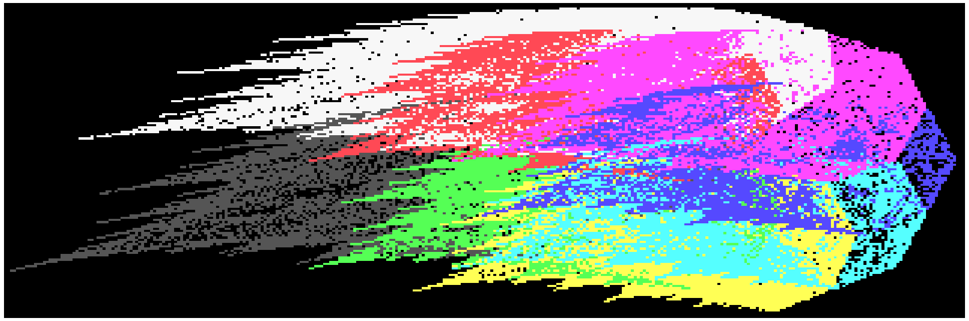
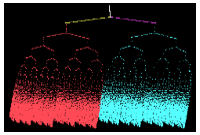


z/VM Virtual Switch
The Basics

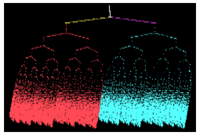
Tim Greer

z/VM System Test



Attribution

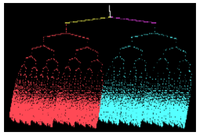
- With contributions from Alan Altmark, IBM's leading Lab Services consultant on z/VM



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 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.



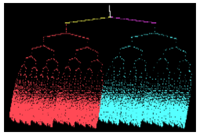
Topics

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

Some z/VM terminology

- Guest LAN (Create with DEFINE LAN)
- Virtual Switch = VSWITCH (Create with DEFINE VSWITCH)
- Virtual NIC (Create with DEFINE NIC)

- VSWITCHes
  - Usually have an UPLINK and/or BRIDGEPORT
  - Allow USERBASED or PORTBASED guest connections
  - Have a CONTROLLER for their UPLINK device



z/VM procedure outline

- Create VSWITCH and permit guests to use it
  - DEFINE VSWITCH …
  - SET VSWITCH GRANT …
- Access the VSWITCH from guests
  - DEFINE NIC …
  - COUPLE …

(Now guests can talk to each other)

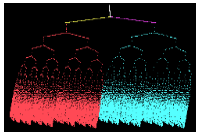
- Connect VSWITCH to the outside world via OSA
  - SET VSWITCH UPLINK …

(Now guests can talk to the world)

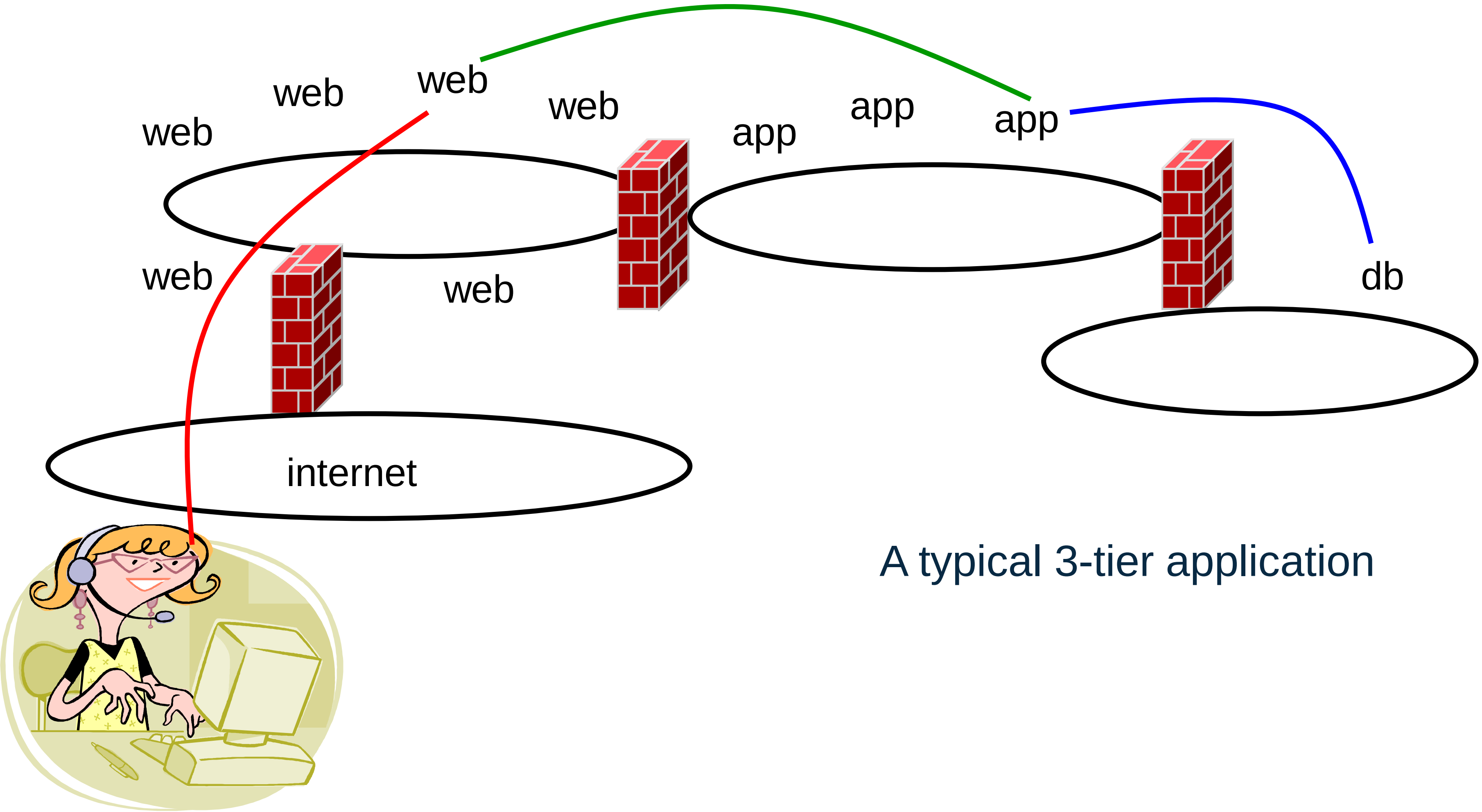
- Connect VSWITCH to hipersocket
  - SET VSWITCH BRIDGEPORT …

(Now guests can talk to/via whatever is on the hipersocket)

Note: There are various ways to invoke the commands.

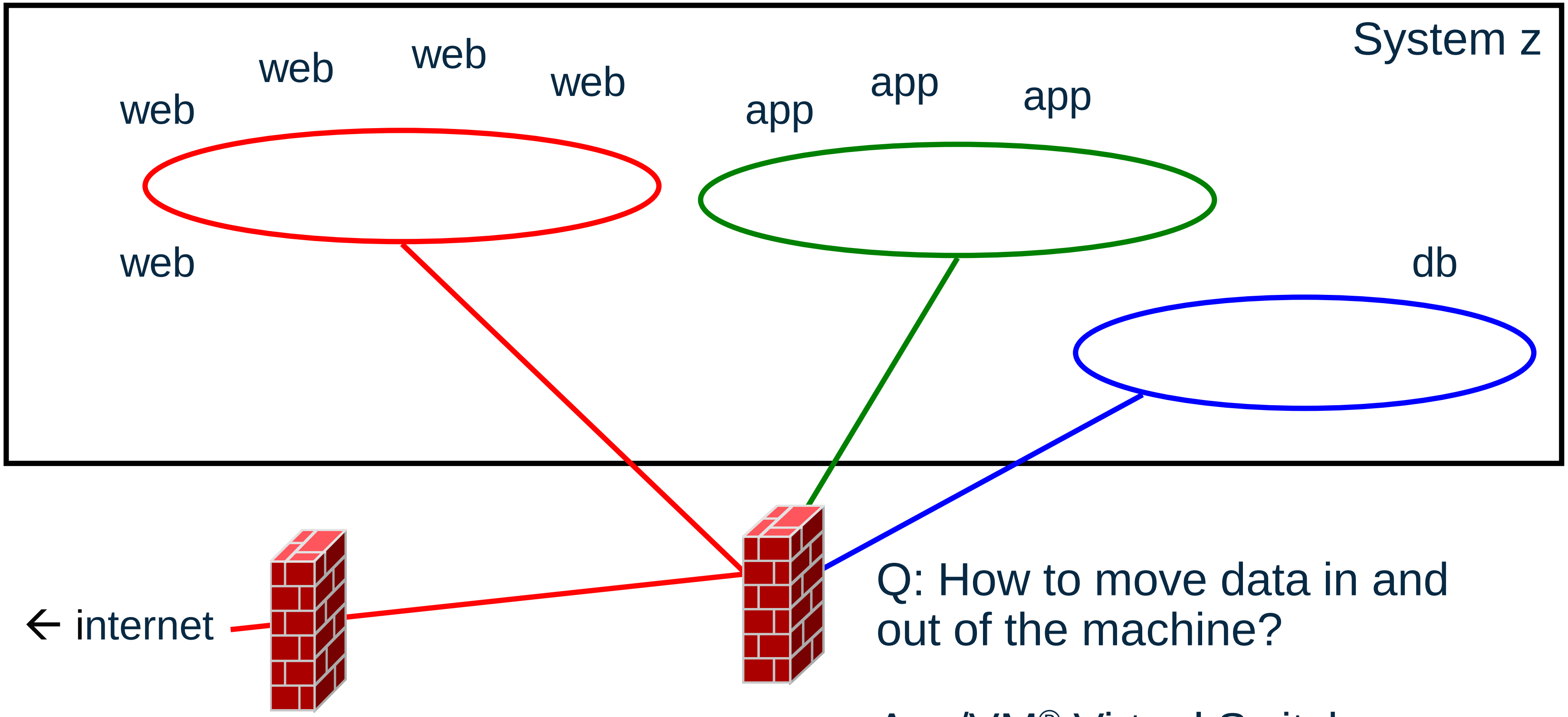


Multi-Zone Network



A typical 3-tier application

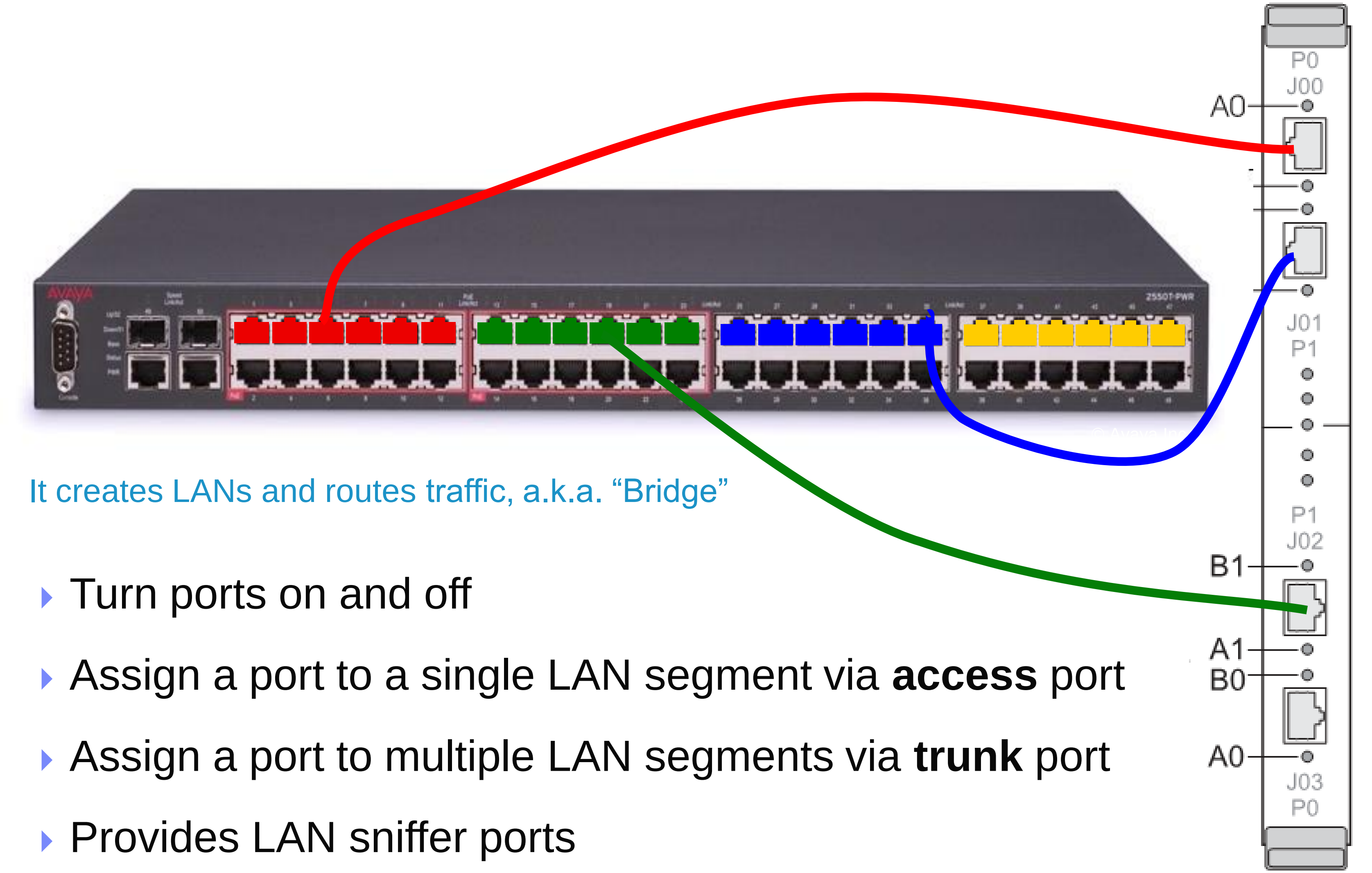
Multi-zone Network on System z
With outboard firewall / router



Q: How to move data in and out of the machine?

A: z/VM® Virtual Switch

What’s a switch?



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

What’s a “VLAN”?

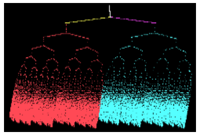
- Defined by IEEE 802.1Q standard (not z/VM!)
  - “A subset of the active topology of a bridged LAN.”

- A bridged LAN is what you get when you use a switch instead of a hub
  - Enables the application of ingress and egress rules to the frames that enter and exit the switch ports

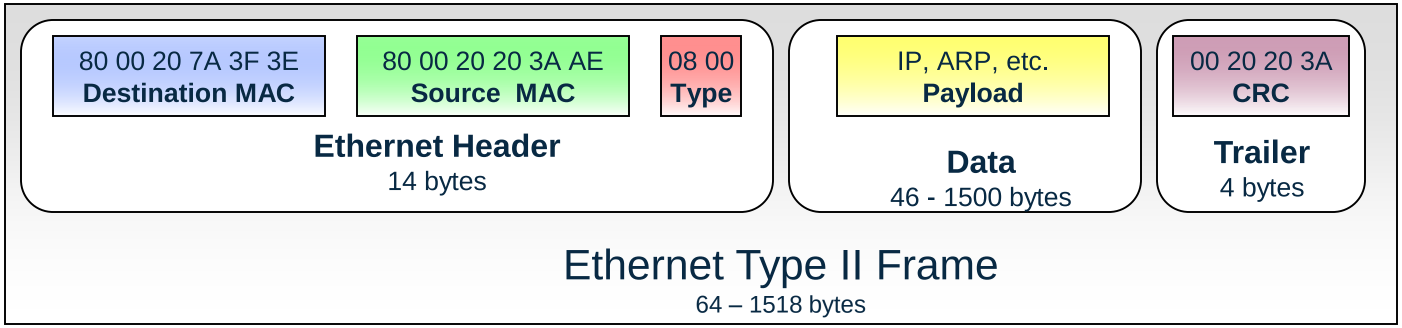
- IEEE 802.1Q establishes a new set of rules and frame formats
  - Associated with each VLAN is a VLAN Identifier (VID).
  - VLAN-tagged frames carry the VID within the frame. Allowed only on trunk ports.
  - Untagged frames do not carry the VID, but are instead associated with a VID by the switch and 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



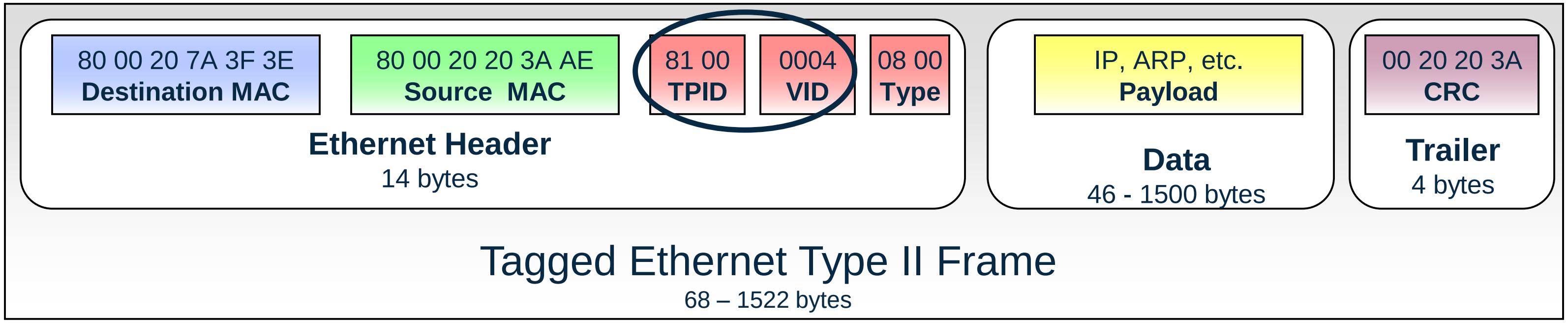
VLAN tags



Access port and Trunk port

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

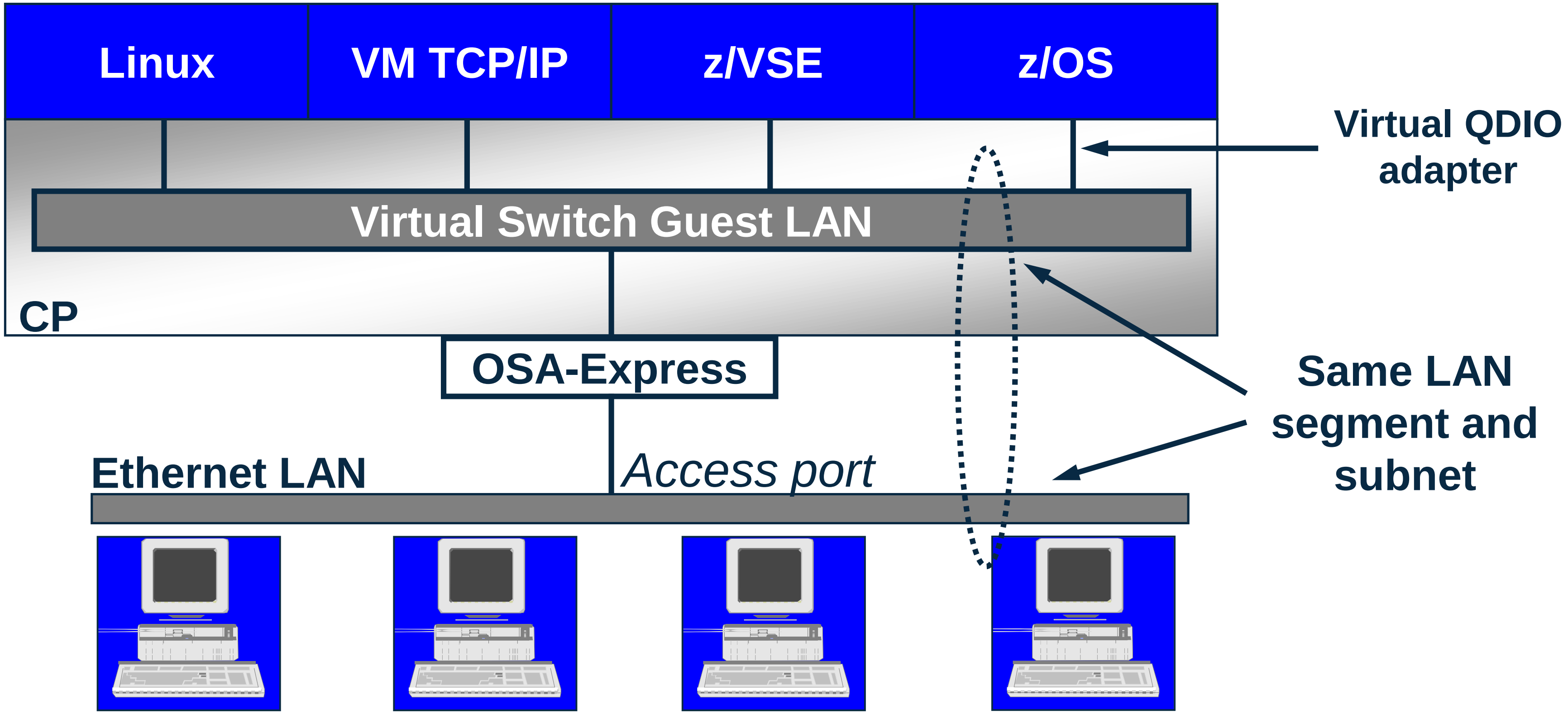
Type/length 0800 means IPv4 (IETF RFC 894)



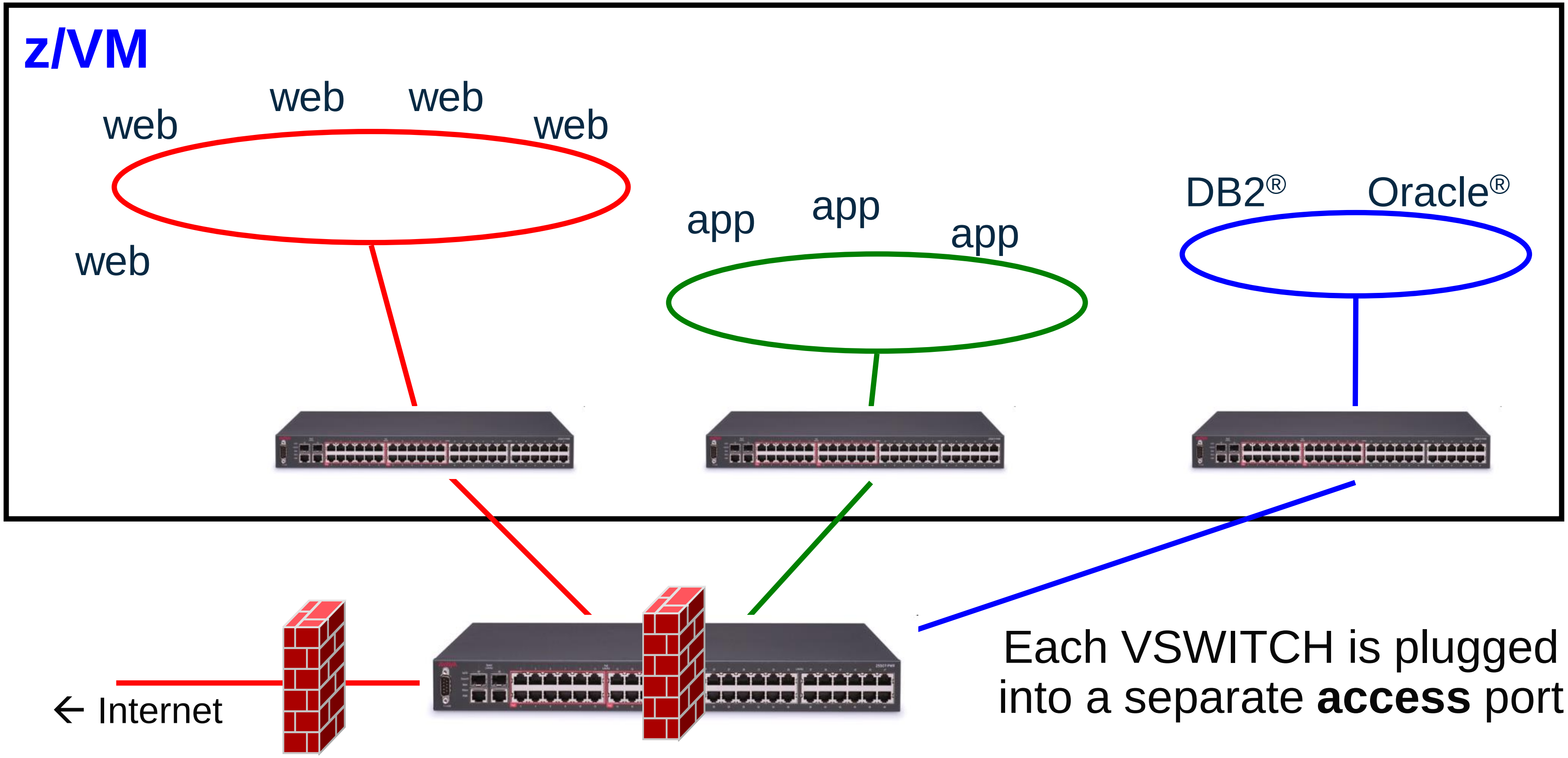
Trunk port only

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

VLAN-unaware Virtual Switch
Sees single LAN segment

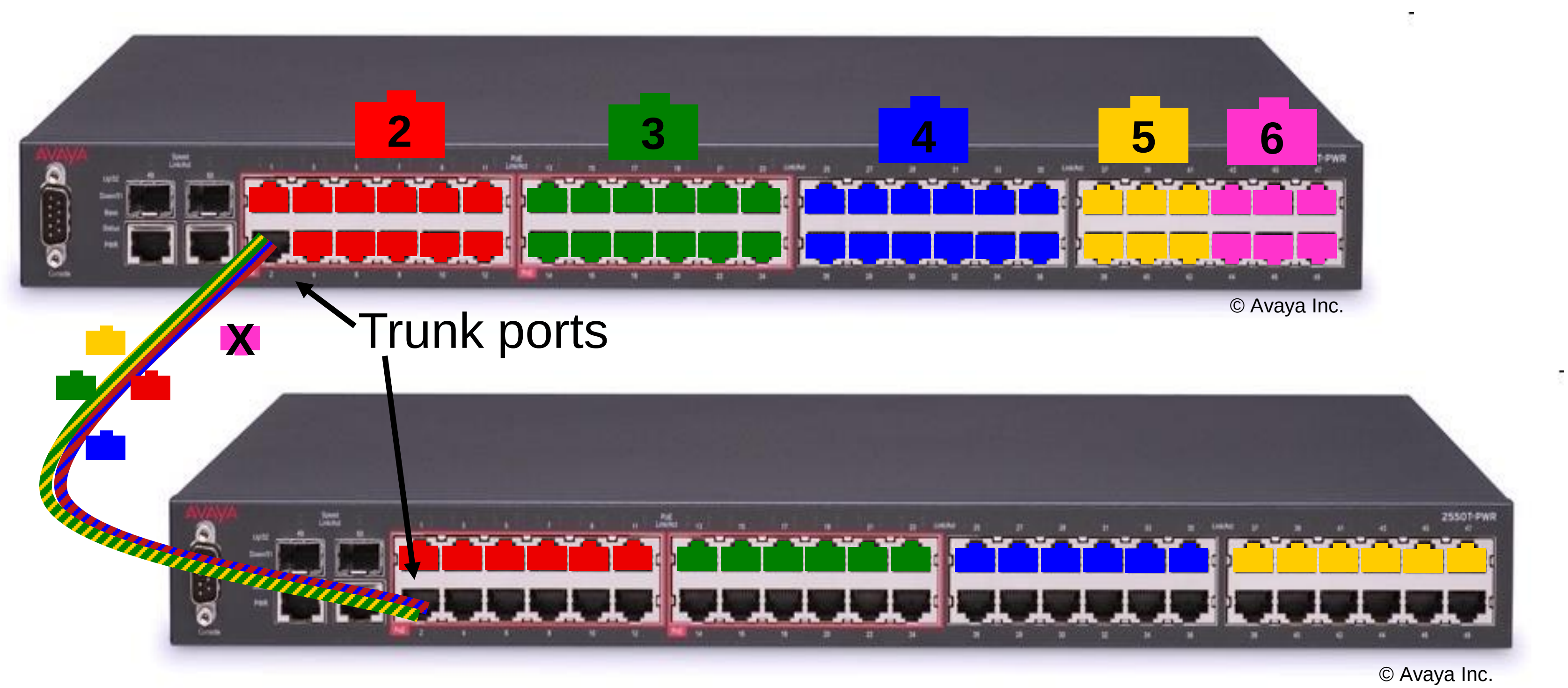


One VSWITCH per LAN segment



Each VSWITCH is plugged into a separate **access** port

IEEE VLANs using a Trunk port

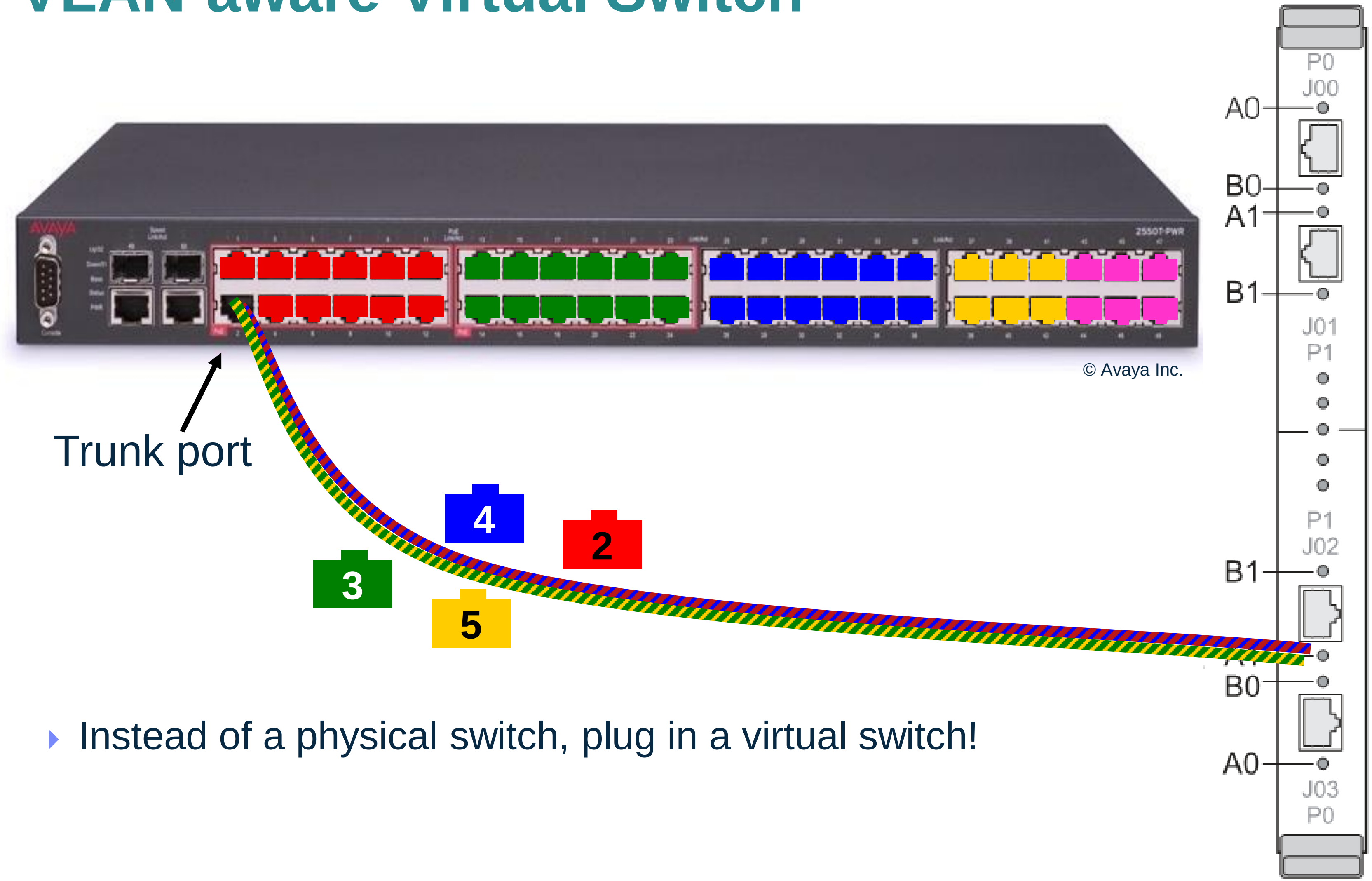


- If you run out of ports, you don’t throw it away, you “trunk” it to another switch
- VLAN tags enable the trunk ports to identify the LAN segment to which a frame belongs.
- Single cable carries frames for multiple LAN segments

What is a native VLAN?

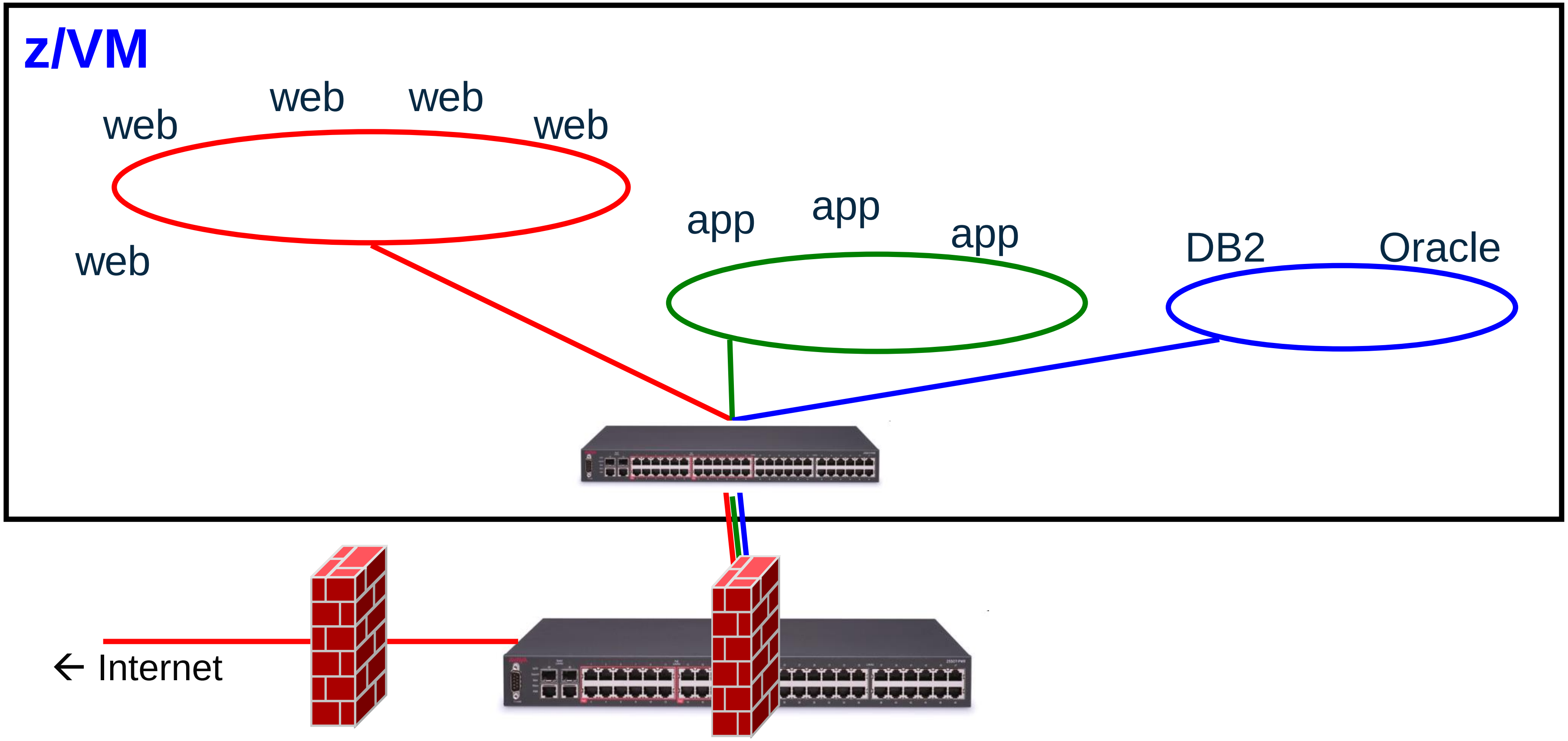
- When an untagged frame is received on a trunk port the switch will associate the frame with the local default or native VID (usually VLAN 1)
  - Identified by the NATIVE keyword on the DEFINE VSWITCH command
  - All inbound and outbound untagged frames are associated with this value for authorization purposes
  - Outbound frames tagged with the native VID will have the tag removed
- All switches should be configured with the same native VID
  - NATIVE NONE is preferred!

VLAN-aware Virtual Switch



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

Multiple LAN segments per VSWITCH



z/VM

web

web

web

web

web

app

app

app

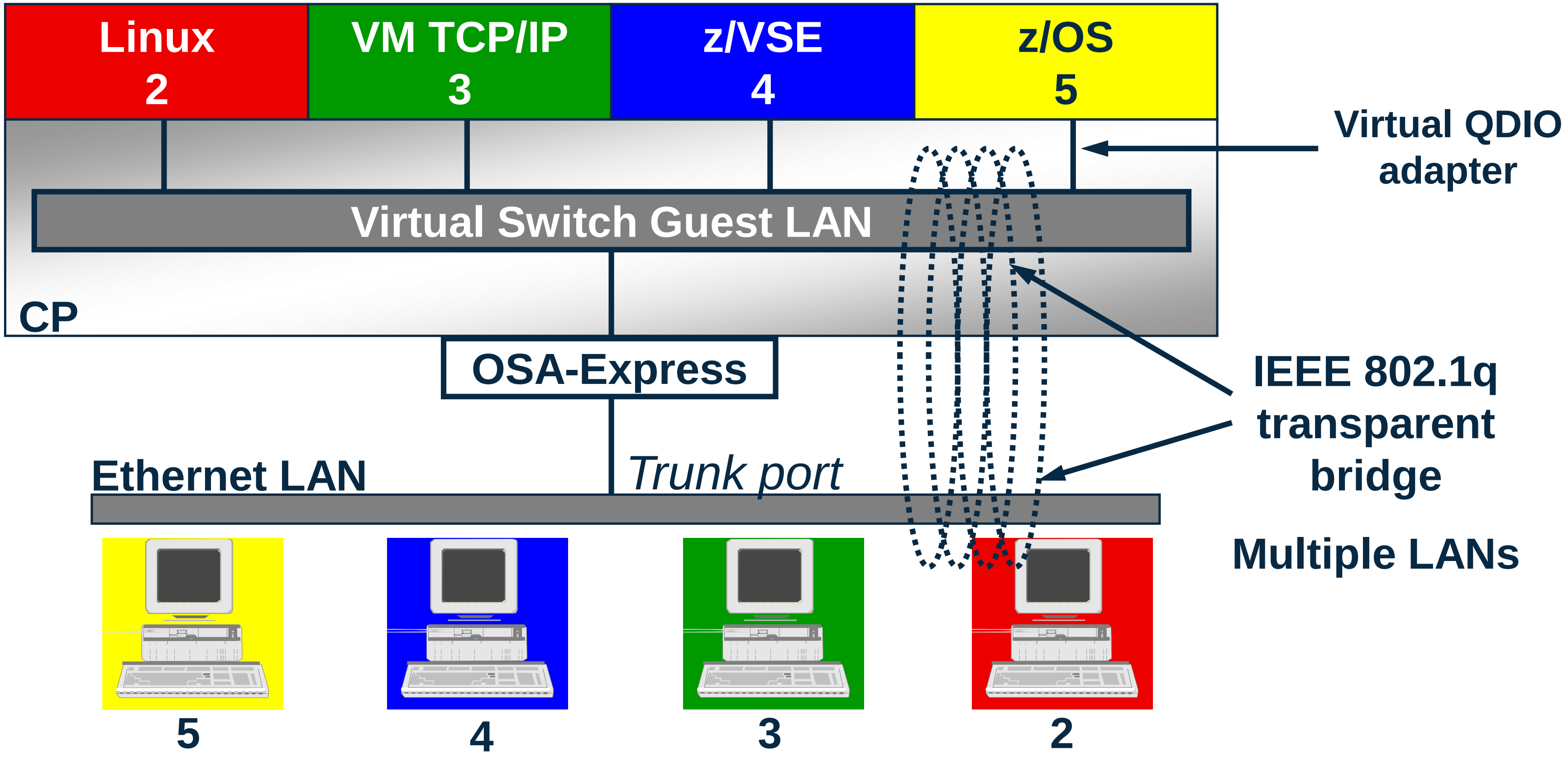
DB2

Oracle

← Internet

Single VSWITCH plugged into a trunk port

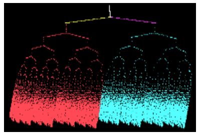
VLAN-aware Virtual Switch
Sees all authorized LAN segments



Primary Virtual Switch Attributes

- An associated controller virtual machine
- Mode of operation: Layer 2 or Layer 3
- Access list
  - Permitted user IDs
  - VLAN assignments
- Associated uplink: individual OSAs, link aggregation port group, virtual NIC, or none (or Hipersockets bridge)

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

Layer 2 and Layer 3
An OSA Point of View

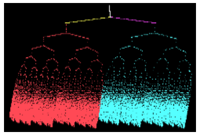
- Layer 2 – Host sends/receives raw ethernet frames to OSA
  - Any protocol: IP, SNA, NETBIOS, AppleTalk, experimental, …
  - CP registers virtual NIC MAC addresses with OSA so it can route inbound frames appropriately
    - Burned-in MAC address not used
  - Guest sends raw frame with its origin and 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
  - Defaults to 02:00:00
  - 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 a Single System Image cluster so that user-defined MAC addresses cannot overlap within the 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]

MODIFY VSWITCH name ISOLATION OFF | ON
SET
```

Create a Layer 3 Virtual Switch

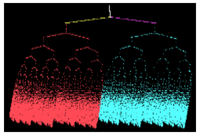
- SYSTEM CONFIG or CP command:

```
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  VLAN  vid …
SET                        [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 (RACF)

- RDEFINE VMLAN SYSTEM.SWITCH12 UACC(NONE)
- RDEFINE VMLAN SYSTEM.SWITCH12.0399 (4 digits)
  - No generics for VLAN IDs

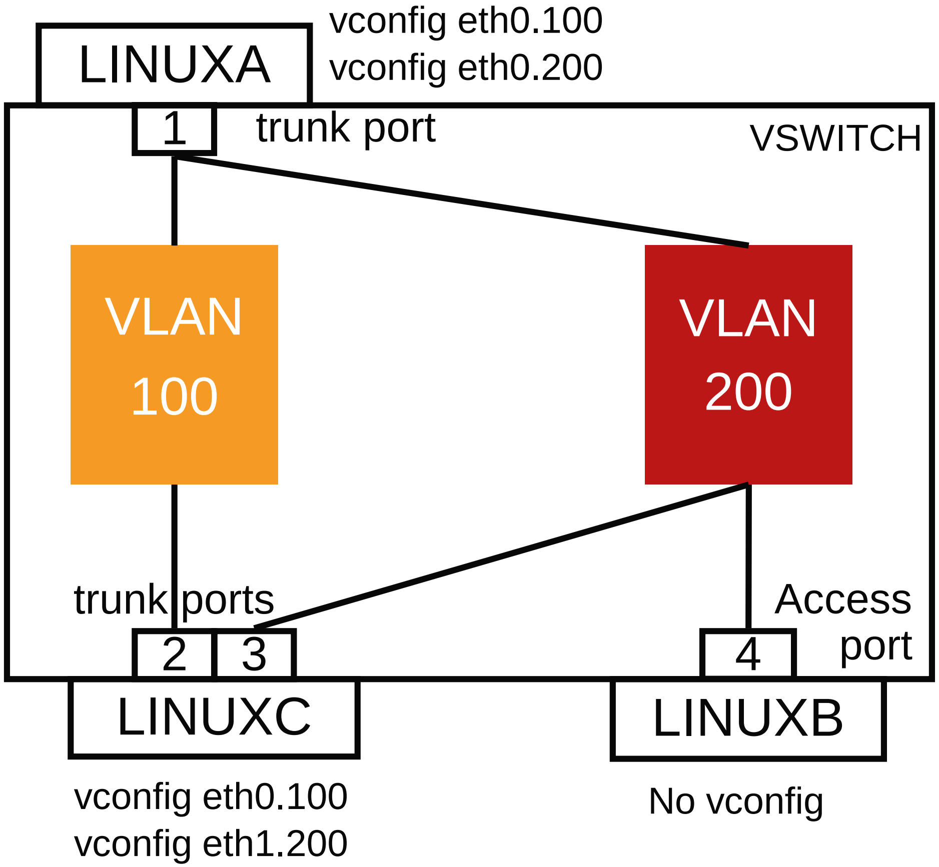
- As virtual machine are on-boarded, connect to a group that has
  PERMIT SYSTEM.vswitch CL(VMLAN) ACCESS(UPDATE)
  PERMIT SYSTEM.vswitch.vlanid CL(VMLAN) ACC(UPDATE)

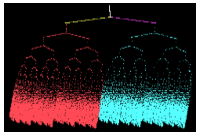
- Normal: UPDATE access
- Sniffer: CONTROL access

- Still use SET VSWITCH GRANT to assign port type
  - Virtual trunk or access port

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 LINUXC 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

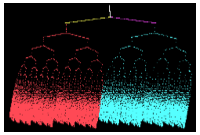
- **LINUXB: NICDEF 4E0 TYPE QDIO LAN SYSTEM VSW1**

- **LINUXC: NICDEF 4E0 TYPE QDIO LAN SYSTEM VSW1**
  NICDEF 5E0 TYPE QDIO LAN SYSTEM VSW1
  + vconfig eth0.100
  + vconfig eth1.200

Additional security controls

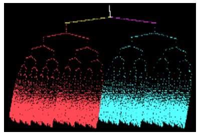
- Virtual Sniffers
  - Guest must be authorized via SET VSWITCH or security server
  - Guest enables promiscuous mode using CP SET NIC or via device driver controls
    - E.g. tcpdump -P
  - Guest receives copies of all frames sent or received for all authorized VLANs
  - Not needed when VEPA is used since frame goes to switch

- 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



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

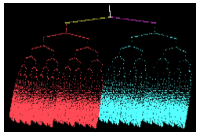


Virtual NIC - User Directory

- One per interface in USER DIRECT file:

```
NICDEF vdev TYPE QDIO
            [LAN SYSTEM switch]
            [DEVICES nn]
            [MACID xxyyzz]

Example:

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

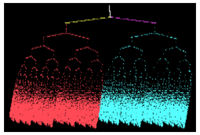
Combined with VMLAN USERPREFIX to create virtual MAC



Virtual NIC - CP Command

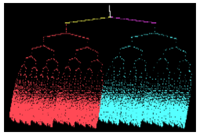
- May be interactive with CP DEFINE NIC and COUPLE commands:

```
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
```



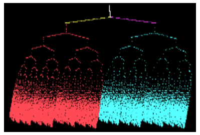
SET NIC

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



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 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

Additional Virtual Switch Technologies

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

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

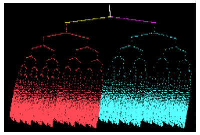
Summary

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

- Use IEEE VLANs to simplify configuration

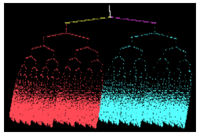
- Use Link Aggregation for best availability

- Integrate into SNMP-based monitoring solutions



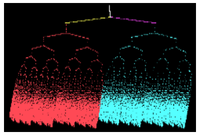
Support Timeline

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 ▪ zEnterprise ensembles (IEDN and INMN) [deprecated] ▪ 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	▪ 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

Tim Greer

z/VM System Test

IBM

1701 North Street
Endicott, NY 13760

Office 607 429 3598

Email: timgreer@us.ibm.com
tim_greer@vnet.ibm.com