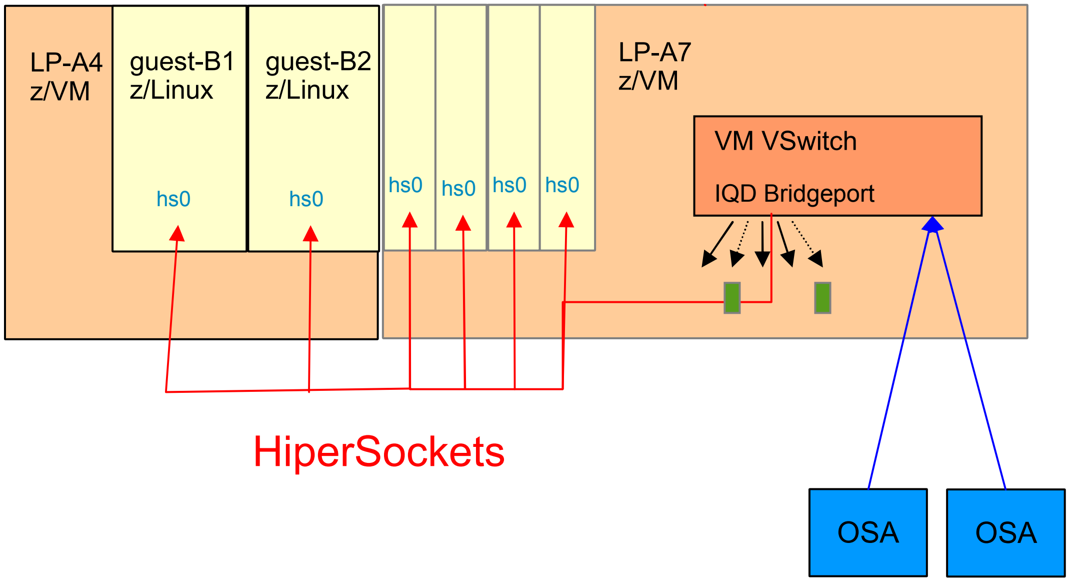


SIGA-r with pending data



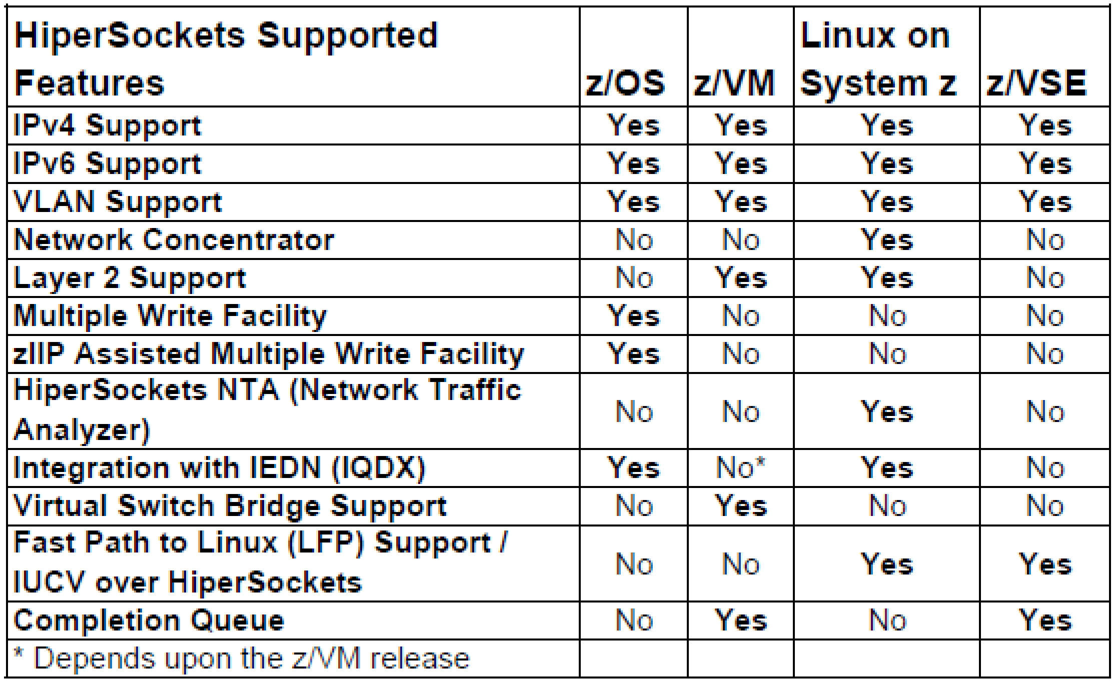
Completion Queue exploitation

- 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 **VM bridge ports**



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



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

THANK YOU