z/VM Virtual Switch
Part 1: The Basics

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

Alan_Altmark@us.ibm.com

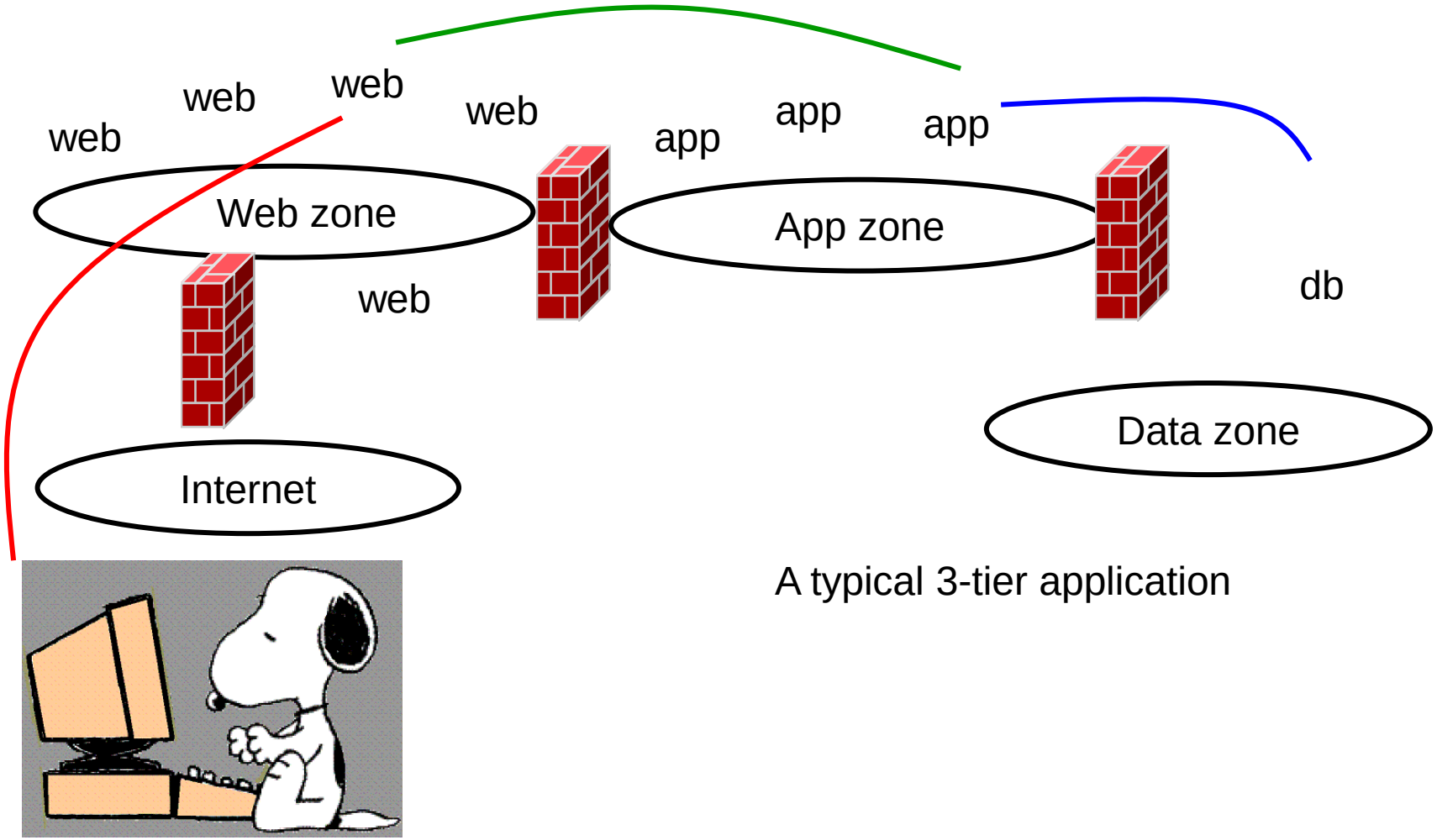
Notes

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.

Topics

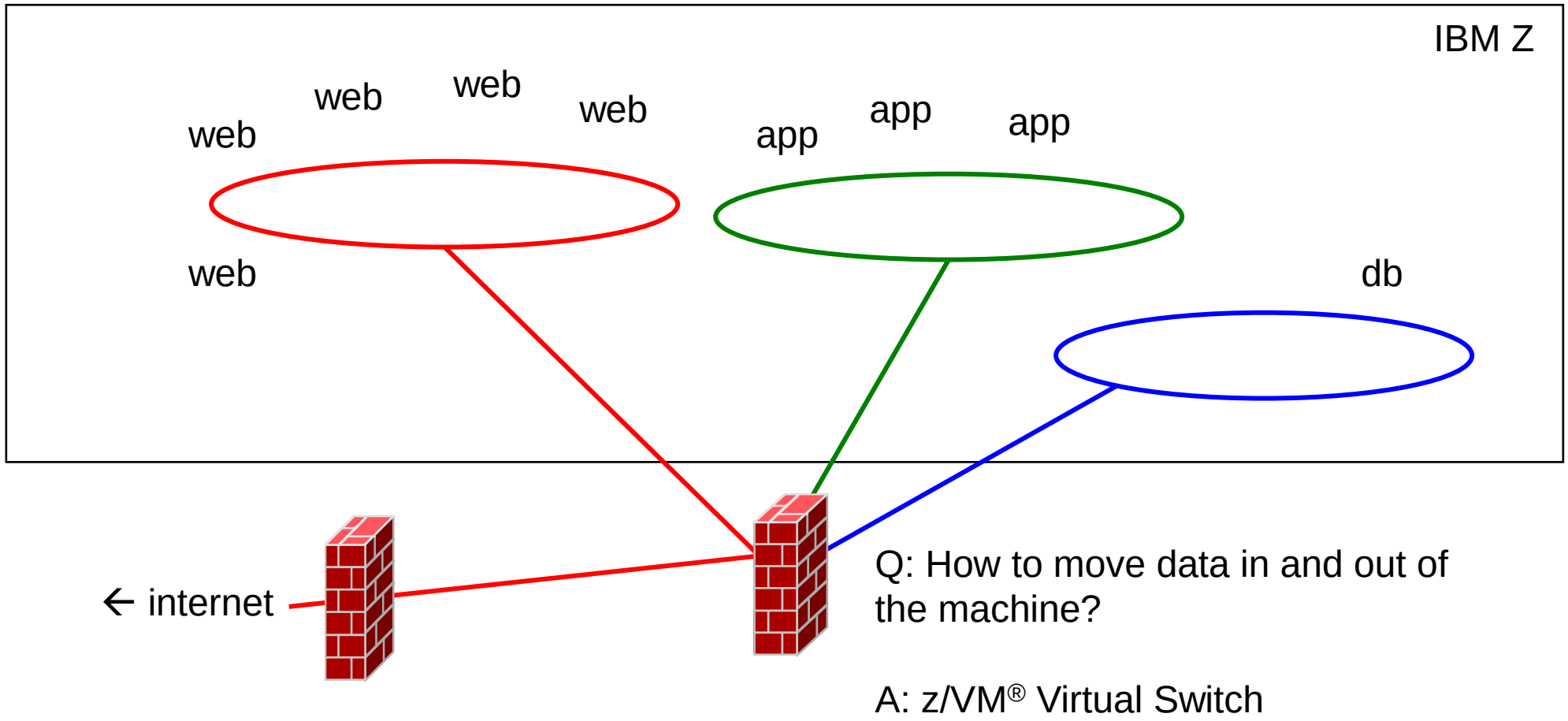
- Overview
- Multi-zone Networks
- Virtual Switch
- Virtual NIC

Multi-Zone Network

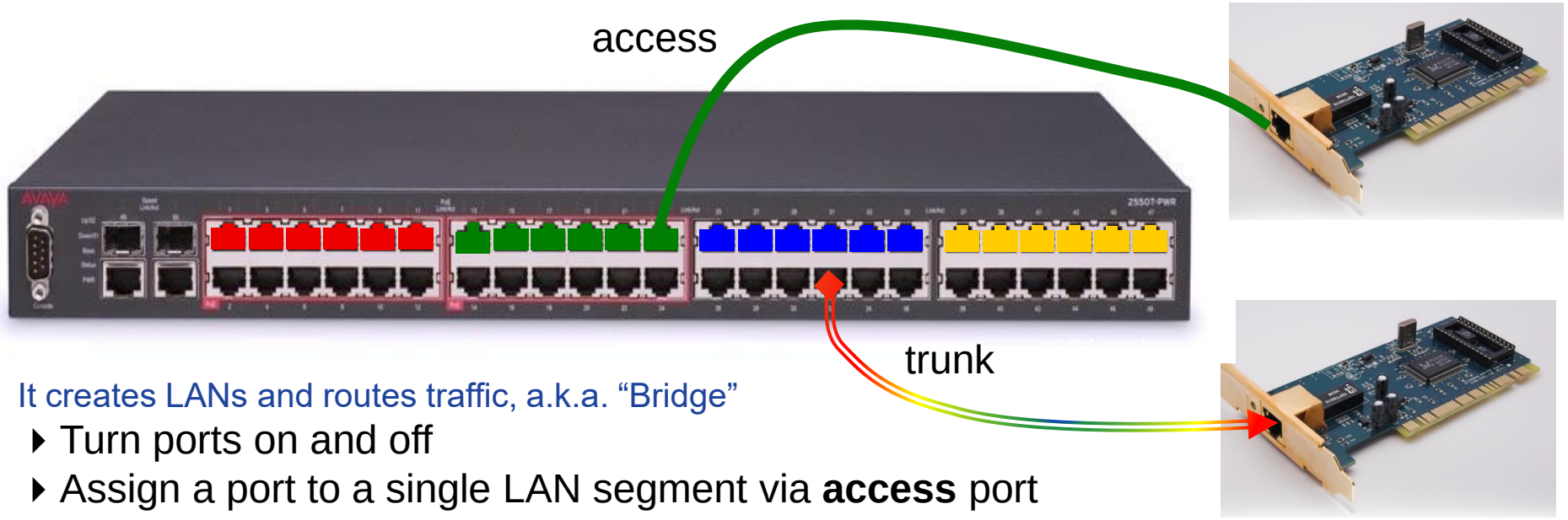


A typical 3-tier application

Multi-zone Network on IBM Z
With outboard firewall / router

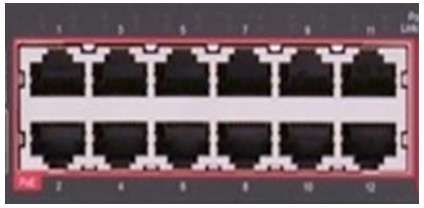


Q: What's a switch?
A: A network device management endpoint

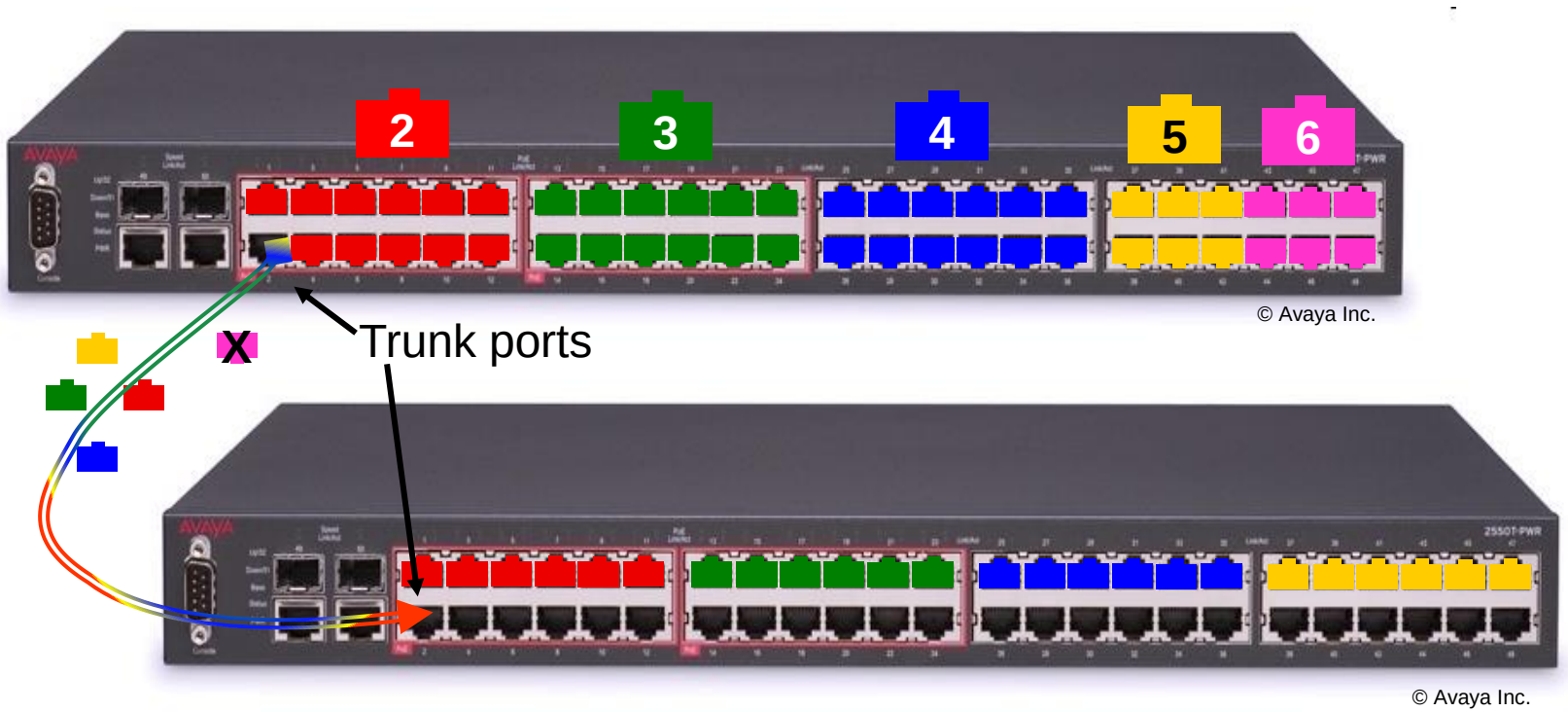


It creates LANs and routes traffic, a.k.a. "Bridge"

- 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
- Ports are numbered for management



Q. What's a Bridge?
A: A way to connect two switches



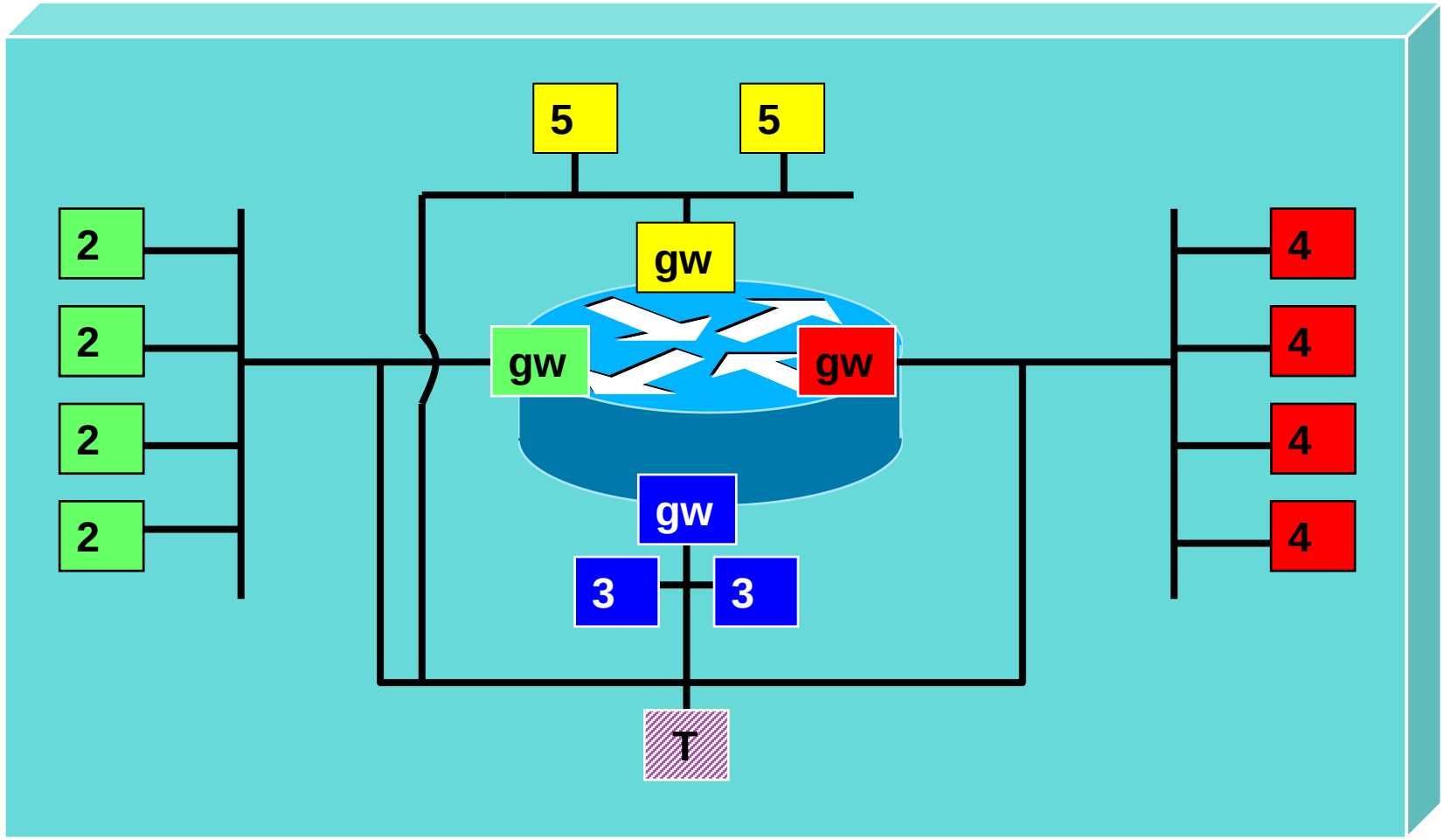
- If you run out of ports, you don't throw it away, you "bridge" or "trunk" it to another switch
- VLAN tags enable the trunk ports to identify the LAN segment to which a frame belongs.
- Single cable carries frames for multiple LAN segments

Layer 2 and Layer 3 Switches
A Network Engineer's Point of View

- Layer 2 Switch
  - Station-to-Station within a single physical LAN segment
  - May implement IEEE Virtual LANs (VLANs)
  - Doesn't care about network protocol, opaque payload
  - May filter based on learned information
    - Which MACs are plugged into which ports
    - Unicast v. multicast v. broadcast MAC addresses

- Layer 3 Switch
  - All functions of layer 2 switch, plus a router
  - Enable Wide Area Network (WAN)
    - Collection of LANs
    - Network addressing awareness: IP, SNA, etc.

Imbedded IP router for Layer 3 Switch



What's a "VLAN"?

- Defined by IEEE 802.1Q standard (not z/VM!)
- IEEE 802.1Q establishes a new set of rules and frame formats
  - Associated with each VLAN is a VLAN Identifier (VID).
  - VLAN-tagged ethernet frames carry the VID within the frame. Allowed only on trunk ports.
  - Untagged frames do not carry the VID, but are instead associated with a VID by the switch and then managed as though they were tagged
- VLAN-aware bridges create logical groups of end stations that can communicate as if they were on the same LAN by associating the physical port used by each of those end stations with the same VID.
- Traffic between VLANs is restricted. Bridges forward unicast, multicast, and broadcast traffic to ports that serve the VLAN to which the traffic belongs.
  - Routers connect to multiple VLANs

IP mode
aka “Layer 3”

- Guest device driver sends/receives IP packets
- CP relays packets to/from other guests or OSA
- IPv4 only

- Guest IP address registered with CP and OSA
  - Inbound packets with unregistered IP addresses are sent to PRIROUTER
    - Default NONROUTER

- OSA builds ethernet frame
- Outbound ethernet frame uses OSA burned-in MAC address
- OSA manages ARP
  - ARP not needed inside VSWITCH

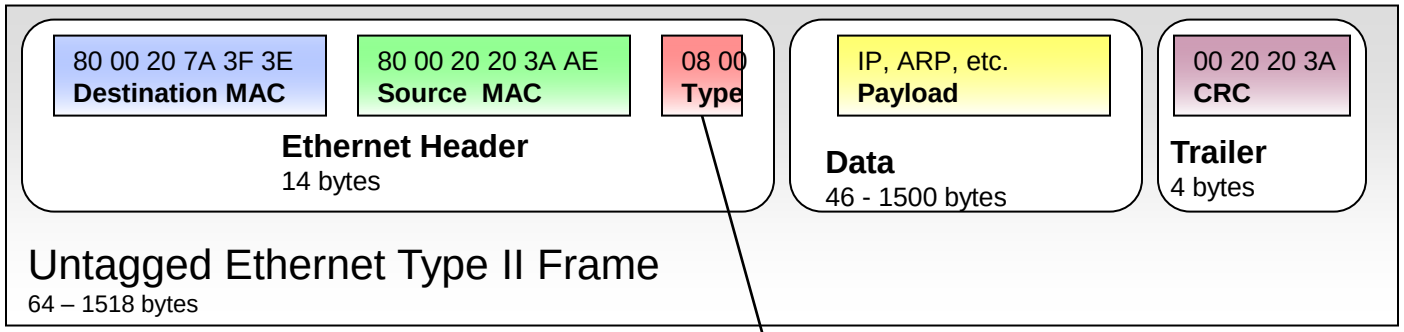
ETHERNET mode
aka “Layer 2”

- Guest device driver sends/receives ethernet frames
- CP relays frames to/from other guests or OSA
- All network protocols, including DHCP

- Guest virtual NIC MAC address registered with OSA
  - Unrecognized inbound MACs are discarded

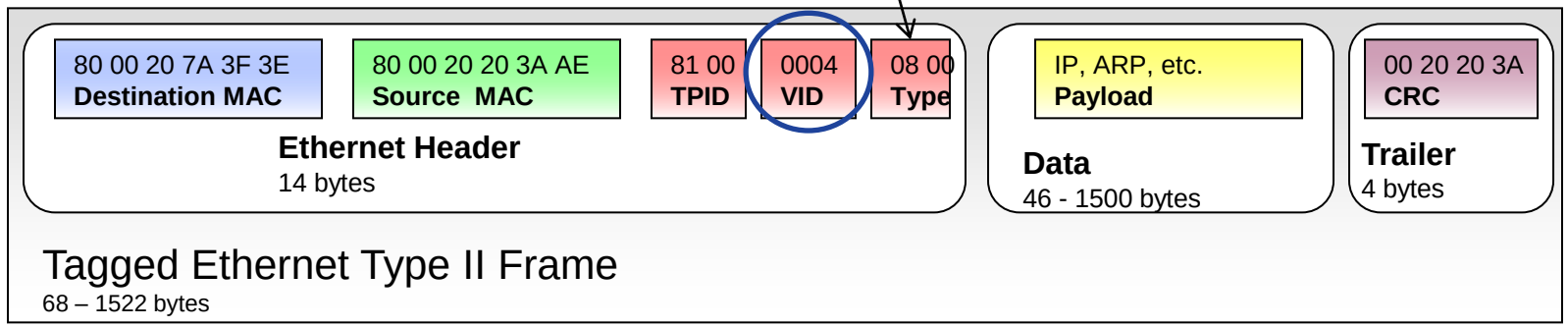
- Guest builds ethernet frame
- Outbound frame uses guest MAC address
- Guest manages ARP
  - CP detects ARP responses to know IP address (Q VSWITCH)

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.
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

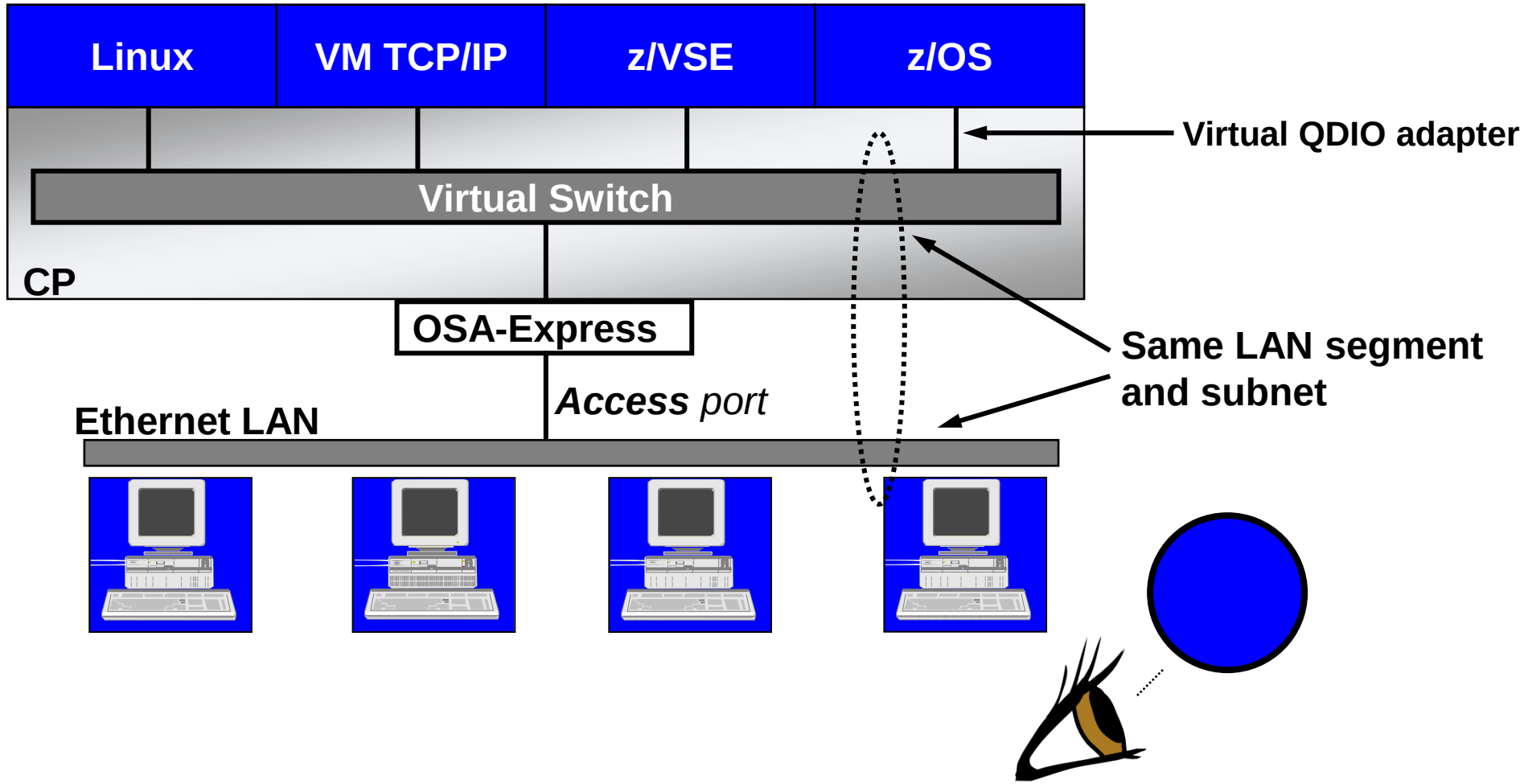
Sidebar: What is a native VLAN?

- When an untagged frame is received on a trunk port the switch will associate the frame with the local default or native VLAN ID (VID), typically VLAN 1
  - Used for switch management traffic

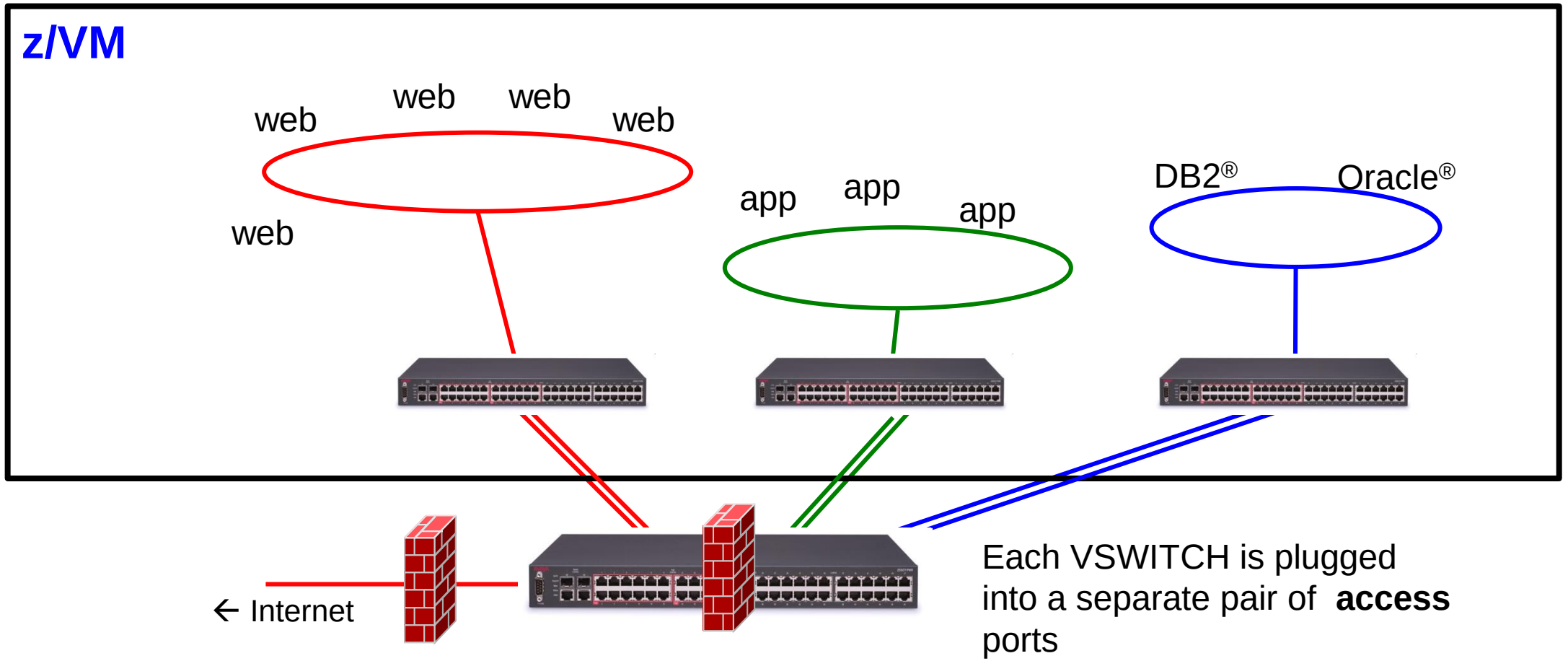
- Identified by the NATIVE keyword on the DEFINE VSWITCH command

Best Practice: Define VSWITCH with “NATIVE NONE”

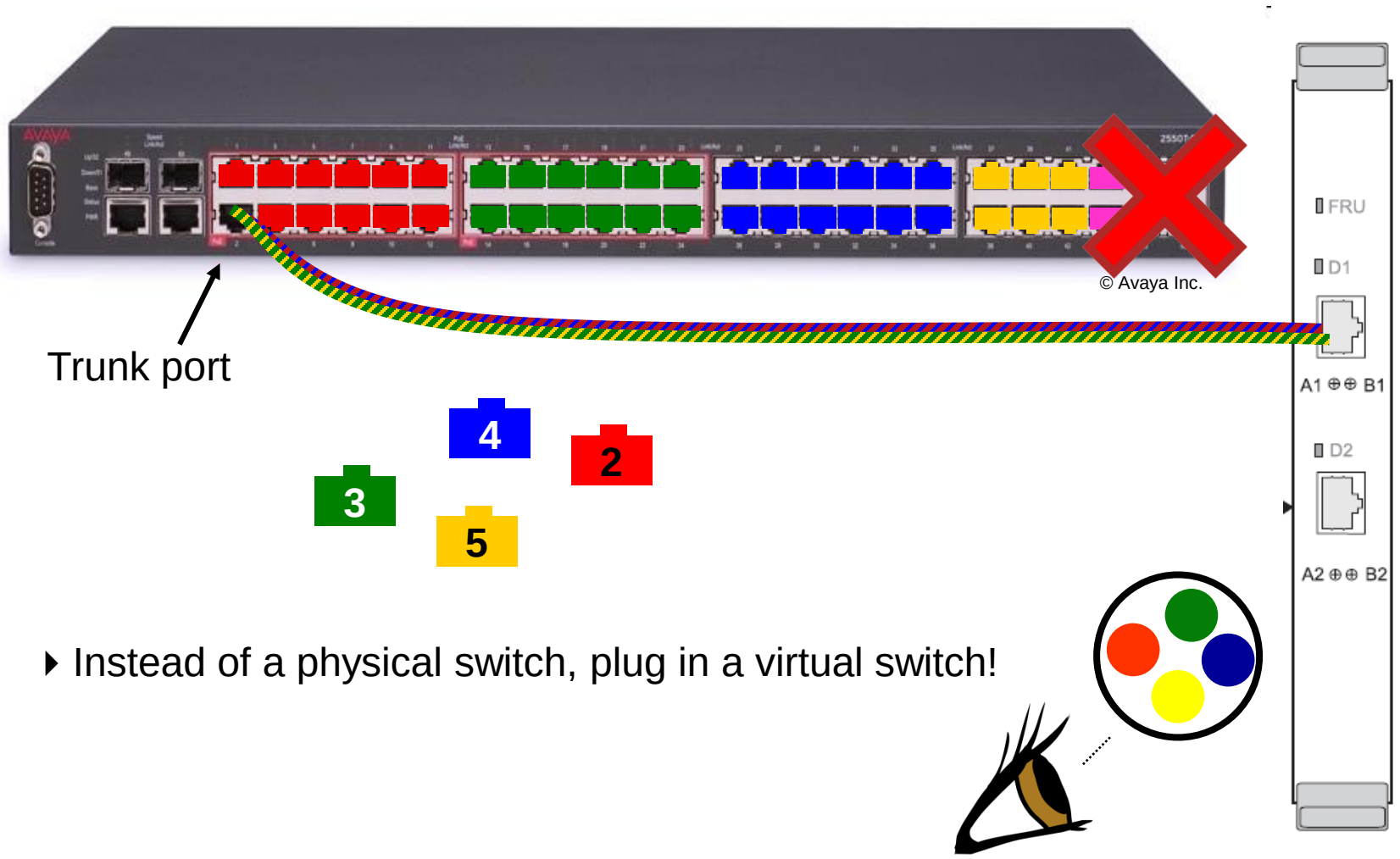
VLAN-unaware Virtual Switch
Sees single LAN segment



One VSWITCH per LAN segment

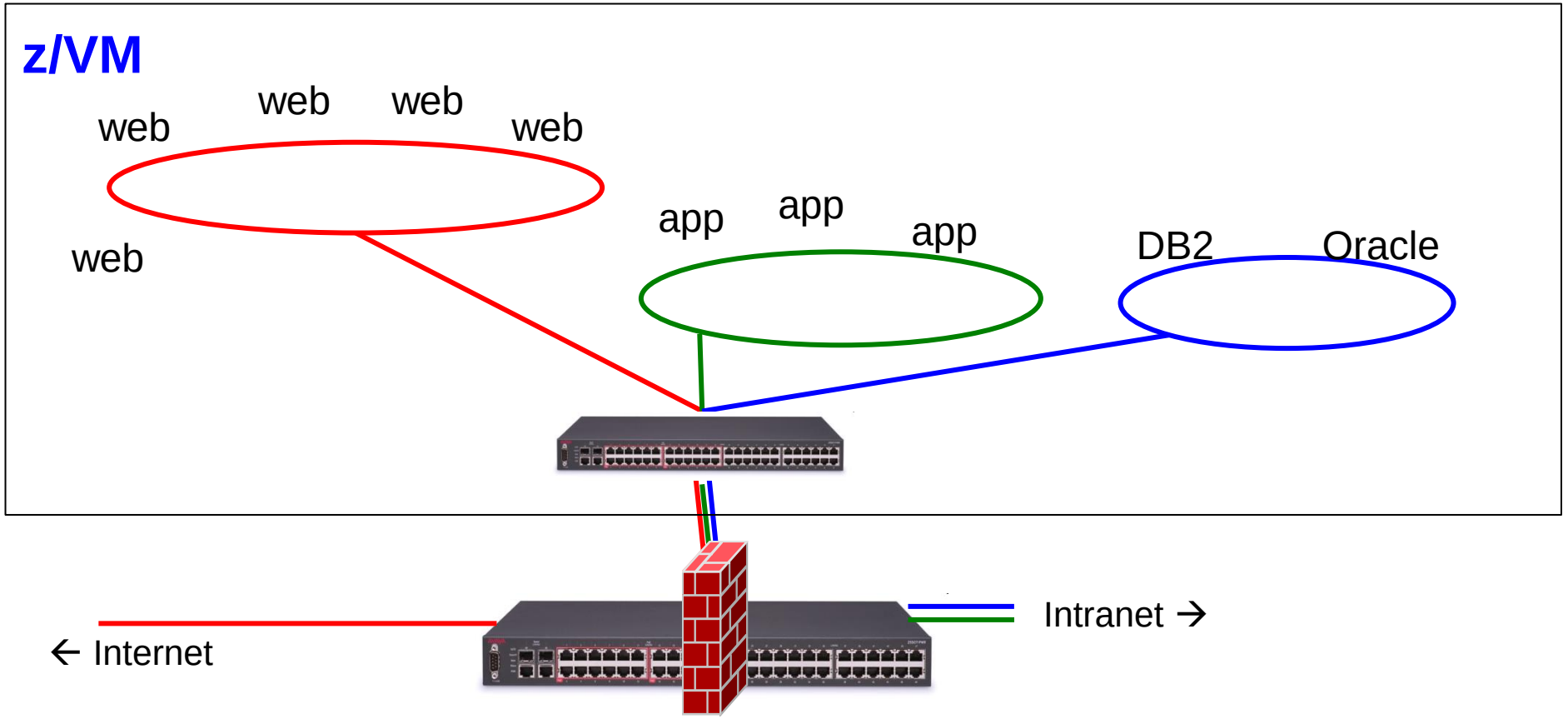


VLAN-aware Virtual Switch



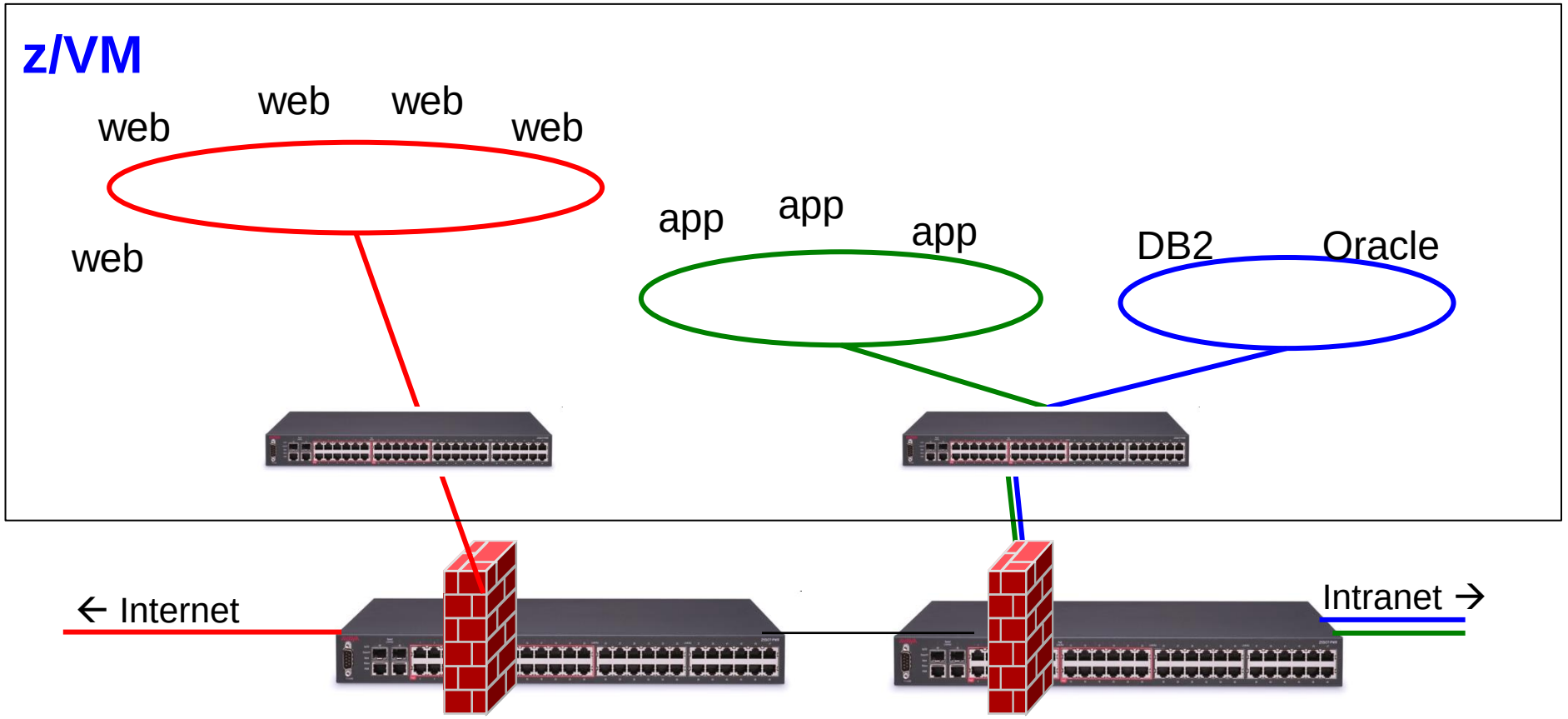
- Instead of a physical switch, plug in a virtual switch!

Multiple LAN segments per VSWITCH



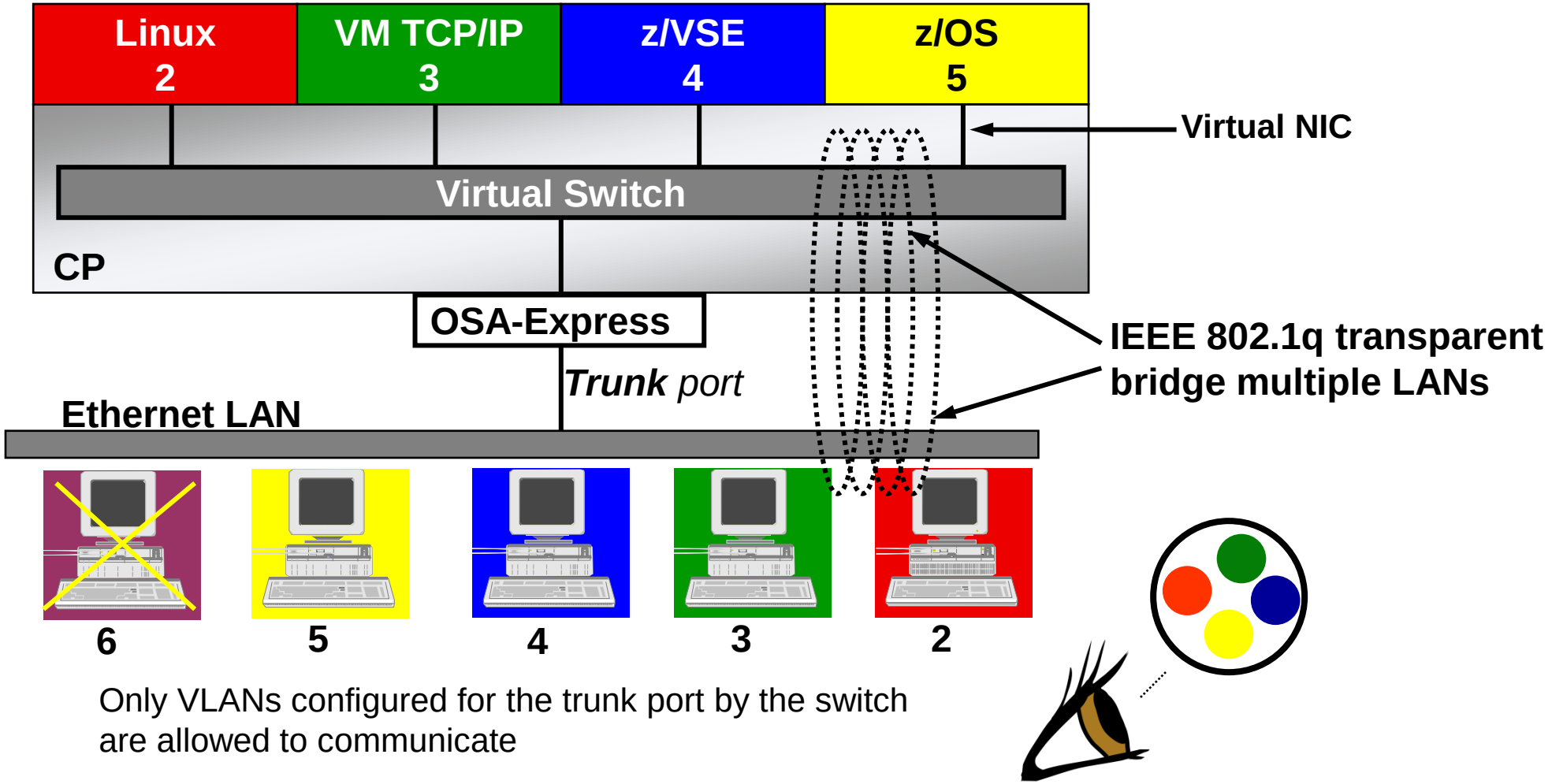
Single VSWITCH plugged into a trunk port

More conservative....



Single VSWITCH plugged into a trunk port

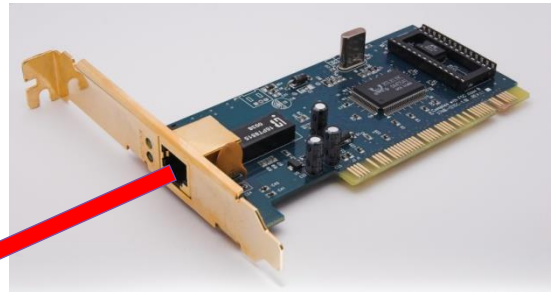
VLAN-aware Virtual Switch
Sees all authorized LAN segments



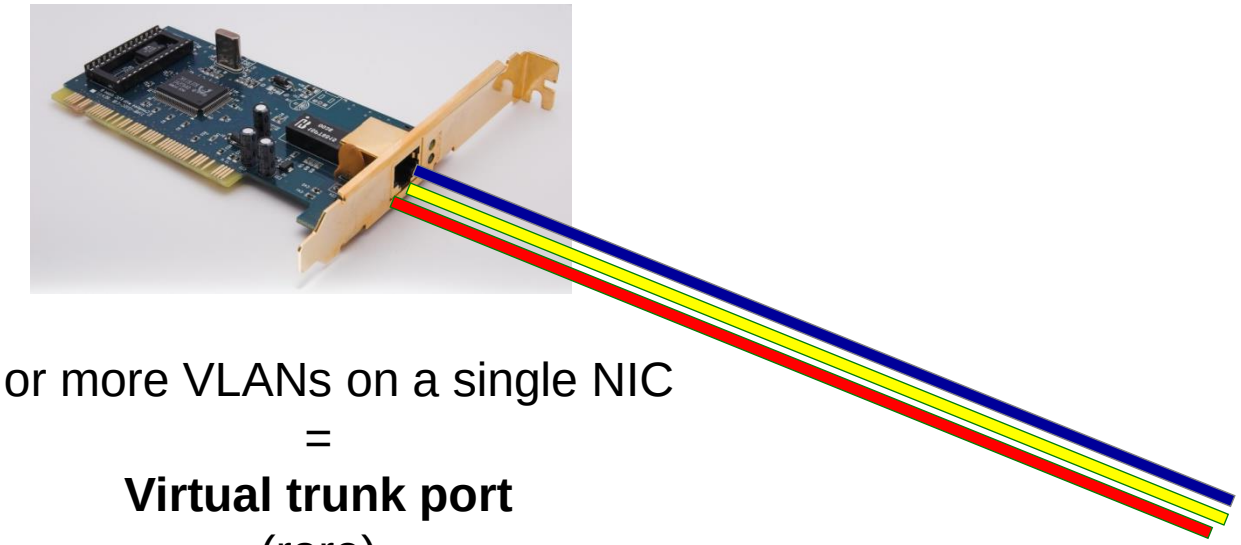
Only VLANs configured for the trunk port by the switch are allowed to communicate

First Look: Virtual NIC

One VLAN on a NIC
=
Virtual access port



What VSWITCH?
What VLAN(s)?
Sniffer authorized?
MAC address?



2 or more VLANs on a single NIC
=
Virtual trunk port
(rare)

First Look: Virtual MAC Addresses

- 6 bytes
  - May appear on physical network

- MAC **PREFIX**
  - From SYSTEM CONFIG
  - High-order 3 bytes: 02xxxx
  - Leading ‘02’ indicates that they are locally-defined addresses

- MAC **ID**
  - Low-order 3 bytes
  - CP can select dynamically
  - Pre-define via NICDEF directory

Virtual Switch primary attributes

- Mode of operation: ETHERNET (preferred) or IP
- Uplink port
- Controller

- Unless otherwise configured, traffic remains as close to the virtual machines as possible
  - Within the VSWITCH
  - Within the OSA
  - Within the physical switch

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-DTCVSW4
  - No need to create more unless directed by Support Center
  - Keep them logged on
    - Monitor with system automation!
  - Automatic failover

- Issues messages to virtual console during error recovery

Uplink Port

- Inbound data from sources not directly coupled to VSWITCH (OSA)
- Outbound packets or frames for unrecognized MAC or IP addresses are placed on the uplink
- Without an uplink, data can move only among coupled guests (more security controls than a Guest LAN)
- For HA, may be a set of 2 or 3 individual ports (failover) or IEEE 802.3ad Link Aggregation port group (port channel)
  - ETHERNET mode only

Setting defaults and limits

— Global attributes in the VMLAN statement in SYSTEM CONFIG:

```
VMLAN
  LIMIT TRANSIENT INFINITE | maxcount

  MACPROTECT OFF | ON

  MACPREFIX prefix1          - For CP-assigned MACs
  USERPREFIX prefix2         - For user-assigned MACs
```

Best Practices

- 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

MACPREFIX 02pppp

- Sets MAC prefix for CP-generated MAC addresses
- Each instance of CP should have a different MACPREFIX
  - Enforced for Single System Image

USERPREFIX 02uuuu

- Sets MAC prefix for NICDEF MACID
- All instances of CP that share a directory should have the same USERPREFIX
  - Enforced for Single System Image
  - Defaults to MACPREFIX value

Best Practice: Do not allow either to default to 02:00:00!
Warning: You must re-IPL to change

Create an Ethernet mode Virtual Switch

— SYSTEM CONFIG or CP command:

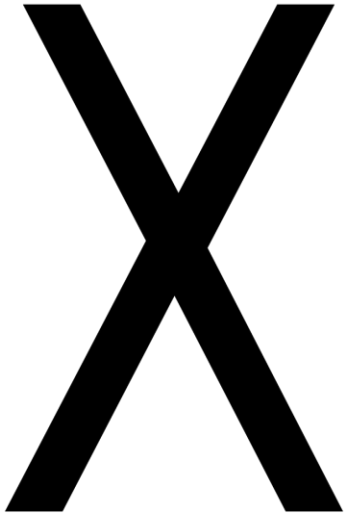
```
DEFINE VSWITCH name ETHERNET    [PORTBASED]
MODIFY
SET             [RDEV NONE | dev1 [dev2 [dev3]] ]

                [GROUP group_name]

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

                [ISOLATION OFF | ON]
```

Best Practice: VLAN AWARE NATIVE NONE

Best Practice: There are other options – don't use them

Create an IP mode Virtual Switch

— SYSTEM CONFIG or CP command:



```
DEFINE VSWITCH name  IP     [PORTBASED]
MODIFY
SET            [RDEV NONE | dev1 [dev2 [dev3]] ]

               [NONROUTER | PRIROUTER]

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

               [ISOLATION OFF | ON]
```

Best Practice: Use ETHERNET mode instead of IP mode

OSA Devices

- RDEV NONE
  - No outside communications
  - Similar to Guest LAN, but with better security
  - Excellent for 2nd level systems

- RDEV *dev1[.port]* [*dev2[.port]* [*dev3[.port]*]]
  - Up to 3 ports
  - P0 (default) or P1
  - Round-robin failover
  - If all dead, wait for signs of life
  - SET VSWITCH SWITCHOVER to manually change

- GROUP *name*
  - IEEE 802.3ad link aggregation (channel bonding)
  - ETHERNET mode only

Virtual NIC - User Directory

— Interface fully configured in the user's directory entry

```
NICDEF vdev TYPE QDIO
               [LAN SYSTEM switch]
               [DEVICES nn]
               [MACID hhhhhh]
               [PORTNUMBER n]
               [PORTTYPE ACCESS|TRUNK]
               [VLAN vidset]
               [PROMISCUOUS|NOPROMISCUOUS]

Example:

NICDEF 1100 TYPE QDIO LAN SYSTEM SWITCH1
NICDEF 1100 MACID B10006
NICDEF 1100 VLAN 57
```

Combined with VMLAN USERPREFIX to create virtual MAC

VSWITCH authorization

- CP authorization **and** configuration in **NICDEF**
- NICDEF overrides SET VSWITCH GRANT

- SET VSWITCH used to change user settings dynamically

```
SET VSWITCH  name GRANT userid VLAN vid
```

  - Immediate effect for PORTTYPE, VLAN, PROMISCUOUS

- Revert to old behavior with

```
VMLAN   DNA DISABLE
SET VMLAN DNA DISABLE
```

  - Results in HCP3224I (NICDEF network configuration ignored)

PORTNUMBER n

- Where on the VSWITCH is the virtual NIC plugged in?
  - Useful for SNMP-based switch monitors
  - “Egad! Port 1 is down!”
  - For USERBASED, CP assumes you don’t care
- If you select a port, must be 1-2048
  - COUPLE will fail if there is a conflict
- If you don’t select a port, CP will choose one 2176-4095
  - Cannot VMRELOCATE to pre-DNA system because port above 2048 not supported

Define and connect to VSWITCH

```
DEFINE VSWITCH VSW1 ETHERNET
                    PORTBASED

                    RDEV E00 F00

                    VLAN AWARE
                    NATIVE NONE
```

```
NICDEF E00 TYPE QDIO LAN SYSTEM VSW1 MACID B10006 VLAN 57
```

Best Practice: Use PORTBASED

RACF-managed VSWITCH access control

- **RDEFINE VMLAN SYSTEM.VSW1 UACC(NONE)**
  RDEFINE VMLAN SYSTEM.VSW1.0057 UACC(NONE)
  - 4-digit VLAN IDs
  - No generics for VLAN IDs
  - COUPLE.G must be CONTROLLED in VMXEVENT
  - VMLAN class must be active

- As virtual machine are on-boarded, connect to a *group* that has
  PERMIT SYSTEM.VSW1 CL(VMLAN) ID(*group*) ACC(UPDATE)
  PERMIT SYSTEM.VSW1.0057 CL(VMLAN) ID(*group*) ACC(UPDATE)
  - Normal access = UPDATE
  - Sniffer access = CONTROL

Sniffers and Port Isolation

- "Promiscuous" mode for sniffers
  - Guest must be authorized
  - Guest enables promiscuous mode using CP SET NIC or via device driver controls
    - E.g. tcpdump –P and download for Wireshark
  - Guest receives copies of all frames sent or received for all authorized VLANs

- 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

Best Practices for all VSWITCHes

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

Best Practices for VLAN-aware VSWITCH

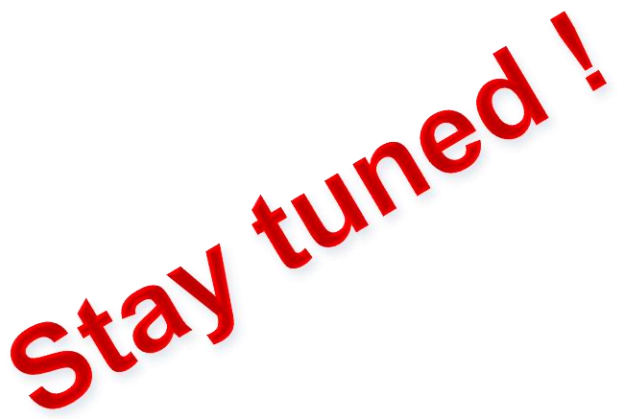
- Use NICDEF to assign VLANs and port numbers

- Define VSWITCH with “VLAN AWARE NATIVE NONE”
  - Guest that has not been given access will get errors
  - No chance of untagged frames escaping from z/VM

- Use ESM and groups to manage VLAN assignments
  - Simplifies VLAN changes
  - Overrides VLAN specification on NICDEF
  - CP will use NICDEF if ESM defers

Additional Virtual Switch Technologies

- Link aggregation
- Cross-LPAR Link Aggregation port group sharing
  - aka “Shared LAG”
- HiperSocket Bridge
- Virtual Ethernet Port Aggregator (VEPA)
- SNMP
- Diagnostics



Support Timeline

z/VM 7.1 2019	▪ Priority queuing
z/VM 6.4 2017	▪ Unified VSWITCH with NICDEF controls CP (VM65925), DIRMAINT (VM65926), RACF(VM65931)
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 ▪ 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 2001	▪ 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 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