

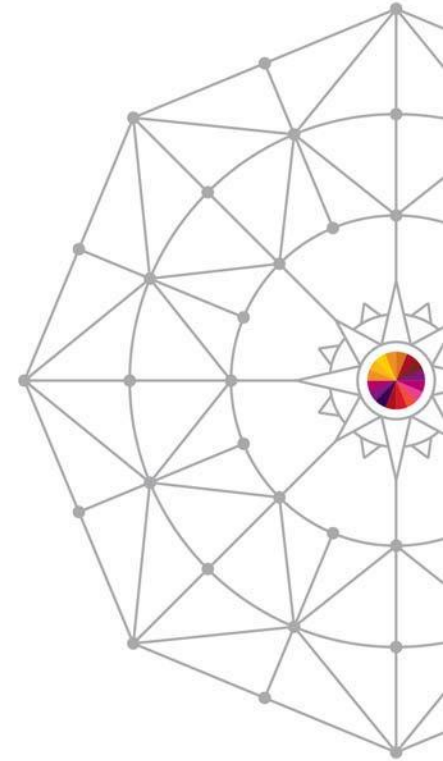
# z/VM Virtual Switch

*Advancing the Art of Virtualization*

*Alan Altmark*

*IBM Senior Managing z/VM Consultant*

*STG Lab Services*



#SHAREorg



SHARE is an independent volunteer-run information technology association that provides **education, professional networking and industry influence.**

# Note

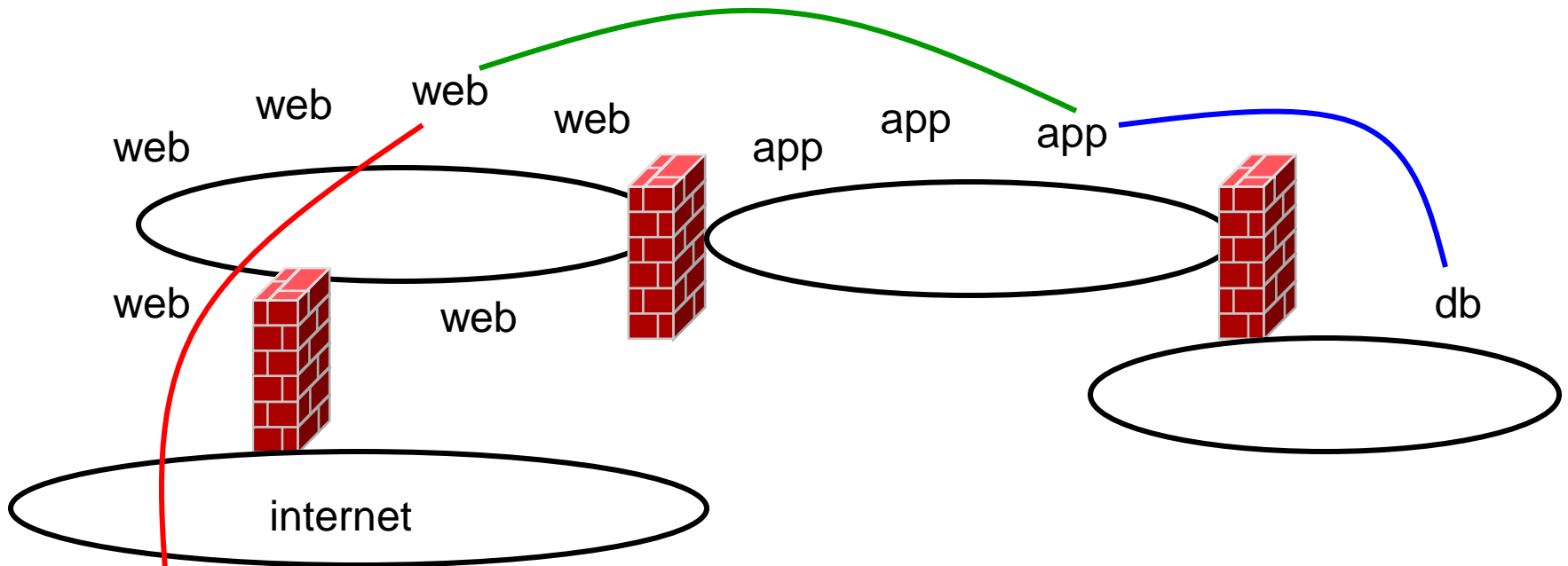
References to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally 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.

IBM, the IBM logo, and [ibm.com](http://ibm.com) are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml).

# Topics

- Overview
- Multi-zone Networks
- Virtual Switch
- Virtual NIC

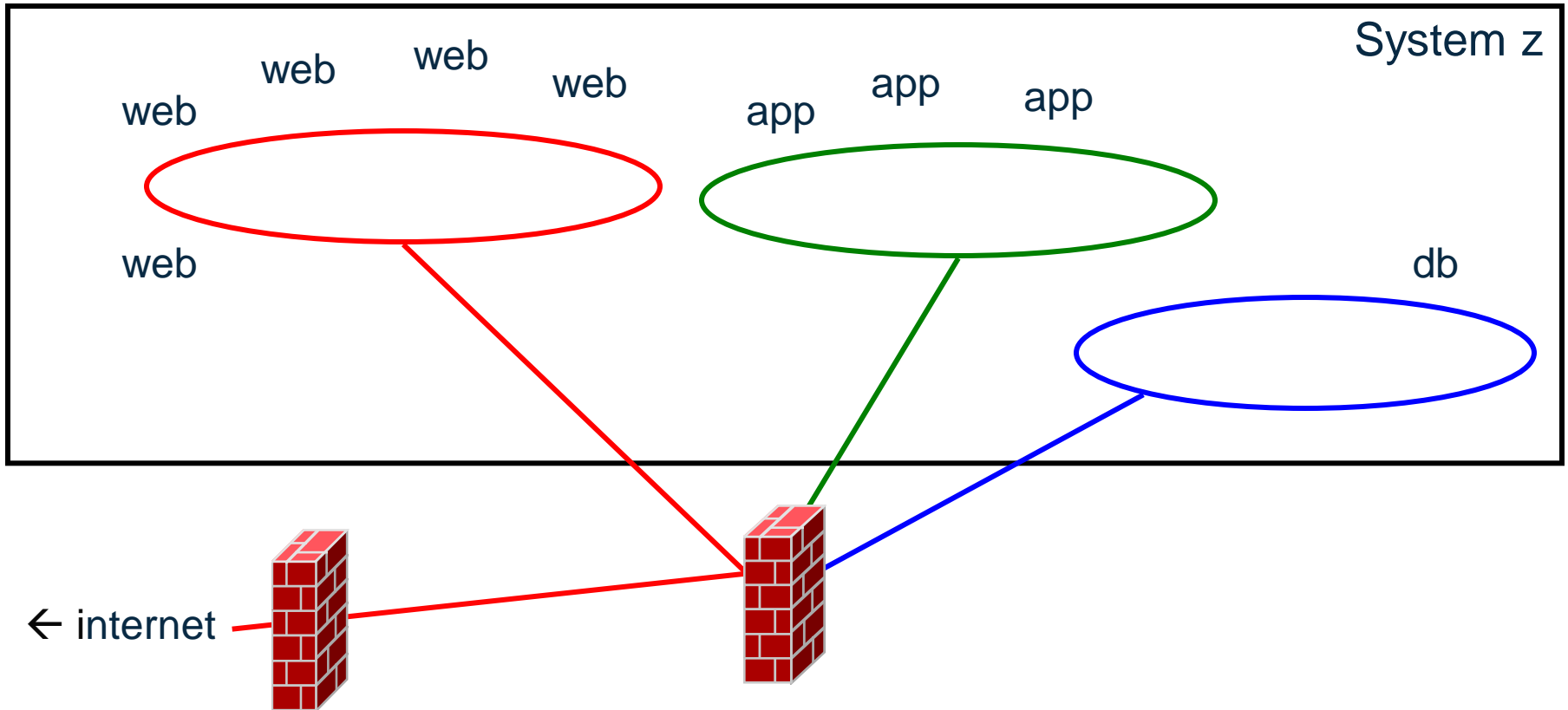
# Multi-Zone Network



A typical 3-tier application

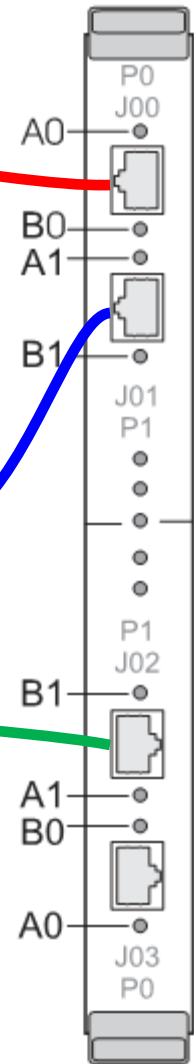


# Multi-zone Network on System z with outboard firewall / router



Q: How to move data in and out of the machine?  
A: z/VM<sup>®</sup> Virtual Switch (VSWITCH)

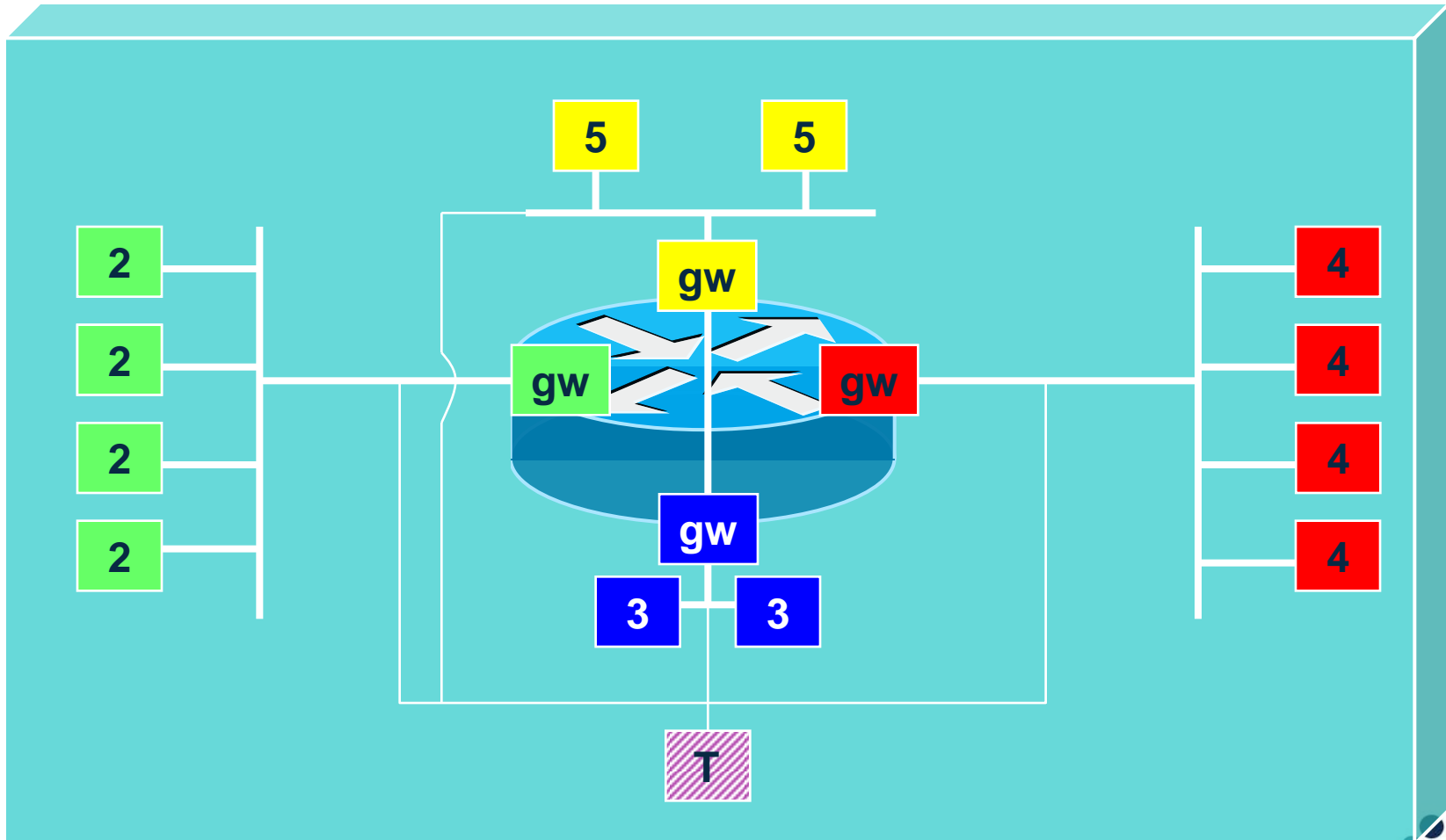
# What's a 'switch' anyway?



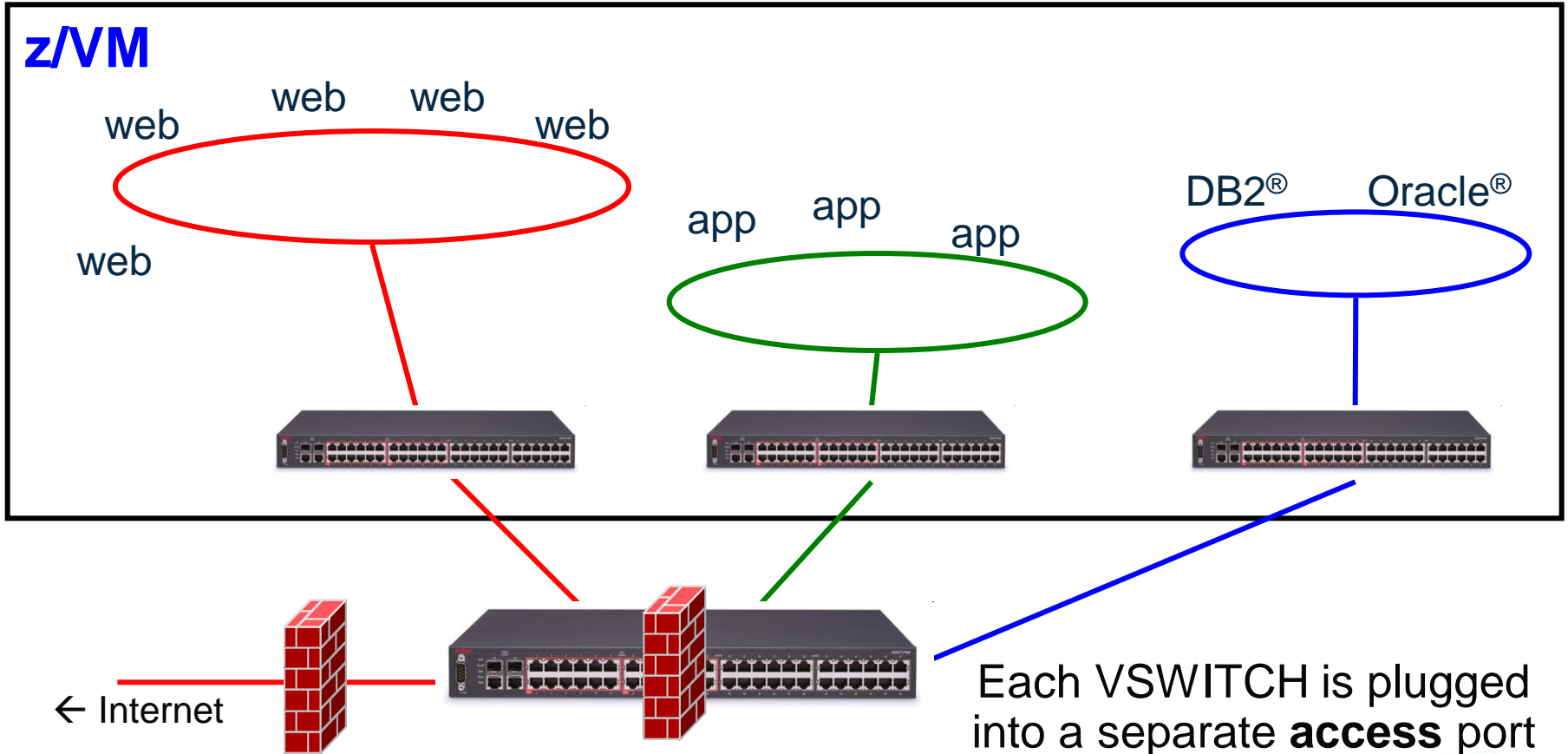
It creates LANs and routes traffic

- ▶ Turn ports on and off
- ▶ Assign a port to a single LAN segment via **access** port
- ▶ Assign a port to multiple LAN segments via **trunk** port
- ▶ Provides LAN sniffer ports

# Imbedded IP router (optional)

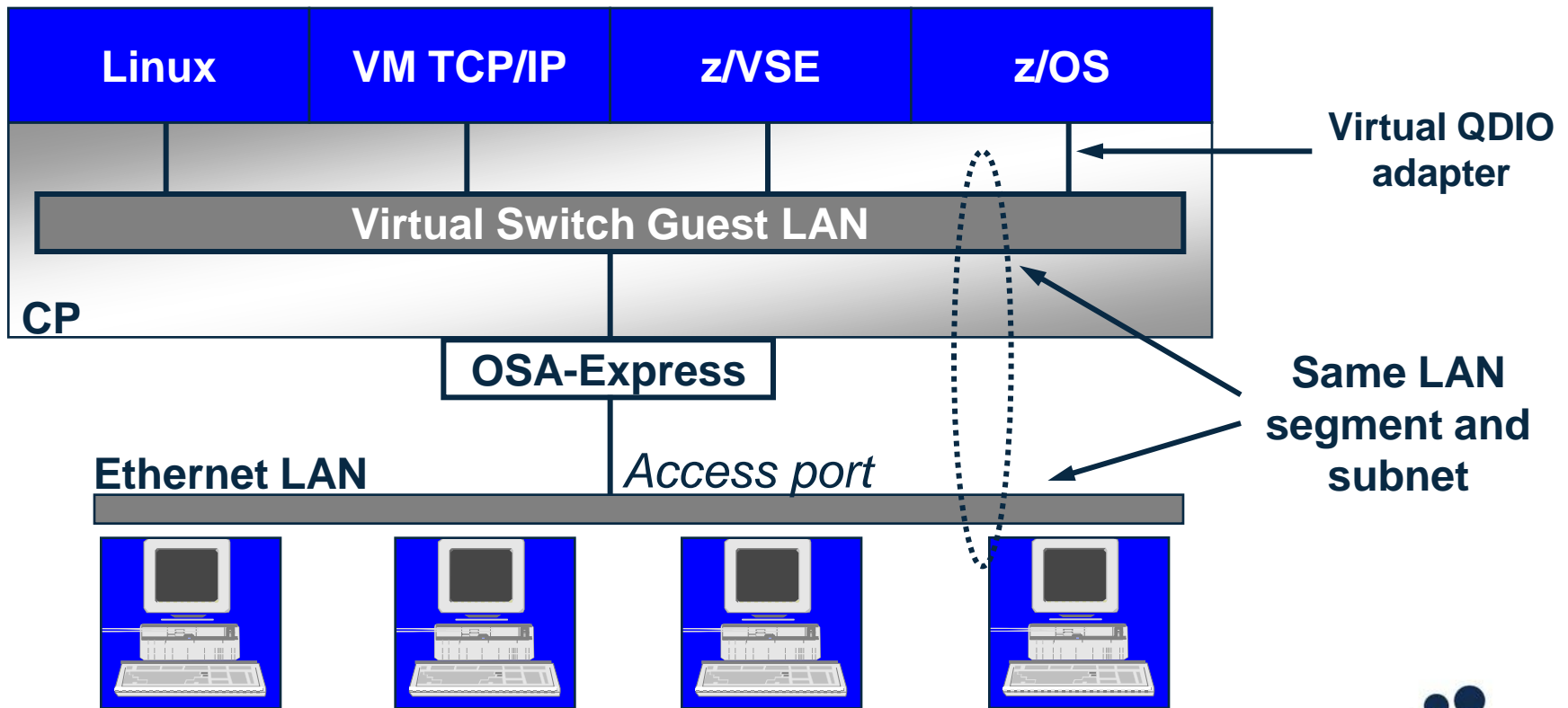


# Option A: VLAN Unaware

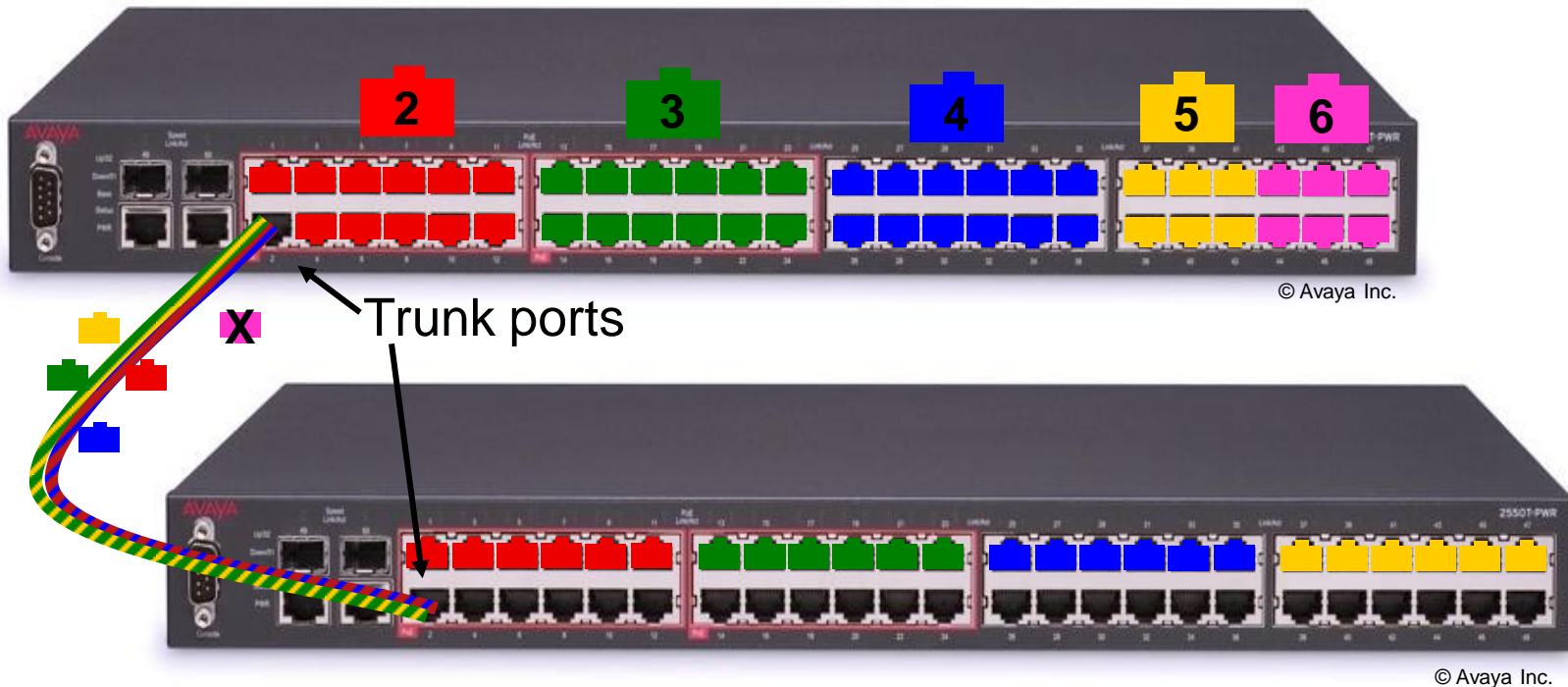




# z/VM Virtual Switch – VLAN unaware Sees only a single LAN segment

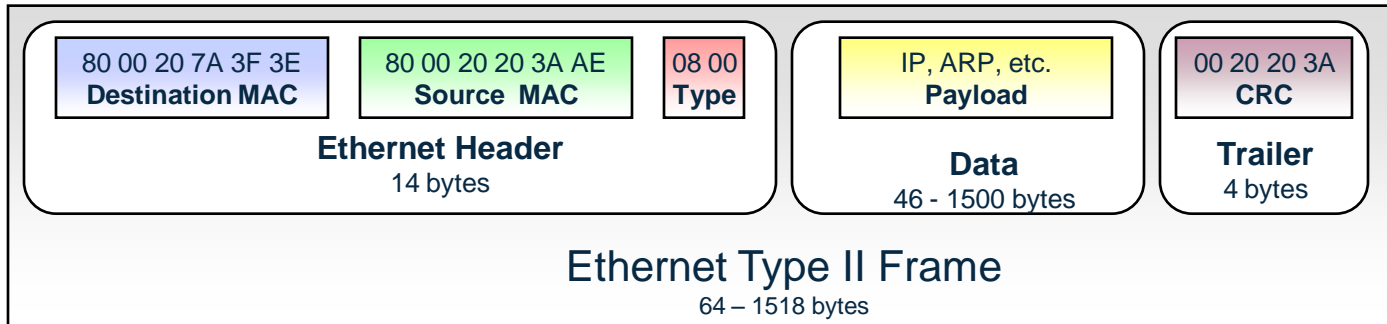


# IEEE VLANs using Trunk port



- ▶ If you run out of ports, you don't throw it away, you "trunk" it to another switch to "bridge" LAN segments together
- ▶ IEEE standards provide a way for trunk ports to exchange data for multiple authorized LAN segments using a single cable.

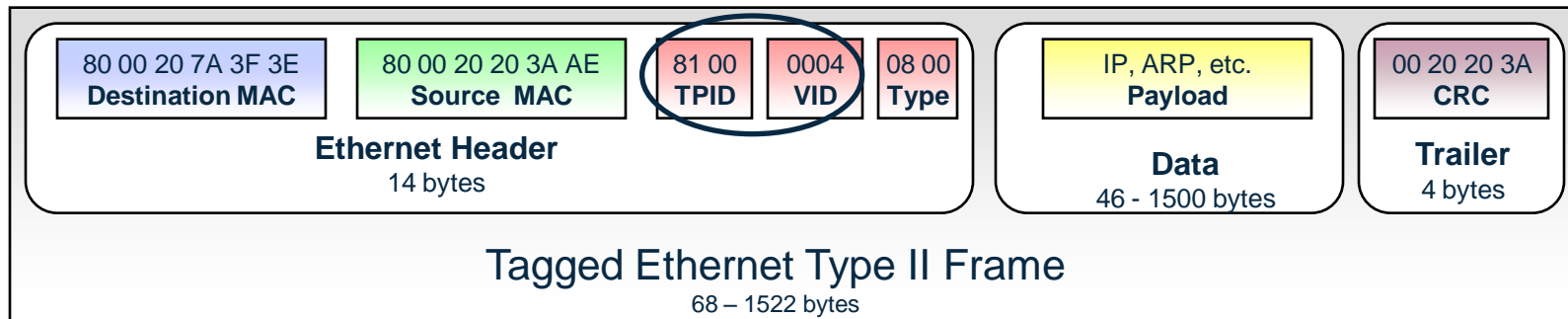
# VLAN tagging



## Access port and Trunk port

When used on a trunk port, the switch will associate (but not tag) it with the **native** VID.

Type/length 0800 means IPv4 (IETF RFC 894)



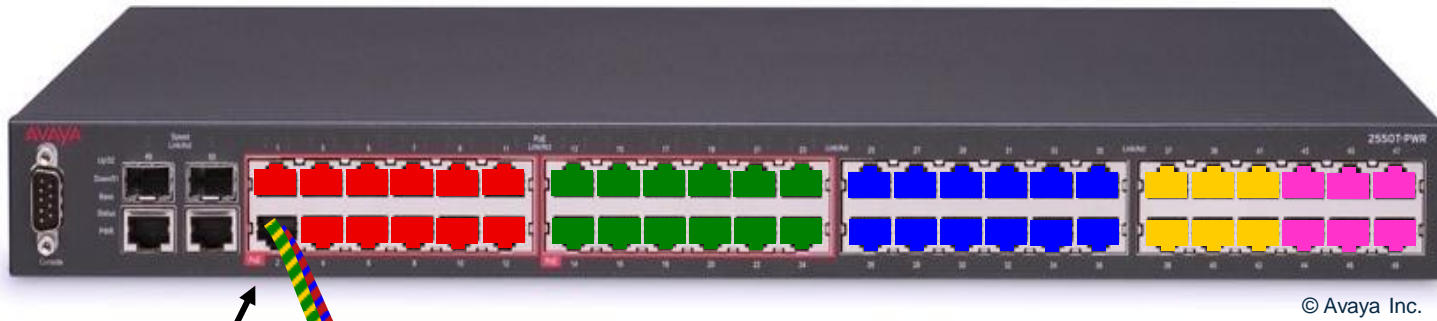
## Trunk port only

Value 8100 in the Type field means a VLAN tag follows, followed by the actual type/length field

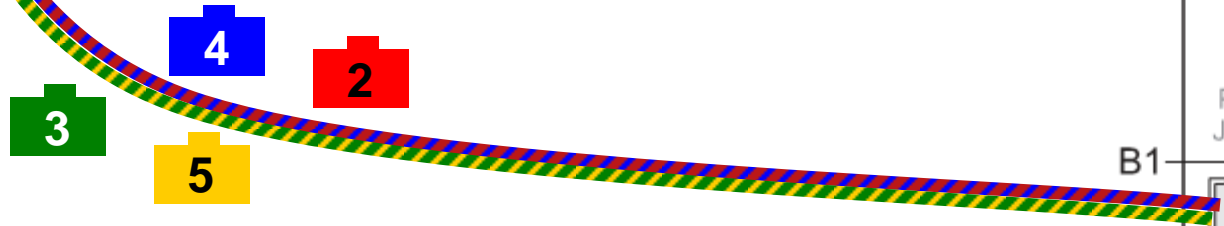
# VLAN Tagging – What is the Native VLAN

- Trunk ports can send and receive tagged and untagged frames
- Tagged frames have an explicit VLAN id contained in them
- Untagged frames have an implicit VLAN id called the “native” VLAN
  - All switches need to be configured with the same VLAN id
  - Default is usually VLAN 1
  - NATIVE value on DEFINE VSWITCH tells CP what VLAN ID inbound untagged frames are to be associated with
  - Guests that generate frames associated with that VLAN ID will be untagged before sending
  - NATIVE NONE is preferred!

# VLAN-aware Virtual Switch

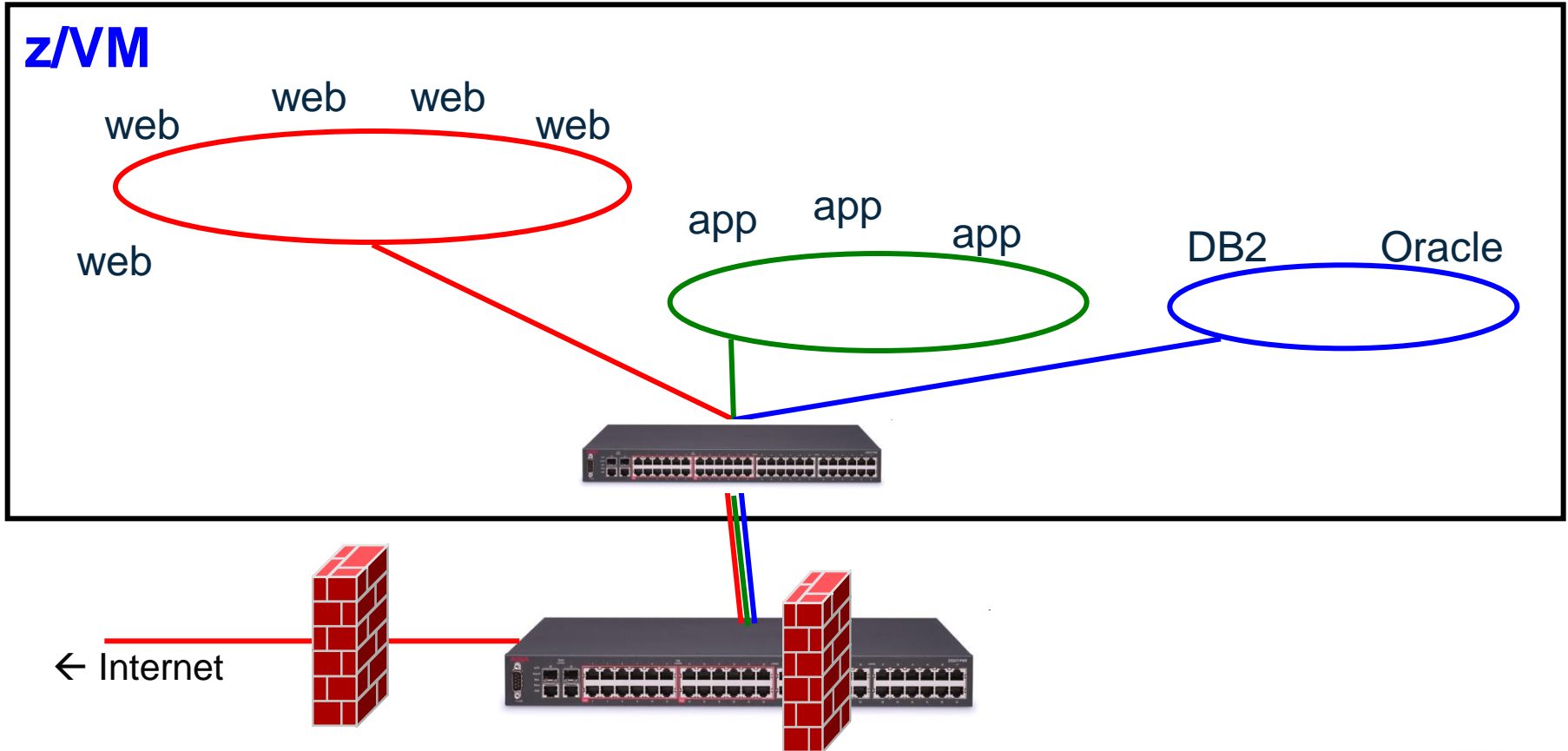


Trunk port



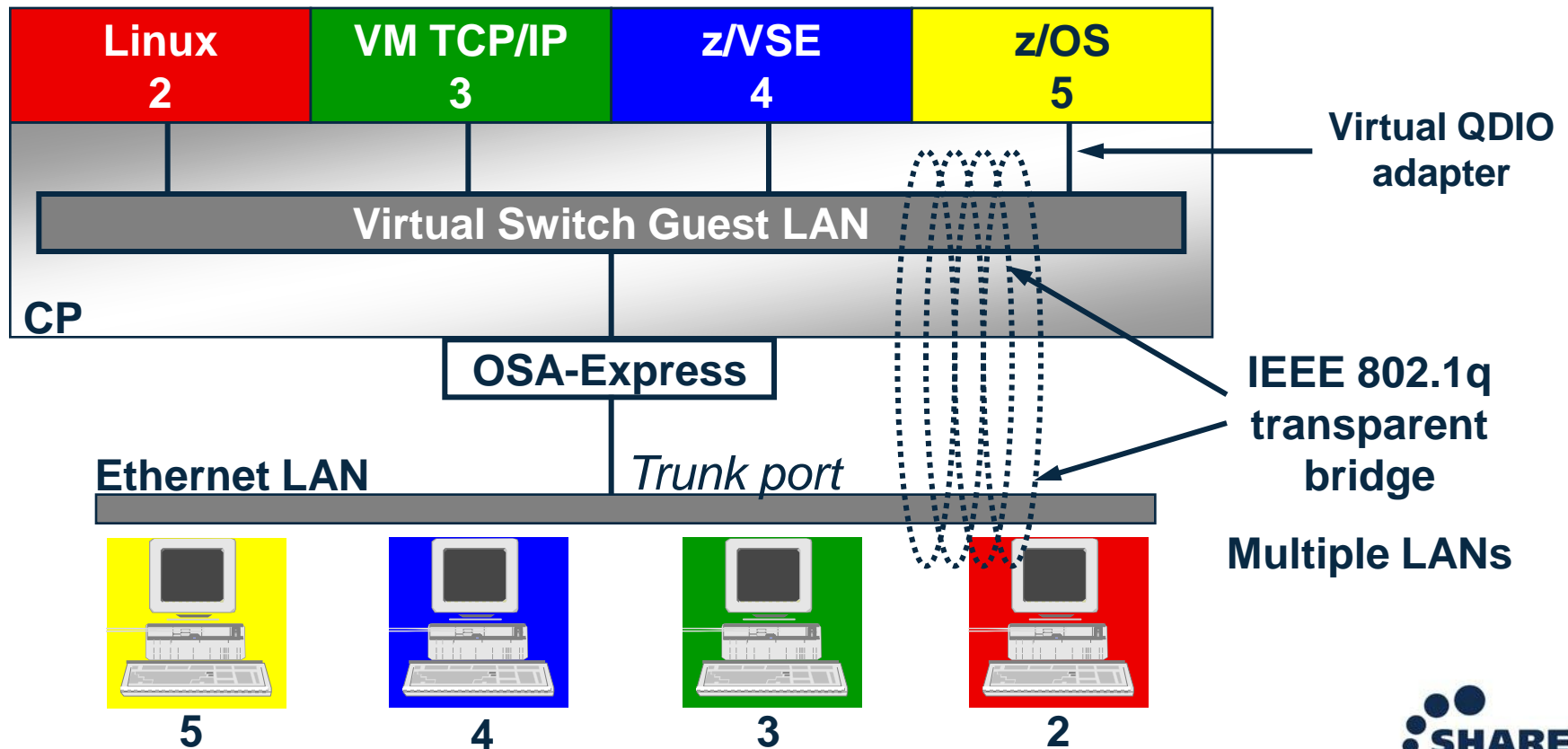
- ▶ Instead of a physical switch, plug in a virtual switch!

# Option B: VLAN Aware



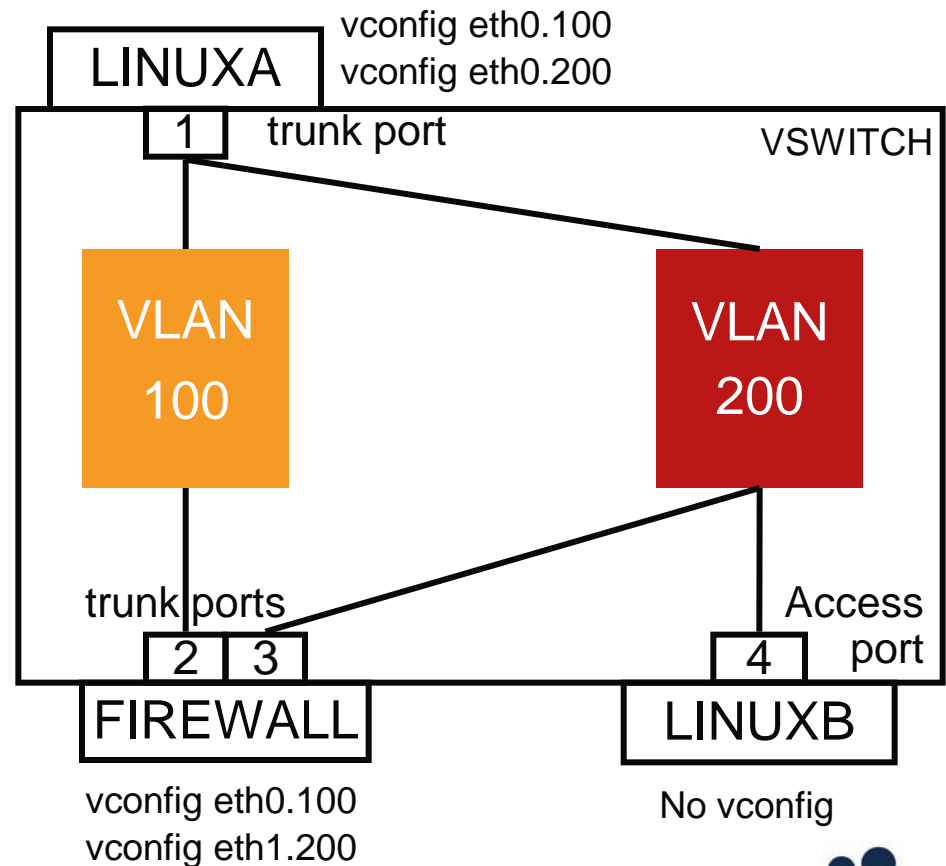
Single VSWITCH plugged into a trunk port

# VLAN-aware Virtual Switch Sees all authorized LAN segments



# User-based VSWITCH access list

- Implicit port definition
  - Ephemeral port number
  - Assigned in order defined
- VLAN assignment applies to all coupled NICs for the authorized user
- Port type applies to all coupled NICs for the authorized user
- SET VSWITCH GRANT
  - ESM controls override CP





# Primary Virtual Switch Attributes

- An associated **controller** virtual machine
- Mode of operation: Layer 2 or Layer 3
- Port-based or user-based access list
  - Permitted user IDs
  - VLAN assignments
- Associated uplink: OSA, virtual NIC, or none

# Layer 2 and Layer 3

## An OSA Point of View

- Layer 2 – Host sends/receives raw ethernet frames to OSA
  - Any protocol: IP, SNA, NETBIOS, AppleTalk, experimental, ...
  - CP registers virtual NIC MAC addresses with OSA so it can route inbound frames appropriately
    - Burned-in MAC address not used
  - Guest sends raw frame with its origin and target MAC address
  - Guest handles ARP
- Layer 3 – Host transfers only IP packets to OSA
  - CP registers guest IP addresses with OSA so it can route inbound packets properly
  - OSA places outbound packet in ethernet frame using burned-in MAC address
  - OSA handles ARP

# Layer 2 and Layer 3


## A Network Engineer's Point of View

- Layer 2 – Ethernet
  - Protocol agnostic
  - Knows which MACs are associated with which ports
    - Filters based on unicast v. multicast v. broadcast
- Layer 3 – Network Protocol
  - All the functions of a layer 2 switch
  - PLUS understands network (not just port-level) addressing
  - PLUS provides interconnect function among attached networks
    - “default gateway”
  - Which means it understands the protocol: IP, SNA, ...

# Setting defaults and limits

- Global attributes in the VMLAN statement in SYSTEM CONFIG:

## VMLAN

LIMIT TRANSIENT INFINITE | *maxcount* 


MACPREFIX *prefix1*

- For CP-assigned MACs

USERPREFIX *prefix2*

- For user-assigned MACs

MACIDRANGE SYSTEM *x-y* [USER *a-b*]

MACPROTECT OFF | ON 

- LIMIT TRANSIENT 0 prevents dynamic definition of Guest LANs by class G users – Don't use Guest LANs
- MACPROTECT ON prevents guests from changing their assigned MAC address

# Virtual MAC Addresses

- MAC prefix = high-order 3 bytes of MAC address
  - 02:00:01
- MAC ID = low-order 3 bytes of MAC address
  - 00:01:23
- Concatenate to create virtual MAC address
  - 02:00:01:00:01:23

# Virtual MAC Addresses

- **VMLAN MACPREFIX** in SYSTEM CONFIG
  - Set MAC prefix for CP-generated MAC addresses
  - Each instance of CP should have a different MACPREFIX
    - **Must** be different for Single System Image (enforced)
    - Avoids duplicate MAC addresses
  
- **VMLAN USERPREFIX** in SYSTEM CONFIG
  - Set MAC prefix for user-defined MAC addresses
  - Defaults to MACPREFIX value
    - **Must** be the same as MACPREFIX in Single System Image (enforced)
      - Ensures that user-defined MAC addresses are unique within the SSI cluster



# Virtual MAC Addresses

- **VMLAN MACIDRANGE** controls allocation of static (USER) and dynamic (SYSTEM) MAC addresses
  - Ensure no conflicts
  - USER range is a subset of SYSTEM range
  - Static MAC IDs must come from USER range
  - Not applicable to SSI
  - Default is entire range
  
- ```
VMLAN MACIDRANGE SYSTEM 000001-002FFF
USER 002000-002FFF
```

# Create a Layer 2 Virtual Switch

- SYSTEM CONFIG or CP command:

```
DEFINE VSWITCH name ETHERNET

    [RDEV NONE | cuu [cuu [cuu]] ]
    [GROUP group name]
    [BRIDGEPORT cuu [PRIMARY] ]
    [USERBASED | PORTBASED]

    [MACPROTECT UNSPECIFIED | ON | OFF]

    [VLAN UNAWARE | VLAN AWARE | VLAN vid]
    [NATIVE 1 | NATIVE vid | NATIVE NONE]

    [CONNECT | DISCONNECT | NOUPLINK]
    [PORTTYPE ACCESS | PORTTYPE TRUNK]

MODIFY VSWITCH name ISOLATION OFF | ON
SET
```





# Create a Layer 3 Virtual Switch

- SYSTEM CONFIG or CP command:

```
DEFINE VSWITCH name IP
MODIFY      [RDEV NONE | cuu [cuu [cuu]] ]
SET        [GROUP group_name]

           [NONROUTER | PRIROUTER]

           [VLAN UNAWARE | VLAN AWARE | VLAN vid]
           [NATIVE 1 | NATIVE vid | NATIVE NONE]

           [ISOLATION OFF | ON]

           [CONNECT | DISCONNECT | NOUPLINK]
           [PORTTYPE ACCESS | PORTTYPE TRUNK]
           [CONTROLLER * | CONTROLLER userid]
```

# User-based Virtual Switch access list

- Specify after DEFINE VSWITCH statement in SYSTEM CONFIG to add users to access list

```
MODIFY VSWITCH name GRANT userid
SET [VLAN vid1 vid2 vid3 vid4]
    [PORTTYPE ACCESS | TRUNK]
    [PROMiscuous | NOPROMiscuous]

SET VSWITCH name REVOKE userid
```

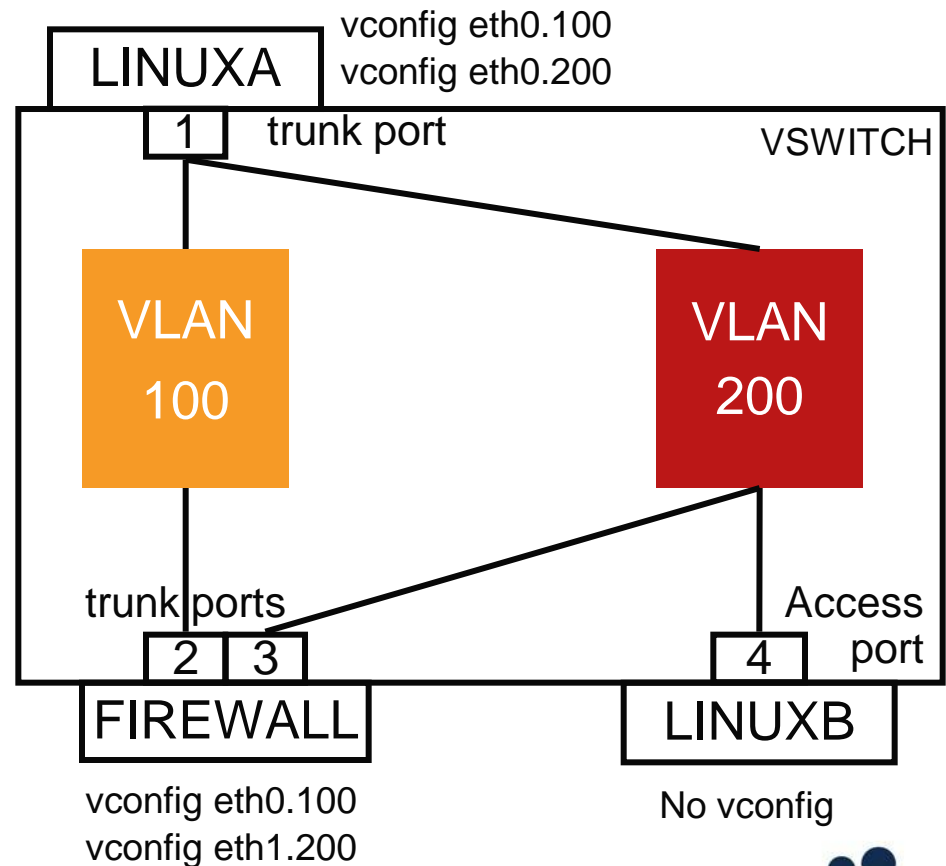
## Examples:

```
MODIFY VSWITCH SWITCH12 GRANT LNX01 VLAN 3
CP SET VSWITCH SWITCH12 GRANT LNX02 PORTTYPE TRUNK
    VLAN 4 20-22 29 302

CP SET VSWITCH SWITCH12 GRANT LNX02 PROMISCUOUS
```

# User-based VSWITCH access list

- Implicit port definition
  - Ephemeral port number
  - Assigned in order defined
- VLAN assignment applies to all coupled NICs for the authorized user
- Port type applies to all coupled NICs for the authorized user
- SET VSWITCH GRANT
  - ESM controls override CP
  - If ESM defers, default VLAN ID will be used!



# User-based VSWITCH access list

```
define vswitch vsw1 vlan aware native none
set vswitch vsw1 grant LINUXA porttype trunk VLAN 100 200
set vswitch vsw1 grant FIREWALL porttype trunk VLAN 100 200
set vswitch vsw1 grant LINUXB VLAN 200
```

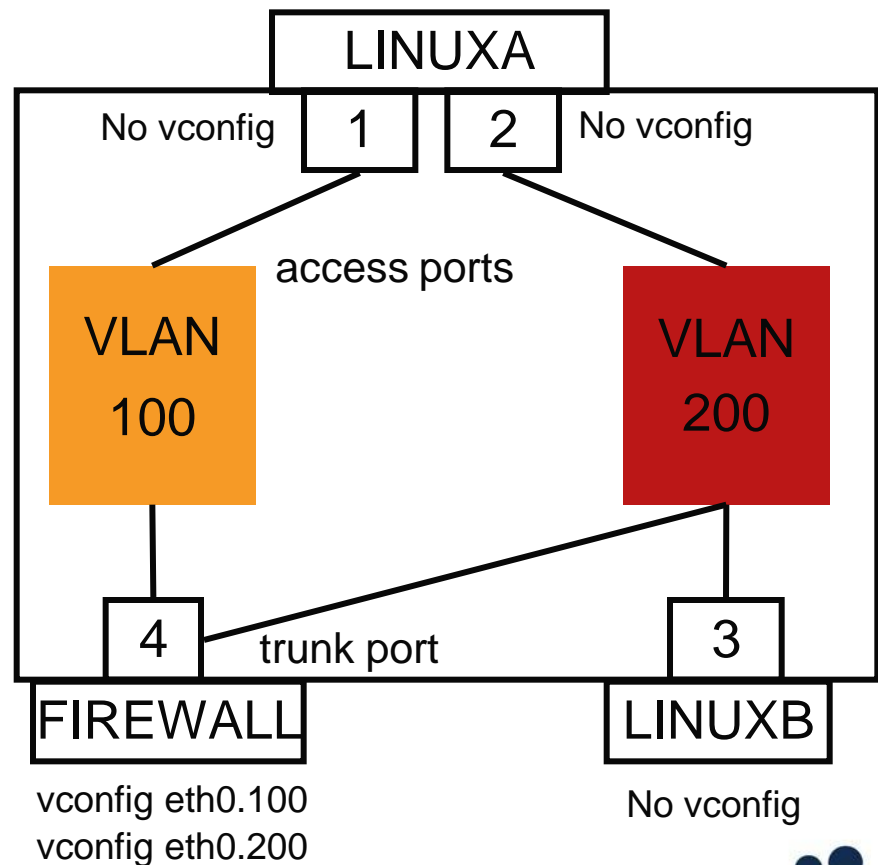
```
LINUXA:  NICDEF 4E0 TYPE QDIO LAN SYSTEM VSW1
          + vconfig eth0.100
          + vconfig eth0.200
```

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

```
FIREWALL:  NICDEF 4E0 TYPE QDIO LAN SYSTEM VSW1
            NICDEF 5E0 TYPE QDIO LAN SYSTEM VSW1
            + vconfig eth0.100
            + vconfig eth1.200
```

# 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 LINUXB
set vswitch vsw1 portnumber 4 userid FIREWALL 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
```

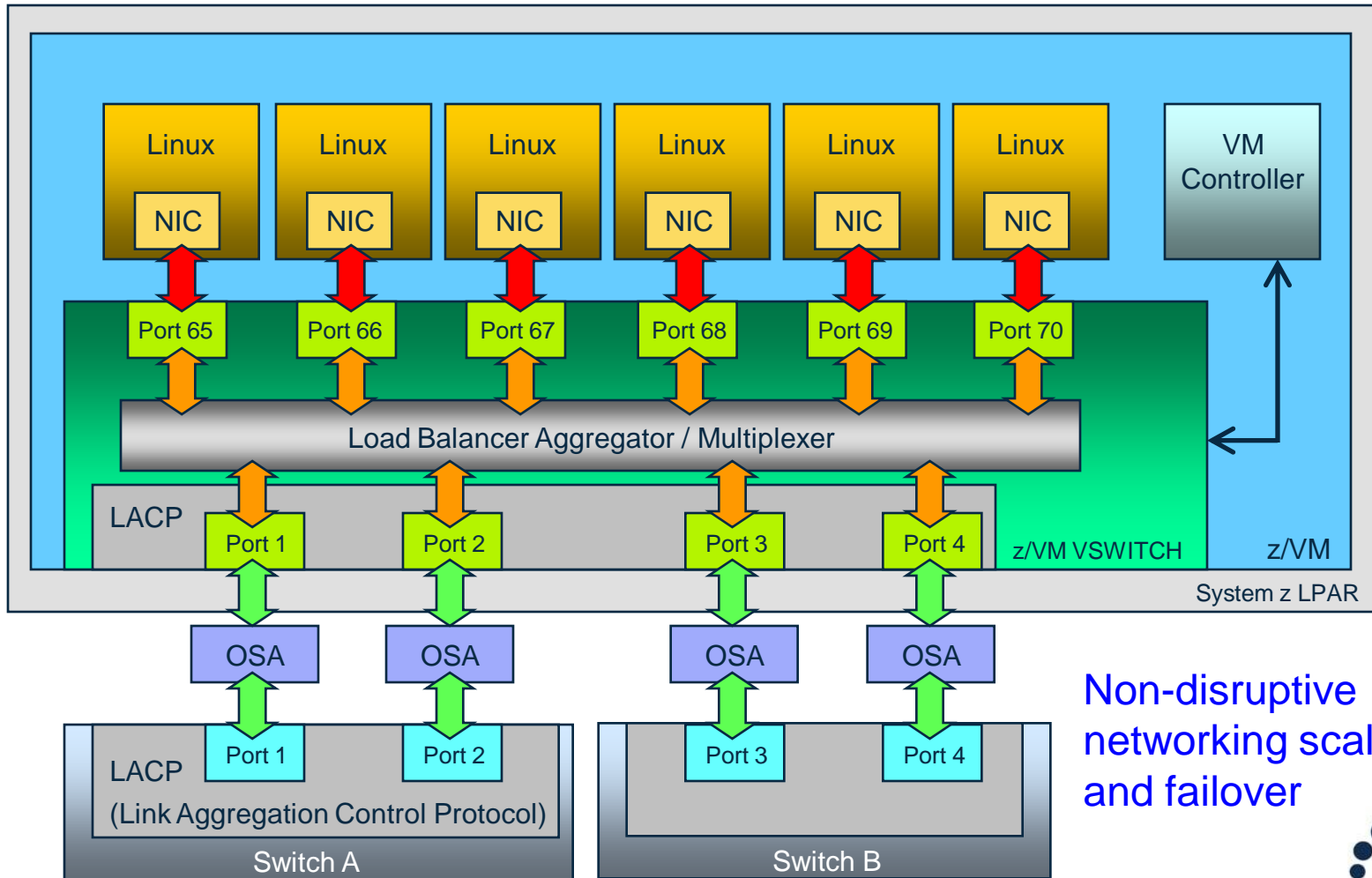
```
LINUXB:  NICDEF 4E0 TYPE QDIO LAN SYSTEM VSW1
```

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

# Additional security controls

- Virtual Sniffers
  - Guest must be authorized via SET VSWITCH or security server
  - Guest enables promiscuous mode using CP SET NIC or via device driver controls
    - E.g. tcpdump -P
  - Guest receives copies of all frames sent or received for all authorized VLANs
  - Not needed when VEPA is used
- Port Isolation (aka “QDIO connection isolation”)
  - Stop guests from talking to each other, even when in same VLAN
  - Shut off OSA “short circuit” to other users (LPARs or guests) of the same OSA port or VSWITCH

# 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

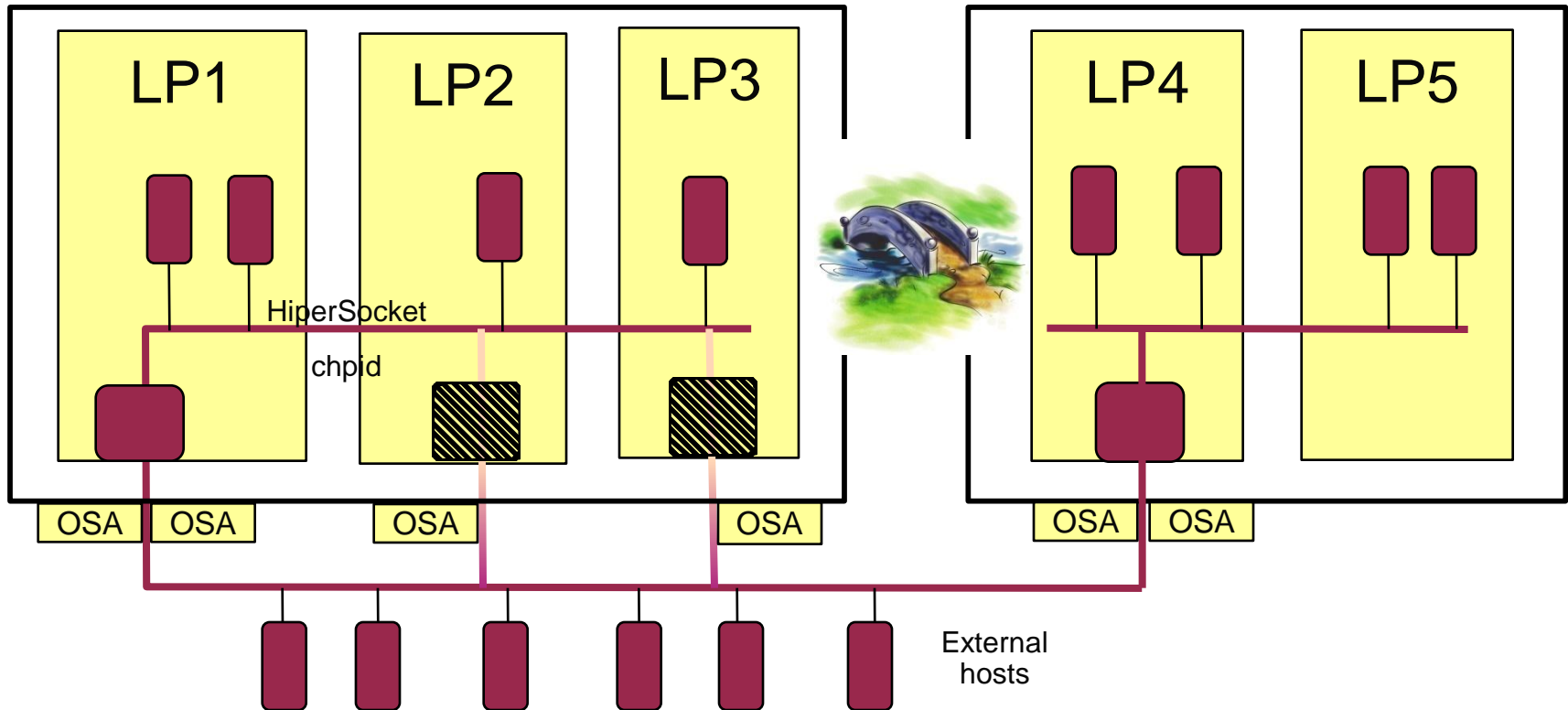
- Define an OSA port group
  - SET PORT GROUP *name* JOIN E100 E200.P1
- DEFINE VSWITCH ... ETHERNET GROUP *name*
- OSA ports cannot be shared with other VSWITCHes or LPARs



# 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

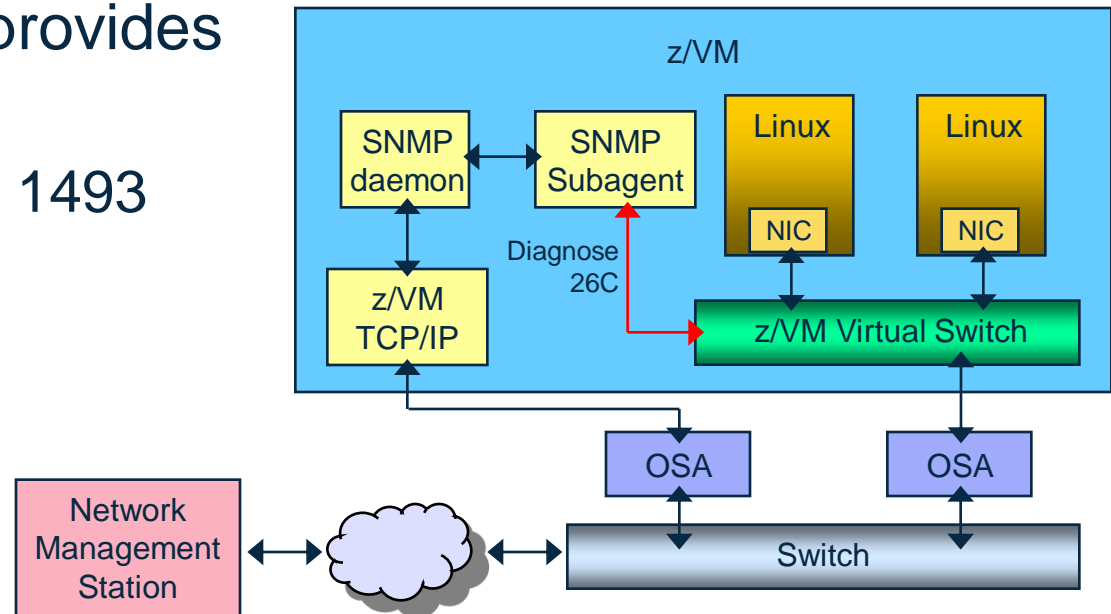


# VEPA - Virtual Ethernet Port Aggregator

- IEEE 802.1Qbg relaxes prohibition on packet reflection
  - Frames now allowed to be "reflected" back to the origin port
  - 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



# Virtual Network Interface Card (NIC)

- A simulated network adapter
- 3 or more devices per NIC
  - More than 3 to simulate port sharing on 2nd-level system or for multiple data channels
- Provides access to Virtual Switch
- Created by NICDEF or CP DEFINE NIC command



# Virtual NIC - User Directory

- One per interface in USER DIRECT file:

```
NICDEF vdev TYPE QDIO  
      [LAN SYSTEM switch]  
      [DEVICES nn]           Combined with VMLAN  
      [MACID xyyzz]         USERPREFIX to create  
                              virtual MAC
```

**Example:**

```
NICDEF 1100 TYPE QDIO LAN SYSTEM SWITCH1 MACID B10006
```

# Virtual NIC - CP Command

- May be interactive with CP DEFINE NIC and COUPLE commands:

```
CP DEFINE NIC vdev TYPE QDIO
```

```
CP COUPLE vdev [TO] owner name
```

**Example:**

```
CP DEFINE NIC 1200 TYPE QDIO
```

```
CP COUPLE 1200 TO SYSTEM SWITCH12
```

# SET NIC

- SET NIC [USER *userid*] *vdev* ...
  - PROMISCUOUS | NOPROMISCUOUS (class G)
  - MACID SYSTEM (class B)
  - MACID USER *hhhhh* (class B)
  - MACPROTECT UNSPECIFIED | OFF | ON (class B)

# VSWITCH Controller

- Virtual machine that handles OSA housekeeping duties
  - Specialized VM TCP/IP stack to start, stop, monitor, and query OSA
  - Not involved in data transfer
- IBM provides DTCVSW1 and DTCVSW2
  - No need to create more unless directed by Support Center
  - Leave them both logged on for redundancy
    - Monitor with system automation!
  - Automatic failover
- Do not ATTACH or DEDICATE devices
  - Handled by CP

# Best Practices

- Use ETHERNET (layer 2) VSWITCH
- Do not specify CONTROLLER
- Do not specify PORTTYPE TRUNK on DEFINE VSWITCH
  - This controls the default **guest** port type, not the OSA!
- Do not put CONTROLLER ON in your own TCP/IP stacks
- Specify MACPROTECT ON and LIMIT TRANSIENT 0 on VMLAN statement in SYSTEM CONFIG

# 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

# Best Practice for VLAN-aware VSWITCH

- DEFINE VSWITCH .....  
    VLAN      AWARE  
    NATIVE    NONE  
    PORTTYPE  ACCESS (or do not specify)
- Explicitly GRANT guest to a particular VLAN ID
- Guest that has not been given access will get errors
- Use ESM and groups to manage VLAN assignments
  - Simplifies VLAN changes

# 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

- VSWITCHes make it easy to control access to the network and simplify server cloning
- 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

# Support Timeline

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

Complete your session evaluations online at [www.SHARE.org/Pittsburgh-Eval](http://www.SHARE.org/Pittsburgh-Eval)

# 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 IT Consultant*

*IBM Systems Lab Services  
and Training*

*z/VM & Linux on System z*

## **IBM**

*1701 North Street  
Endicott, NY 13760*

*Mobile 607 321 7556*

*Fax 607 429 3323*

*Email: [alan\\_altmark@us.ibm.com](mailto:alan_altmark@us.ibm.com)*



Mailing lists: [IBMTCP-L@vm.marist.edu](mailto:IBMTCP-L@vm.marist.edu)  
[IBMVM@listserv.uark.edu](mailto:IBMVM@listserv.uark.edu)  
[LINUX-390@vm.marist.edu](mailto:LINUX-390@vm.marist.edu)

<http://ibm.com/vm/techinfo/listserv.html>