



# The z/VM Virtual Switch: Advancing the Art of Virtual Networking

## Session 12479

Alan Altmark  
Senior Managing z/VM and Linux Consultant  
IBM Lab Services  
*Alan\_Altmark@us.ibm.com*



# 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.

The following terms are trademarks of the International Business Machines Corporation in the United States or other countries or both:

IBM                      IBM logo                      DB2                      z/OS                      z/VM

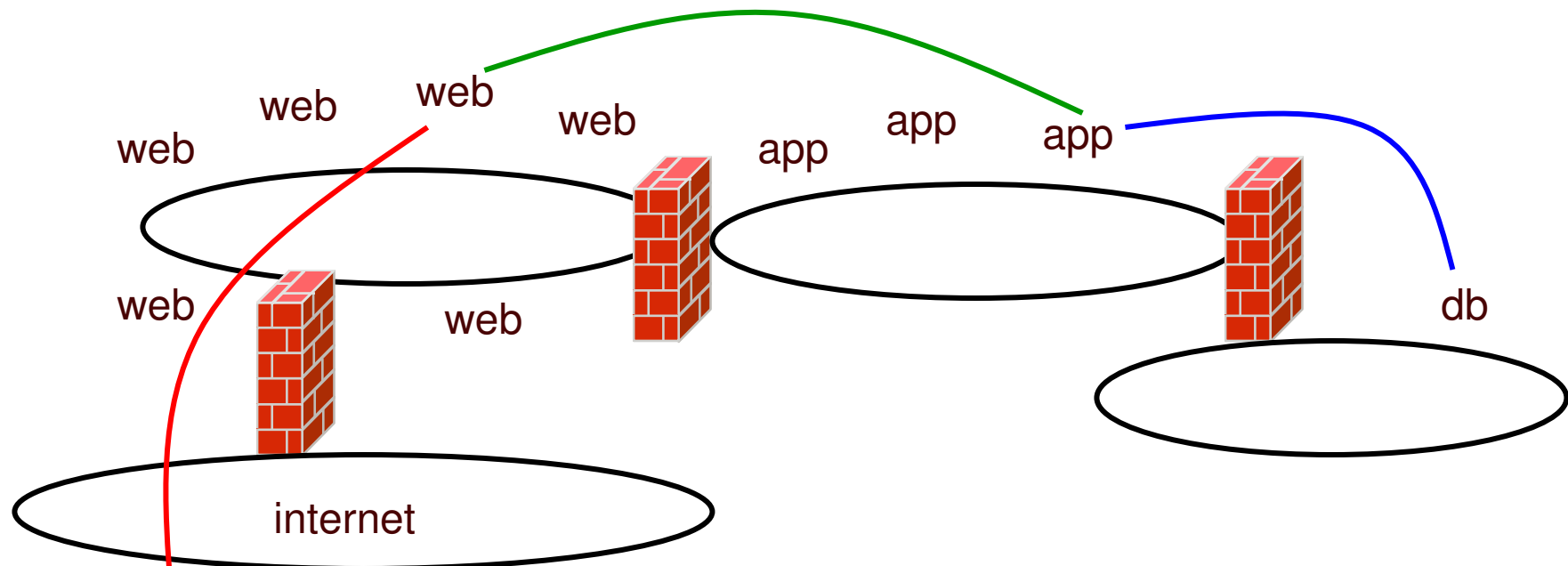
Other company, product, and service names may be trademarks or service marks of others.

# Topics



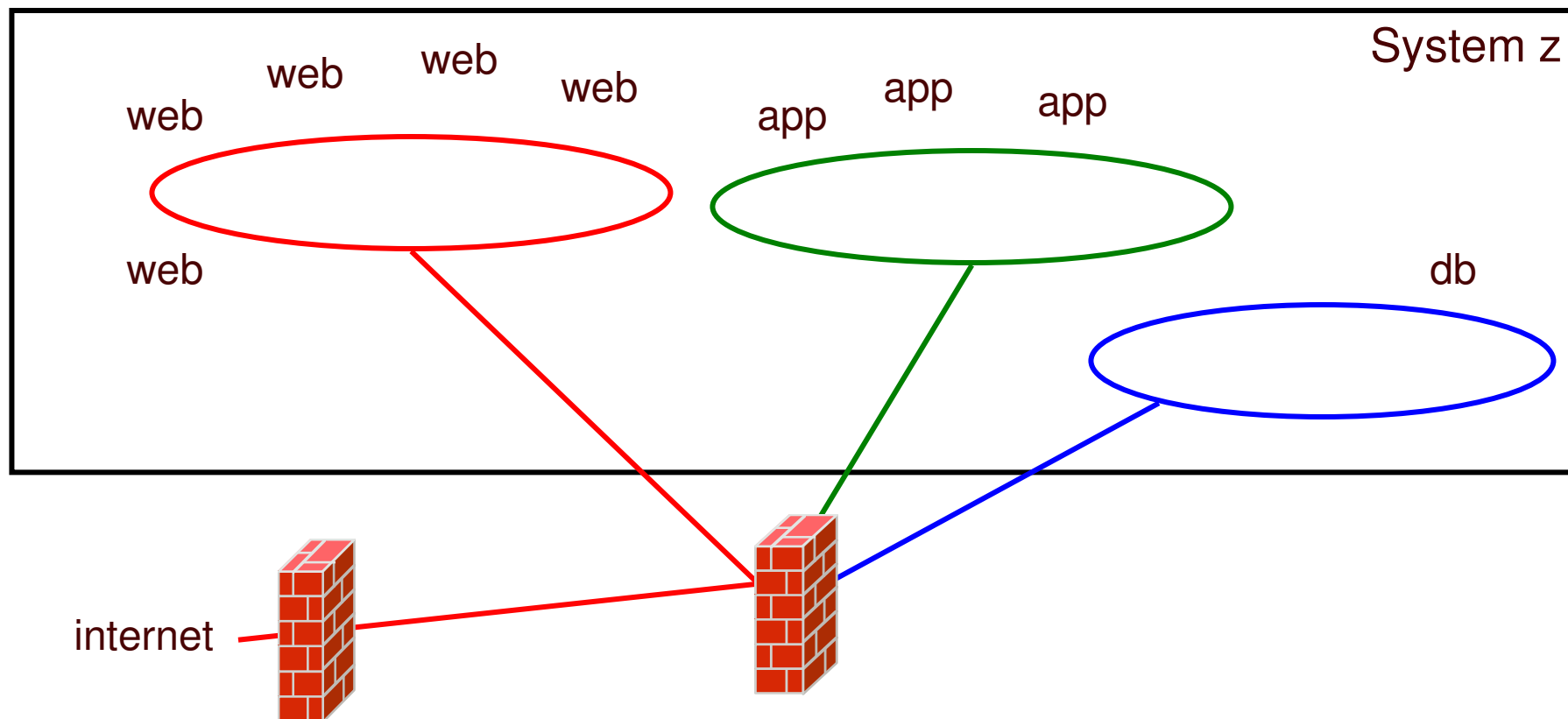
- Overview
- Multi-zone Networks
- Virtual Switch
- Virtual NIC

# Multi-Zone Network



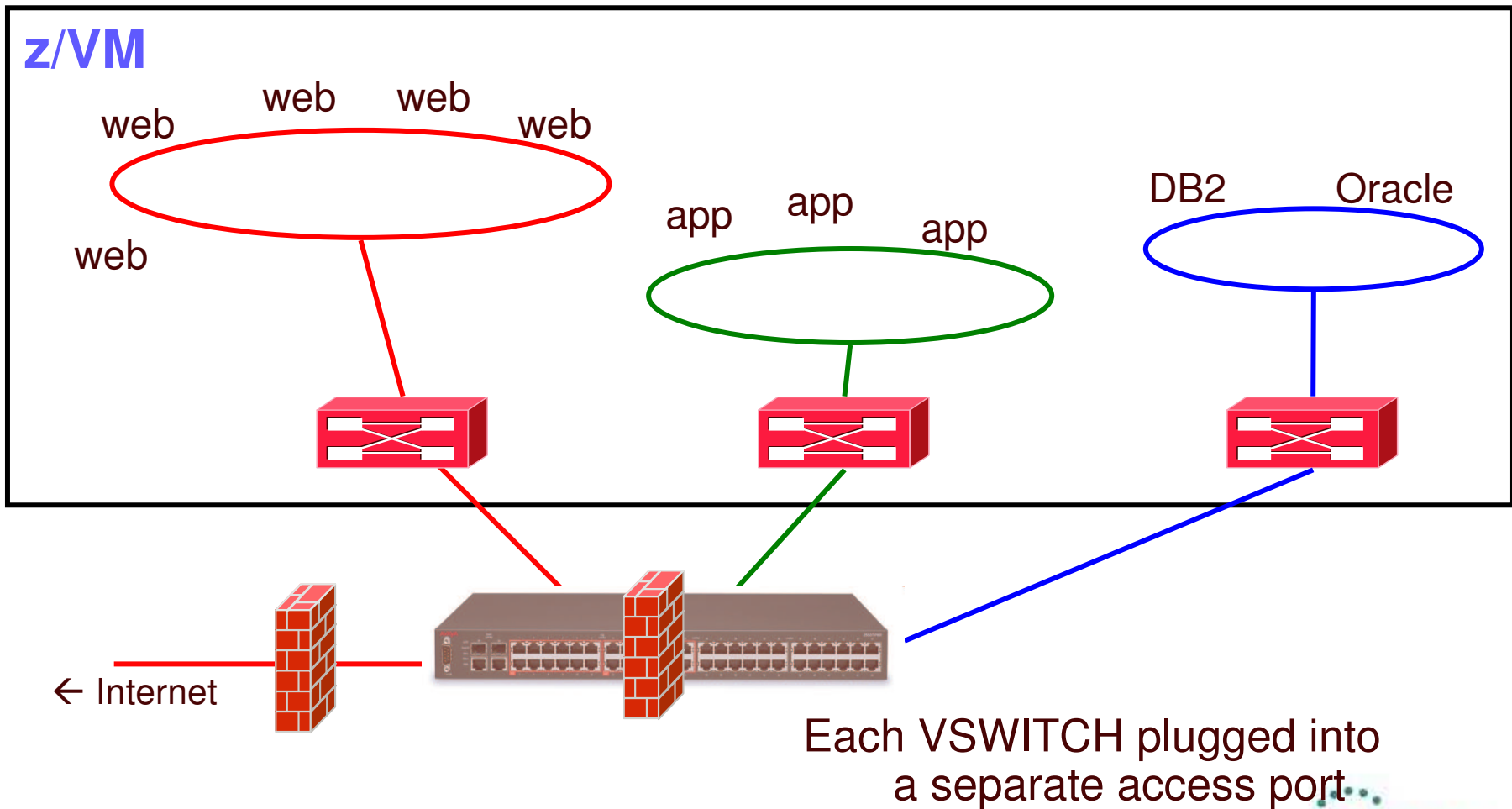
A typical 3-tier application

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



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

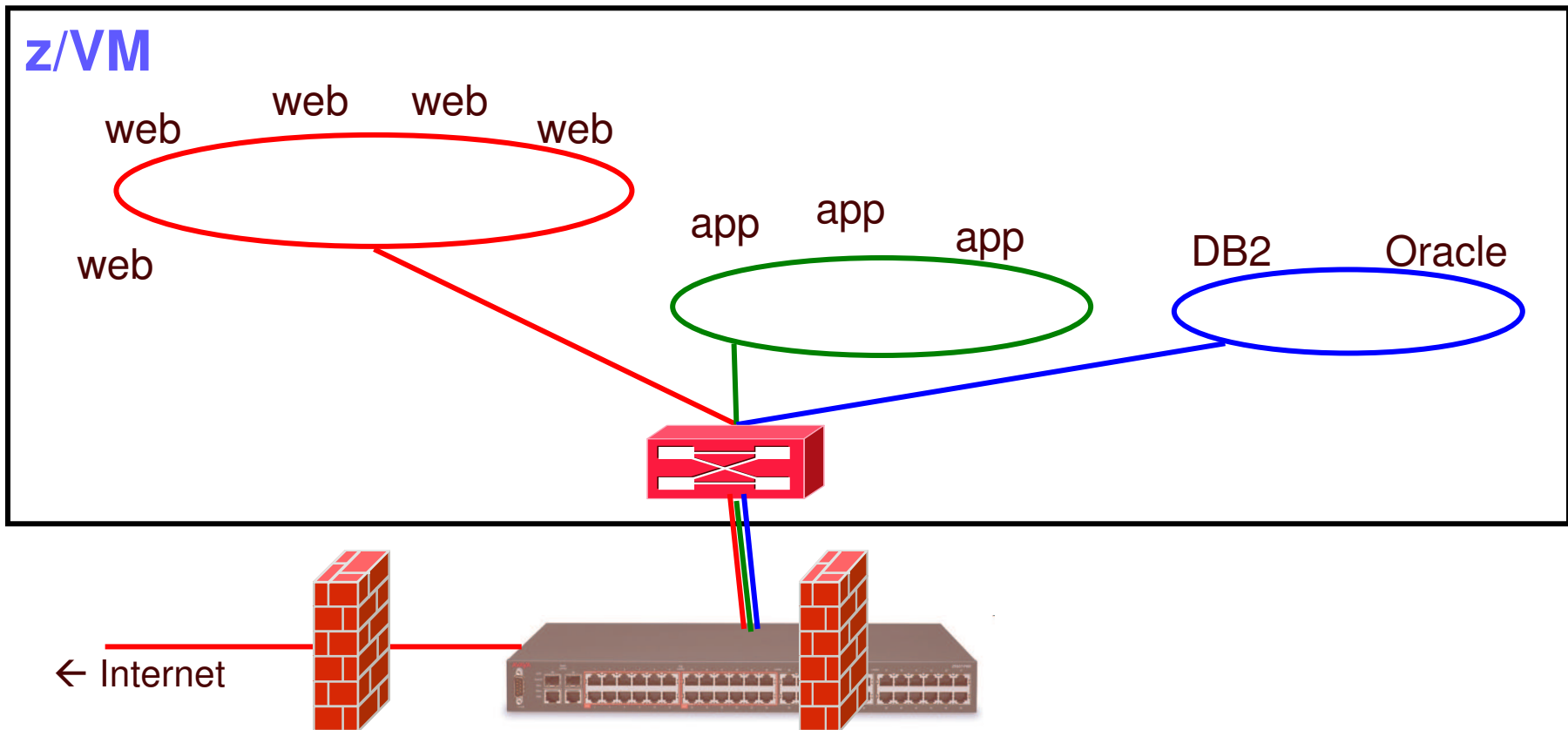
# Option A: VLAN Unaware



Each VSWITCH plugged into a separate access port.



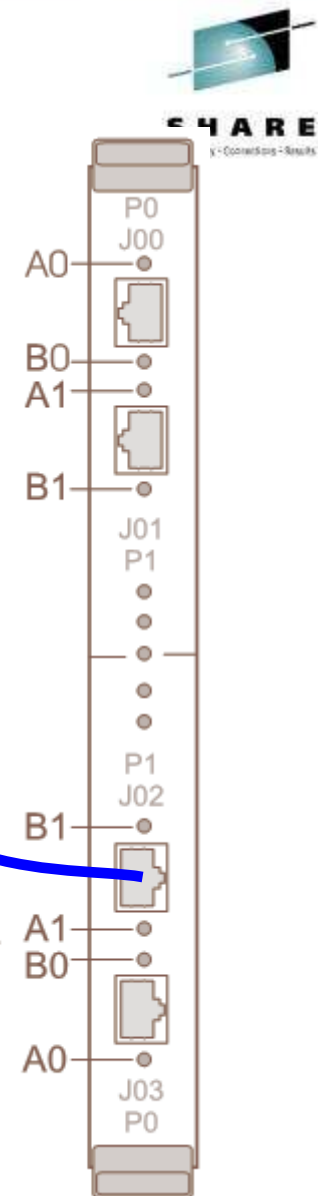
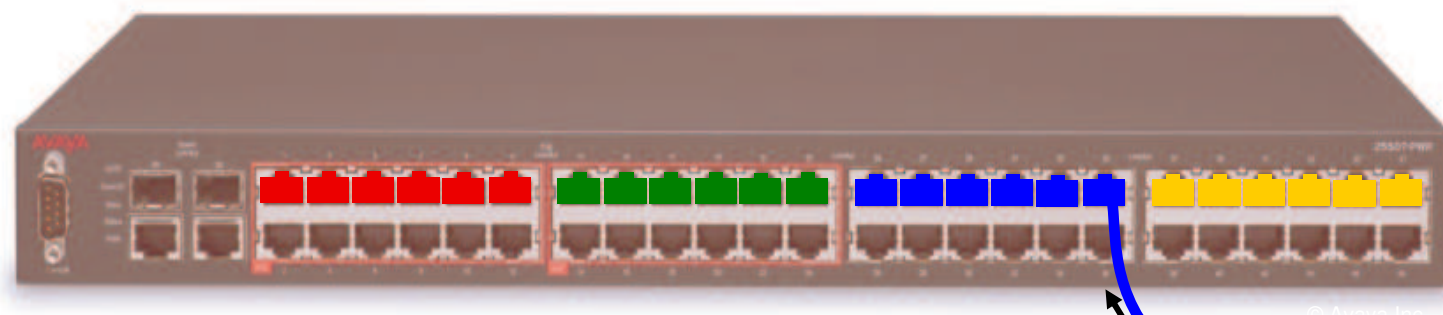
# Option B: VLAN Aware



Single VSWITCH plugged into a trunk port



# What's a 'switch' anyway?

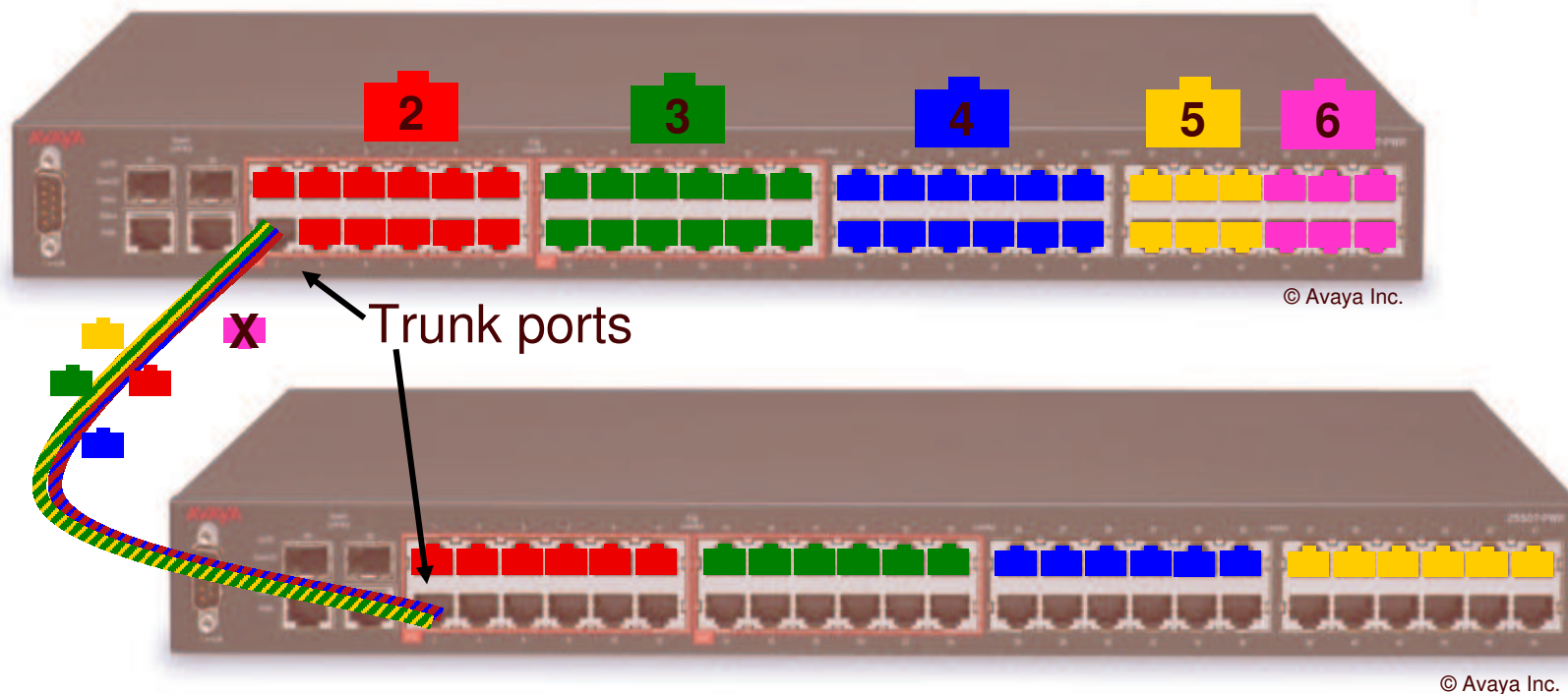


## It creates LANs and routes traffic

- ▶ 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

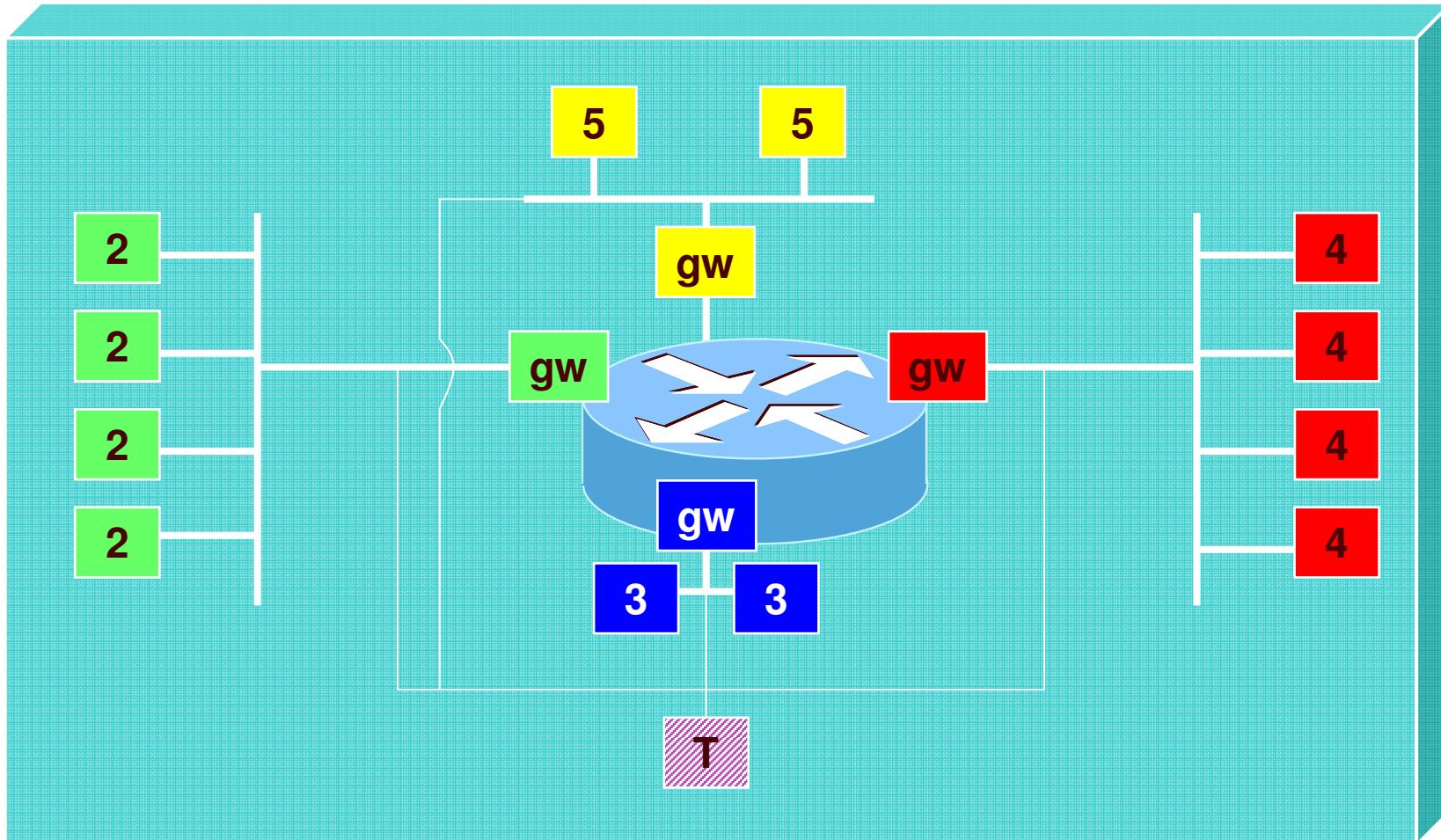


# IEEE VLANs using Trunk port

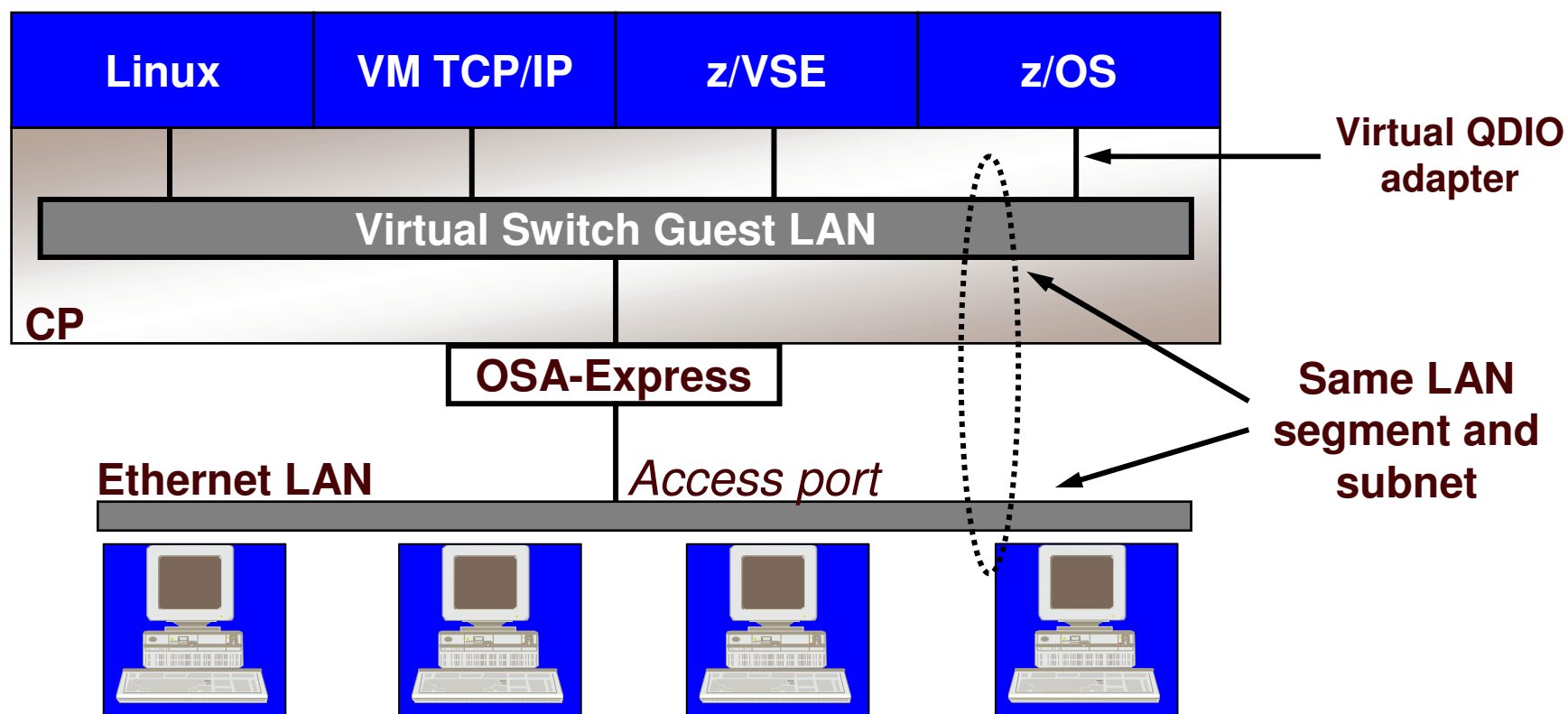


- ▶ If you run out of ports, you don't throw it away, you “trunk” it to another switch to “bridge” LAN segments together
- ▶ IEEE standards provide a way for trunk ports to exchange data for multiple authorized LAN segments using a single cable.

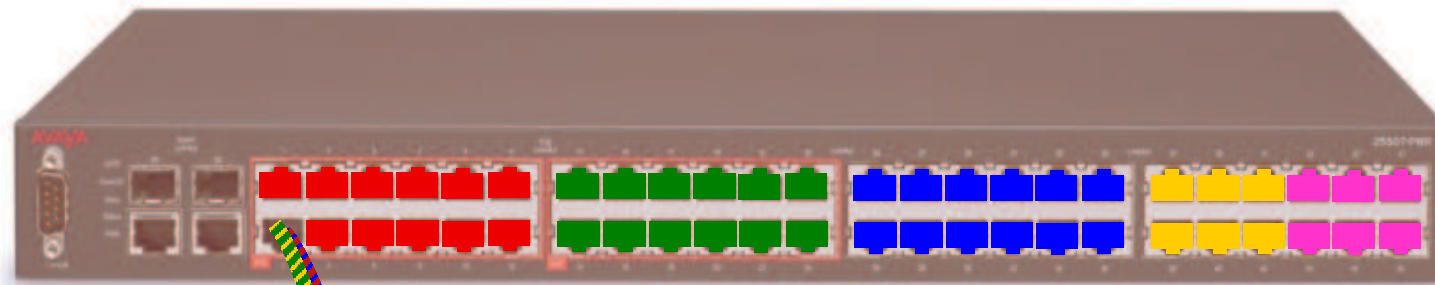
# Imbedded IP router (optional)



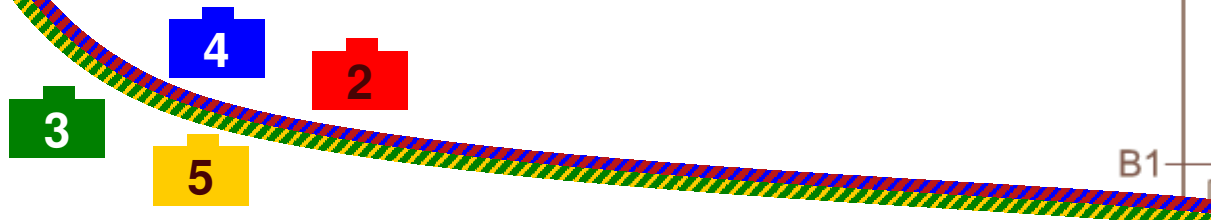
# z/VM Virtual Switch – VLAN unaware Sees only a single LAN segment



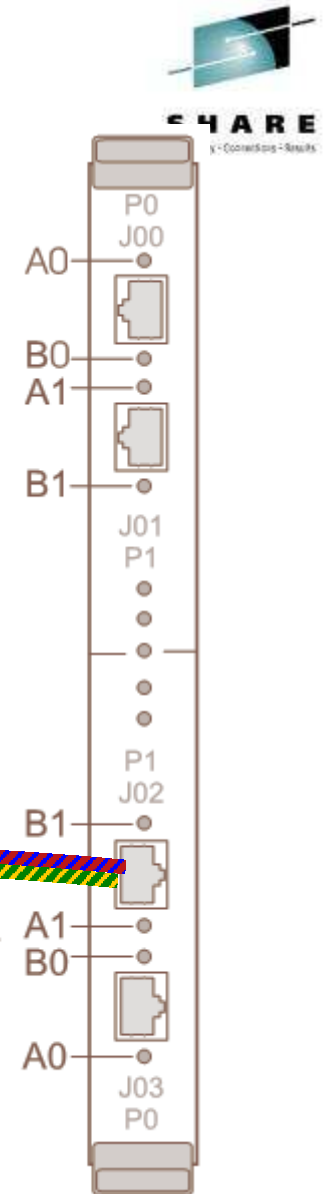
# VLAN-aware Virtual Switch



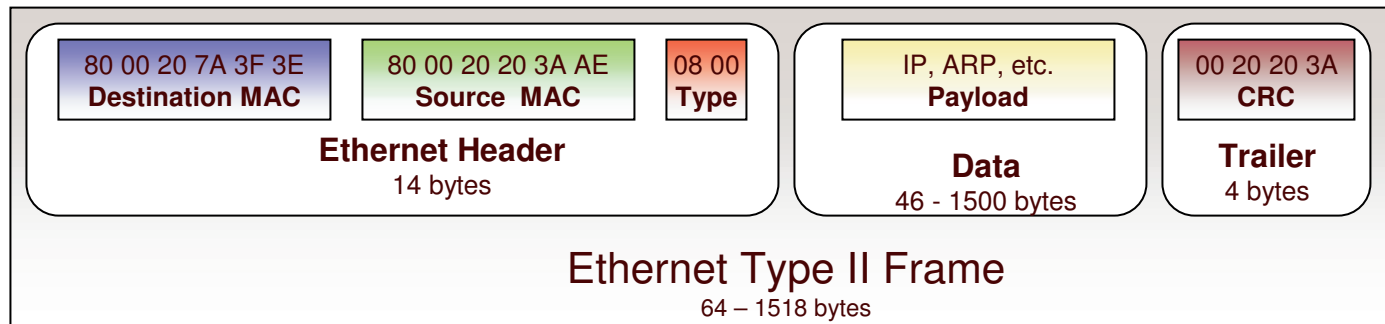
Trunk port



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



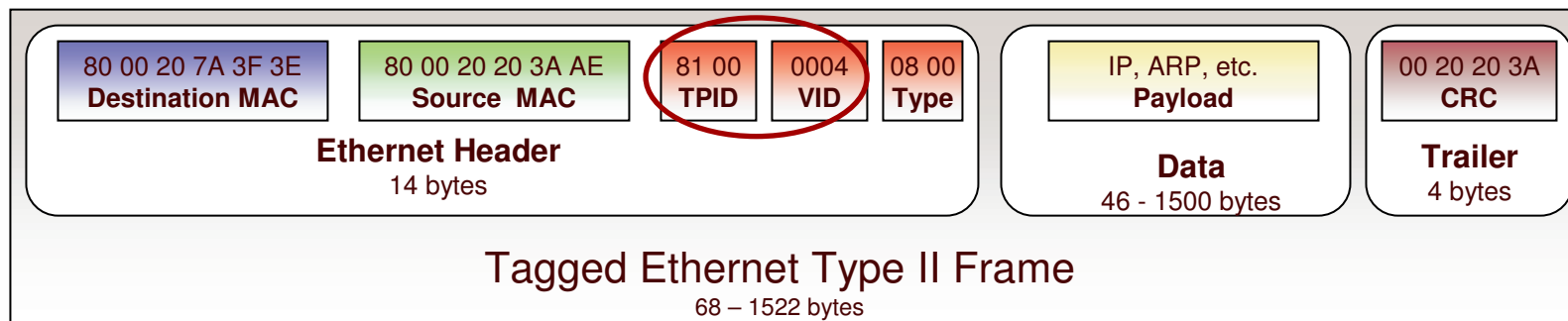
# VLAN tagging



## Access port and Trunk port

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

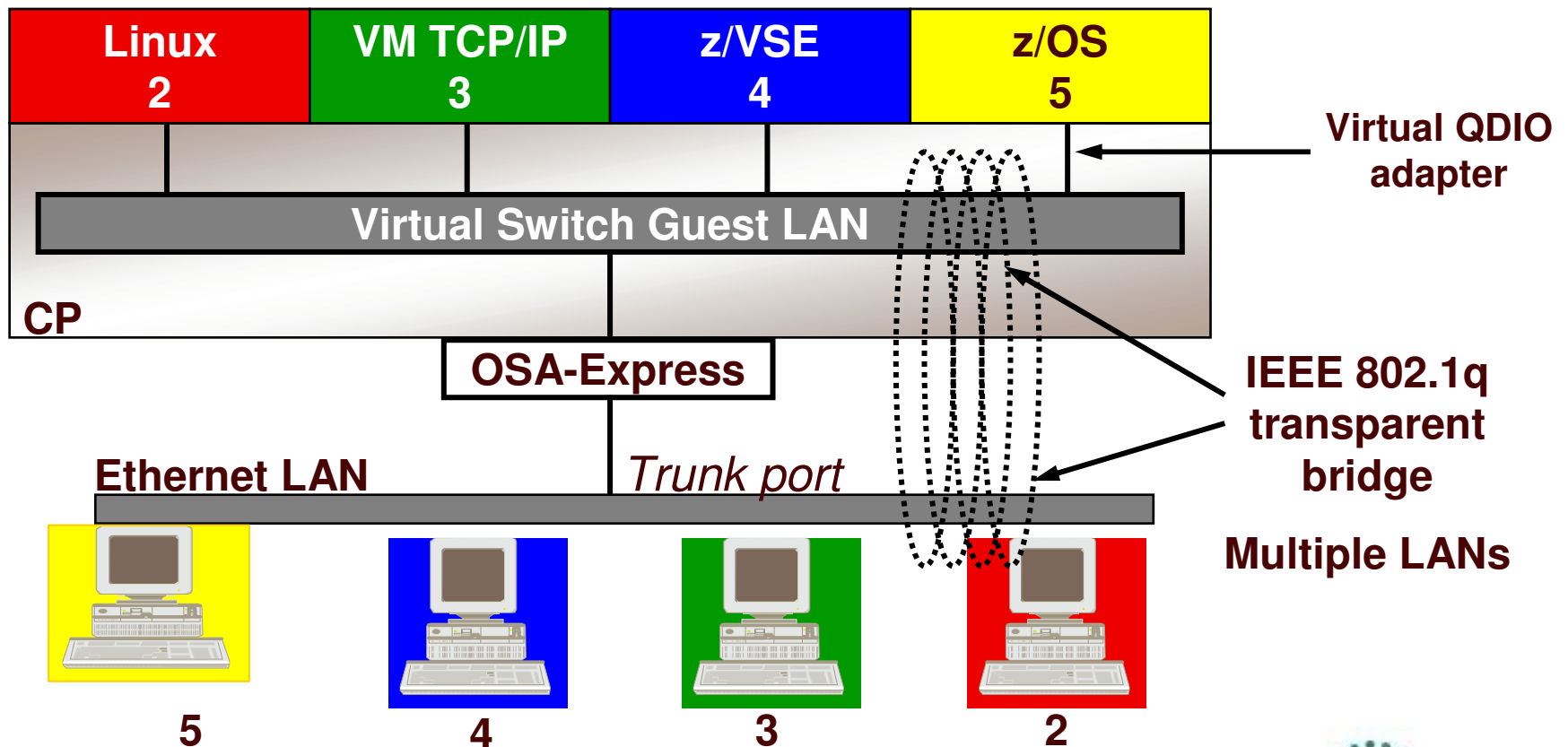
Type/length 0800 means IPv4 (IETF RFC 894)



## Trunk port only

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

# VLAN-aware Virtual Switch Sees all authorized LAN segments



# 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  

- VMLAN 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

Complete your sessions evaluation online at [SHARE.org/SFEval](http://SHARE.org/SFEval)



# 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 unique MACPREFIX
    - Enforced for Single System Image
  
- **VMLAN USERPREFIX** in SYSTEM CONFIG
  - Set MAC prefix for user-defined MAC addresses
  - Can be the same or unique (default to MACPREFIX)
    - Must be the same for Single System Image

# 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

```
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:

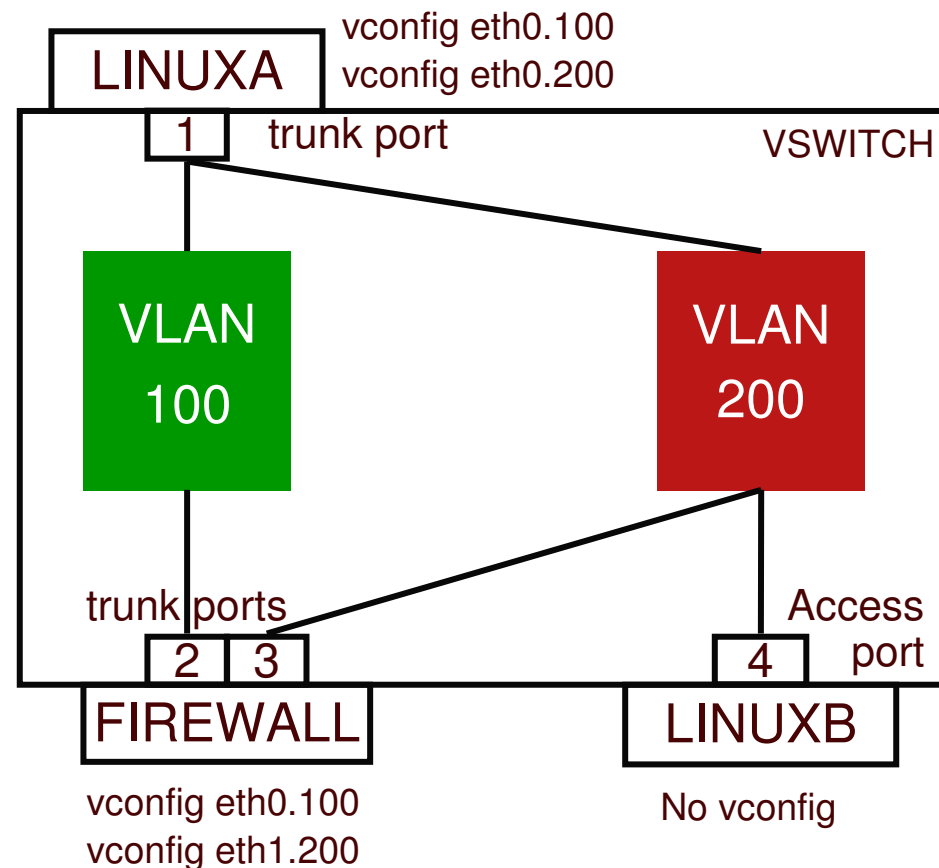
```
MODIFY VSWITCH SWITCH12 GRANT LNX01 VLAN 3
CP SET VSWITCH SWITCH12 GRANT LNX02 PORTTYPE TRUNK
VLAN 4 20-22 29 302
```

```
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



# 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
```

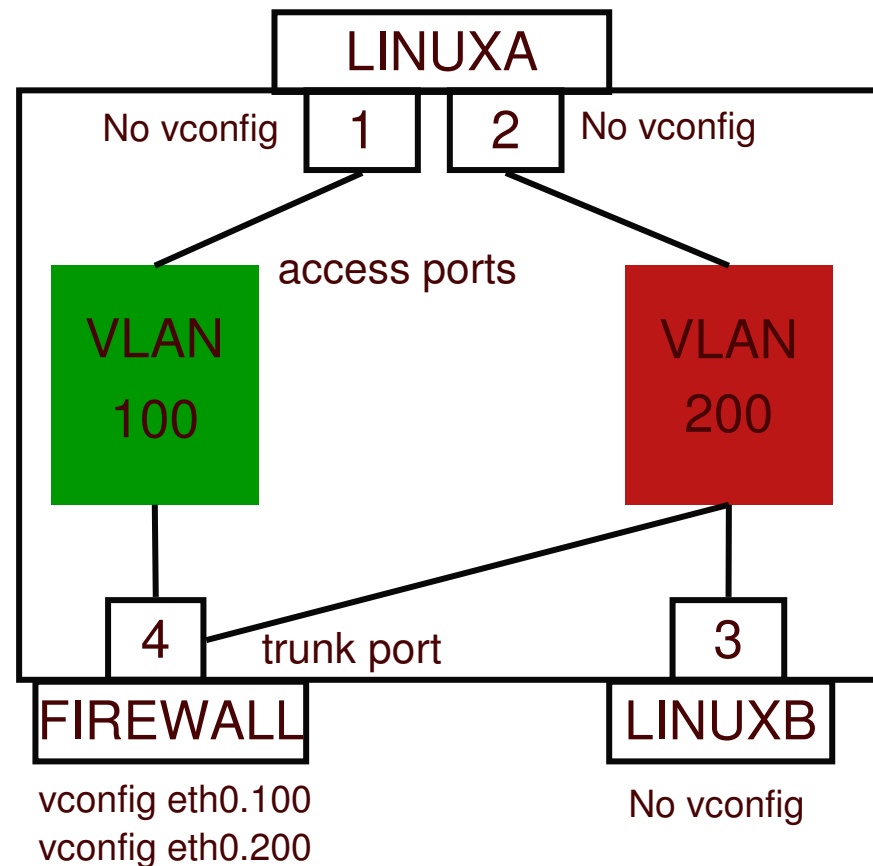
```
LINUXB:  NICDEF 4E0 TYPE QDIO LAN SYSTEM VSW1
```

```
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

6.2

- 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

6.2

```
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

## ■ Port 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

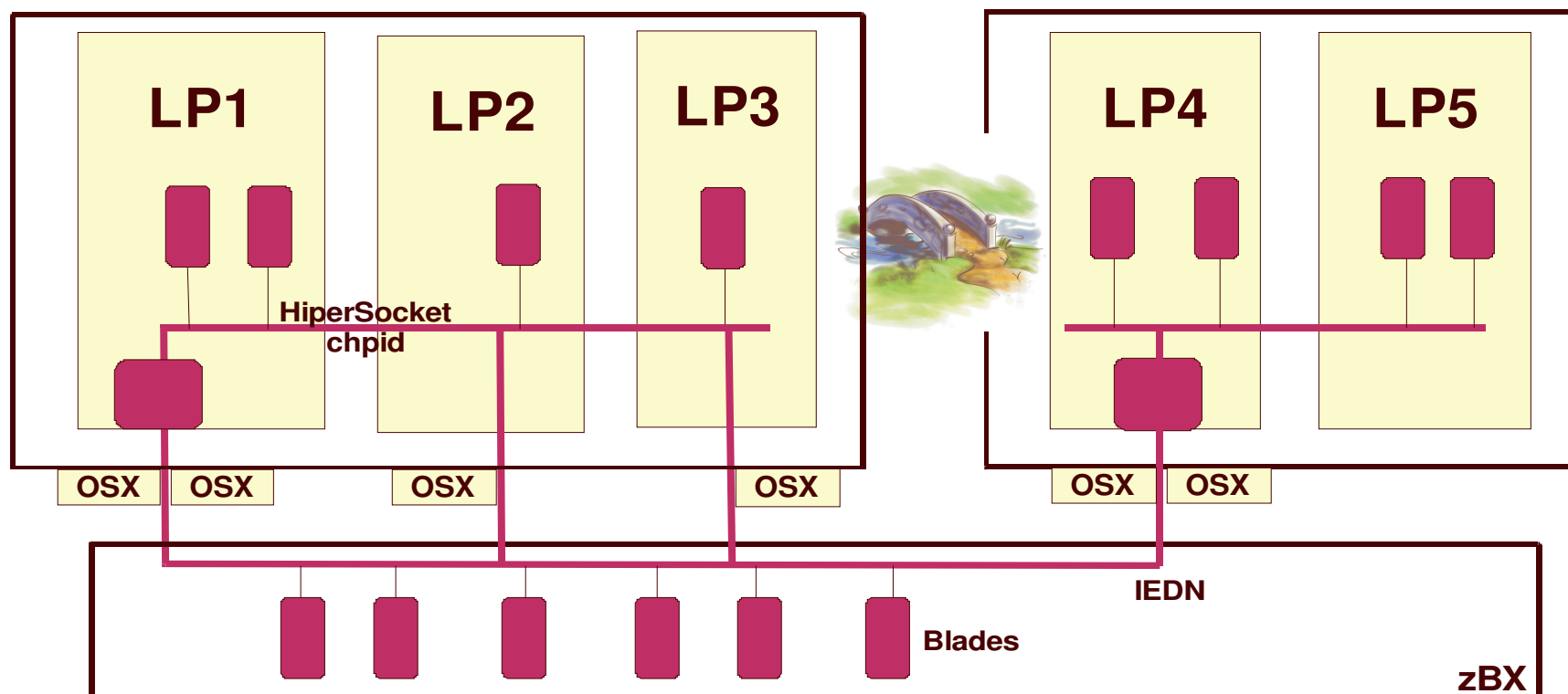
# HiperSocket Virtual Switch Bridge

6.2

- Ethernet – HiperSocket Bridge managed by VSWITCH
  - zEnterprise IEDN bridge
    - TYPE=OSX to TYPE=IQD
    - One IEDN HiperSocket bridge per CEC
    - VLAN aware
  - External ethernet bridge
    - Bridge OSTYPE=OSD to TYPE=IQD HiperSockets
    - Up to 5 bridges
  
- Full redundancy
  - Up to 5 bridges per CEC (CPC)
  - One bridge per LPAR
  - Automatic takeover
  - Optionally designate one “primary”
    - Primary will perform “takeback” when it comes up
  - Each bridge can have more than one OSA uplink

# HiperSocket Virtual Switch Bridge

6.2



- ▶ Built-in failover and failback
- ▶ Bridge to OSX or OSD chipid
- ▶ CHPARM=x4
- ▶ Same or different LPAR
- ▶ One active bridge per CEC
- ▶ PMTU simulation

Complete your sessions evaluation online at [SHARE.org/SFEval](http://SHARE.org/SFEval)



# HiperSocket Virtual Switch Bridge

6.2

```
DEFINE VSWITCH switch
```

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

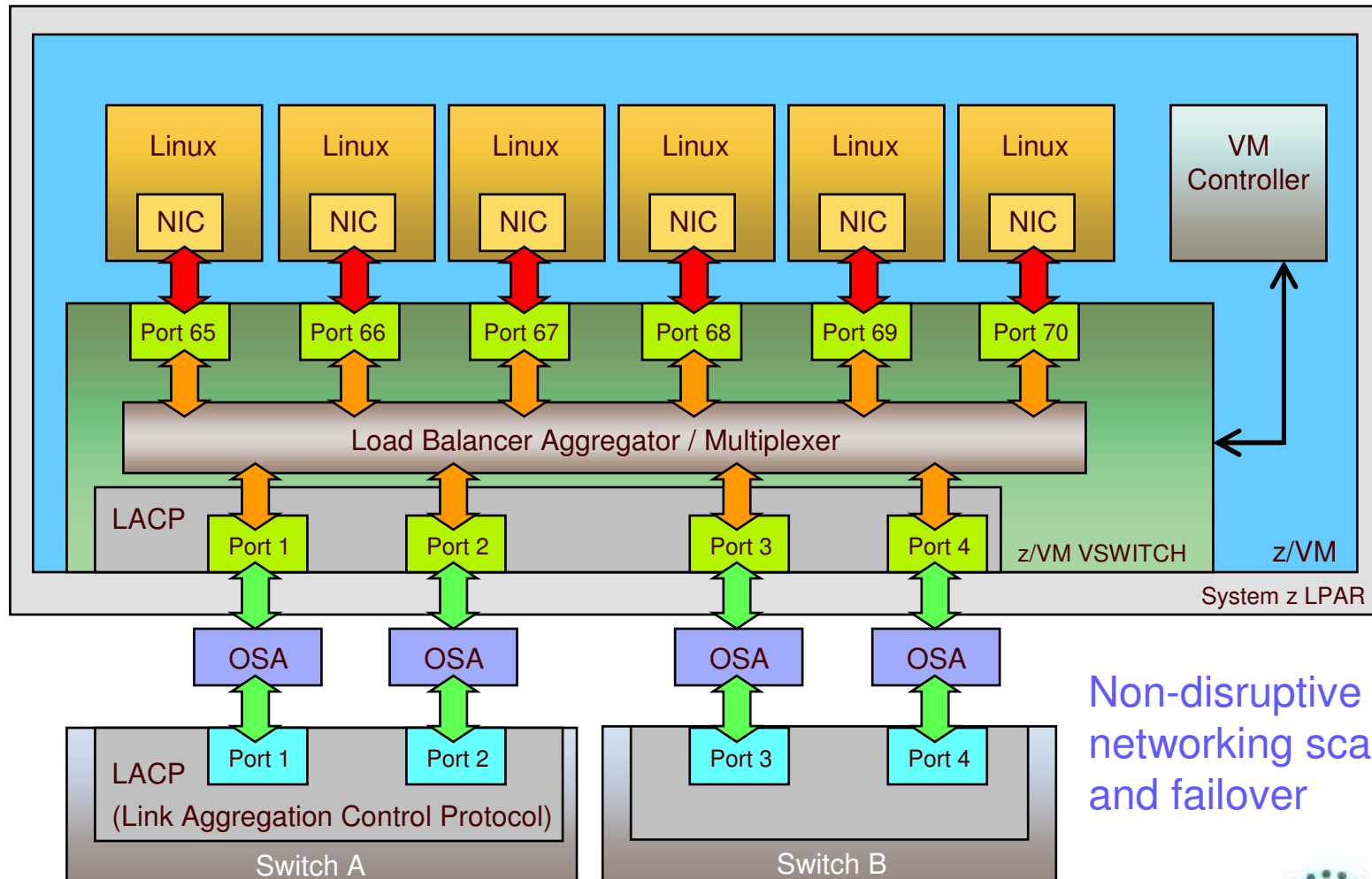
```
ETHERNET
```

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

- The HiperSocket device must be on a CHPID defined in the IOCP with **CHPARM=x4**
  - **DEFINE CHPID .... EXTERNAL\_BRIDGED** and **IEDN\_BRIDGED** is available for dynamic I/O



# 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 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 Layer 2 VSWITCHes
  
- With “virtual chassis” support from switch vendor, can even handle physical switch outage

# IEEE 802.3ad Link Aggregation

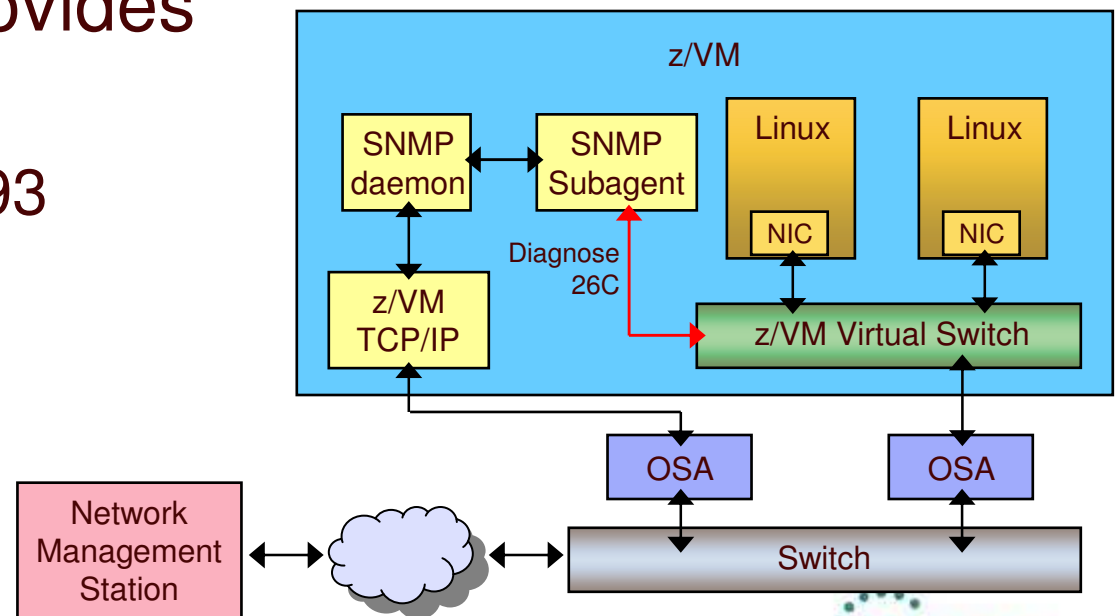


- Define an OSA port group
  - SET PORT GROUP name JOIN E100 E200.P1
  
- DEFINE VSWITCH ... ETHERNET GROUP name
  
- OSA ports cannot be shared with other VSWITCHes or LPARs

# 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

# 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]
          [MACID xyyyzz]
```

Combined with VMLAN  
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 Controllers



- Virtual machines that handle OSA housekeeping duties
  - Specialized VM TCP/IP stacks that start, stop, monitor, query
  - Not involved in data transfer
  
- IBM provides DTCSVSW1 and DTCSVSW2
  - No need to create more
  - Leave them both logged on for redundancy
  - Automatic failover
  
- Do not ATTACH or DEDICATE devices
  - Handled by CP

# Summary



- VSWITCHes make it easy to control access to the network and simplify server cloning
- Support for IEEE VLANs
- Support for Link Aggregation
- Support for SNMP-based monitoring (Switch MIB)
- Port-based or User-based

Complete your sessions evaluation online at [SHARE.org/SFEval](http://SHARE.org/SFEval)



# Built-in 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

# Support Summary



z/VM 6.2	<ul style="list-style-type: none"> <li>▪ Port-based configuration</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 5.4	<ul style="list-style-type: none"> <li>▪ Port isolation</li> <li>▪ Native VLAN id defaults to 1</li> <li>▪ z/VM TCP/IP support for Layer 2</li> </ul>
z/VM V5.3	<ul style="list-style-type: none"> <li>▪ Link aggregation</li> <li>▪ Separation of default VLAN id from native VLAN id</li> <li>▪ SNMP monitor</li> </ul>
z/VM V5.2	<ul style="list-style-type: none"> <li>▪ Virtual SPAN ports for sniffers</li> </ul>
z/VM V5.1	<ul style="list-style-type: none"> <li>▪ Virtual trunk and access port controls</li> <li>▪ Removal of VLAN ANY</li> <li>▪ Layer 2 (MAC) frame transport</li> <li>▪ Improved virtual switch error detection &amp; recovery</li> <li>▪ External security manager access control</li> </ul>
z/VM V4	<ul style="list-style-type: none"> <li>▪ IPv4 Virtual Switch with IEEE VLANs</li> <li>▪ IPv4 HiperSocket Guest LAN</li> <li>▪ IPv4 and IPv6 QDIO Guest LAN</li> </ul>

Complete your sessions evaluation online at [SHARE.org/SFEval](http://SHARE.org/SFEval)



# 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 STG Lab Services*

*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*



Session 12479

- 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>