# z/VM Virtual Switch
# Part 2: Advanced Topics

Alan Altmark
Senior Managing z/VM Consultant
IBM Systems Lab Services

Alan_Altmark@us.ibm.com

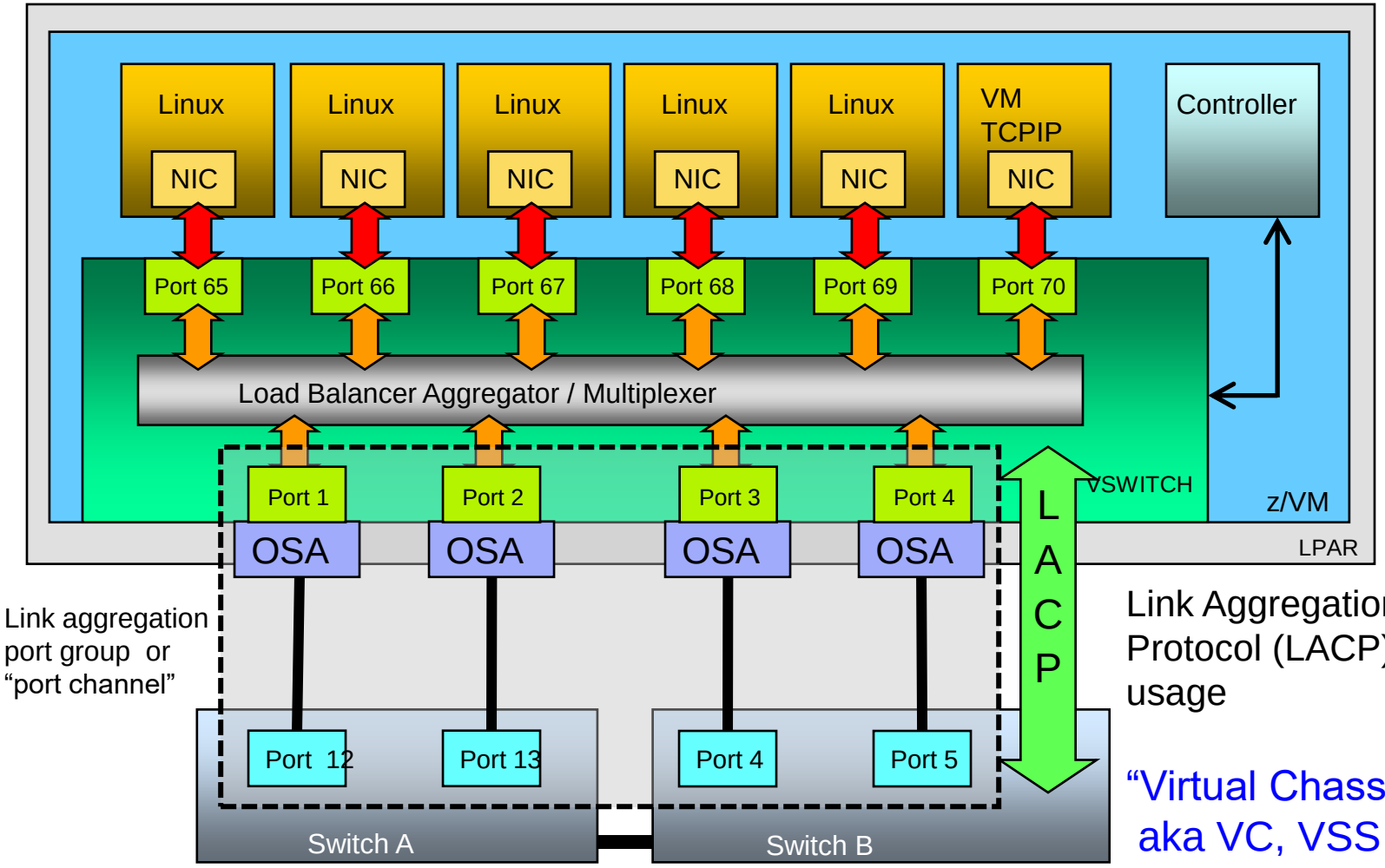
# Notes

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

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

# IEEE 802.3ad Link Aggregation




Non-disruptive networking scalability and failover

Link Aggregation Control Protocol (LACP) controls port usage

"Virtual Chassis" aka VC, VSS

# 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 bandwidth
  - Provides seamless failover in the event of a failed OSA, switch port, cable, or switch
  - Only supported for ETHERNET VSWITCHes
  - Virtual NIC is still limited to bandwidth of single OSA
  - Also called a **port channel**

- 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 PCHNL01 JOIN F100 F200.P1`**

- Create a VSWITCH that references to group
  - **`DEFINE VSWITCH … ETHERNET GROUP PCHNL01`**

- Note: OSA ports cannot be shared with other VSWITCHes or LPARs unless using **shared port groups**

# 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

# Shared Link Aggregation Port Groups

- Every link aggregation port group uses at least two OSA ports per port group
  - A four-member SSI cluster will use at least 8 ports.
  - Four clusters (dev, test, prod, sandbox) will use 32 ports
  - Large capital investment: OSAs, switch ports, cables, connectors (SFPs)
  - Limit of 48 OSD channel paths

- New OSA Express capability on **IBM z13 and later** provides ability to share OSA ports in link aggregation mode to be shared across **z/VM** LPARs

# Shared Link Aggregation Port Groups

- Multiple VSWITCHes can share a single OSA link aggregation port group.
  - Same or different LPAR

- Two new system constructs
  - **Inter-VSWITCH Link** (IVL) - Provides management and data communications between participating members of a Global VSWITCH.
    - Data communication only if LPAR loses connection to OSA still operable from another sharing LPAR (rare)
  - **Global VSWITCH** - Provides the mechanism for a Virtual Switch to span multiple z/VM LPARs within a CPC.

# Shared Link Aggregation Port Groups

- VSWITCHes are in communication with each other using a registered multicast group (not IP)
- 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
- z/VM equivalent of “virtual chassis”

# IVL Domain

- Provides control and error recovery functions for all global VSWITCHes in the domain
  - All z/VM hypervisors sharing the same physical port group must be members of the same IVL domain

- The IVL domain is established through an IVL VSWITCH
  - Global VSWITCH definition is deferred until domain is established

- In rare cases, may be required to forward guest traffic to another LPAR

# IVL VSWITCH

- **DEFINE VSWITCH *name* TYPE IVL DOMAIN *d* [VLAN *vid*]**
  - DOMAIN A through H  (max 8 domains per CPC)
  - Identify VLAN

- Maximum 16 z/VM LPARs in an IVL domain
  - A z/VM LPAR can be in only one IVL domain

- Conventional RDEV list or exclusive port GROUP
  - Remember to provide OSA port redundancy!
  - **No, you cannot use the same OSAs that the global VSWITCHes are using!**

# IVL Dynamic 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
- CP will select the device numbers to be used on the target adapter.
- CP propagates changes to the port group configuration to all active members of the IVL domain

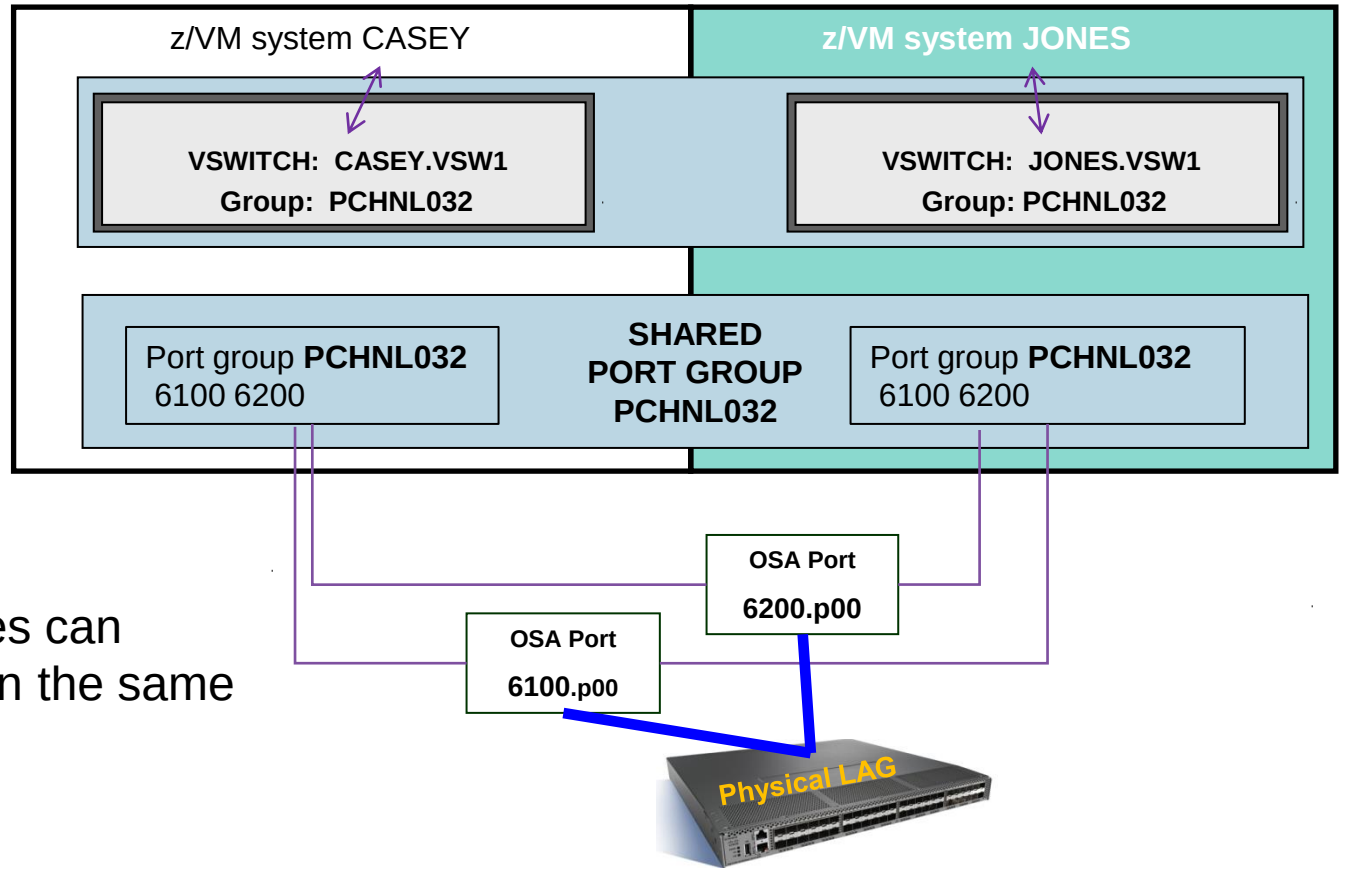
# Define a Global VSWITCH

**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
- 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).

# Define a Global VSWITCH

- **SET PORT GROUP PCHNL032 LACP ACTIVE SHARED**
  **SET PORT GROUP PCHNL032 JOIN 6100 6200**
  **DEFINE VSWITCH VSW1 GLOBAL ETHERNET GROUP PCHNL032**




Up to 4 VSWITCHes can share a group within the same LPAR

# Asynchronous Global VSWITCH Initialization

- Guests cannot connect to a VSWITCH until it is defined
- A Global VSWITCH cannot be defined until port group has been formed
- Port group cannot form until IVL is up and has discovered other members of the IVL domain

- **Placing in SYSTEM CONFIG is not sufficient!**
  - If you bring guests up before your global VSWITCH is defined, guests will get NIC errors
  - Defer guest startup to automation (e.g. IBM Operations Manager) which waits for CP messages
    - Or add polling logic to AUTOLOG1 that delays dependent guests until Global VSWITCH is up.

# OSA Priority Queuing

- OSA Express enables the host to provide an ordered set of outbound data queues that OSA will service in order, but without queue starvation.
  - Fair share

- CP creates four queues (in order):
  - System
  - High priority guest
  - Normal priority guest (default)
  - Low priority guest

- You assign a virtual NIC to a queue

# OSA Priority Queuing

- I/O configuration controls the availability of priority queuing in the OSA
  - Enabled by default
  - Use CHPARM to disable – see IOCP book

- You must turn on priority queuing in your VSWITCHes

  DEFINE VSWITCH ... UPLINK PRIQUEUING ON
  SET VSWITCH ... UPLINK PRIQUEING ON

  - SET requires that VSWITCH be DISCONNECTed

- IVLs will automatically use priority queuing, if available

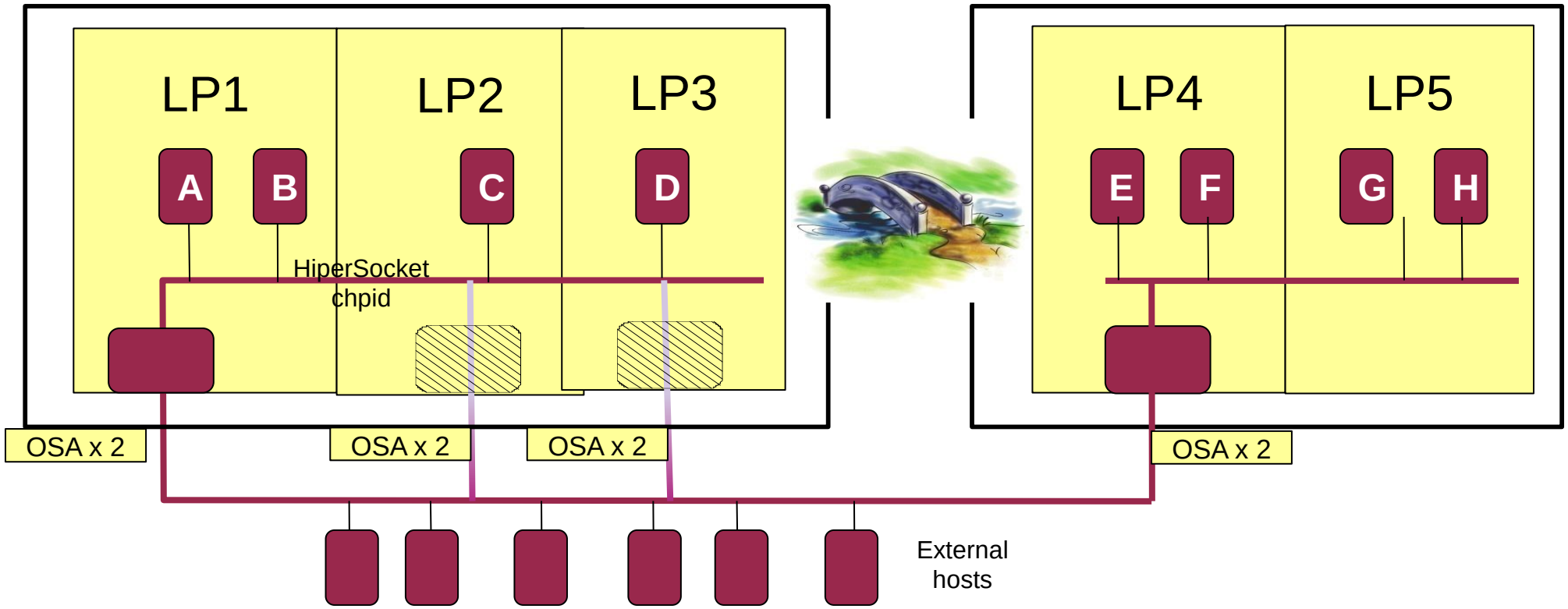
# OSA Priority Queuing

- By default, a virtual NIC is “normal” priority
- Applies only to guest outbound transmissions
- Use NICDEF to specify priority

  NICDEF 800 ....  PQUPLINKTX  LOW | NORMAL | HIGH

  - Can also use SET VSWITCH GRANT, but stick with NICDEF!

# HiperSocket Virtual Switch Bridge




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

# HiperSocket VSWITCH 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)

- Enables cross-CPC Live Guest Relocation
  - Does not work with z/OS LPARs!

# HiperSocket VSWITCH Bridge

- **DEFINE VSWITCH**
  - all the traditional keywords
  - ETHERNET BRIDGEPORT RDEV hs_devaddr [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

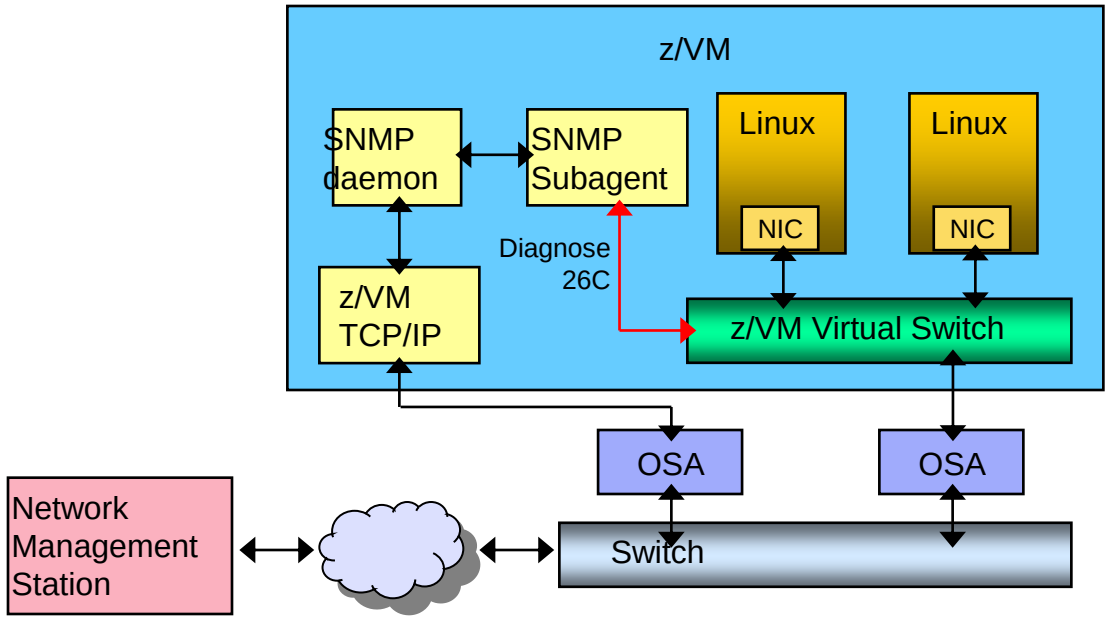
- IEEE 802.1Qbg relaxes prohibition on packet reflection
  - Frames now allowed to be "reflected" back to the origin port
  - Physical switch receives all guest-to-guest traffic
  - Enables use of external packet filtering and monitoring

- SET VSWITCH ... VEPA ON | OFF
  - VEPA and ISOLATE are mutually exclusive
    - VEPA implies isolation
  - VSWITCH will verify external switch support
  - Negotiated use - no hardware configuration required

# z/VM Virtual Switch SNMP MIB

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

- SNMP subagent provides bridge MIB data
  - Defined by RFC 1493
  - Version 1



# 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: Discard Counters

| Discard Counter | Uplink: QUERY VSWITCH ACTIVE | Guest NIC: QUERY NIC USER userid vdev |
|---|---|---|
| RX > 0 inbound | VSWITCH definition mismatch<br>• Unused VLAN ID<br>• VLAN UNAWARE on trunk | Packets are arriving faster than the guest can consume them |
| TX > 0 outbound | Overrun on the physical OSA.<br>• Link is too slow compared to guests<br>• Use faster OSA or link aggregation | • Unauthorized VLAN ID on virtual trunk port<br>• Untagged frame on virtual trunk with NATIVE NONE<br>• Guest configured as VLAN-aware with virtual access port<br>• Overrun target guest |
| To reset | CP SET VSWITCH COUNTERS CLEAR | Resets when NIC is detached |

# Support Timeline

| | |
|---|---|
| z/VM 7.1<br>2019 | ▪ Priority queuing |
| z/VM 6.4<br>2017 | ▪ Unified VSWITCH with NICDEF controls  CP (VM65925), DIRMAINT (VM65926), RACF(VM65931) |
| z/VM 6.3 | ▪ Shared link aggregation  port groups<br>▪ VEPA<br>▪ SET VSWITCH SWITCHOVER |
| z/VM 6.2 | ▪ Port-based configuration provides separate VLAN per virtual access port<br>▪ HiperSocket bridge |
| z/VM 6.1 | ▪ Uplink port can be OSA or guest<br>▪ VLAN UNAWARE, NATIVE NONE |
| z/VM V5 | ▪ Virtual and physical port  isolation<br>▪ z/VM TCP/IP support for Layer 2<br>▪ Link aggregation<br>▪ SNMP monitor<br>▪ Virtual SPAN ports for sniffers<br>▪ Virtual trunk and access port controls<br>▪ Layer 2 (MAC) frame transport<br>▪ External security manager access control |
| z/VM V4<br>2001 | ▪ Layer 3 (IPv4 only) Virtual Switch with IEEE VLANs<br>▪ 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 Altmark**
*Senior Managing z/VM Consultant*

IBM Systems Lab Services
z Systems Delivery Practice

IBM Systems Hardware Client Technical Team
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*