V23

Virtual Networking with z/VM Guest LANs and the Virtual Switch

Tracy Adams

IBM System z Expo
September 17-21, 2007
San Antonio, TX

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:

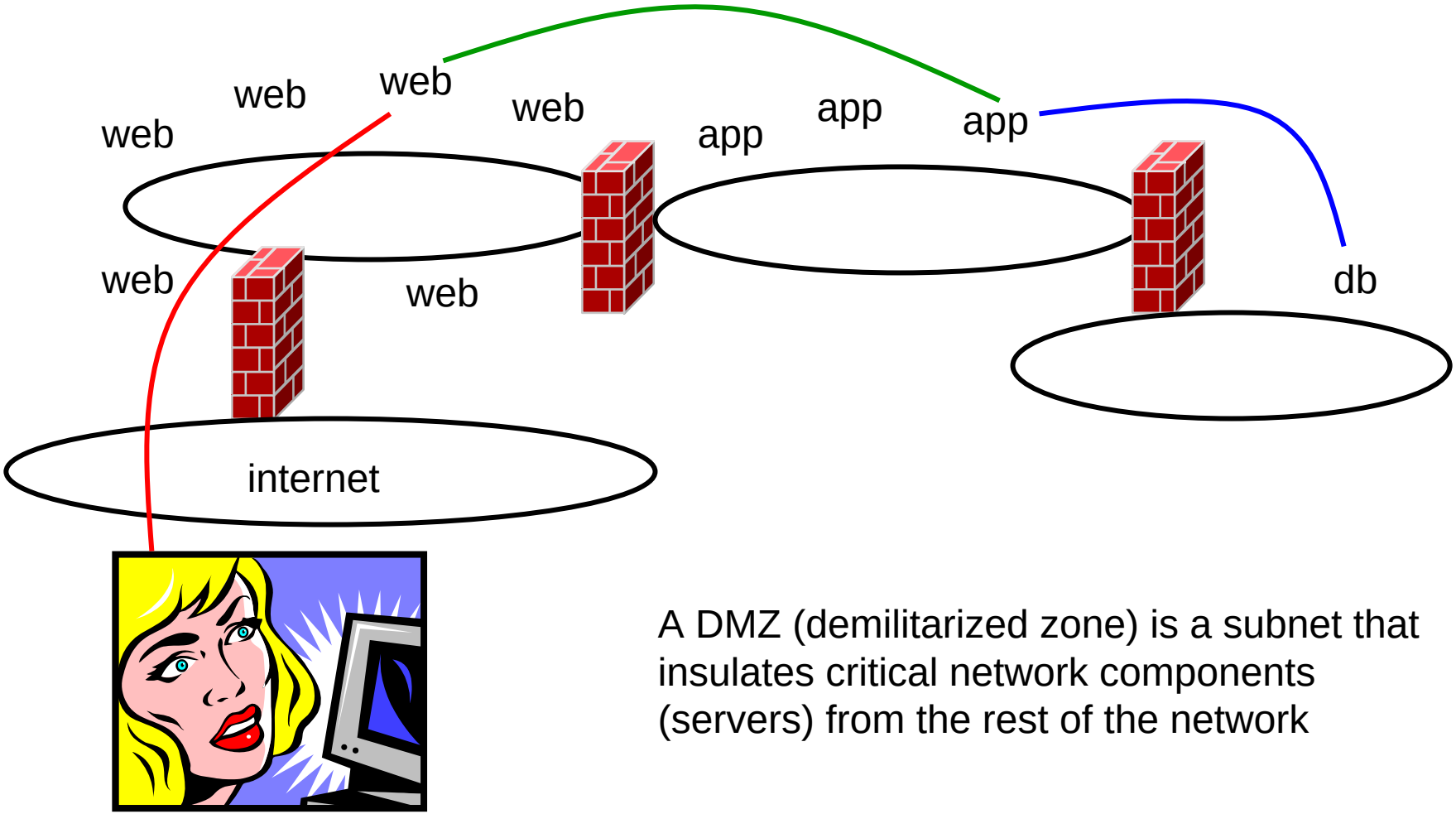
IBM	IBM logo	eServer	zSeries
System z9	DB2	z/OS	z/VM

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

Topics

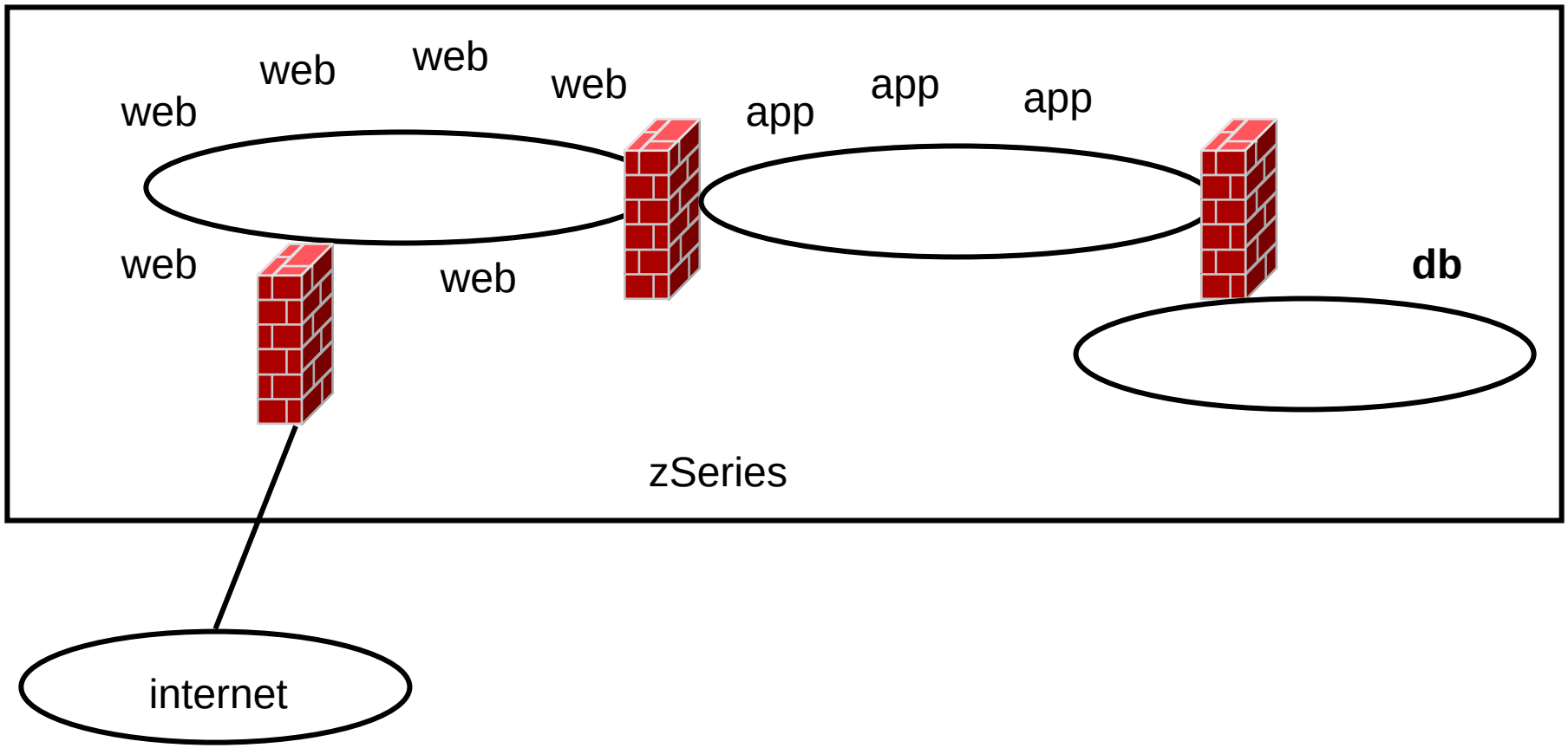
- Overview
- Guest LANs
- Virtual Network Interface Card
- Virtual Switch
- Virtual Switch Failover
- What’s new 5.2 and 5.3

Multi-DMZ Network

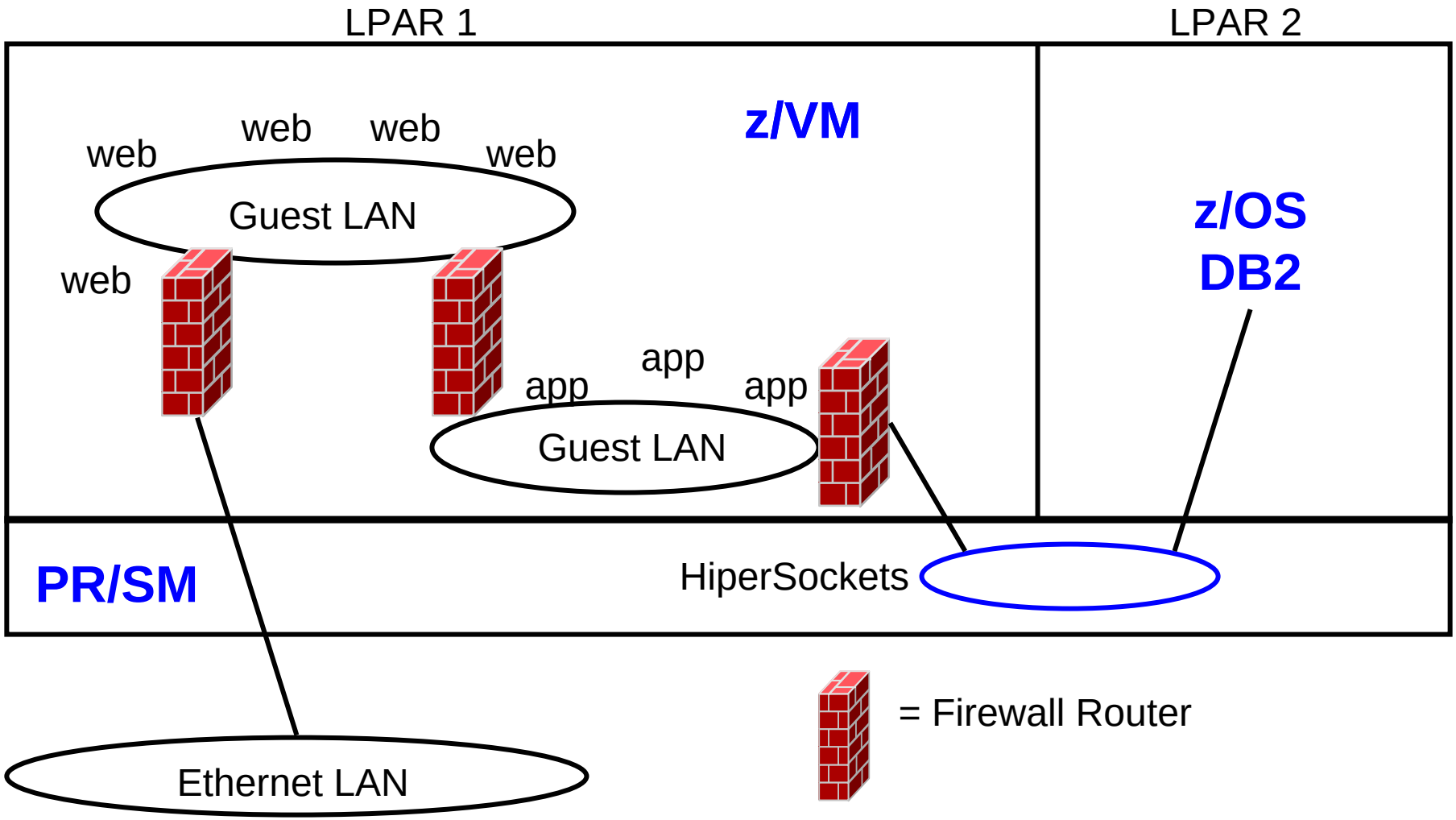


A DMZ (demilitarized zone) is a subnet that insulates critical network components (servers) from the rest of the network

Multi-DMZ Network on zSeries



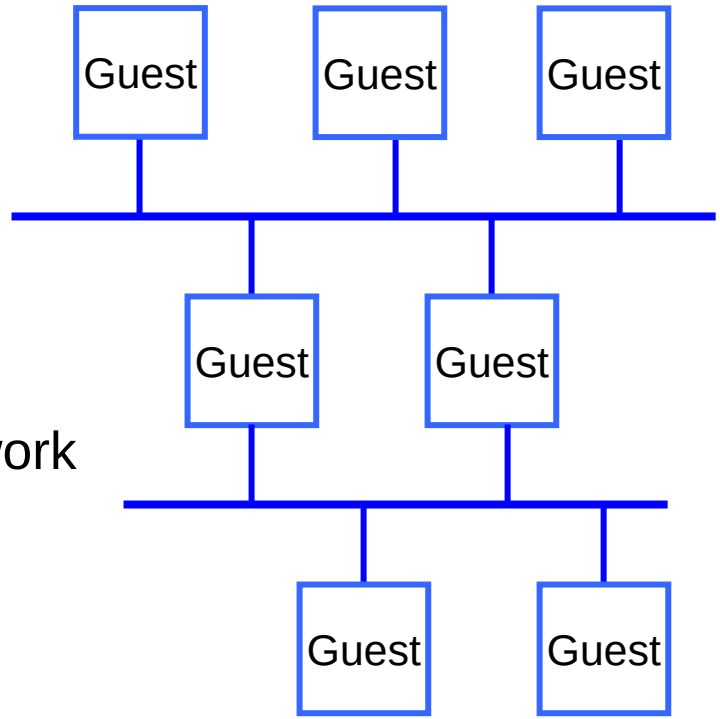
Multi-DMZ Network with Guest LANs



Guest LANs

z/VM Guest LAN

- A simulated LAN segment
  - QDIO: IPv4 and IPv6
  - Ethernet: Lots of protocols
  - HiperSockets: IPv4 and IPv6
  - No built-in connection to outside network
- As many as you want
- Created in SYSTEM CONFIG, directory, or by CP DEFINE LAN command



Primary Guest LAN Attributes

- Name & Owner
- Type
- Access list
- IP/Ethernet (QDIO only)
- Maximum frame size (HiperSockets only)

- Some attributes can be changed after the LAN is defined

- There are some others not discussed here
  - Maximum number of connections
  - Accounting

LAN Name and Owner

- The LAN name is a simple 1-8 character token
- The LAN owner is a VM user ID or “SYSTEM”
- (name, owner) is unique within the system
- A Class G LAN owner can
  - modify the LAN access list
  - delete the LAN
- A Class B user can create, modify, or detach any LAN

HiperSockets or Ethernet

TYPE HIPERsockets | QDIO [IP | ETHERNET]

- HiperSockets
  - Synchronous
  - Low latency
  - Slightly smaller path length in CP (less CPU time)
- QDIO
  - OSA-Express in QDIO mode
  - Asynchronous
  - Higher latency than HiperSockets
  - Higher CPU cost
  - IP = Layer 3, ETHERNET = Layer 2

Access list

- Unrestricted
  - Any user can connect (couple) to this LAN
  - Hint: CP QUERY LAN can show you who is connected
- Restricted
  - Only users in the access list can connect (couple) to this LAN
  - LAN owner uses CP SET LAN to GRANT or REVOKE access
  - CP QUERY LAN can show you the current access list
  - CP QUERY LAN can show you who is connected
- **External Security Manager**
  - RACF/VM support for Guest Lan and Virtual Switch

Persistent vs. Transient LAN

- Persistent / Transient is inferred from other attributes
  - Any LAN owned by user “SYSTEM” is *persistent*
  - Any LAN created by SYSTEM CONFIG is *persistent*
  - All other LANs are *transient*
- A *persistent* LAN must be explicitly deleted by CP DETACH LAN
- A *transient* LAN is automatically deleted when the last user uncouples from the LAN

Setting Guest LAN defaults and limits

- Set global VM LAN attributes in the SYSTEM CONFIG file:

```
VMLAN LIMit PERSistent INFinite|maxcount
VMLAN LIMit TRANSient INFinite|maxcount
VMLAN ACNT|ACCOUNTing SYSTEM ON|OFF
VMLAN ACNT|ACCOUNTing USER ON|OFF
VMLAN MACPREFIX 020000-02FFFF
VMLAN MACIDRANGE SYSTEM x-y [USER a-b]
```

- *Maxcount* of 0 prevents dynamic definition
- SET VMLAN to change dynamically

Virtual MAC Addresses

- Each instance of CP should have a unique VMLAN MACPREFIX
- Virtual MAC = MACPREFIX || MACID
- VMLAN MACIDRANGE
  - SYSTEM – The range of MACIDs from which CP will select a dynamically defined MAC
  - USER – The range of MACIDs reserved by CP for NICDEF. All MACIDs on NICDEFs must be in this range.
  - USER is a subset of SYSTEM

Create a Guest LAN

- DEFINE LAN in SYSTEM CONFIG

```
DEFINE LAN name [OWNERid ownerid]
           [TYPE HIPERsockets|QDIO]
           [MAXCONN INFinite|nnnn]
           [MFS 16K|24K|40K|64K]
           [ACCOUNTing ON|OFF]
           [UNRESTricted|RESTricted]
           [GRANT userlist]

Examples:

DEFINE LAN QDIO5 OWNER SYSTEM TYPE QDIO
```

- CP DEFINE LAN to create dynamically

```
DEFINE LAN NET9 OWNER SYSTEM RESTRICTED TYPE QDIO
```

Grant Guest LAN Access

- DEFINE LAN and MODIFY LAN in SYSTEM CONFIG

```
MODIFY LAN  name
            [OWNERid ownerid | OWNERID SYSTEM]
            [GRANT userid]

Example:

DEFINE LAN HIPER1 OWNER SYSTEM RESTRICTED
MODIFY LAN HIPER1 OWNER SYSTEM GRANT LINUX01
MODIFY LAN HIPER1 OWNER SYSTEM GRANT LINUX02
```

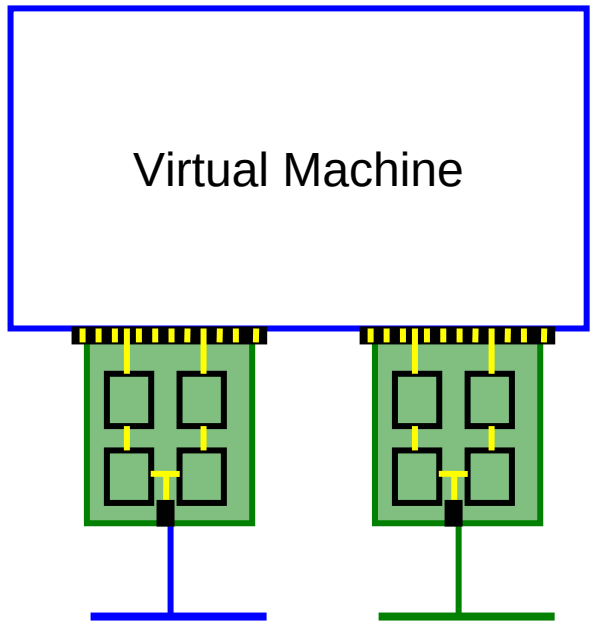
- CP SET LAN to change dynamically

```
CP SET LAN HIPER1 OWNER SYSTEM GRANT LINUX03
```

Virtual Network Interface Card

Virtual Network Interface Card (NIC)

- A simulated network adapter
  - OSA-Express QDIO
  - HiperSockets
  - Must match LAN type
- 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 directory or CP DEFINE NIC command



Virtual NIC - User Directory

- May be automated with USER DIRECT file:

```
NICDEF vdev [TYPE HIPERS | QDIO]
            [DEVices devs]
            [LAN owner name]
            [CHPID xx]
            [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 xx]

CP COUPLE vdev [TO] owner name

Example:

CP DEFINE NIC 1200 TYPE QDIO
CP COUPLE 1200 TO SYSTEM CSC201
```

NIC CHPID parameter

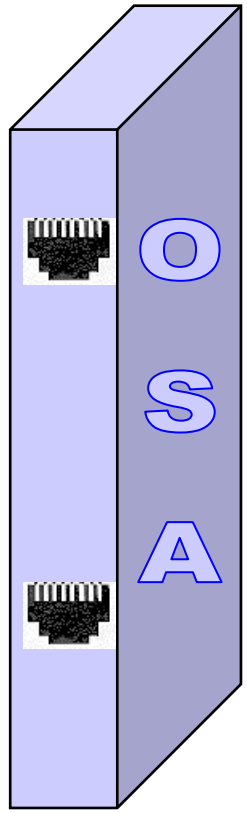
CHPID xx

- Specifies the Channel Path ID number (in hex) to use for this NIC
- Needed for z/OS guest because HiperSockets are managed by CHPID number
- **This is a virtual CHPID number**

Virtual Switch

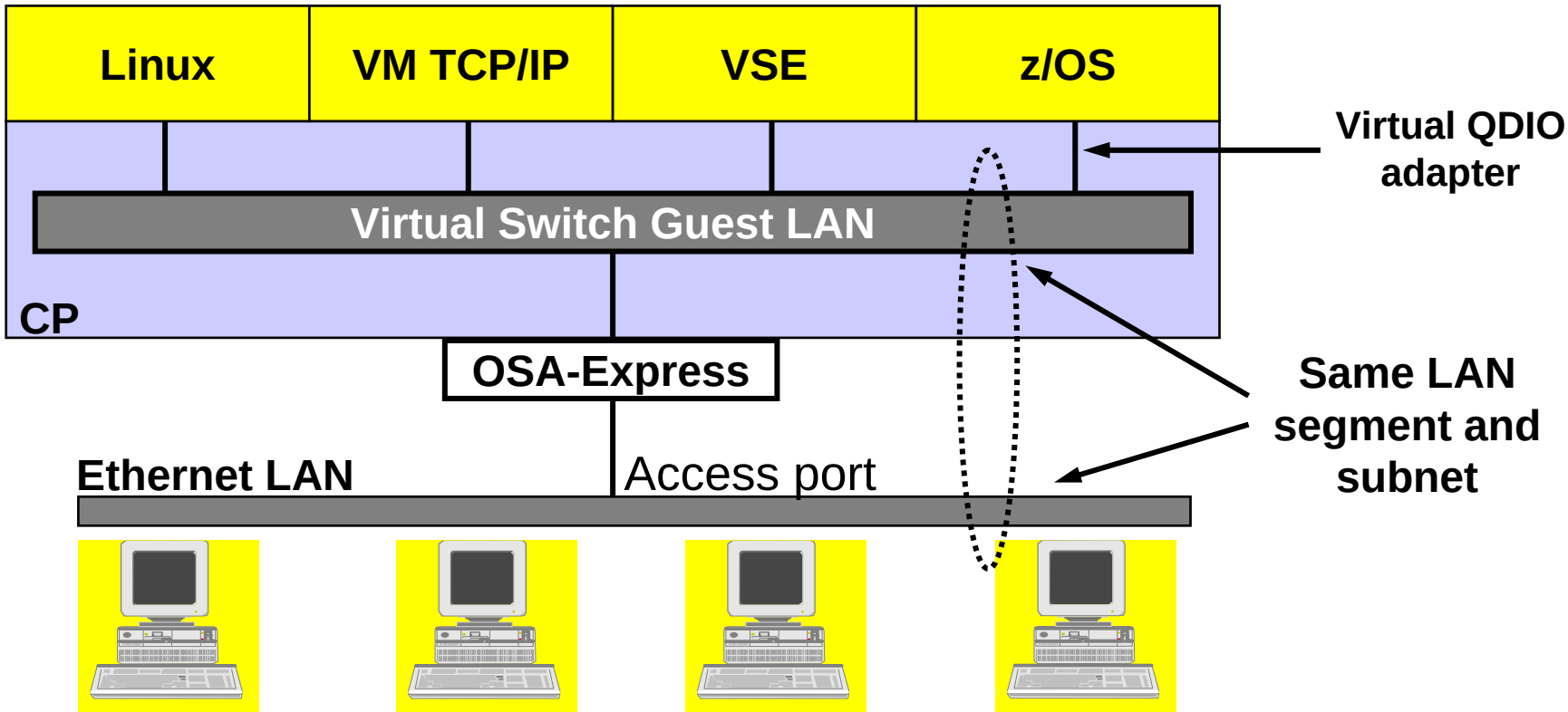
What's a 'switch' anyway?

© Cisco Corp

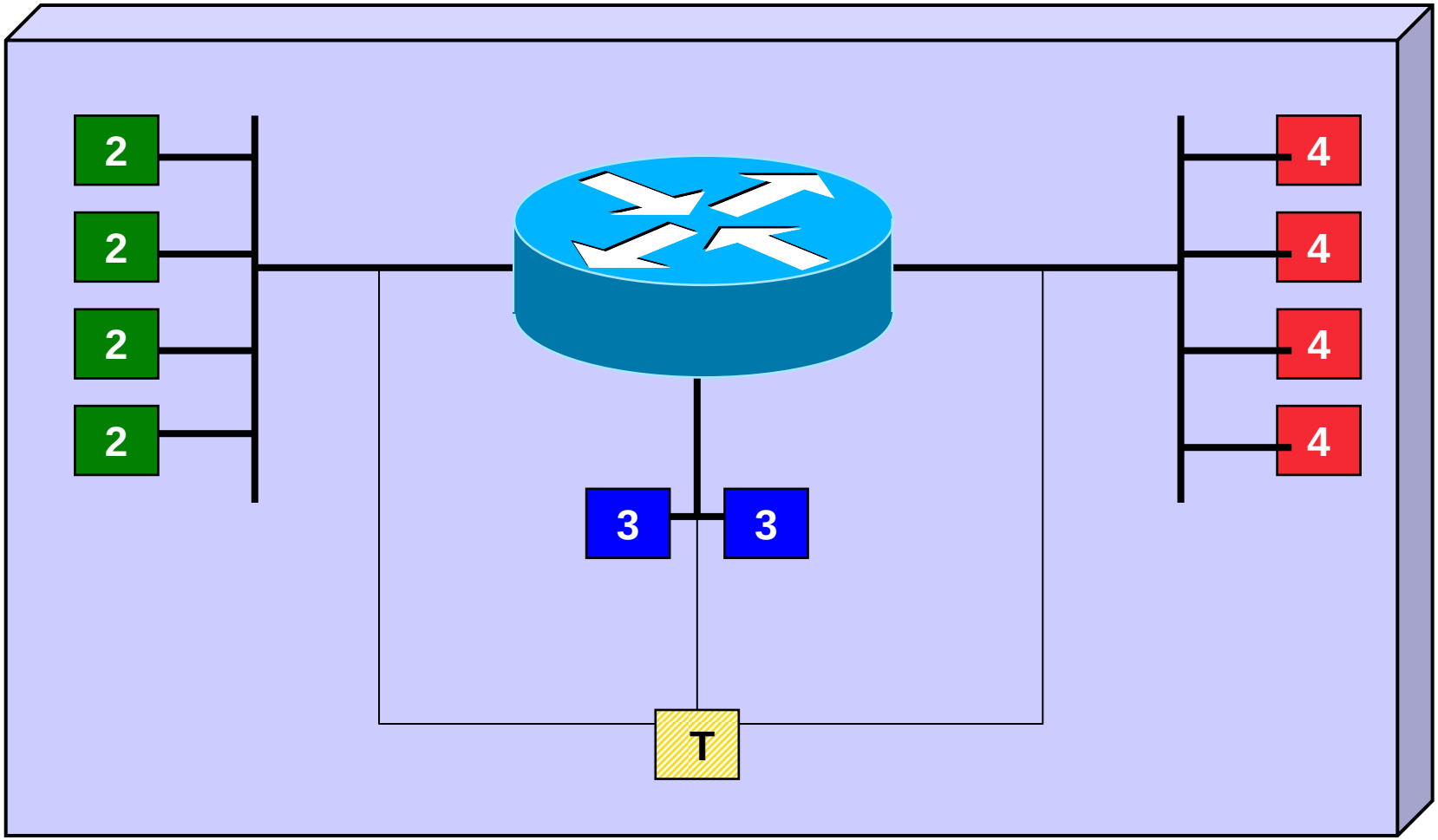


- A box that creates a LAN
- It can be remotely configured
  - E.g. Turn ports on and off
- Similar to a home router

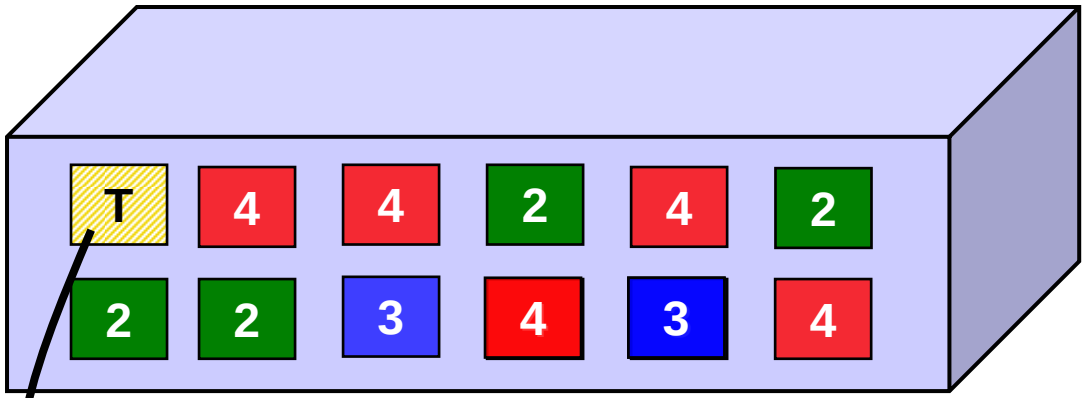
z/VM Virtual Switch – VLAN unaware



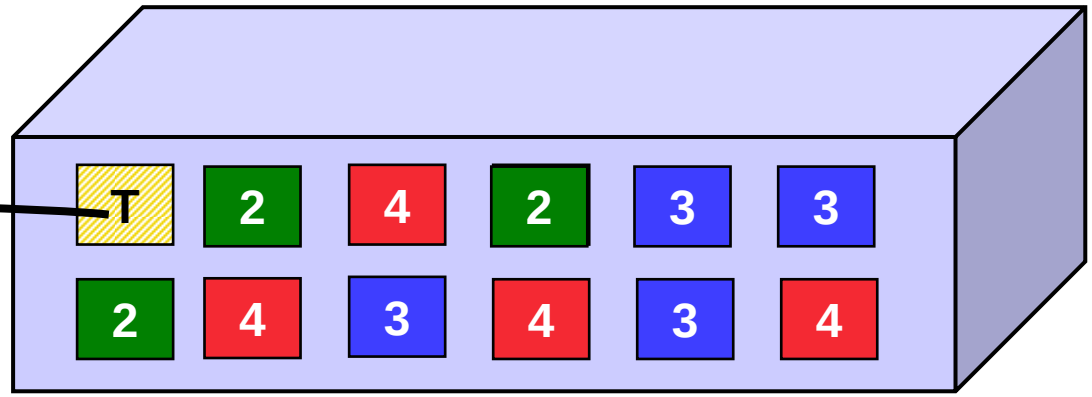
A VLAN-aware switch: An inside look



Trunk Port vs. Access Port

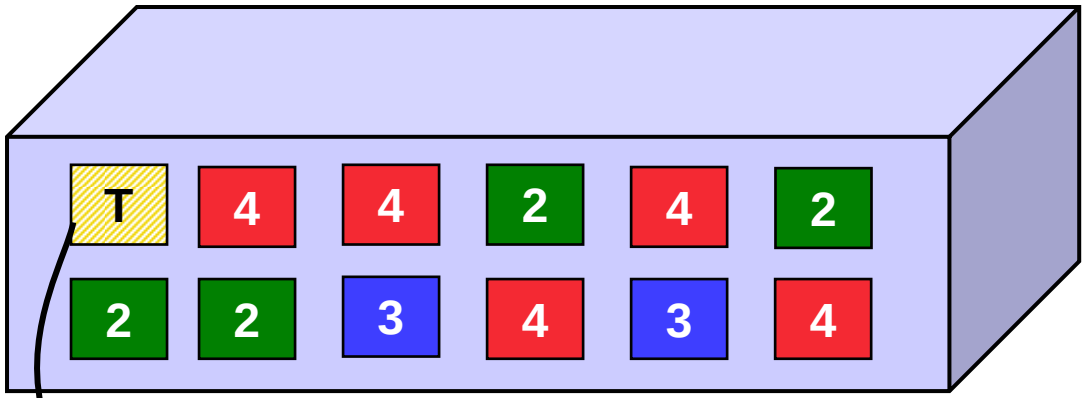


- Access port carries traffic for a single VLAN
- Host not aware of VLANs



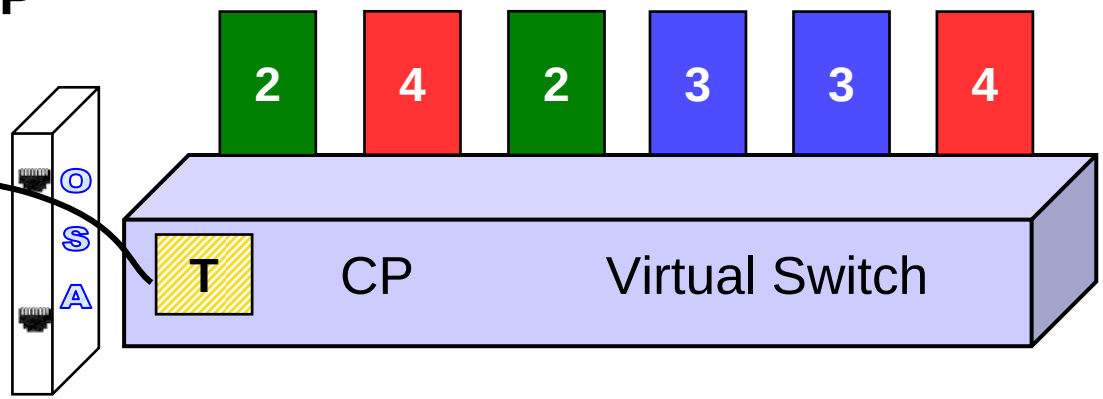
- Trunk port carries traffic from all VLANs
- Every frame is tagged with the VLAN id

Physical Switch to Virtual Switch

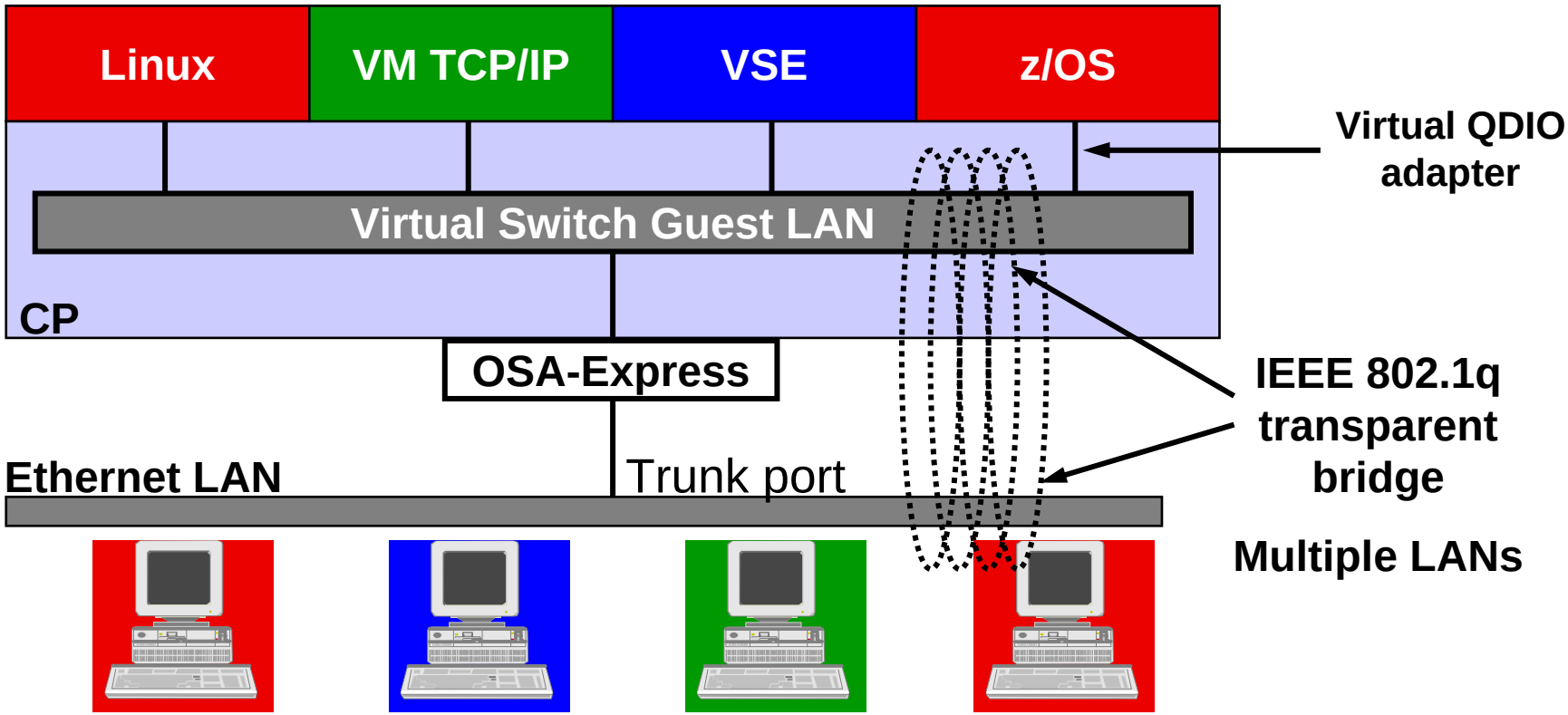


- **Trunk port carries traffic between CP and switch**

- **Each guest can be in a different VLAN**



z/VM Virtual Switch – VLAN aware

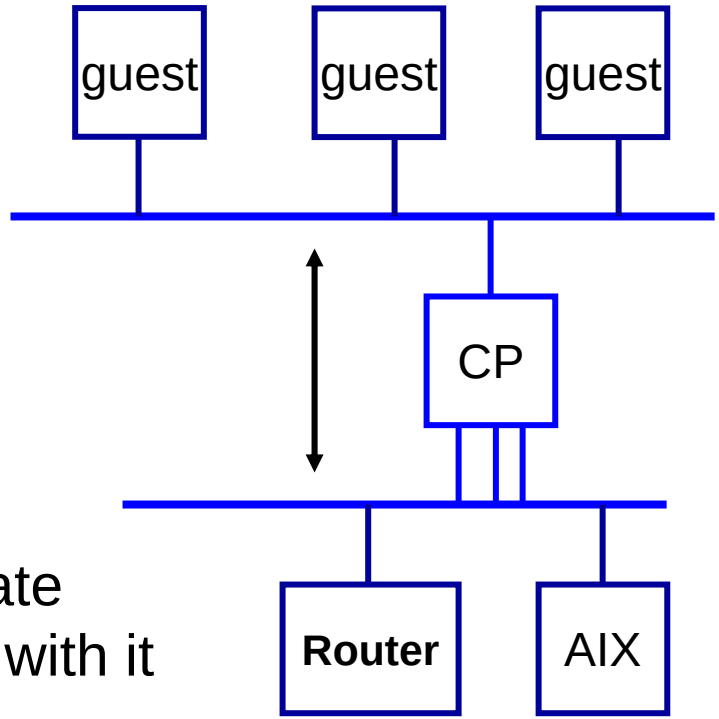


z/VM Virtual Switch

- A special-purpose Guest LAN
  - Ethernet IPv4
  - Built-in IEEE 802.1q bridge to outside network
  - IEEE VLAN capable

z/VM 5.3

- Each Virtual Switch has up to 8 separate OSA-Express connections associated with it
- Created in SYSTEM CONFIG or by CP DEFINE VSWITCH command



Virtual Switch Attributes

- Name
- Associated OSAs
- One or more controlling virtual machines (minimal VM TCP/IP stack servers)
  - Controller not involved in data transfer
  - Do not ATTACH or DEDICATE
  - User pre-configured DTCVSW1 and DTCVSW2
- Similar to Guest LAN
  - Owner SYSTEM
  - Type QDIO
  - Persistent
  - Restricted

Create a Virtual Switch

- SYSTEM CONFIG or CP command:

```
DEFINE VSWITCH name
              [RDEV NONE | cuu [cuu [cuu]] ]
              [CONNECT | DISCONNECT]
              [CONTROLLER * | userid]
              [IP IPTIMEOUT 5 NONROUTER | ETHERNET]
              [NOGroup | GROup groupname]

              [VLAN UNAWARE | VLAN native_vid]
              [PORTTYPE ACCESS | PORTTYPE TRUNK]

Example:

DEFINE VSWITCH SWITCH12 RDEV 1E00 1F04 CONNECT
```

z/VM 5.3

Change the Virtual Switch access list

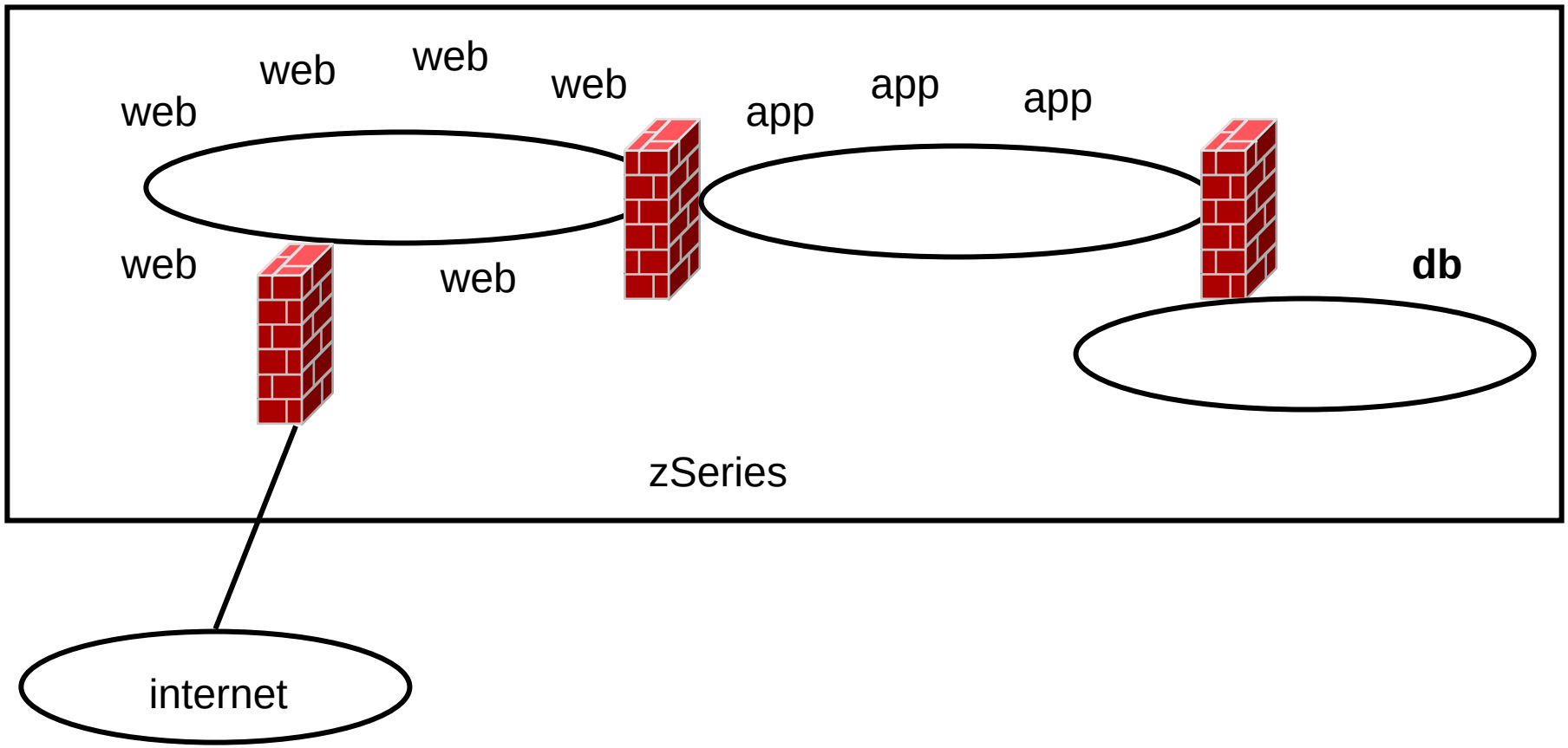
- Specify after DEFINE VSWITCH statement in SYSTEM CONFIG to add users to access list

```
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 7 105
CP SET VSWITCH SWITCH12 GRANT LNX02 PORTTYPE TRUNK
                                    VLAN 4-20 22-29


CP SET VSWITCH SWITCH12 GRANT LNX03 PRO
```

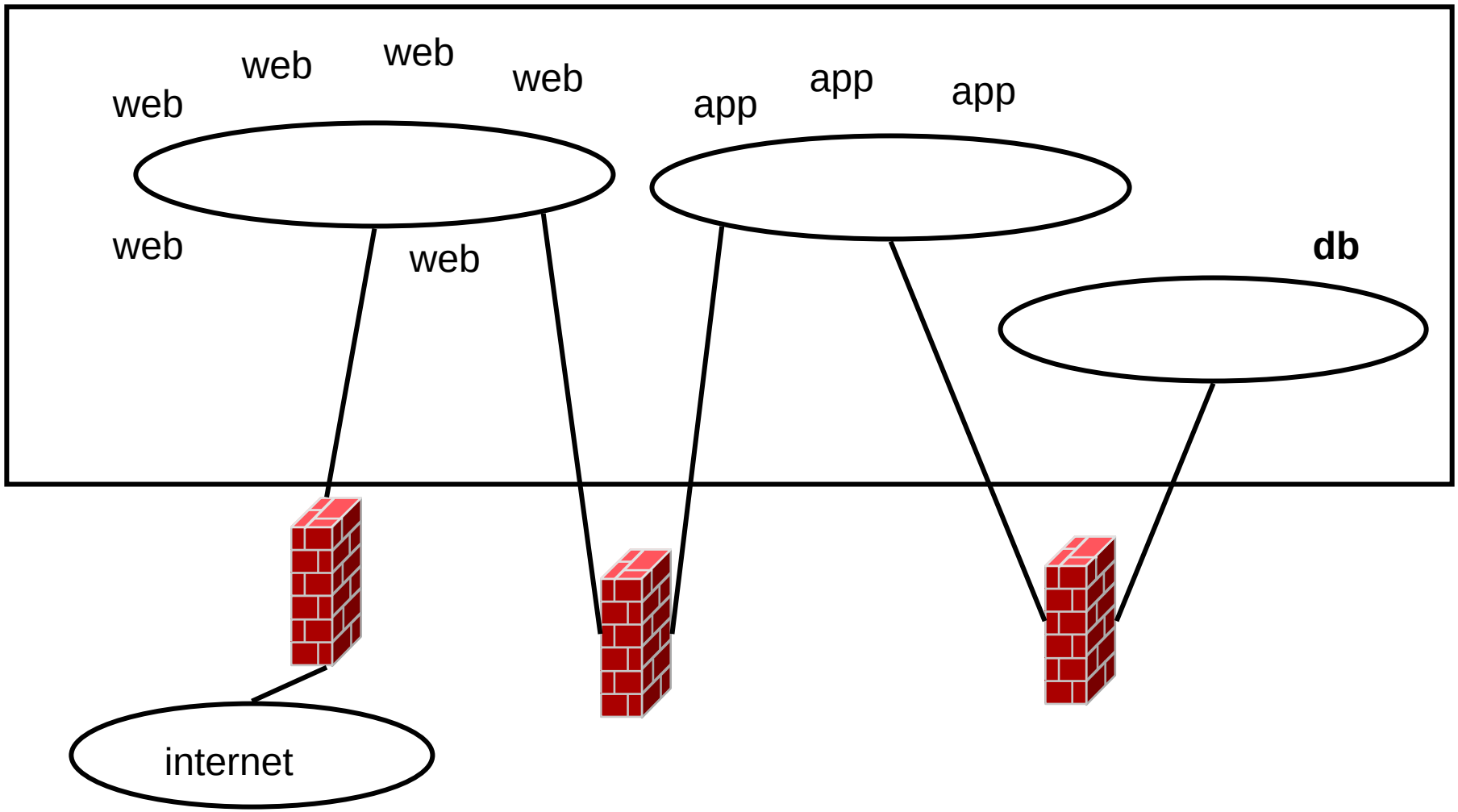
z/VM 5.2

- z/VM 4.4 supported “VLAN ANY”, but it’s removed in z/VM5.1!

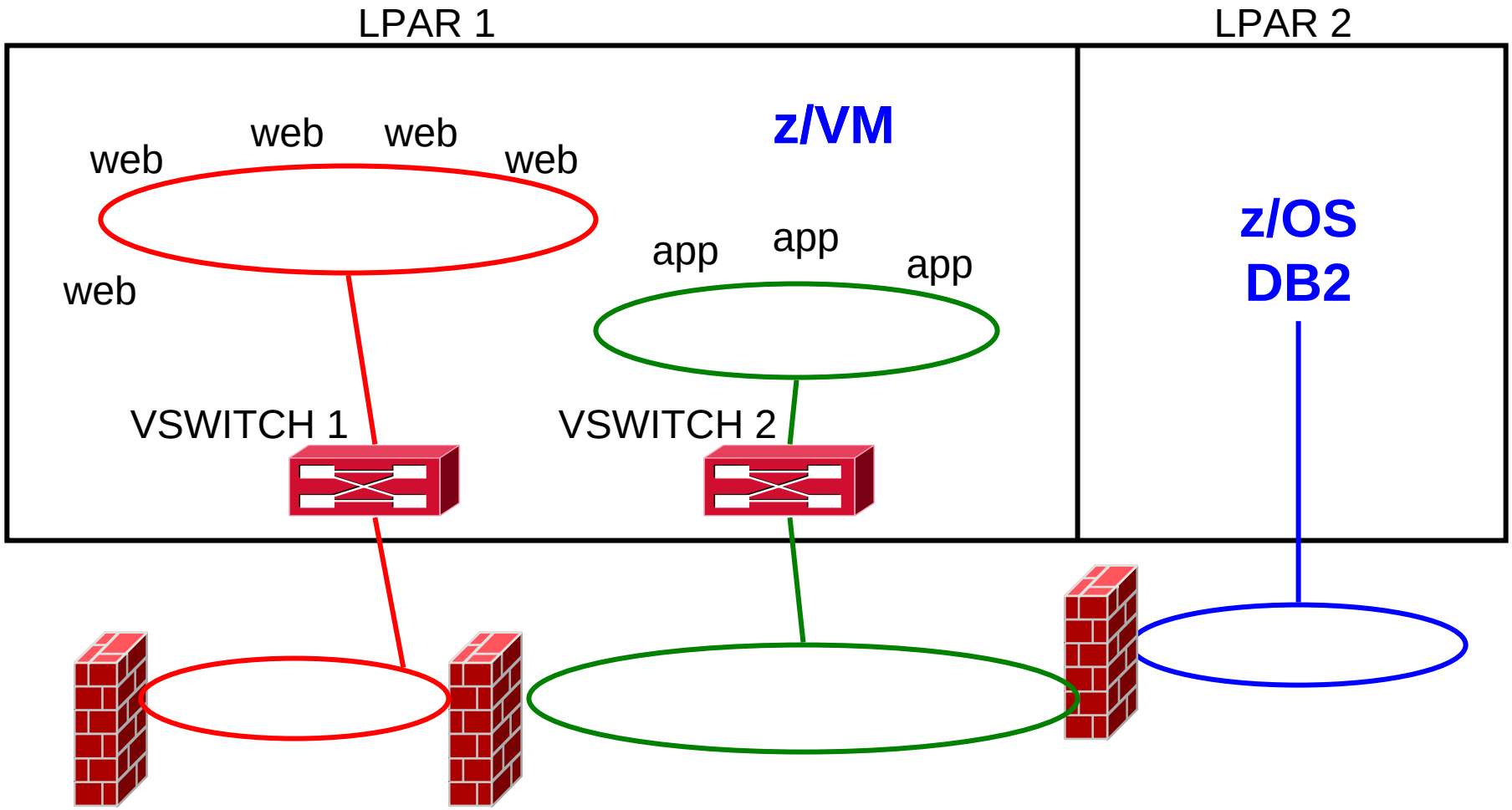
Multi-DMZ Network on zSeries - Reloaded



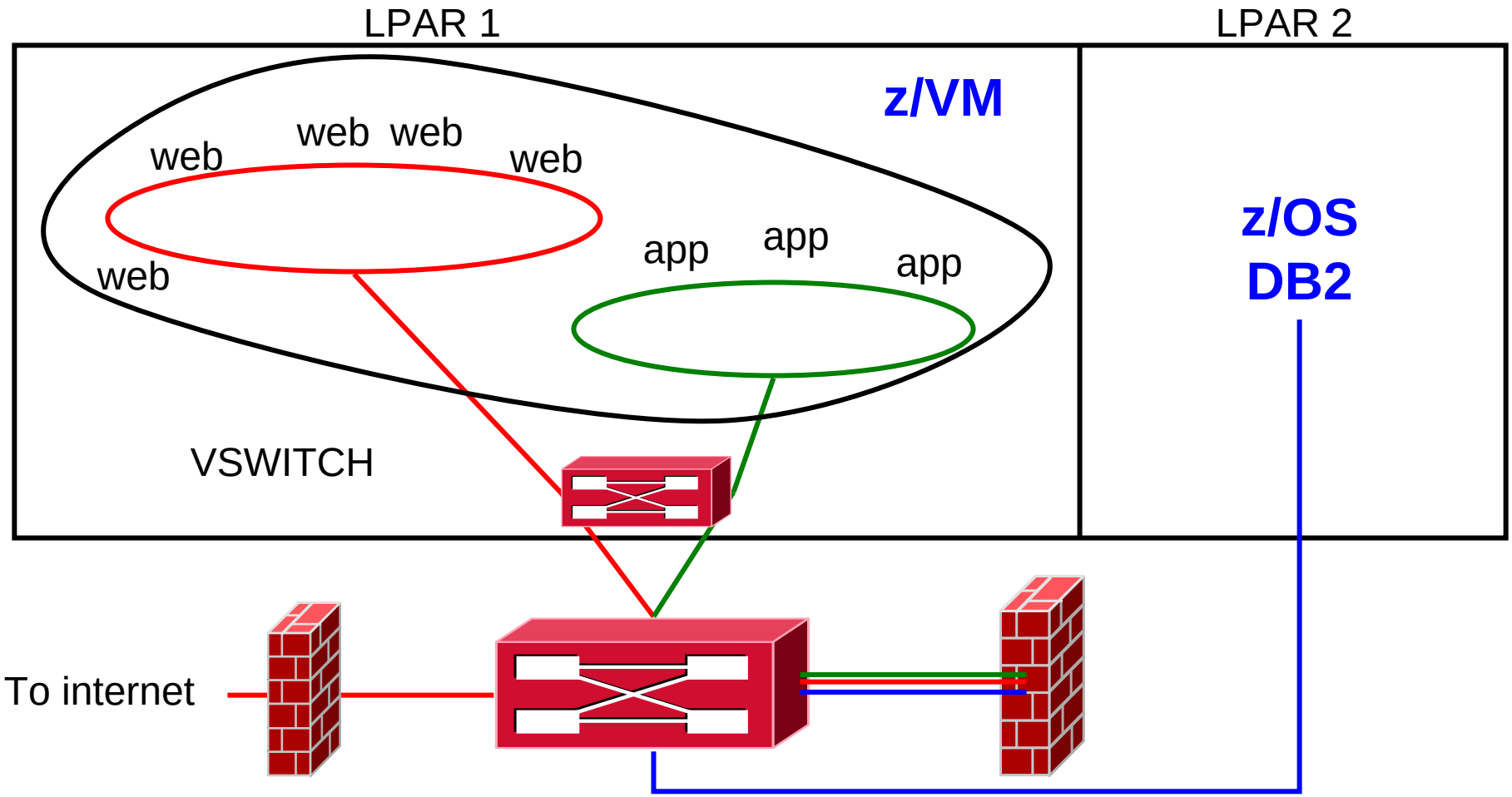
Multi-DMZ Network on zSeries with outboard firewall



Multi-DMZ Network with VSWITCH (A)



Multi-DMZ Network with VSWITCH (B)



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

VSWITCH: An Example

SYSTEM CONFIG

```
/*********************************************************/
/*                  VSWITCH CONFIG                       */
/*********************************************************/
DEFINE VSWITCH VSWTCH1 RDEV C004 C100
MODIFY VSWITCH VSWTCH1 GRANT LINUX001
MODIFY VSWITCH VSWTCH1 GRANT LINUX002
MODIFY VSWITCH VSWTCH1 GRANT LINUX003
```

In z/VM V5.1 you can use ESM to control access to a Guest LAN or VSWITCH!

RACF/VM

- RDEFINE VMLAN SYSTEM.VSWTCH1 UACC(NONE)
- PERMIT SYSTEM.VSWTCH1 CLASS(VMLAN)
  ID(LINUX002 LINUX003 LINUX004)
  ACCESS(UPDATE)
- VMLAN class must be active and COUPLE.G command must be controlled

AUTOLOG1 PROFILE EXEC

```
/*****************************/
/*  Autolog1 Profile Exec  */
/*****************************/
ADDRESS COMMAND CP XAUTOLOG PERFSVM
ADDRESS COMMAND CP XAUTOLOG VMRTM
ADDRESS COMMAND CP AUTOLOG VMSERVS VMSERVS
ADDRESS COMMAND CP AUTOLOG VMSERVU VMSERVU
ADDRESS COMMAND CP AUTOLOG VMSERVR VMSERVR
ADDRESS COMMAND CP AUTOLOG TCPIP TCPIP
ADDRESS COMMAND CP SLEEP 5 SEC
ADDRESS COMMAND CP XAUTOLOG DTCVSW1
ADDRESS COMMAND CP XAUTOLOG DTCVSW2
```

Linux directory entry

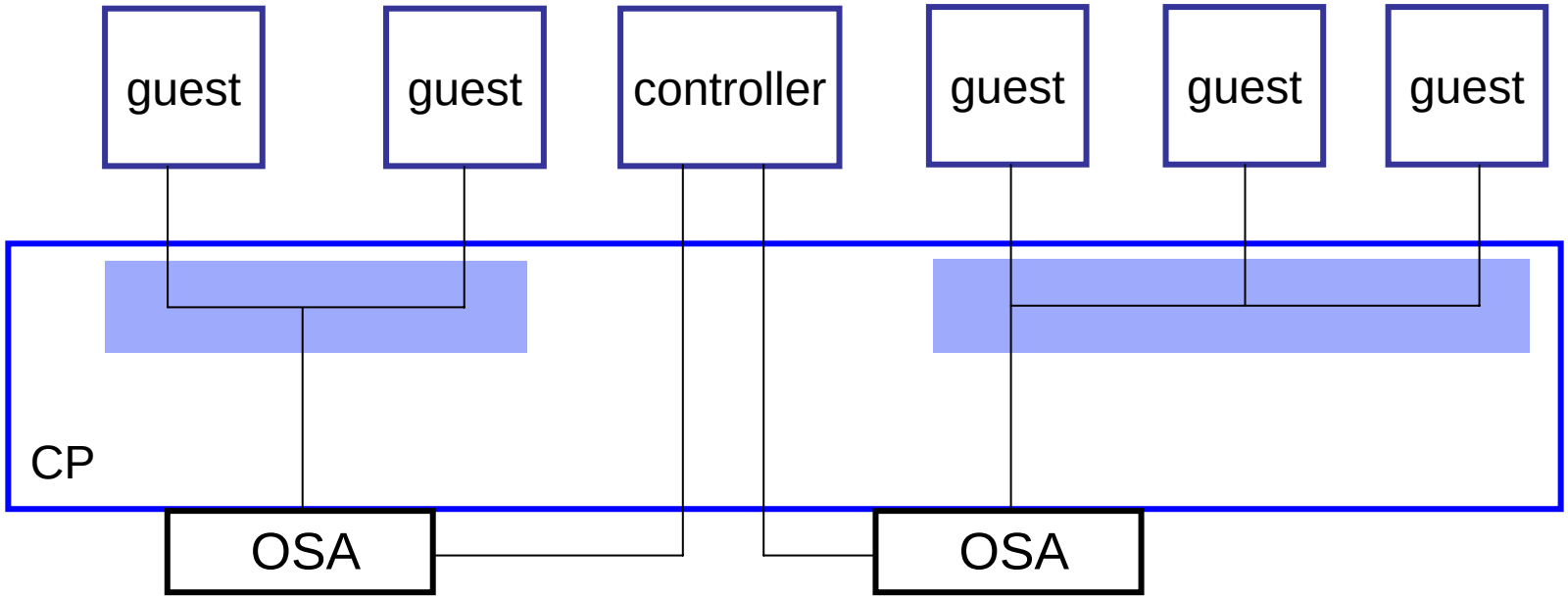
```
*************************************************************
*        LINUX002 - SLES8 SP2                               *
*           using VSWITCH  VSWTCH1                          *
*************************************************************
USER LINUX002 XXXXXXXX 128M 2048M G
INCLUDE LINDFLT
NICDEF C204 TYPE QDIO DEVICES 3 LAN SYSTEM VSWTCH1
 MDISK 191 3390 0001 0010 V2LX11 MR
 MDISK 200 3390 0011 0100 V2LX11 MR
 MDISK 201 3390 0111 3228 V2LX11 MR
 MDISK 202 3390 0001 3338 V2LX10 MR
 MDISK 203 FB-512 V-DISK 8000  WV
```

Virtual Switch Failover

New in z/VM 5.2.0

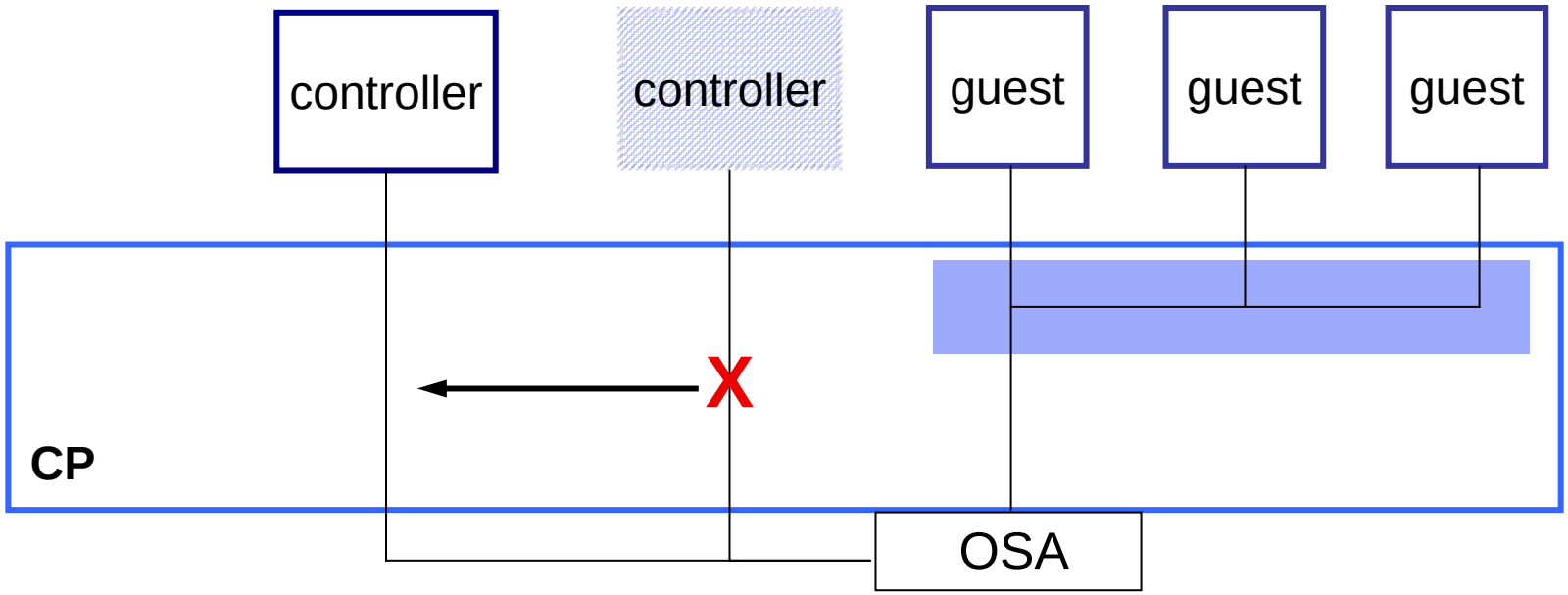
- Pre-defined VSWITCH controllers
  - DTCVSW1 and DTCVSW2
  - Same as shown in Getting Started with Linux
    - Add them to AUTOLOG1
    - Remove “VSWITCH CONTROLLER ON” from PROFILE TCPIP in your production stacks

VSWITCH Controller



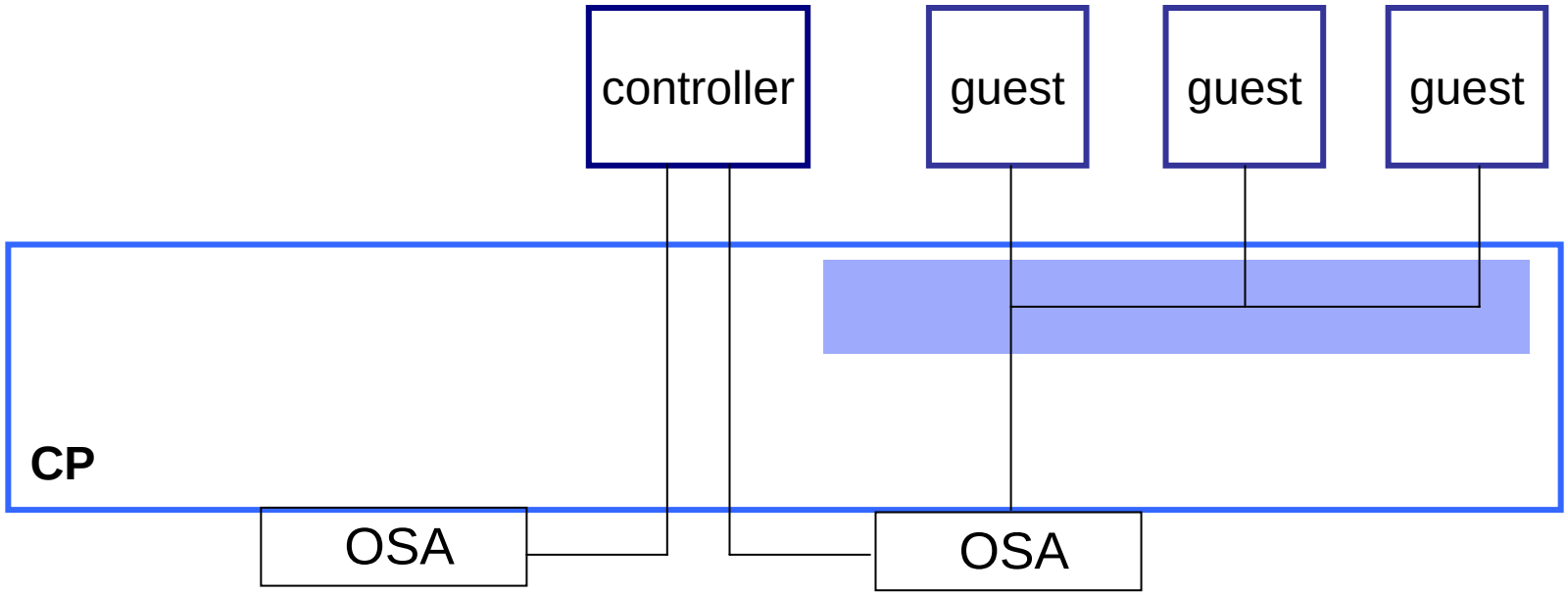
- **A controller is a VM TCP/IP stack, but it doesn't have to be your production stack. Use a predefined one.**
- **Not involved in data transfer; only handles OSA housekeeping.**

VSWITCH Controller Failover



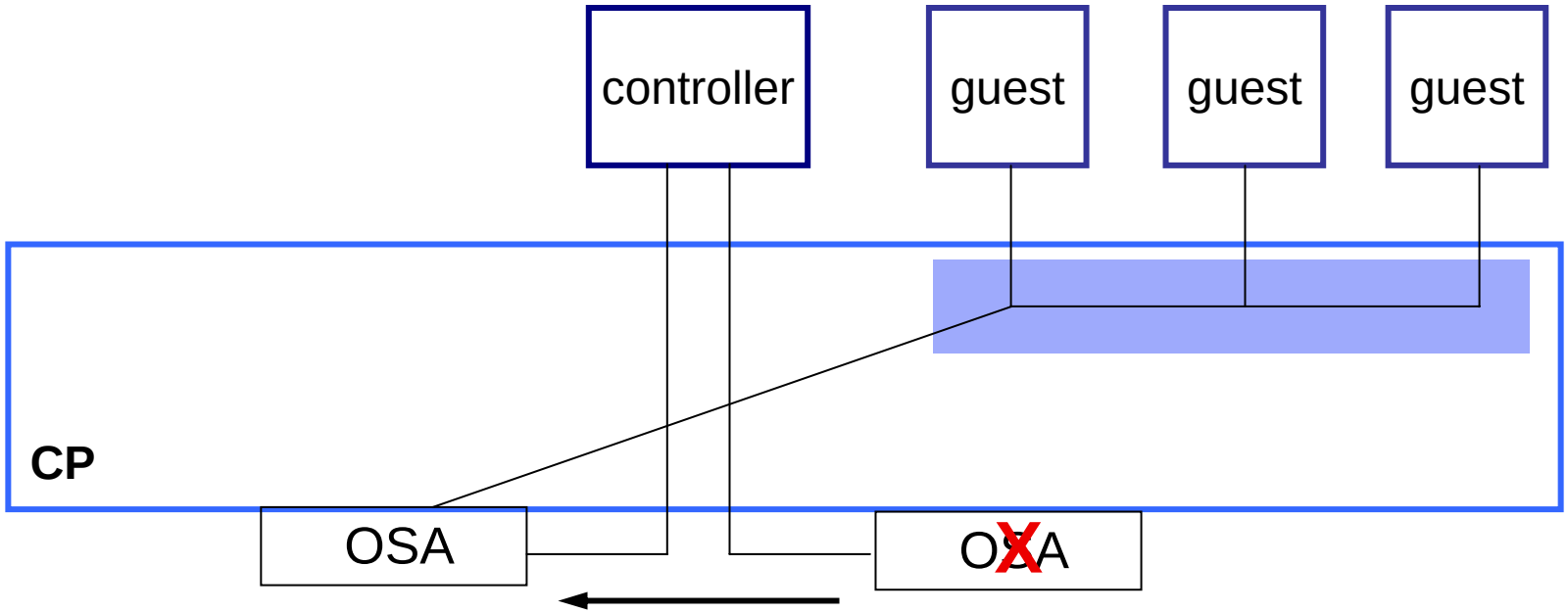
- **In case a controller fails or is forced off, CP will find another, if available.**
- **A VSWITCH can be limited to a specific controller, but is not recommended.**
- **If no controller, VSWITCH external connection is deactivated.**

OSA Failover



- **Up to 8 OSAs per VSWITCH**
- **Automatic failover**

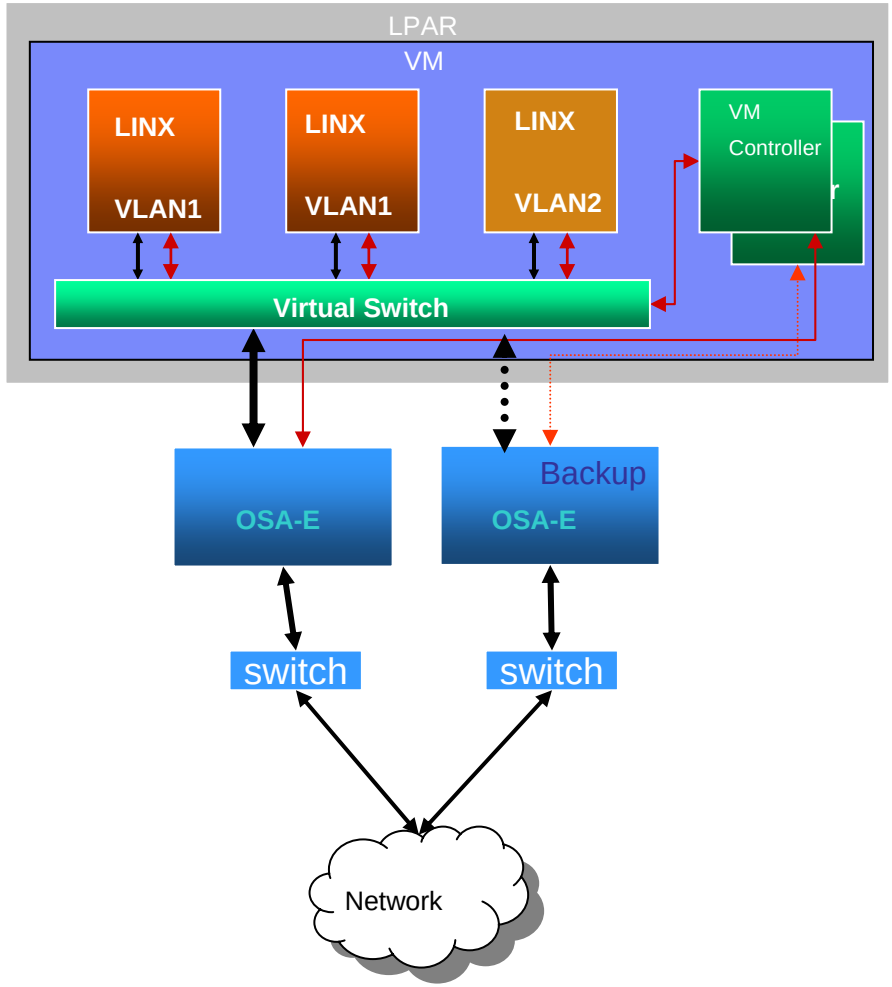
OSA Failover



- **If OSA dies or stalls, controller will detect it and switch to backup OSA**

Recovery from an OSA adapter, switch or Controller failure

- Upon detection of an OSA failure another OSA adapter takes over and data transfer is resumed.
- Upon detection of a Controller outage or non-responsive control connection another Controller takes over the control connection and data transfer is resumed.
- Port to port transfers are unaffected by trunk failures
- Dependent on planned redundancy for physical network connections and VM Controller allocation.
- Layer 2 is non-disruptive to network



Initial state

```
q osa
OSA  C004 ATTACHED TO DTCVSW1  C004
OSA  C005 ATTACHED TO DTCVSW1  C005
OSA  C006 ATTACHED TO DTCVSW1  C006
OSA  C100 ATTACHED TO DTCVSW1  C100
OSA  C101 ATTACHED TO DTCVSW1  C101
OSA  C102 ATTACHED TO DTCVSW1  C102
OSA  C20C ATTACHED TO TCPIP    C20C
OSA  C20D ATTACHED TO TCPIP    C20D
OSA  C20E ATTACHED TO TCPIP    C20E

q controller
Controller DTCVSW1   Available: YES   VDEV Range: *          Level 510
  Capability: IP ETHERNET VLAN_ARP
   SYSTEM VSWTCH1    Primary          Controller: *          VDEV: C004
   SYSTEM VSWTCH1    Backup           Controller: *          VDEV: C100
Controller DTCVSW2   Available: YES   VDEV Range: *          Level 510
  Capability: IP ETHERNET VLAN_ARP

q vswitch
VSWITCH SYSTEM VSWTCH1  Type: VSWITCH Connected: 1    Maxconn: INFINITE
  PERSISTENT  RESTRICTED    NONROUTER                 Accounting: OFF
  VLAN Unaware
  State: Ready
  IPTimeout: 5          QueueStorage: 8
  Portname: UNASSIGNED RDEV: C004 Controller: DTCVSW1  VDEV:  C004
  Portname: UNASSIGNED RDEV: C100 Controller: DTCVSW1  VDEV:  C100 BACKUP
```

Simulate failures

- Controller failure
  - FORCE a controller off the system
- OSA failure
  - Configure the OSA offline from the HMC

FORCE DTCVSW1

```
force DTCVSW1
USER DSC   LOGOFF AS  DTCVSW1  USERS = 15    FORCED BY MAINT

HCPSWU2843E The path was severed for TCP/IP Controller DTCVSW1.
HCPSWU2843E It was managing device C004 for VSWITCH SYSTEM VSWTCH1.

HCPSWU2830I VSWITCH SYSTEM VSWTCH1 status is in error recovery.
HCPSWU2830I DTCVSW2 is new VSWITCH controller.

HCPSWU2830I VSWITCH SYSTEM VSWTCH1 status is ready.
HCPSWU2830I DTCVSW2 is VSWITCH controller.

q controller
Controller DTCVSW2   Available: YES   VDEV Range: *         Level 510
  Capability: IP ETHERNET VLAN_ARP
   SYSTEM VSWTCH1    Primary         Controller: *         VDEV: C004
   SYSTEM VSWTCH1    Backup          Controller: *         VDEV: C100
```

Configure OSA offline

```
HCPSWU2830I VSWITCH SYSTEM VSWTCH1 status is devices attached.
HCPSWU2830I DTCVSW2 is VSWITCH controller.

HCPSWU2830I VSWITCH SYSTEM VSWTCH1 status is in error recovery.
HCPSWU2830I DTCVSW2 is new VSWITCH controller.

HCPSWU2845W Backup device C004 specified for VSWITCH VSWTCH1 is
  not initialized.

HCPSWU2830I VSWITCH SYSTEM VSWTCH1 status is ready.
HCPSWU2830I DTCVSW2 is VSWITCH controller.
```

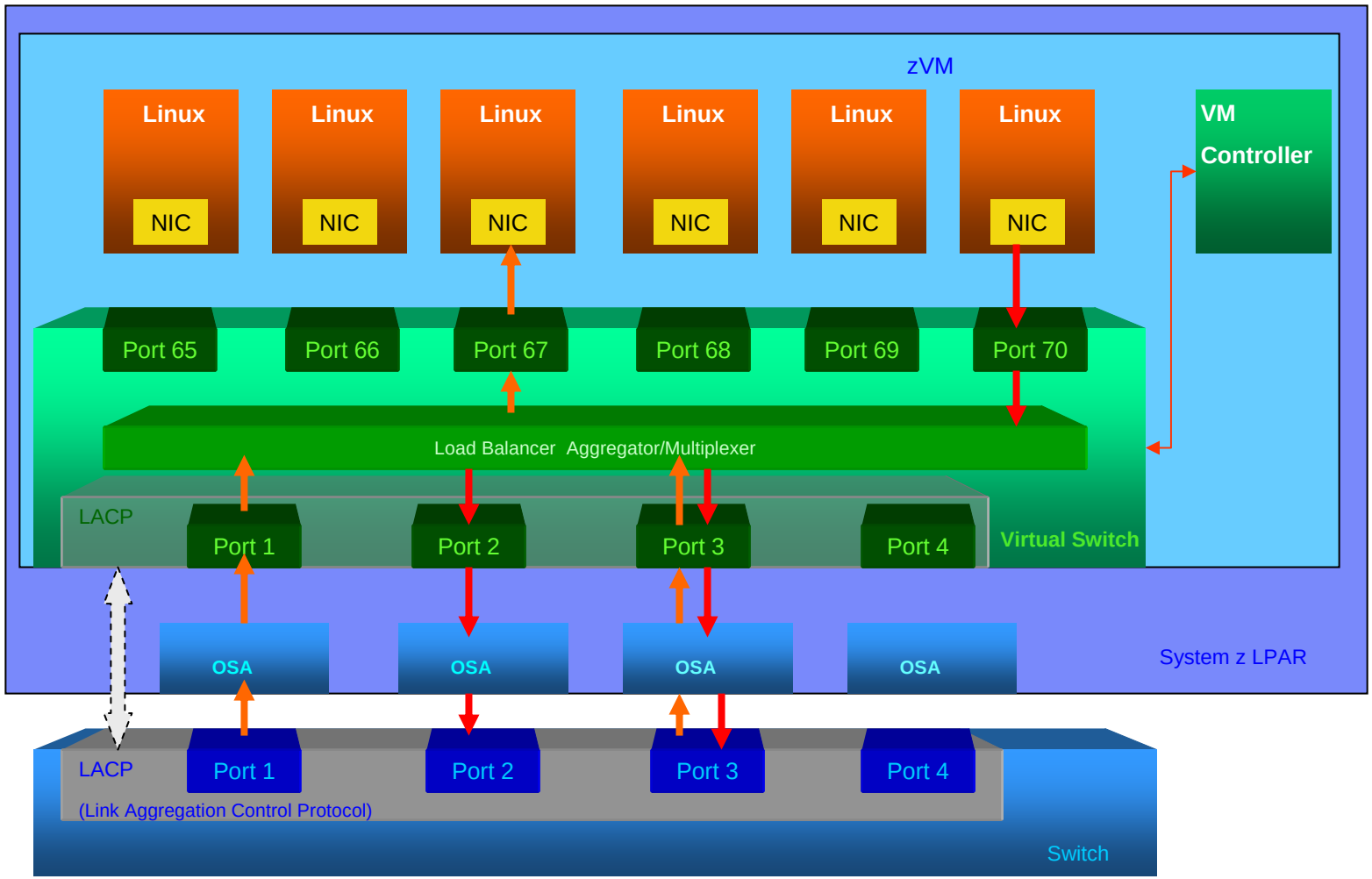
Configure OSA offline

```
q vswitch
VSWITCH SYSTEM VSWTCH1  Type: VSWITCH Connected: 1    Maxconn: INFINITE
  PERSISTENT  RESTRICTED    NONROUTER                 Accounting: OFF
  VLAN Unaware
  State: Ready
  IPTimeout: 5          QueueStorage: 8
  Portname: UNASSIGNED RDEV: C004 Controller: DTCVSW2  Error: No RDEV
  Portname: UNASSIGNED RDEV: C100 Controller: DTCVSW2  VDEV:  C100
```

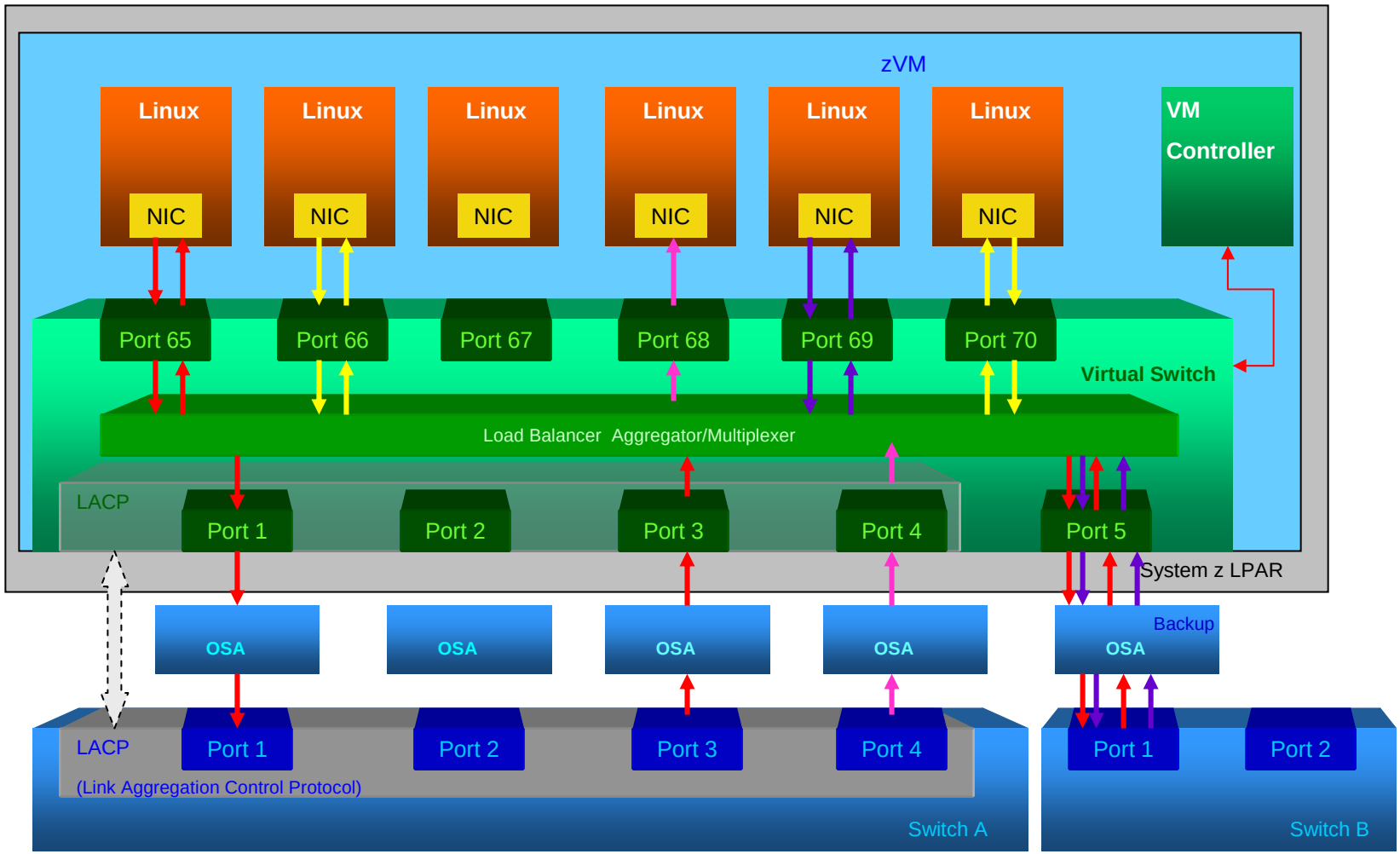
Failover in z/VM 5.3 with Link Aggregation

- New 802.3ad Link Aggregation Support
  - New GROUP option for VSWITCH
    - Multiple OSAs per group
    - Multiple Controllers per VSWITCH
  - Non-disruptive failover
    - Communications will continue if a hardware link in the group experiences a non-recoverable failure.
  - Can manually take a link up or down
  - Learn more at the z/VM Link Aggregation presentation on Thursday (V?)

Recovery of a failed link in a Link Aggregation Configuration



Recovery of a failed switch in a Link Aggregation Configuration



What's new?

New in z/VM 5.3.0 …

- **Usability enhancements**
  - **Dynamic authorization**
  - **Native VLAN**
  - **New Monitor Domain – Virtual Networking Domain 8**
- **Virtual Switch Management (SNMP)**
- **IEEE 802.3ad - Link Aggregation**
  - **Hint: Link Aggregation presentation on Thursday (V26)**

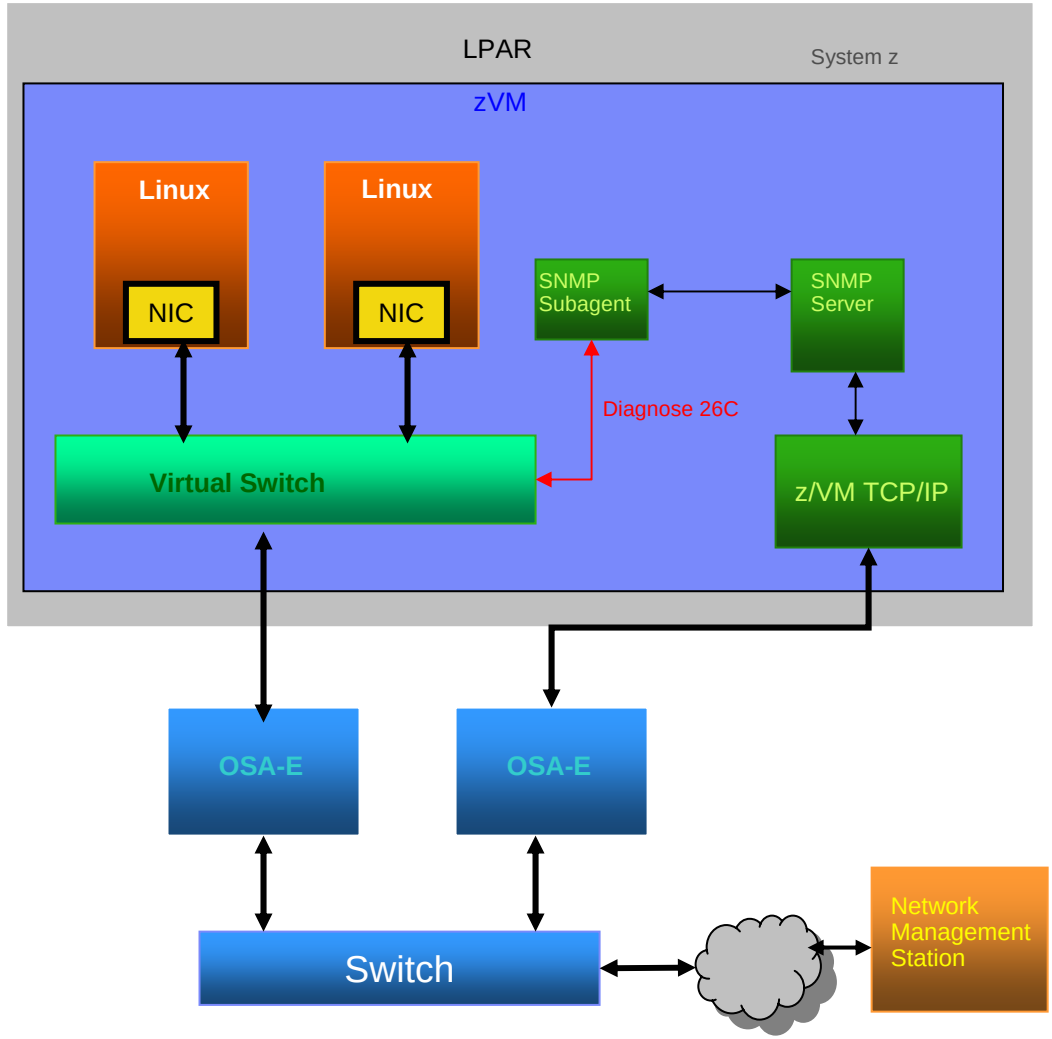
Virtual Switch Management (SNMP)

- **Enhanced Diag 26C API provides QUERY NIC, QUERY VSWITCH, QUERY LAN, QUERY CONTROLLER, QUERY VMLAN equivalent information. Used by SNMP for MIB construction**

•Virtual Switch's IP address is serviced by the z/VM TCP/IP stack.
• SNMP subagent is a TCP/IP application that serves Virtual Switch dot1dBridge MIB (RFC1493).

•Support GET and GETNEXT SNMP commands.

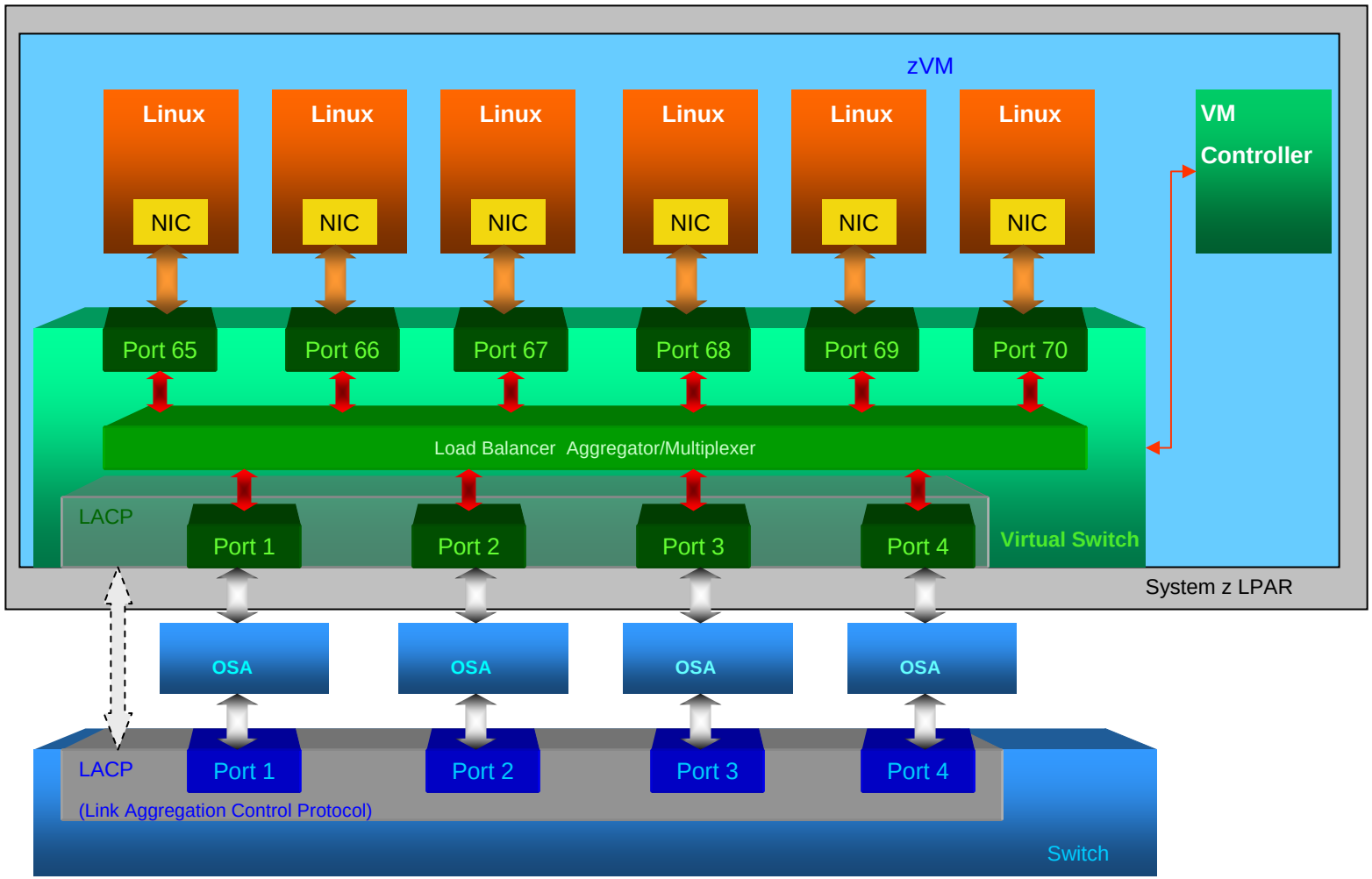
•SNMP TRAP notifications;
- **VSWITCH guest ports up/down transitions**
- **VSWITCH OSA-E port up/down transitions**



Virtual Switch – Link Aggregation

- IEEE 802.3ad compliant including support of active LACP (Link Aggregation Control Protocol (switch to switch only)
  - No support for aggregation of virtual NICs.
- Deploy up to 8 OSA adapters.
- OSA Adapters that are part of the aggregated group are not sharable with other hosts on z/VM or LPAR.
- Non-disruptive failover
  - Communications will continue if a hardware link in the group experiences a non-recoverable failure.
- Improved bandwidth over link aggregate group
- Workload balanced across aggregated links

VSWITCH Link Aggregation Support



z/VM 5.2 Post-GA Support

- Hipersockets IPv6 support (VM63850)
- VSWITCH GRVP support (VM63784)
  - GARP (Generic Attribute Registration Protocol) VLAN Registration Protocol
  - Provides VLAN pruning in conjunction with Physical Switch
  - VLAN Aware only

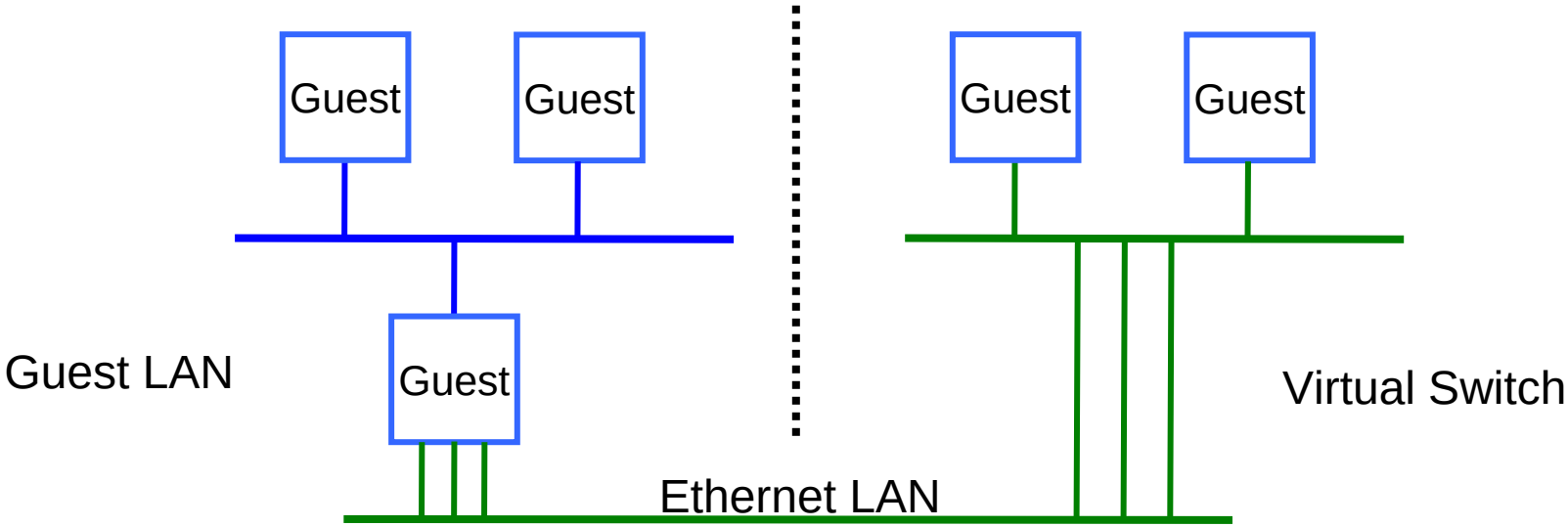
New in z/VM 5.2…

- Support for LAN Sniffers
  - CP command or device driver control (“promiscuous mode”)
    - SET VSWITCH GRANT, SET LAN GRANT, SET NIC
  - External security manager
    - RACF/VM CONTROL access to VMLAN profile
  - Guest receives copies of all frames sent or received

- Pre-defined VSWITCH controllers
  - DTCVSW1 and DTCVSW2
  - Same as shown in Getting Started with Linux
    - Add them to AUTOLOG1
    - Remove “VSWITCH CONTROLLER ON” from PROFILE TCPIP in your production stacks

Some Final Thoughts...

Guest LAN vs. Virtual Switch



- Virtual router is required
- Different subnet
- External router awareness
- Guest-managed failover

- No virtual router
- Same subnet
- Transparent bridge
- CP-managed failover

Network Configuration

- In general, configure a Guest LAN network like any other network
  - Subnet routing
- Use the VSWITCH whenever possible
  - Exploit IEEE VLAN if you can
- By having virtual and real configurations be the same, you can easily test network configuration before deployment with real hardware

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 LAN 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
- **Diagnose x'26C'**
  - provides API for this info (subcode x'08' = Q VMLAN, x'18' = Q LAN, x'24' = Q NIC)

Support Summary

z/VM 5.3	▪Usability Enhancements ▪Virtual Switch Management (SNMP) ▪Link Aggregation ▪API to retrieve virtual networking information (Diagnose x'26C')
Post z/VM 5.2	▪Hipersockets IPv6 Support ▪GVRP Support
z/VM V5.2	▪Virtual SPAN ports for sniffers

References

- Publications:
  - z/VM CP Planning and Administration
  - z/VM CP Command and Utility Reference
  - z/VM TCP/IP Planning and Customization
  - z/VM Connectivity Planning, Administration and Operation
- Links:
  - http://www.ibm.com/servers/eserver/zseries/os/linux/
  - http://www.linuxvm.org/
  - http://www.vm.ibm.com/virtualnetwork/

Contact Information

- By e-mail: bolinda@us.ibm.com
- In person: USA 607.429.5469
- Mailing lists: IBMVM@listserv.uark.edu
  LINUX-390@vm.marist.edu

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

Thanks for Listening!