#SHAREorg

SHARE
Technology • Connections • Results

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


## Alan Altmark

### IBM STG Lab Services
### Senior Managing z/VM and Linux Consultant


### Session 10312

#SHAREorg

SHARE
Technology • Connections • Results

# 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:

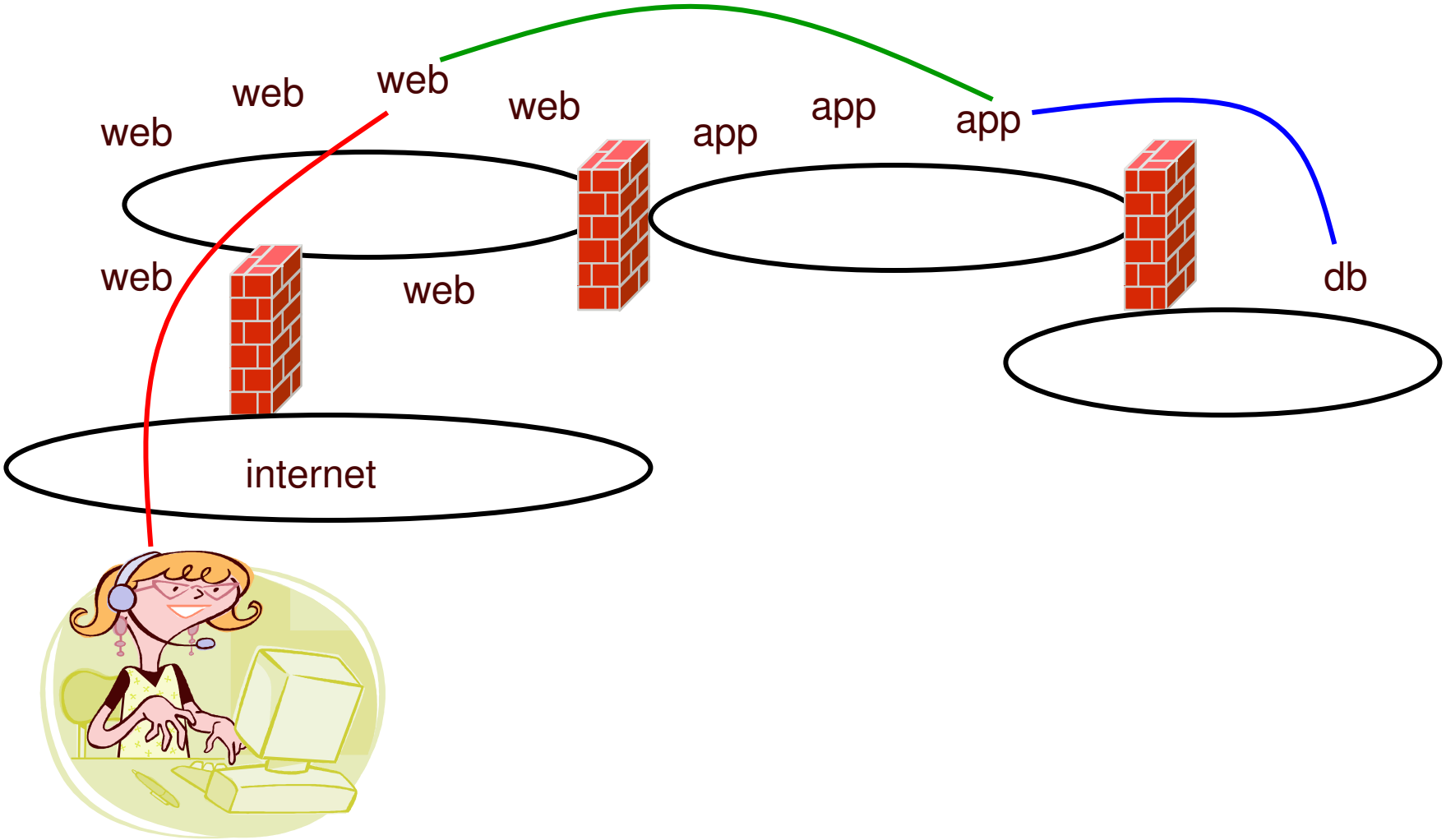
IBM IBM logo DB2 z/OS z/VM

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

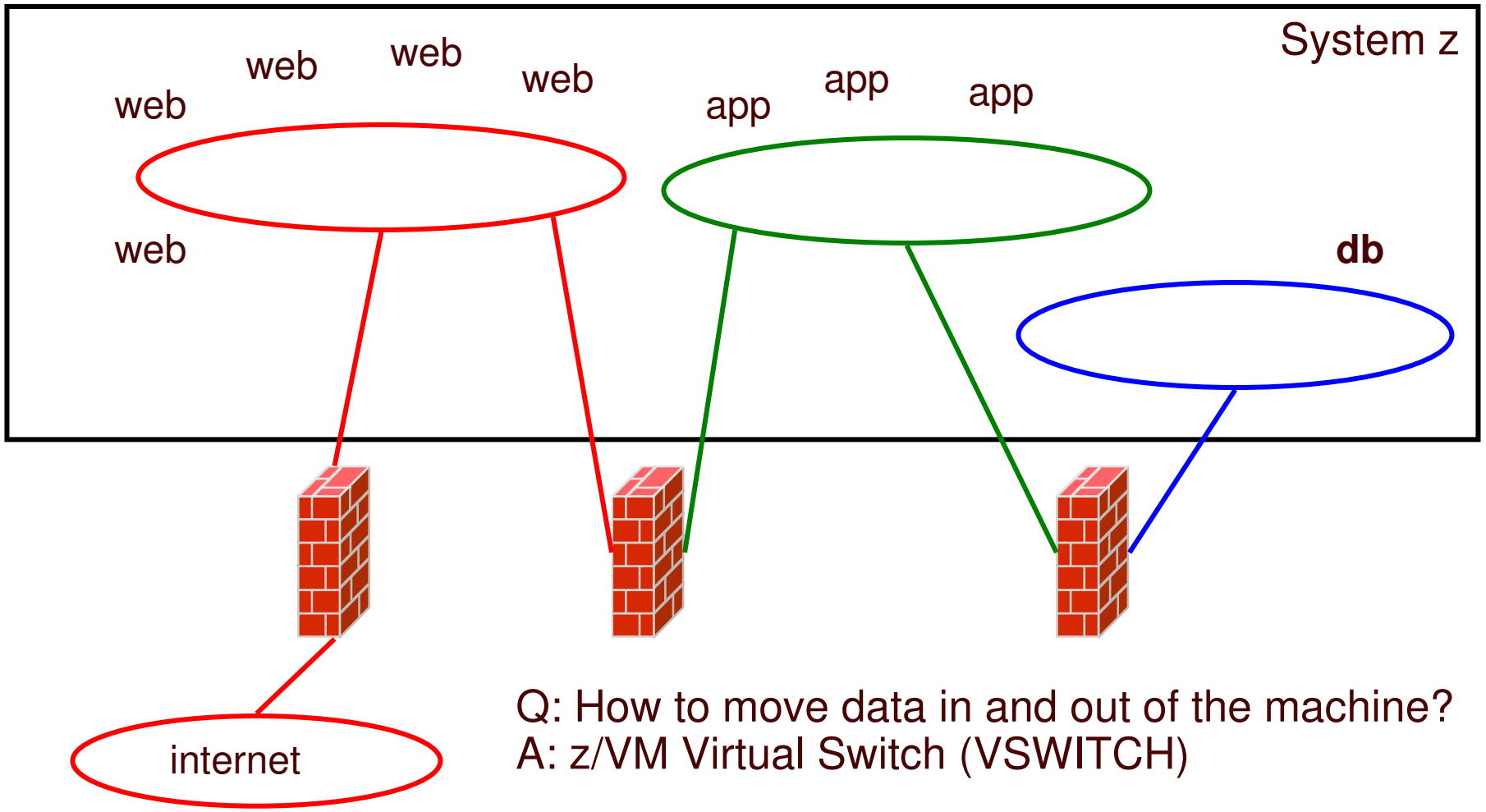
# Topics

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

# Multi-Zone Network

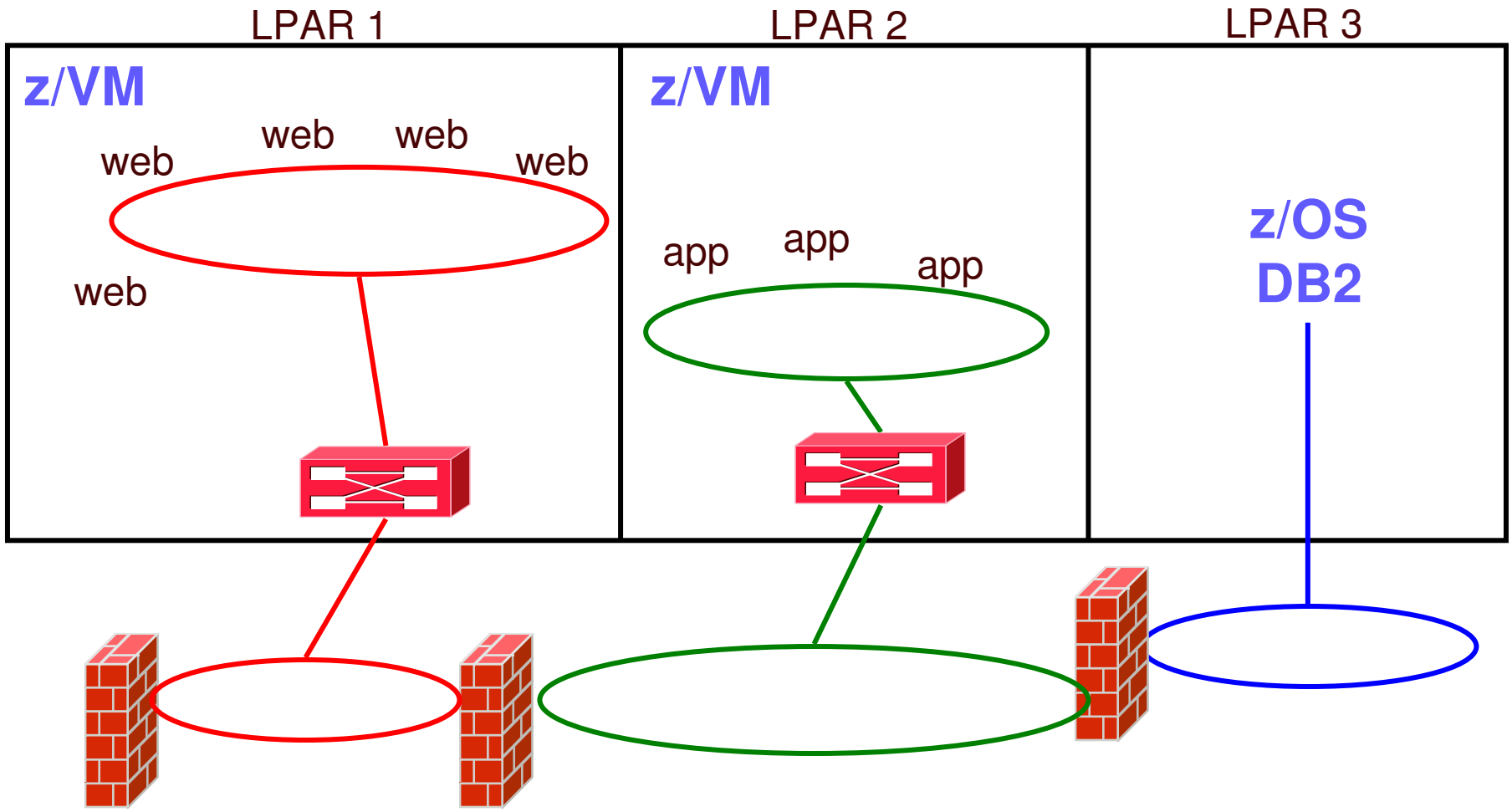


# Multi-zone Network on System z with outboard firewall

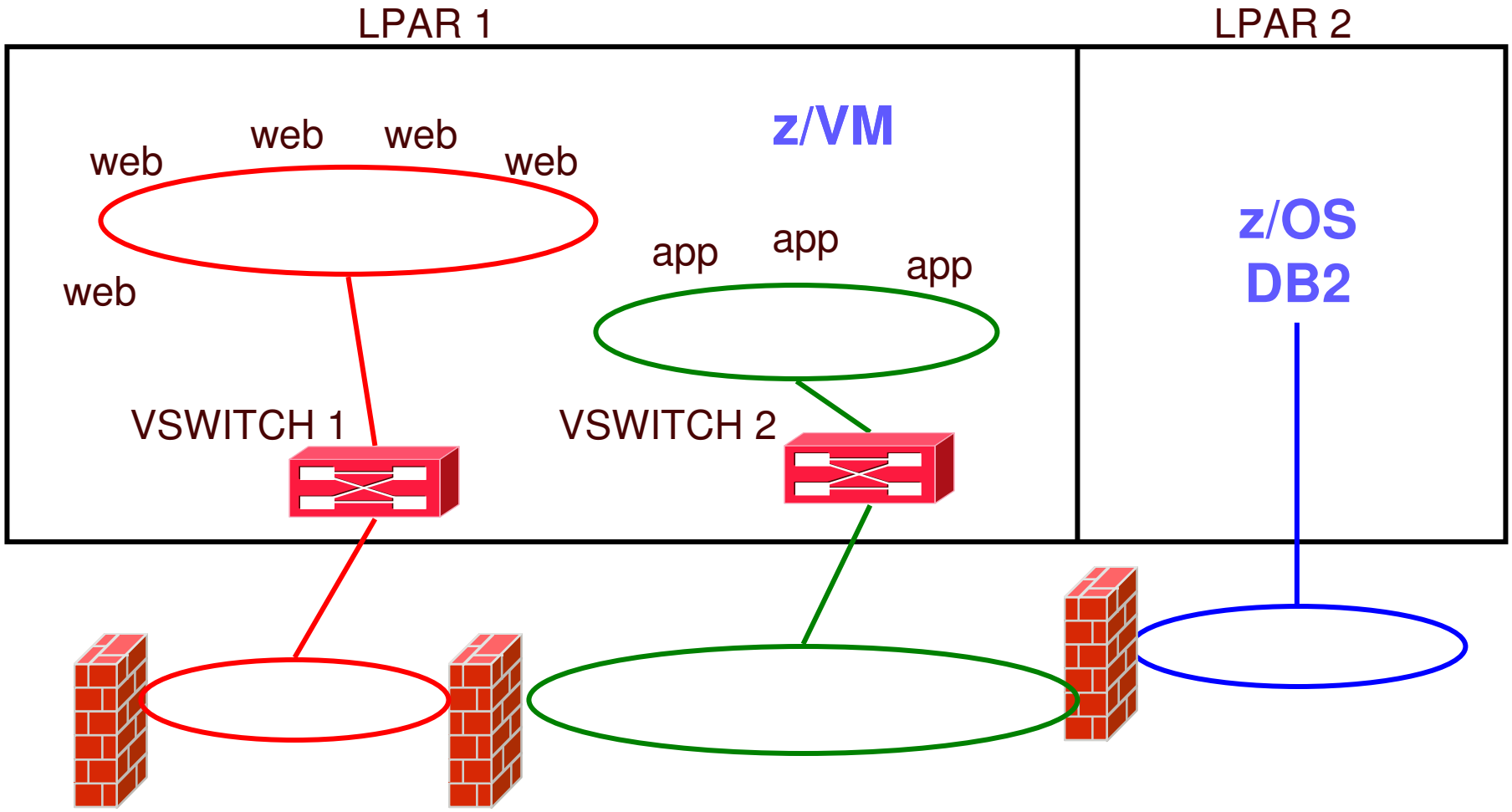



Q: How to move data in and out of the machine?
A: z/VM Virtual Switch (VSWITCH)

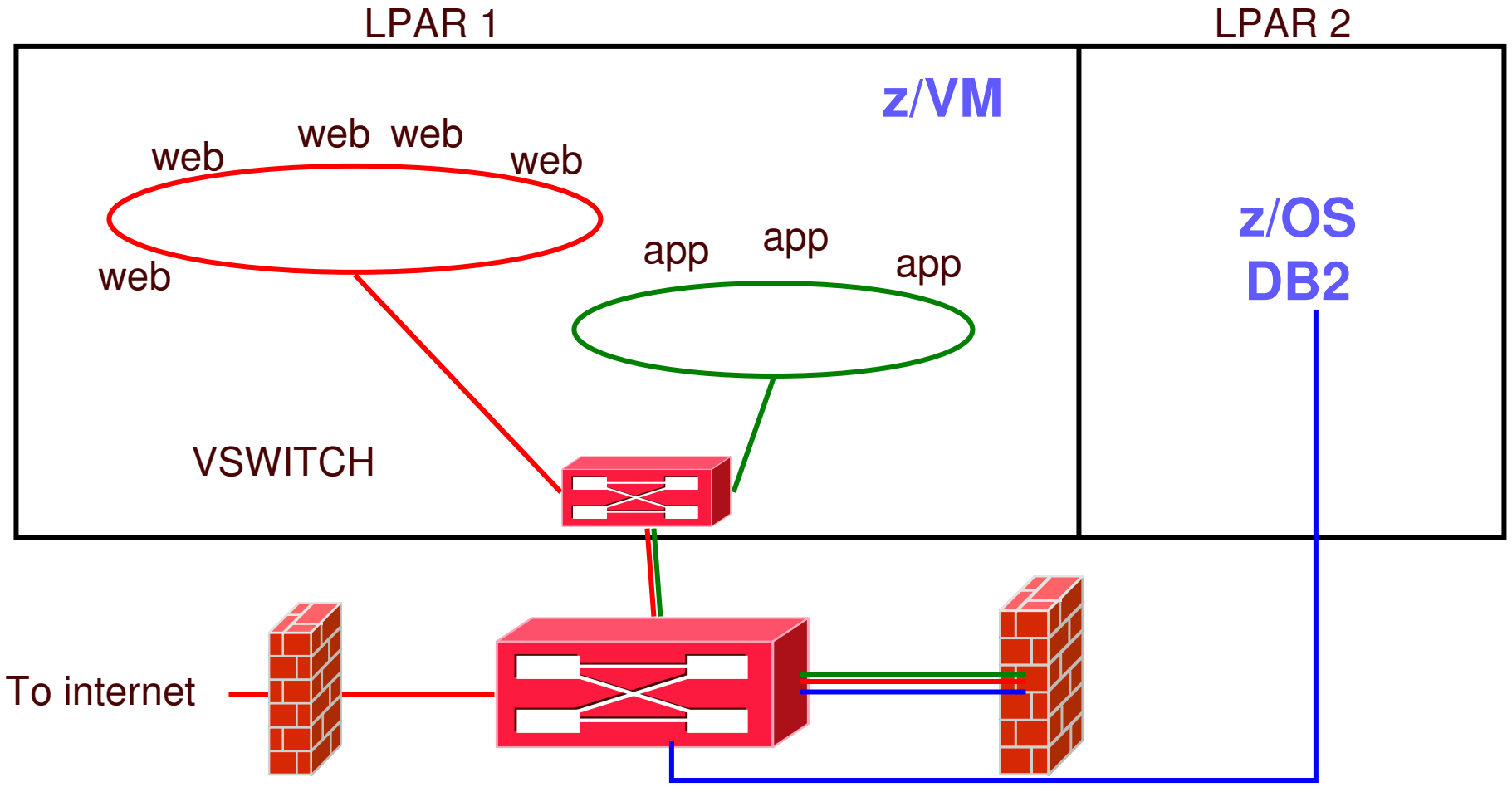
# Option A: Two z/VM LPARs, one VSWITCH each



# Option B: One LPAR, two VLAN unaware VSWITCHes

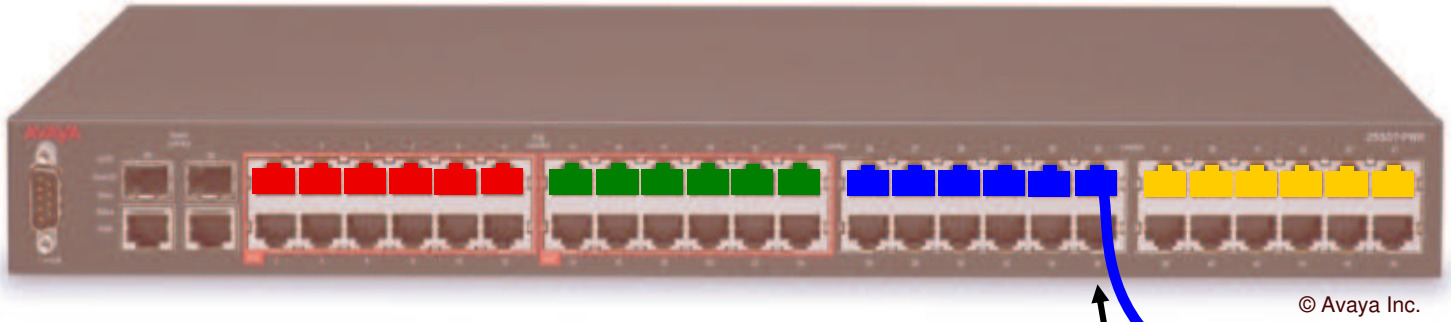


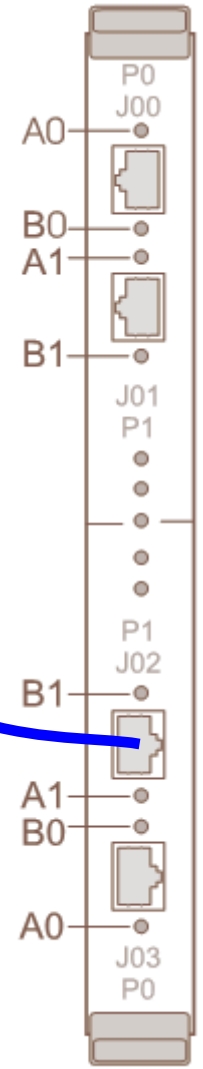
# Option C: One LPAR, one VLAN aware VSWITCH




With 1 VSWITCH, 3 VLANs, and a multi-domain firewall

# What's a 'switch' anyway?

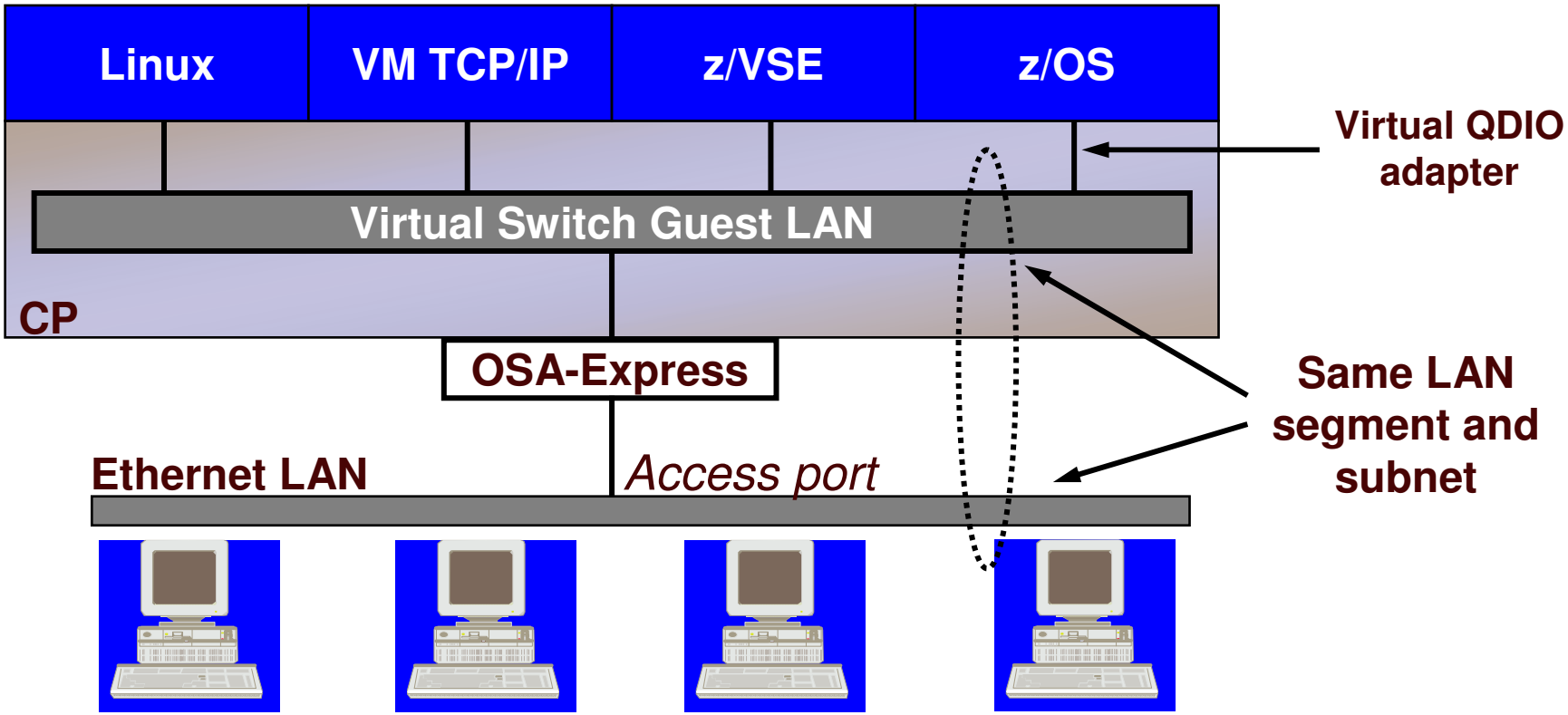


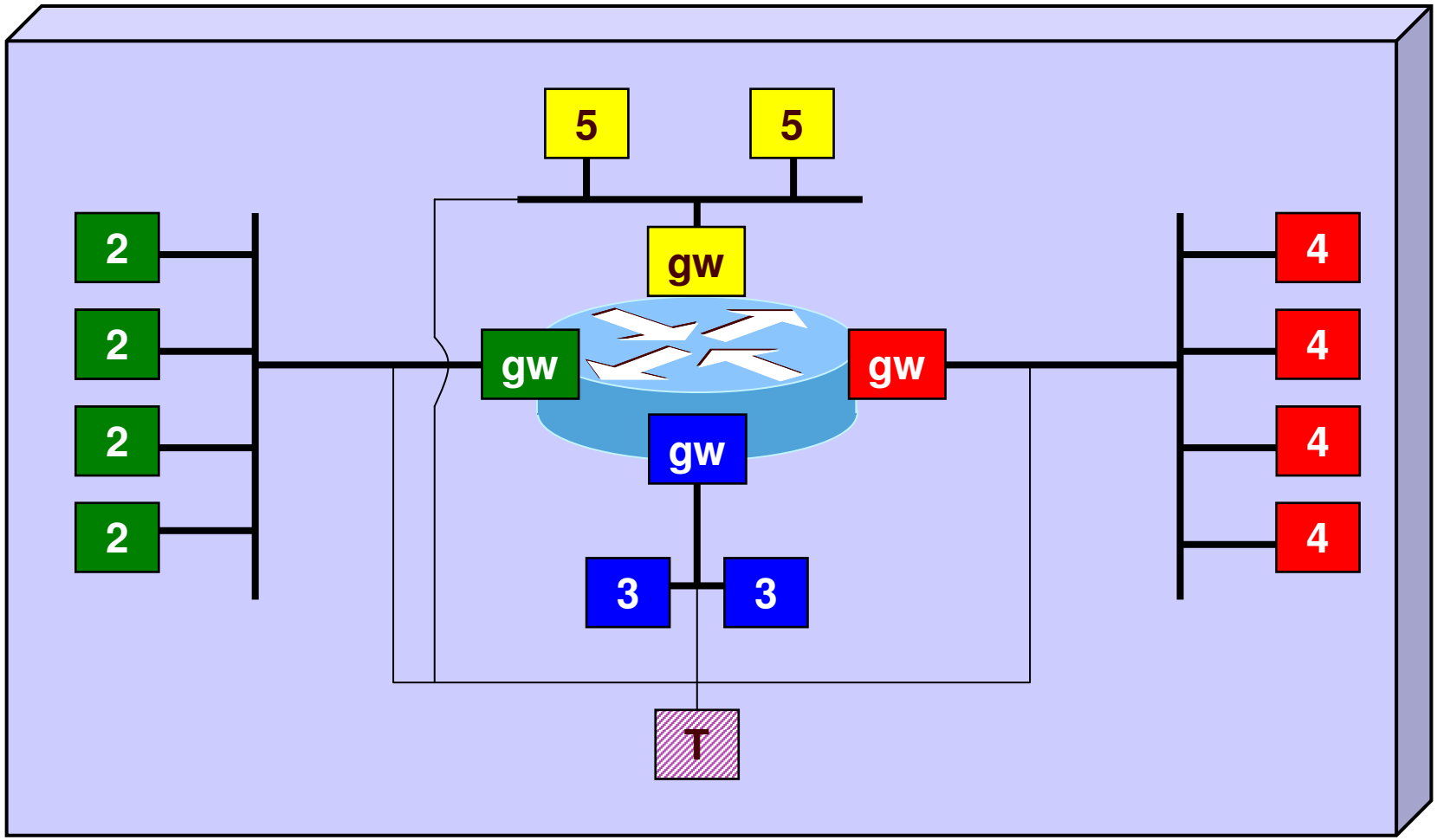

## It creates LANs and routes traffic

- 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

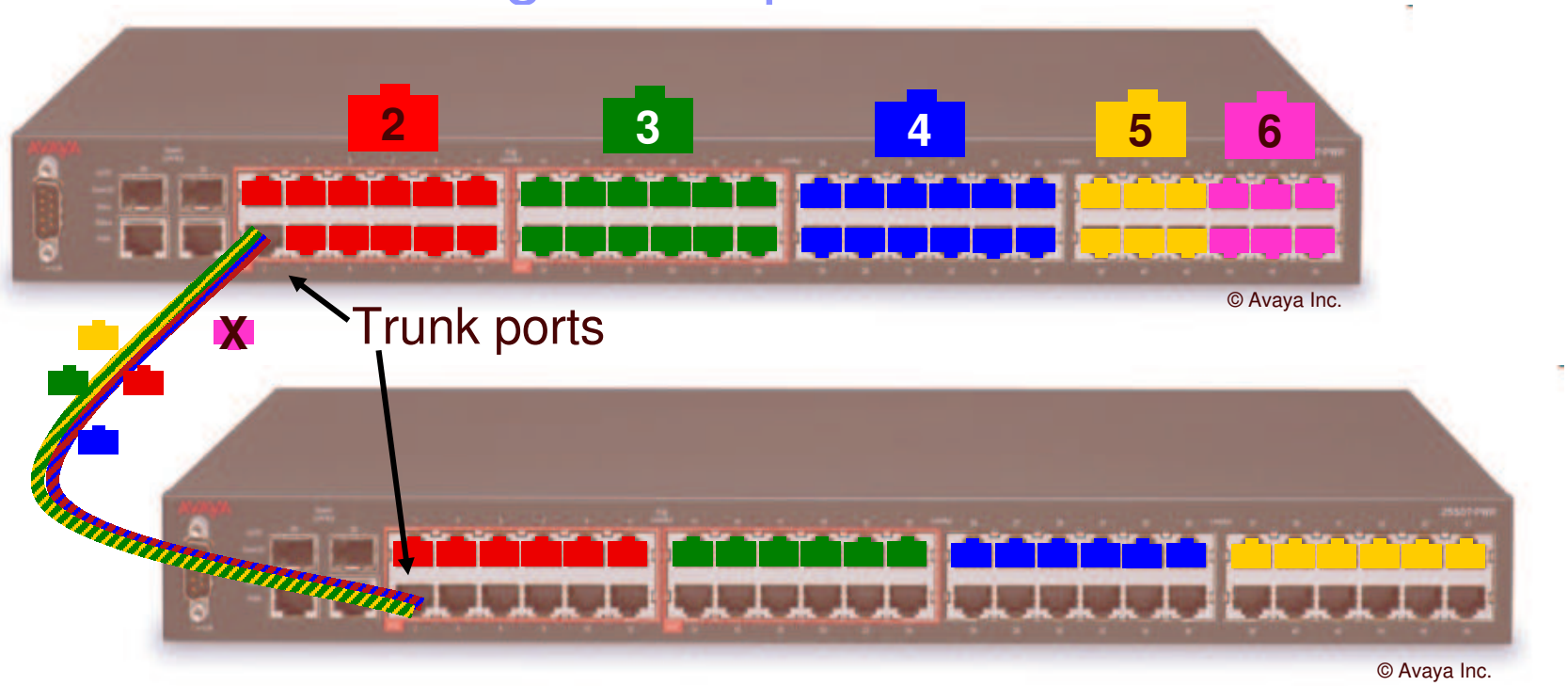
# z/VM Virtual Switch – VLAN unaware
# Sees only a single LAN segment



# Internal routing function

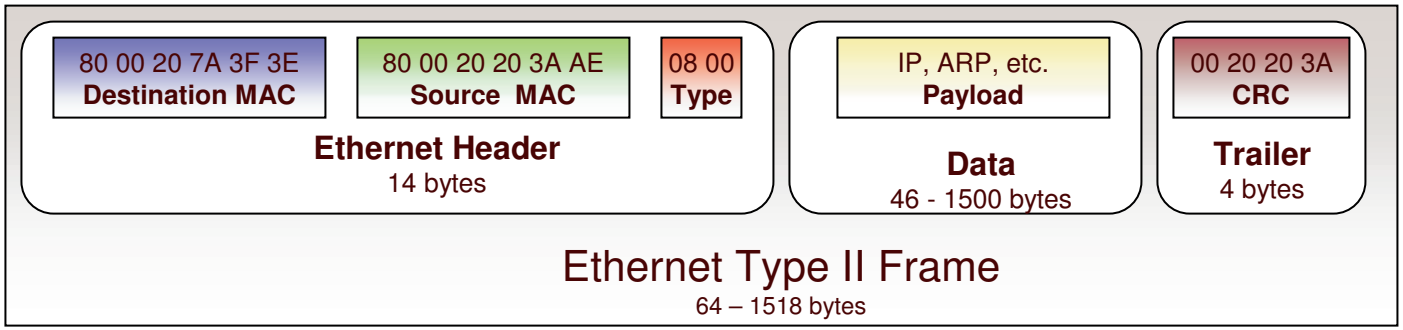


# IEEE VLANs using Trunk port



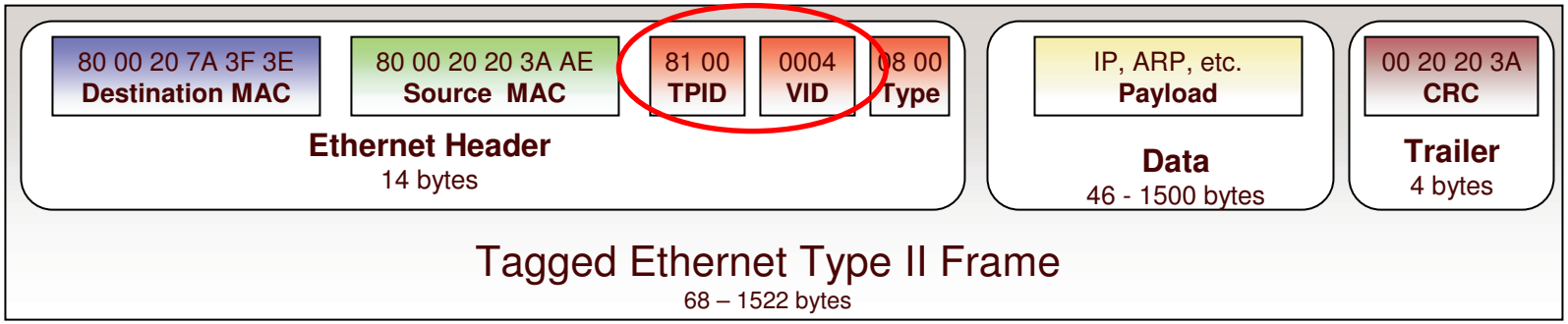

- If you run out of ports, you don't throw it away, you "trunk" it to another switch to "bridge" LAN segments together
- IEEE standards provide a way for trunk ports to exchange data for multiple authorized LAN segments using a single cable.

# VLAN “tags”



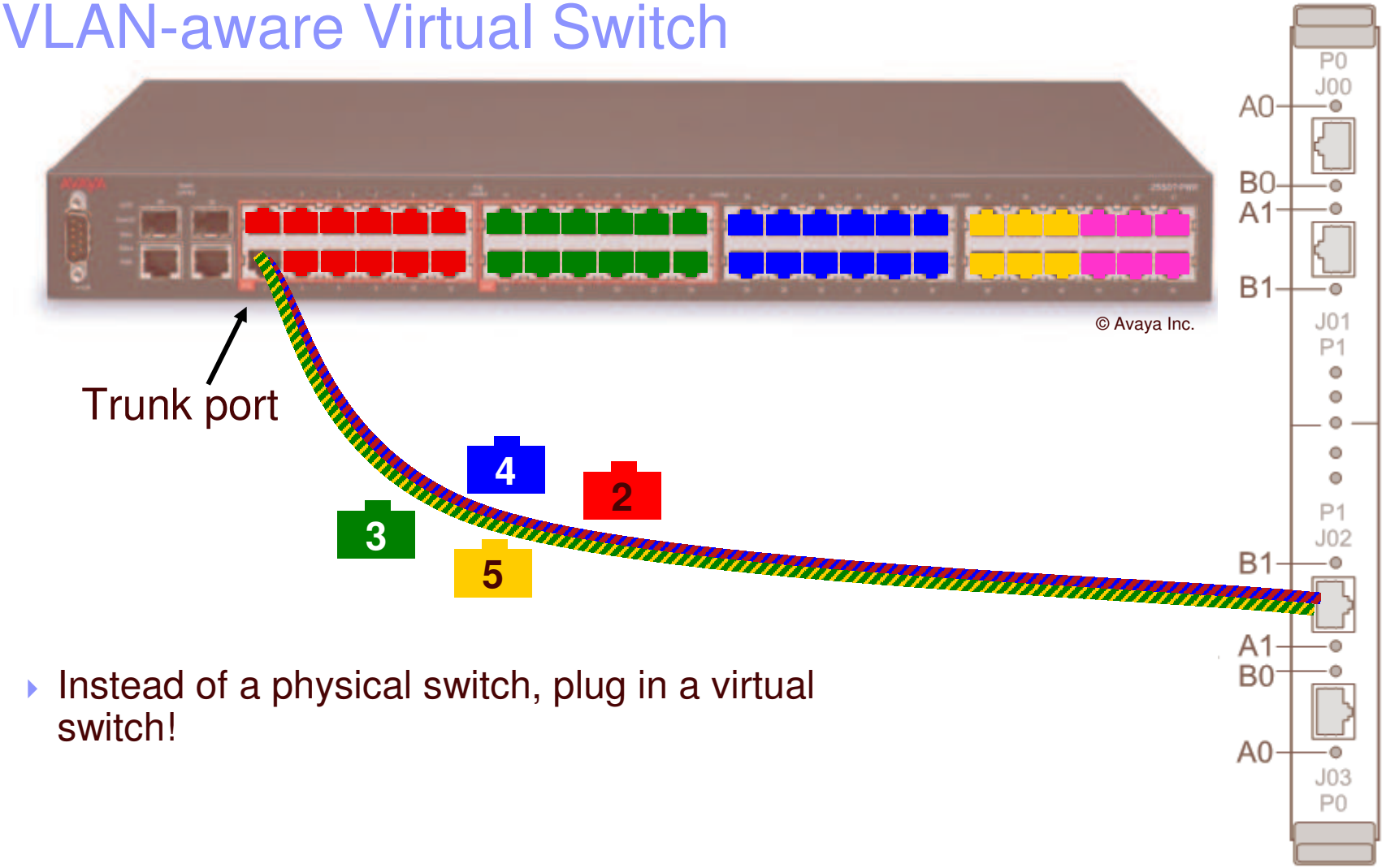

Access port and Trunk port

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



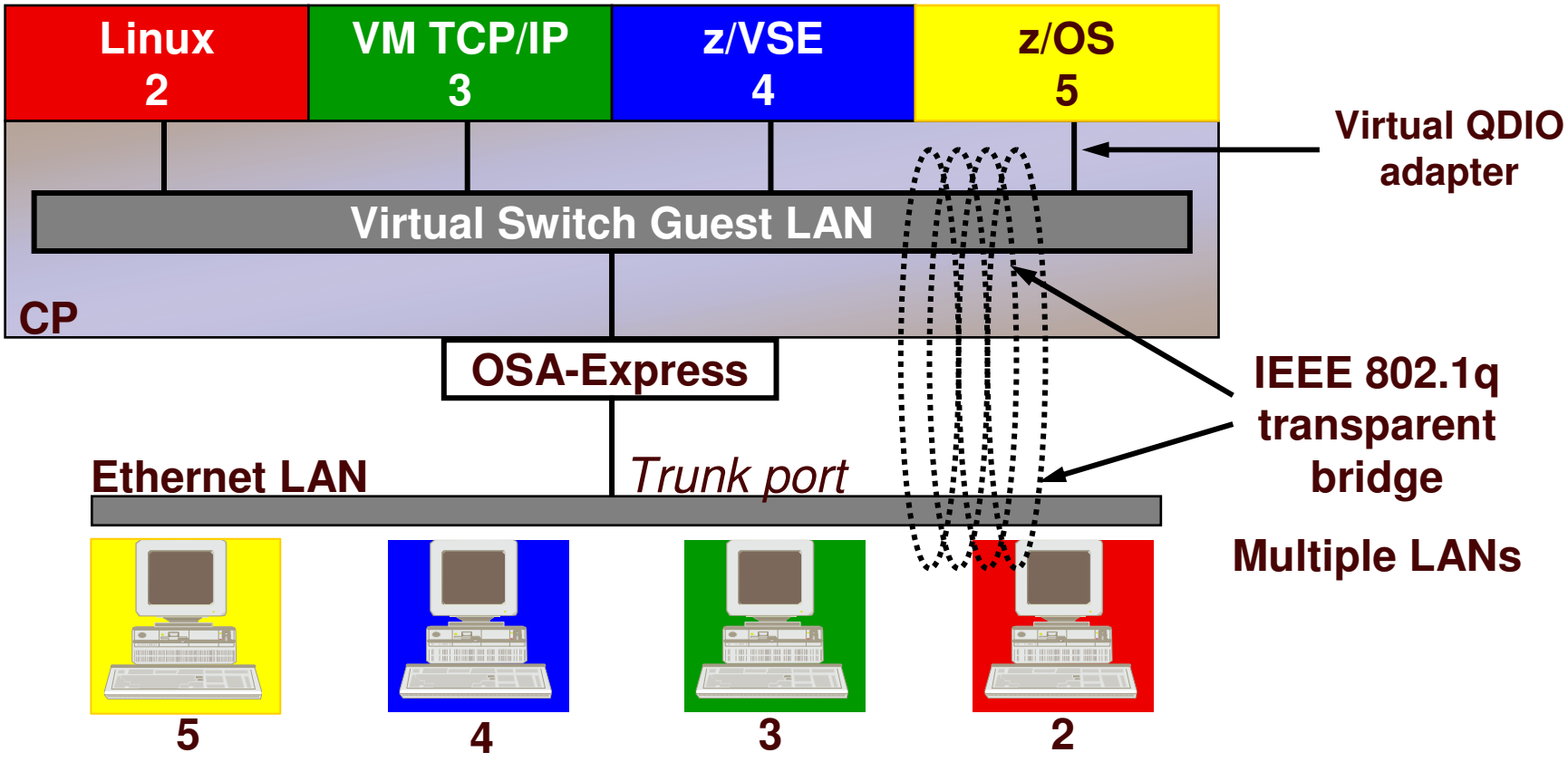

Trunk port only

# VLAN-aware Virtual Switch




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

# 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

# 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 DTCVSW1 and DTCVSW2
  - No need to create more
  - Leave them both logged on for redundancy
  - Automatic failover
- Do not ATTACH or DEDICATE devices
  - Handled by CP

# zTerminology: Layer 2 and Layer 3 OSA

- Layer 2 - All protocols: SNA, IP, NETBIOS, ….
  - Host registers virtual MAC addresses with OSA
    - Burned-in MAC address not used
  - Host sends ethernet frame with registered MAC address
  - Host handles ARP
- Layer 3 - IP only
  - Host registers IP addresses with OSA
  - Host sends IP packet
  - OSA places packet in ethernet frame using burned-in MAC address
  - OSA handles ARP

# Network Terminology: Layer 2 and Layer 3 Switches

- Layer 2 – Physical connectivity
  - Protocol agnostic
  - Knows which MACs are associated with which ports
    - Filters based on unicast v. multicast v. broadcast
- Layer 3 – Network connectivity
  - A layer 2 switch
  - PLUS understands local network topology
  - PLUS provides interconnect function among attached networks
    - “gateway”

# Create a Layer 2 Virtual Switch

- SYSTEM CONFIG or CP command:

```
DEFINE VSWITCH name ETHERNET

              [RDEV NONE | cuu [cuu [cuu]] ]
              [GROUP group_name]

              [USERBASED | PORTBASED]                    6.2
              [MACPROTECT UNSPECIFIED | ON | OFF]        6.1

              [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]
```

# Setting Guest LAN and VSWITCH 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
```

6.1

- `VMLAN LIMIT TRANSIENT 0` prevents dynamic definition of Guest LANs by class G users

# Virtual MAC Addresses 02:00:01:00:01:23

- Virtual MAC address
  - MAC prefix = high-order 3 bytes of MAC address
  - MAC ID = low-order 3 bytes of MAC address
  - Concatenate to create virtual MAC address
- Each instance of CP should have a unique MACPREFIX
  - VMLAN MACPREFIX for CP-assigned MAC IDs
    - Reserve 020000 (the default) to recognize a misconfigured system
  - VMLAN USERPREFIX for user-assigned MAC IDs

# 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

```
VMLAN MACIDRANGE SYSTEM  000001-002FFF
                 USER    002000-002FFF
```

- MACPROTECT ON prevents guests from changing their assigned MAC address

# Create a Layer 3 Virtual Switch

- SYSTEM CONFIG or CP command:

```
DEFINE VSWITCH name  IP
              [RDEV NONE | cuu [cuu [cuu]] ]
              [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]
```

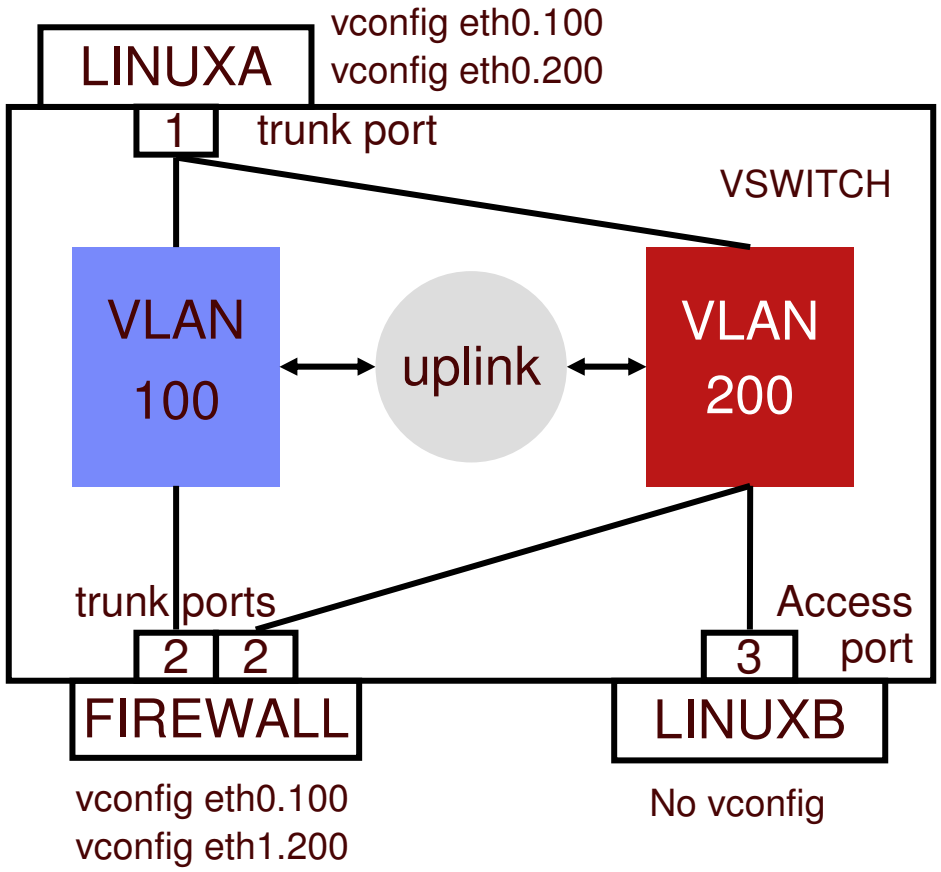
# User-based Virtual Switch access list

- 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
  - CP-assigned port number
  - Applies to user, not NIC
- 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

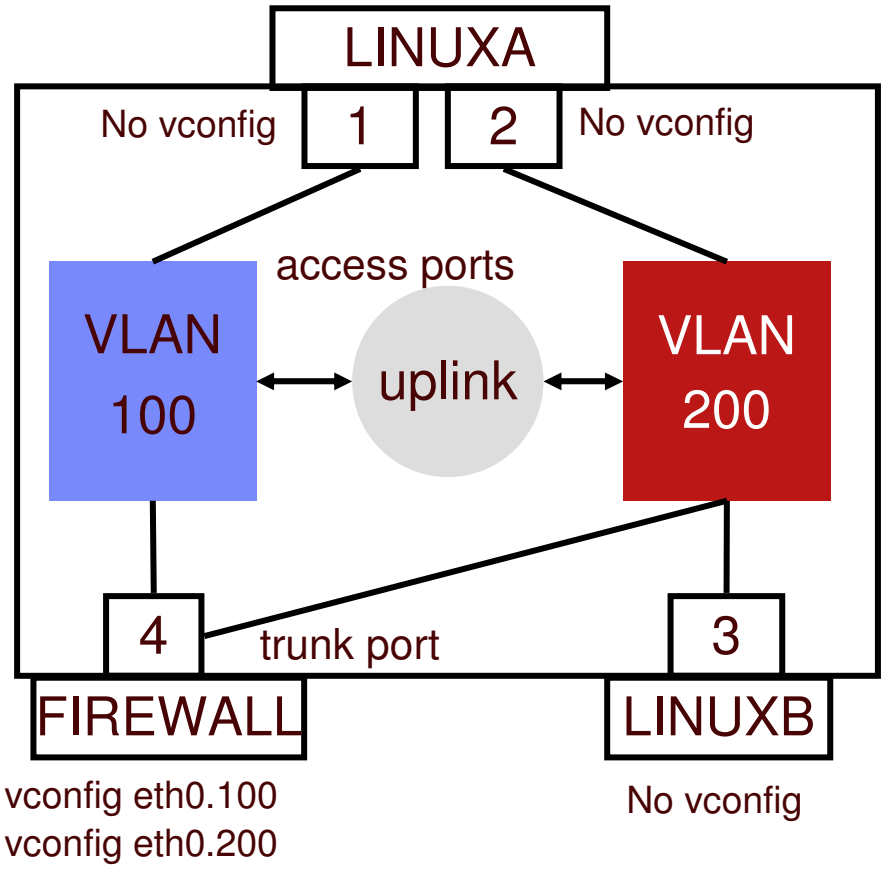


# User-based VSWITCH access list

```
define vswitch vswitch1 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
  - 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 vswitch1 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

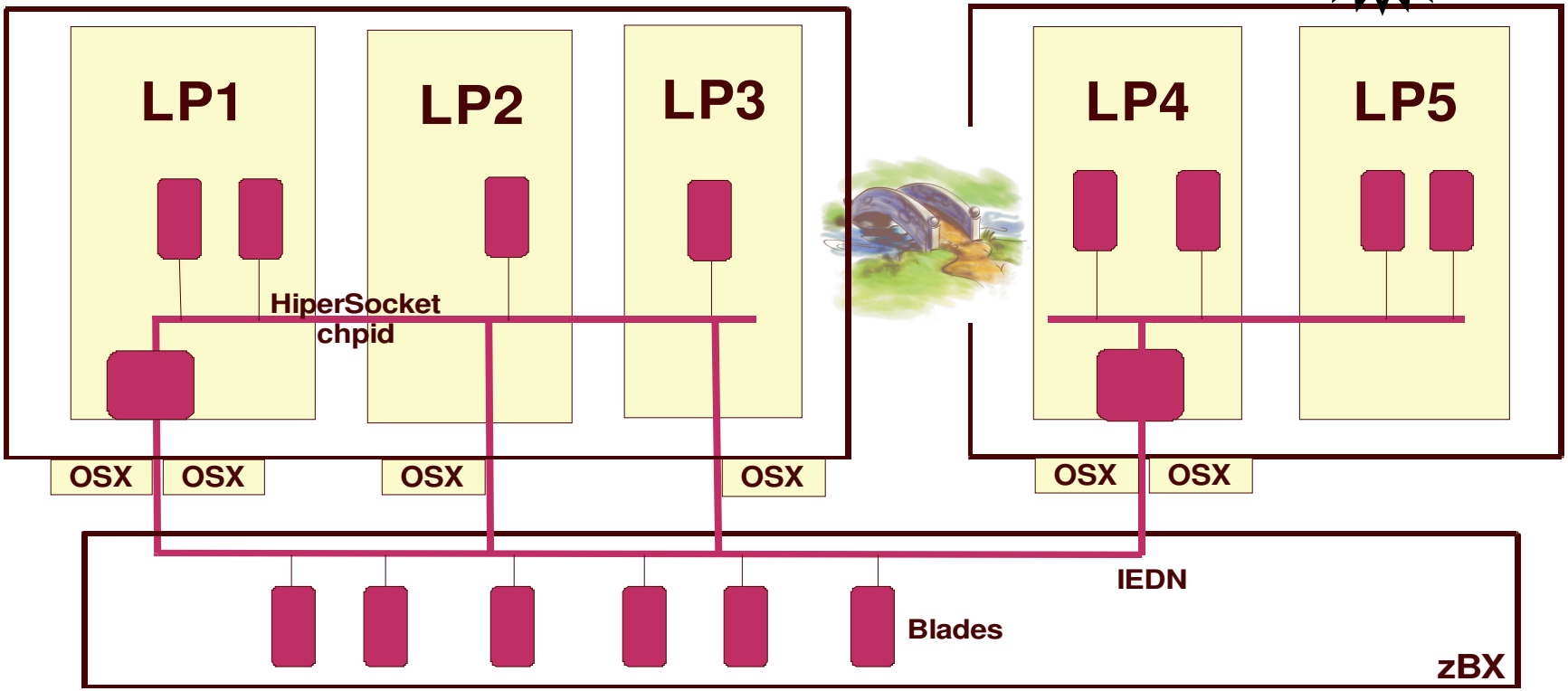
- Virtual Sniffers
  - Guest must be authorized via SET VSWITCH or security server
  - Guest enables promiscuous mode using CP SET NIC or via device driver controls
    - E.g. tcpdump -P
  - Guest receives copies of all frames sent or received for authorized VLANs
- Port Isolation
  - Stop guests from talking to each other, even when in same VLAN
  - Shut off OSA “short circuit” to other users of the same OSA port

# HiperSocket Virtual Switch Bridge


- Ethernet – HiperSocket Bridge managed by VSWITCH
  - zEnterprise IEDN bridge
    - TYPE=OSX to TYPE=IQDX (new) HiperSockets
    - One IEDN HiperSocket (IQDX) chpid per CEC
    - VLAN aware
  - External ethernet bridge
    - TYPE=OSD to TYPE=IQD HiperSockets
- Full redundancy
  - Up to 5 bridges per CEC
  - 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

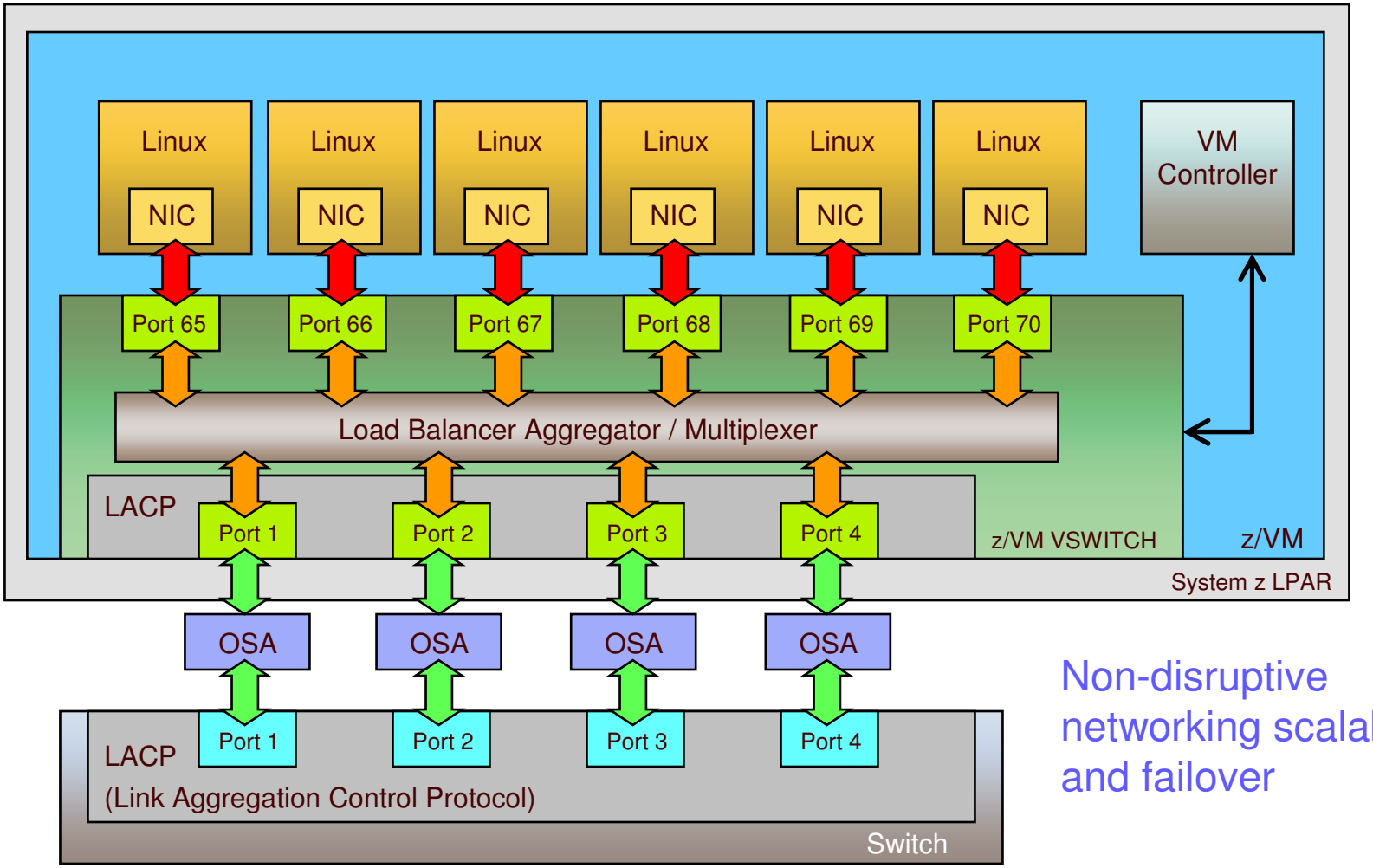
6.2




- Built-in failover and failback
- Bridge new IQDX chpid to OSX chpid
- Also works for IQD to OSD

- Same or different LPAR
- One active bridge per CEC
- PMTU simulation

# IEEE 802.3ad Link Aggregation




Non-disruptive networking scalability and failover

# IEEE 802.3ad Link Aggregation

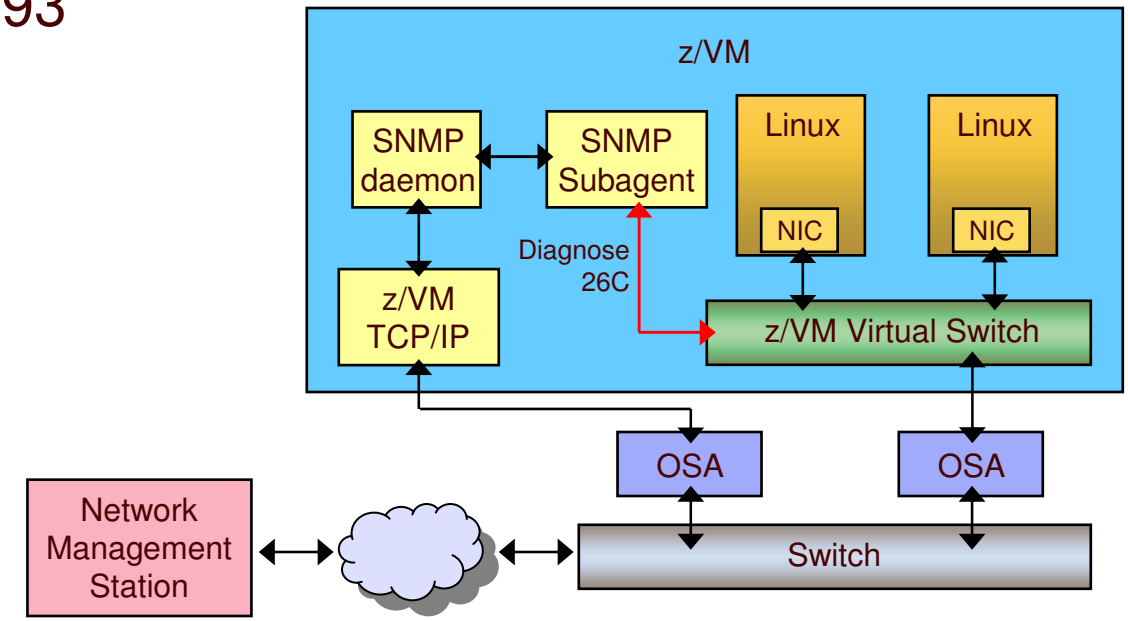
- **Binds multiple OSA-Express ports into a single "pipe"**
  - Up to 8 ports per virtual switch
  - Increases Virtual Switch bandwidth and provides nearly seamless failover in the event of a failed controller, link or switch
  - Only supported for Layer 2 VSWITCHes
- **Includes support to recover from a failed external switch**

# IEEE 802.3ad Link Aggregation

- Define an OSA port group
  - SET PORT GROUP *name* JOIN E100 E200.P1
- DEFINE VSWITCH … ETHERNET GROUP *name*
- OSAs **cannot** be shared

# z/VM Virtual Switch SNMP MIB

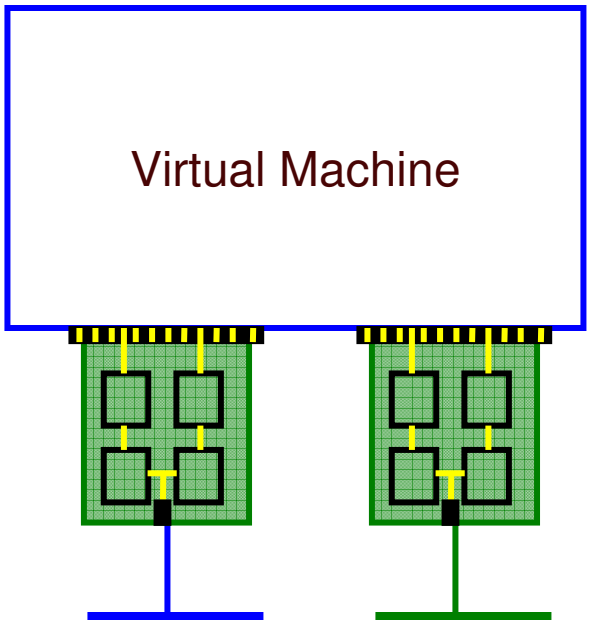
- Integrates VSWITCH into standards-based switch management and monitoring tools
- SNMP subagent provides Bridge MIB data
  - Defined by RFC 1493



# 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 Guest LAN or Virtual Switch
- Created by NICDEF or CP DEFINE NIC command



# Virtual NIC - User Directory

- One per interface in USER DIRECT file:

```
NICDEF vdev [TYPE HIPERS | QDIO]
            [LAN owner name]
            [DEVICES nn]
            [CHPID cc]
            [MACID xxyyzz]

Example:

NICDEF 1100 LAN SYSTEM SWITCH1 CHPID B1 MACID B10006
```

Combined with VMLAN MACPREFIX to create virtual MAC

# Virtual NIC - CP Command

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

```
CP DEFINE NIC vdev
        [[TYPE] HIPERsockets|QDIO]
        [DEVices devs]
        [CHPID cc]

CP COUPLE vdev [TO] owner name

Example:

CP DEFINE NIC 1200 TYPE QDIO
CP COUPLE 1200 TO SYSTEM SWITCH12
```

# NIC CHPID parameter

**CHPID cc**

- Specifies the Channel Path ID number (in hex) to use for this NIC
- Needed for z/OS guests only when connecting to HiperSockets Guest LAN
- **This is a virtual CHPID number**
  - Default is any available unused real CHPID number

# SET NIC

- SET NIC [USER *userid*] *vdev …*
  - PROMISCUOUS | NOPROMISCUOUS (class G)
  - MACID SYSTEM (class B)
  - MACID USER *hhhhhh* (class B)
  - MACPROTECT UNSPECIFIED | OFF | ON (class B) 6.1

# 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

# 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 | ▪Port-based configuration<br>▪HiperSocket bridge |
| z/VM 6.1 | ▪Uplink port can be OSA or guest<br>▪zEnterprise Ensemble (IEDN and INMN)<br>▪VLAN UNAWARE, NATIVE NONE |
| z/VM 5.4 | ▪Port isolation<br>▪Native VLAN id defaults to 1<br>▪z/VM TCP/IP support for Layer 2 |
| z/VM V5.3 | ▪Link aggregation<br>▪Separation of default VLAN id from native VLAN id<br>▪SNMP monitor |
| z/VM V5.2 | ▪Virtual SPAN ports for sniffers |
| z/VM V5.1 | ▪Virtual trunk and access port controls<br>▪Removal of VLAN ANY<br>▪Layer 2 (MAC) frame transport<br>▪Improved virtual switch error detection & recovery<br>▪External security manager access control |
| z/VM V4 | ▪IPv4 Virtual Switch with IEEE VLANs<br>▪IPv4 HiperSocket Guest LAN<br>▪IPv4 and IPv6 QDIO Guest LAN |

# References

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

# Contact Information

Session 10312

- By e-mail: Alan_Altmark@us.ibm.com
- In person: USA 607.429.3323
- On the Web: http://ibm.com/vm/devpages/altmarka
- Mailing lists: IBMTCP-L@vm.marist.edu
  IBMVM@listserv.uark.edu
  LINUX-390@vm.marist.edu

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