Session 17525

z/VM Virtual Switch: Advanced Topics

Alan Altmark

Senior Managing z/VM Consultant

IBM Systems Lab Services

Note

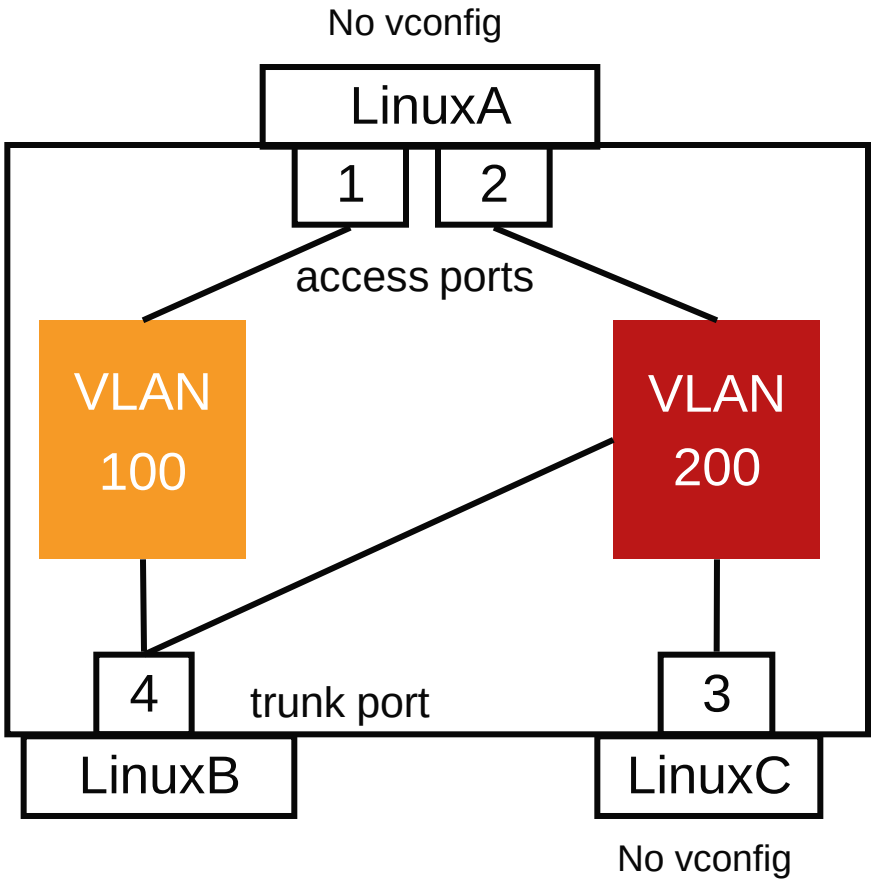
References to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe on any of the intellectual property rights of IBM may be used instead. The evaluation and verification of operation in conjunction with other products, except those expressly designed by IBM, are the responsibility of the user.

Agenda

- Port-based authorization
- Link aggregation (channel bonding)
- Shared Link Aggregation port groups
- HiperSocket Bridge
- Virtual Ethernet Port Aggregator (VEPA)
- SNMP MIB

Port-based VSWITCH access list

- Explicit port definitions
  - Admin-assigned port number
  - Each is associated with one or more VLAN ids
  - Each is reserved for a specific user ID
  - Port type
  - SET VSWITCH GRANT not used
- If user has more than one reserved port, must select via PORTNUM on COUPLE command



Port-based VSWITCH access list

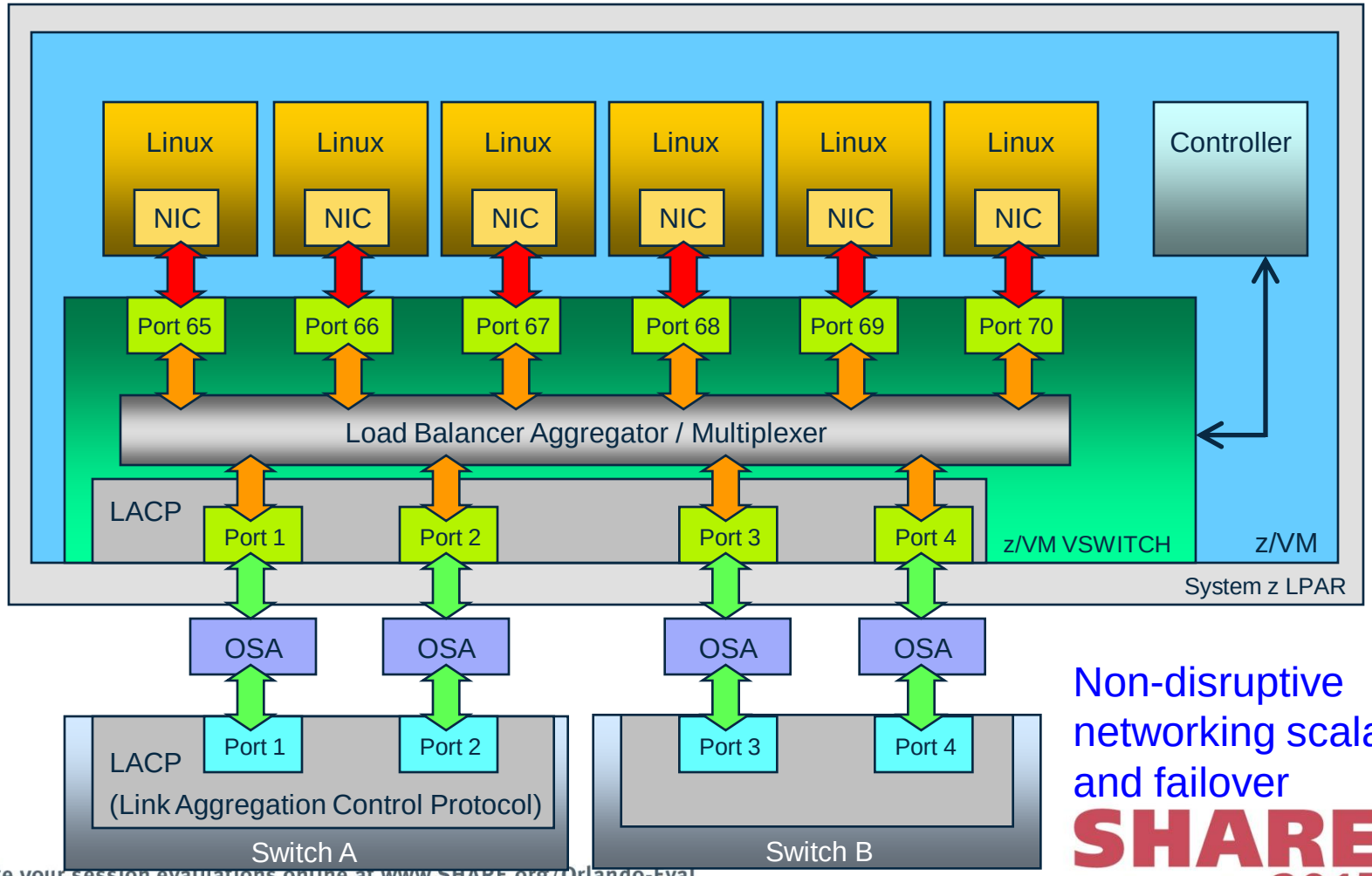
```
define vswitch vsw1 portbased vlan aware native none
set vswitch vsw1 portnumber 1 userid LINUXA
set vswitch vsw1 portnumber 2 userid LINUXA
set vswitch vsw1 portnumber 3 userid LINUXC
set vswitch vsw1 portnumber 4 userid LINUXB porttype TRUNK
set vswitch vsw1 vlanid 100 add 1     4
set vswitch vsw1 vlanid 200 add   2 3 4

LINUXA:  NICDEF 4E0 TYPE QDIO
         NICDEF 5E0 TYPE QDIO
         COMMAND COUPLE 4E0 TO SYSTEM VSW1 PORTNUM 1
         COMMAND COUPLE 5E0 TO SYSTEM VSW1 PORTNUM 2

LINUXB:  NICDEF 4E0 TYPE QDIO LAN SYSTEM VSW1
           + vconfig eth0.100
           + vconfig eth0.200

LINUXC:  NICDEF 4E0 TYPE QDIO LAN SYSTEM VSW1
```

IEEE 802.3ad Link Aggregation



Non-disruptive networking scalability and failover

IEEE 802.3ad Link Aggregation

- Binds multiple OSA-Express ports into a single pipe
  - Up to 8 OSA ports per virtual switch
  - Increases Virtual Switch total bandwidth
  - Provides seamless failover in the event of a failed OSA, switch port, cable, or switch
  - Only supported for Layer 2 VSWITCHes
  - Virtual NIC is limited to bandwidth of single OSA

- With “virtual chassis” support from switch vendor, can even handle physical switch outage

IEEE 802.3ad Link Aggregation

- Define an OSA port group
  - SET PORT GROUP name JOIN E100 E200.P1
- DEFINE VSWITCH … ETHERNET GROUP name
- OSA ports cannot be shared with other VSWITCHes or LPARs unless using IBM z13 and z/VM 6.3

Multi-VSWITCH Link Aggregation Port Groups

- IBM z13 exclusive!

- Provides a single point of control for OSA Port management across multiple VSWITCHes sharing the same physical port group.

- Requires two new system constructs
  - Global VSWITCH - Provides the mechanism for a Virtual Switch to span multiple z/VM LPARs within a CPC.
  - Inter-VSWITCH Link (IVL) - Provides management and data communications between Global VSWITCHes within the same or other z/VM instances.

Shared Link Aggregation Port Groups

- VSWITCHes are in communication with each other using a registered multicast group
- Port group can be used by different VSWITCHes
- Configuration changes are propagated to all z/VM systems sharing the port group
- You can manage the port group from any z/VM system connected to it
- Systems cooperate to balance traffic flow

The IVL Domain

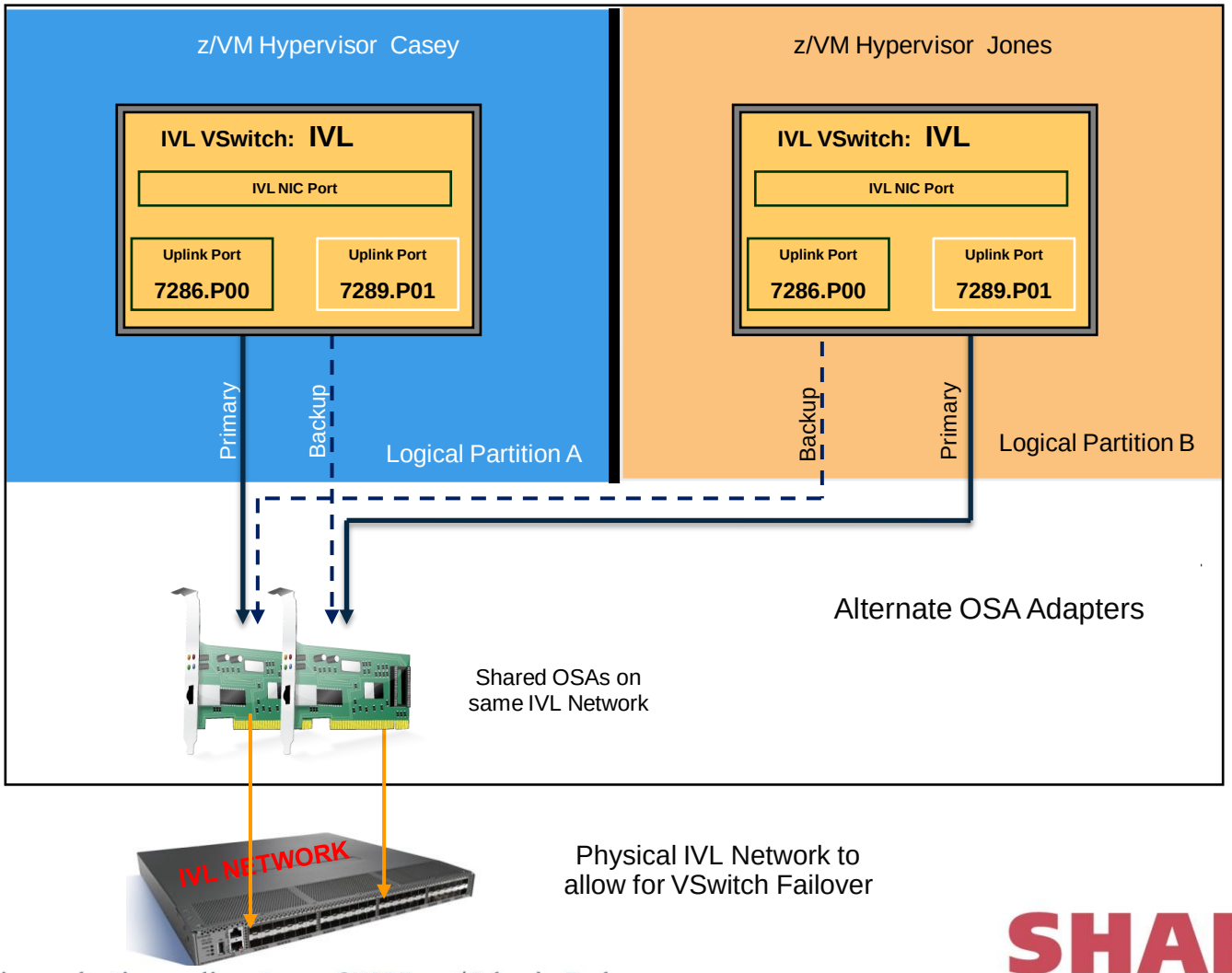
- An IVL domain is a group of up to 16 z/VM LPARs on a CPC
- All z/VM Hypervisors sharing the same physical port group must be members of the same IVL domain
- A z/VM LPAR can be a member of exactly one IVL domain
- The IVL domain is established through an IVL VSWITCH
  - One per z/VM LPAR
- Up to 8 IVL Domains can share a single LAN segment
- The bandwidth required by the IVL is minor, consisting of management and LAG data recovery communications.

IVL VSWITCH

- **DEFINE VSWITCH name {options}**
  - TYPE IVL
  - DOMAIN A through P
  - VLAN vid
  - Conventional RDEV list or exclusive port GROUP

- Remember to provide OSA port redundancy for IVL!

IVL Network Configuration Domain B VLAN 8



IVL Controls

- **SET VSWITCH name IVLPORT {option}**
  - VLAN - Change the VLAN ID associated with the IVL
  - RESET - Terminate and recreate the IVL port connection
  - PING - Tests connectivity between z/VM hypervisors in the same IVL domain
    - SET VSWITCH IVL IVLPORT PING ALL
  - HEARTBEAT TIMEOUT - Adjusts how often the local z/VM system confirms connectivity with the other domain members

Create the Shared Port Group

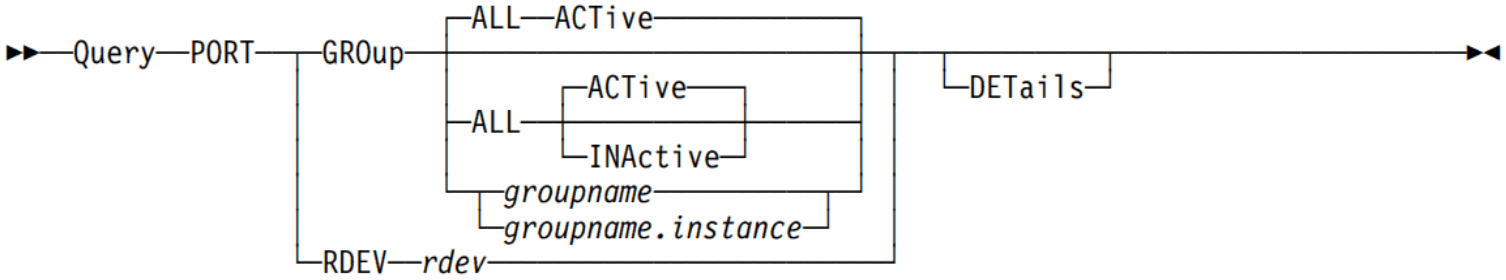
- **SET PORT GROUP name LACP ACTIVE SHARED**
- **SET PORT GROUP name JOIN rdev1.port rdev2.port**

- Device numbers can be any device number on the chpid

- The z/VM Control Program will select the device numbers to be used on the target adapter.

- z/VM will automatically propagate Shared Port Group information to all active IVL Members in the same IVL domain (B, in this example)

Port Group Verification

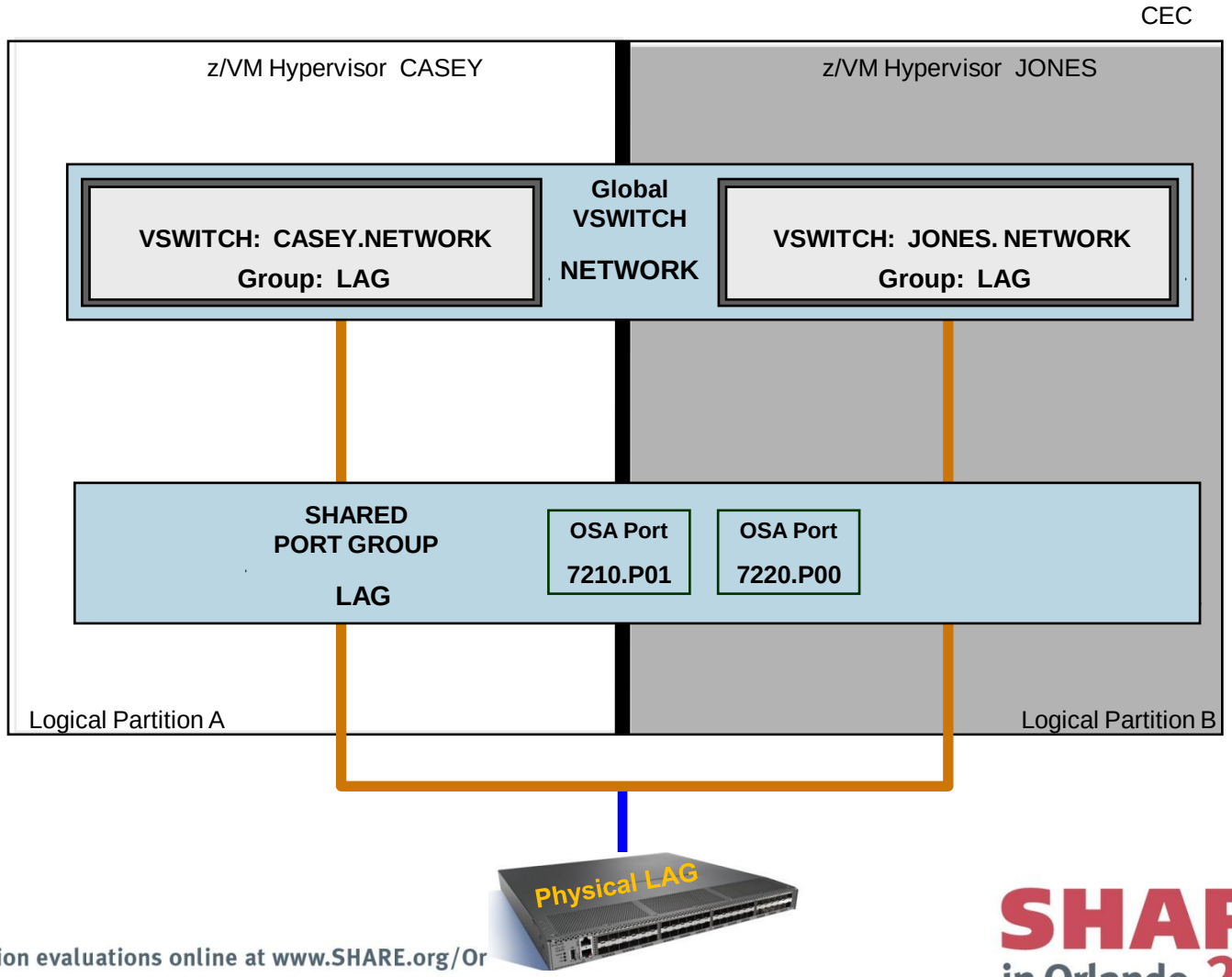


- ALL
  Return all active port groups defined in the system
- ACTIVE
  Return only those port groups associated with a virtual switch
- INACTIVE
  Return only those port groups NOT associated with a virtual switch
- GROUP groupname
  Return only the specified port group
- GROUP groupname.instance
  Return only the specified port group instance
- RDEV
  Return only information for the specified real device
- DETAILS
  Return additional information

Define a Global VSWITCH

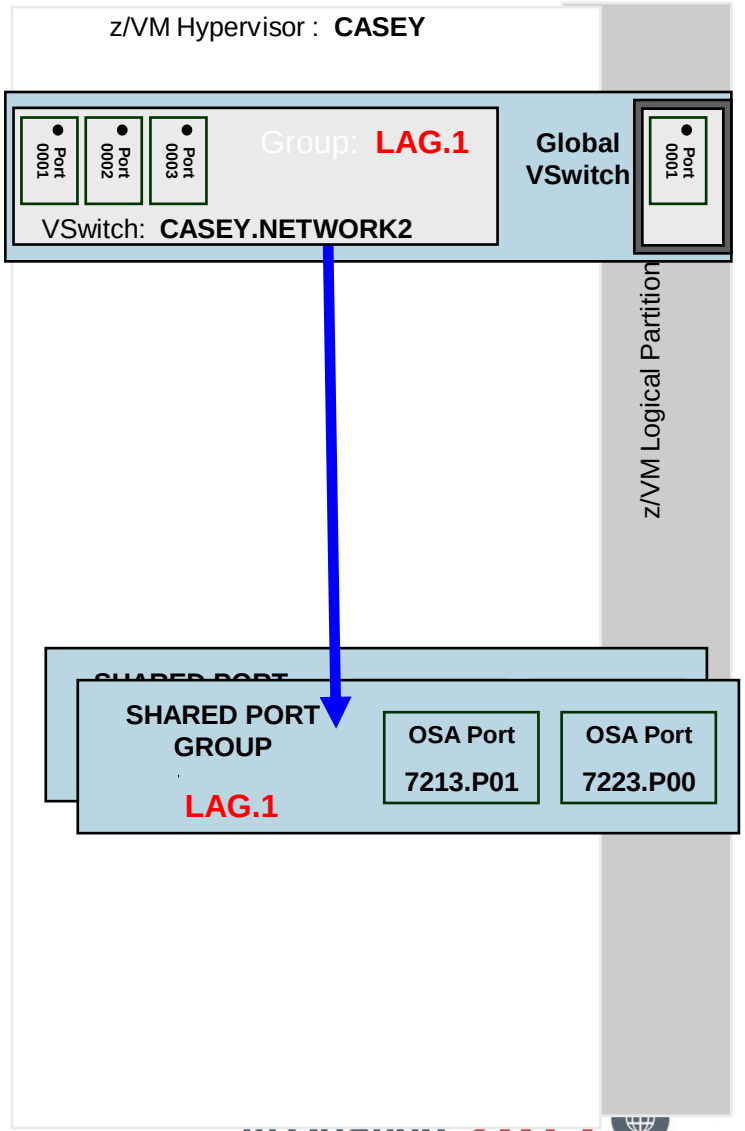
- **DEFINE VSWITCH name GLOBAL ETHERNET GROUP group**
- A Global VSWITCH is a virtual switch which can span multiple z/VM instances through the IVL Network and which shares the same physical port group.
- Must be defined with the same name in all sharing LPARs
- A Global ID (systemid.vsw_name) is generated by the control program
- Multiple Global VSWITCHes can be defined per z/VM LPAR
- An instance of a Shared Port Group is created when it is configured to a virtual switch (group.0).

Multi-VSwitch LAG Configuration



Sharing A LAG within the Same z/VM Hypervisor

- DEFINE VSWITCH NETWORK2
  GLOBAL ETHERNET GROUP LAG
- LAG.0 is the base instance of a Shared Port Group and is the only instance propagated to other IVL Members within the same domain.
- A second instance of the shared Port Group is created (LAG.1) when it is configured to a second vswitch. It remains local to the defining system.
- Up to four port group instances can be defined within an LPAR.
- The only difference between the base and its other instances are the device numbers allocated for each adapter within the LAG.
- z/VM will automatically allocate an OSA triplet for each adapter within in the group from the available devices in the LPAR.



Best Practices for Link Aggregation

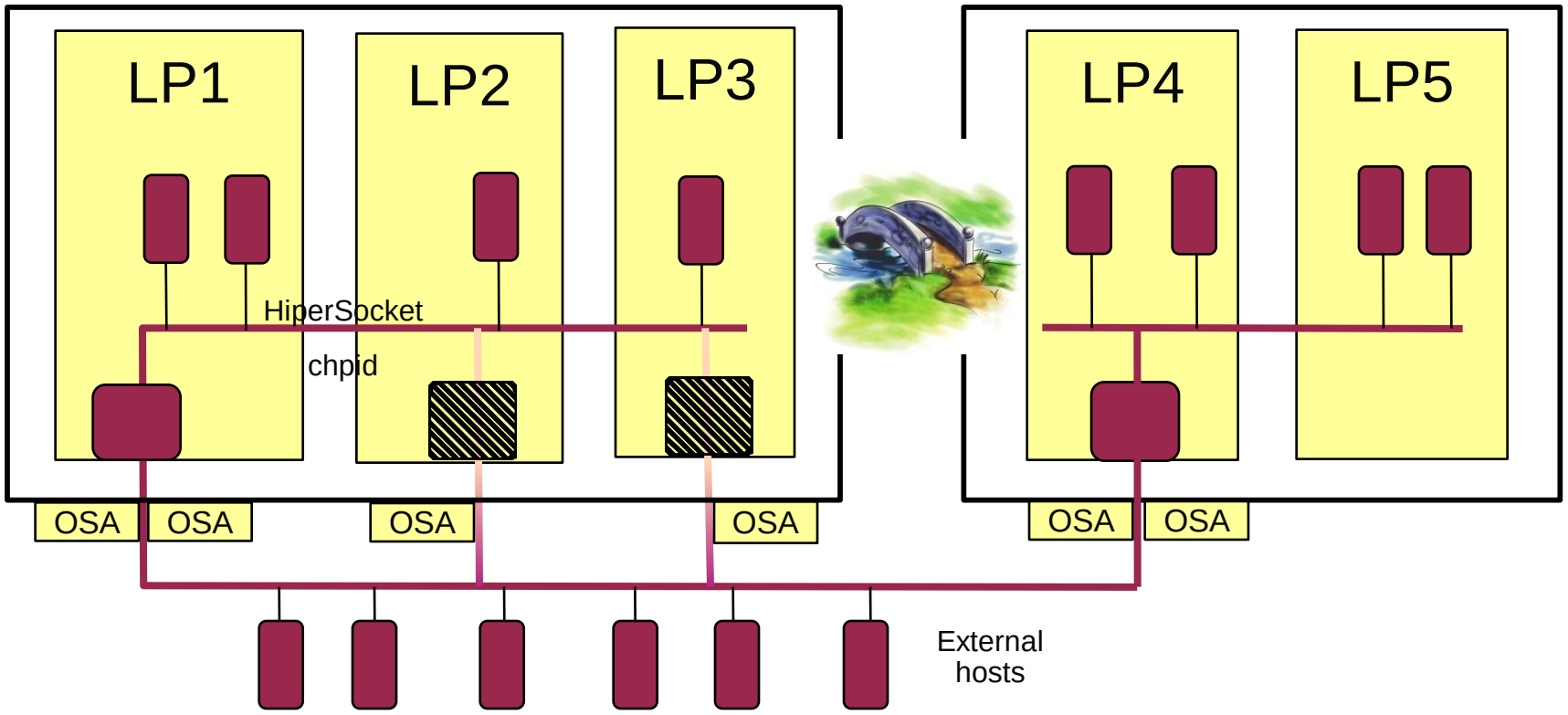
- Use a pair of switches that support “virtual chassis”
  - Provides cross-switch link aggregation port group
  - Plug each switch into separate power source
- Use two OSA ports on different PCHIDs
  - Each one plugged into one of the two switches
  - Separate back-planes to ensure separate power supply
- Provides continuous operation in case of
  - Single-source power failure
  - Switch reboot (e.g. maintenance)
  - Switch port failure
  - OSA port failure
  - OSA firmware upgrade
  - Cable failure

HiperSocket Virtual Switch 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)

HiperSocket Virtual Switch Bridge



- One active bridge per LPAR
- Path MTU discovery support
  - Large frames inside
  - Smaller frames outside

HiperSocket Virtual Switch 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

VEPA - Virtual Ethernet Port Aggregator

6.3

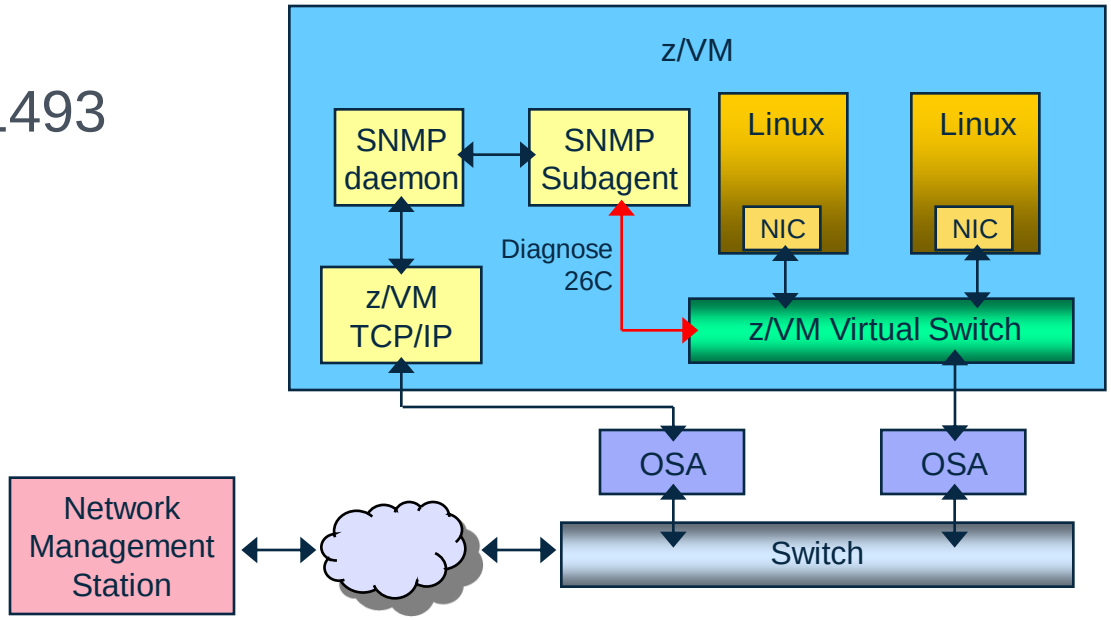
- IEEE 802.1Qbg relaxes prohibition on packet reflection
  - Frames now allowed to be "reflected" back to the origin port
  - Physical switch receives all guest-to-guest traffic
  - Enables use of external packet filtering and monitoring
  - No hardware configuration required

- SET VSWITCH … VEPA ON | OFF
  - VEPA and ISOLATE are mutually exclusive
    - VEPA implies isolation
  - VSWITCH will verify external switch support

z/VM Virtual Switch SNMP MIB

- Integrates VSWITCH into standards-based switch management and monitoring tools

- SNMP subagent provides bridge MIB data
  - Defined by RFC 1493



Diagnostics

- CP QUERY VMLAN
  - to get global VM LAN information (e.g. limits)
  - to find out what service has been applied
- CP QUERY VSWITCH ACTIVE
  - to find out which users are coupled
  - to find out which IP addresses are active
- CP QUERY NIC DETAILS
  - to find out if your adapter is coupled
  - to find out if your adapter is initialized
  - to find out if your IP addresses have been registered
  - to find out how many bytes/packets sent/received

Diagnostics – Discarded packets

- Uplink port (CP's perspective)
  - QUERY VSWITCH ACTIVE
  - RX: VSWITCH definition does not match physical port definition (trunk vs, access)
  - TX: Overrun on the OSA. Link is too slow. Use faster OSA or link aggregation.

- Virtual NIC (guest perspective)
  - QUERY NIC USER <userid> <vdev>
  - RX: Packets are arriving faster than the guest can consume them
  - TX: Packet cannot be delivered to destination
    - Unauthorized VLAN ID on virtual trunk port
    - Untagged frame on virtual trunk with NATIVE NONE
    - Guest configured as VLAN-aware (vconfig), but has virtual access port
    - Overrun target guest

Summary

- Use IEEE VLANs to simplify configuration
- Use Link Aggregation for best availability
- Integrate into SNMP-based monitoring solutions
- Port-based or User-based configuration style
- The latest technologies

Support Timeline

z/VM 6.3	▪ Shared link aggregation port groups ▪ VEPA ▪ SET VSWITCH SWITCHOVER
z/VM 6.2	▪ Port-based configuration provides separate VLAN per virtual access port ▪ HiperSocket bridge
z/VM 6.1	▪ Uplink port can be OSA or guest ▪ zEnterprise Ensemble (IEDN and INMN) ▪ VLAN UNAWARE, NATIVE NONE
z/VM V5	▪ Virtual and physical port isolation ▪ z/VM TCP/IP support for Layer 2 ▪ Link aggregation ▪ SNMP monitor ▪ Virtual SPAN ports for sniffers ▪ Virtual trunk and access port controls ▪ Layer 2 (MAC) frame transport ▪ External security manager access control
z/VM V4	▪ Layer 3 (IPv4 only) Virtual Switch with IEEE VLANs ▪ Guest LAN with OSA and HiperSocket simulation

References

- Publications:
  - z/VM CP Planning and Administration
  - z/VM CP Command and Utility Reference
  - z/VM Connectivity

Contact Information

Alan C. Altmark

Senior Managing z/VM Consultant

z Systems Delivery Practice
IBM Systems Lab Services

IBM

1701 North Street

Endicott, NY 13760

Mobile 607 321 7556

Fax 607 429 3323

Email: alan_altmark@us.ibm.com

- Mailing lists: IBMTCP-L@vm.marist.edu
- IBMVM@listserv.uark.edu
  LINUX-390@vm.marist.edu

- See http://ibm.com/vm/techinfo/listserv.html for details.

IBM Systems Lab Services

We make cloud and big data work for you

From business strategy to technical deployment, our experienced consultants can help you plan and deliver your new cloud and big data analytics solutions faster.

IBM Systems Lab Services

ibm.com/systems/services/labservices
stgls@us.ibm.com