IBM TRAINING

Session V26

High Availability and Automatic Network Failover of the z/VM VSWITCH

Tracy Adams

IBM
SYSTEM z9 AND zSERIES EXPO
October 9 - 13, 2006

Orlando, FL

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 z/VM

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

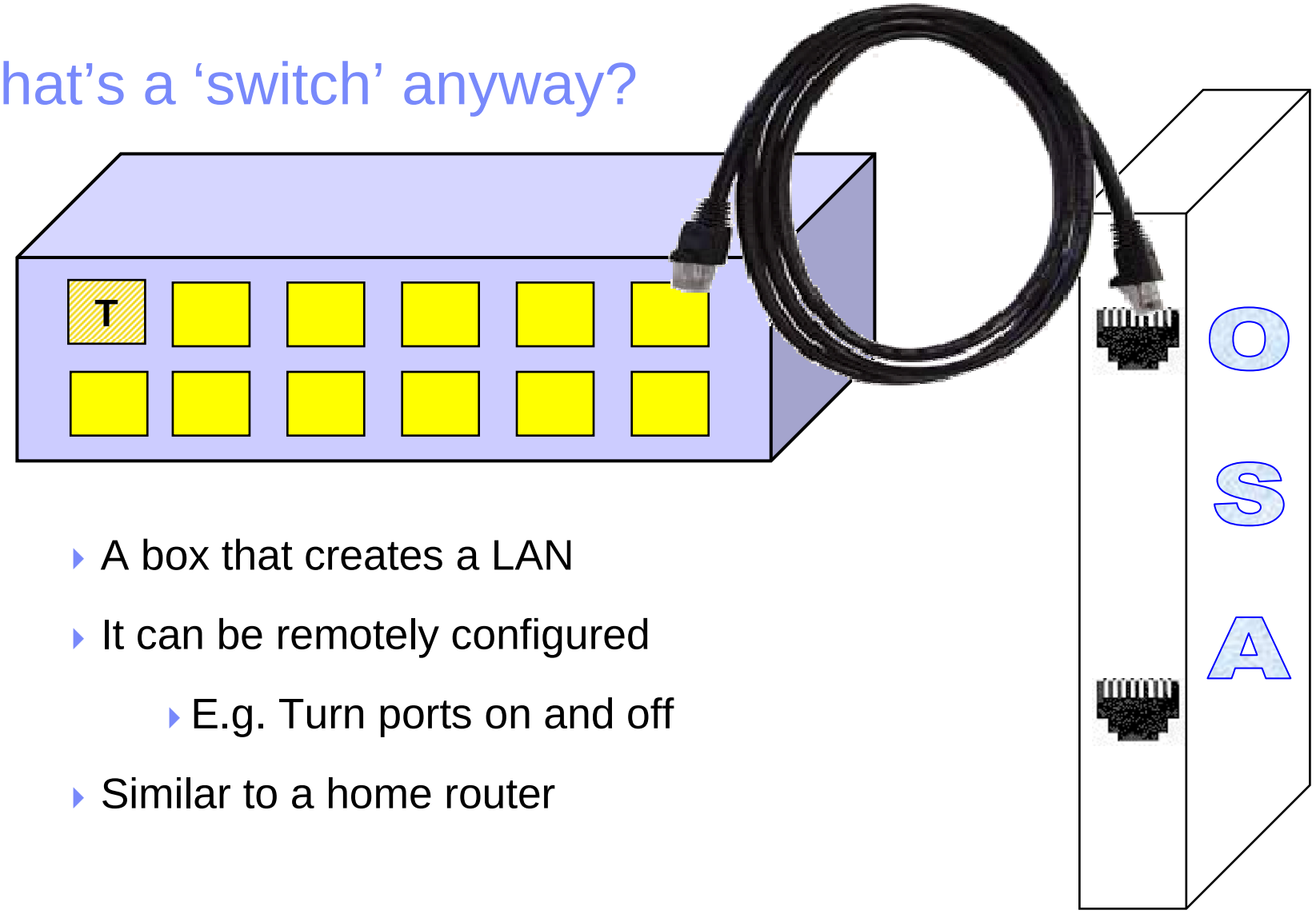
Topics

- What is a Virtual Switch?
- How do I set it up for High Availability and Automatic Failover?
  - System Configuration file or command
  - Controllers
  - Authorize user access
- Testing High Availability & Automatic Failover
  - How can I test this stuff is really going to work

z/VM Virtual Switch

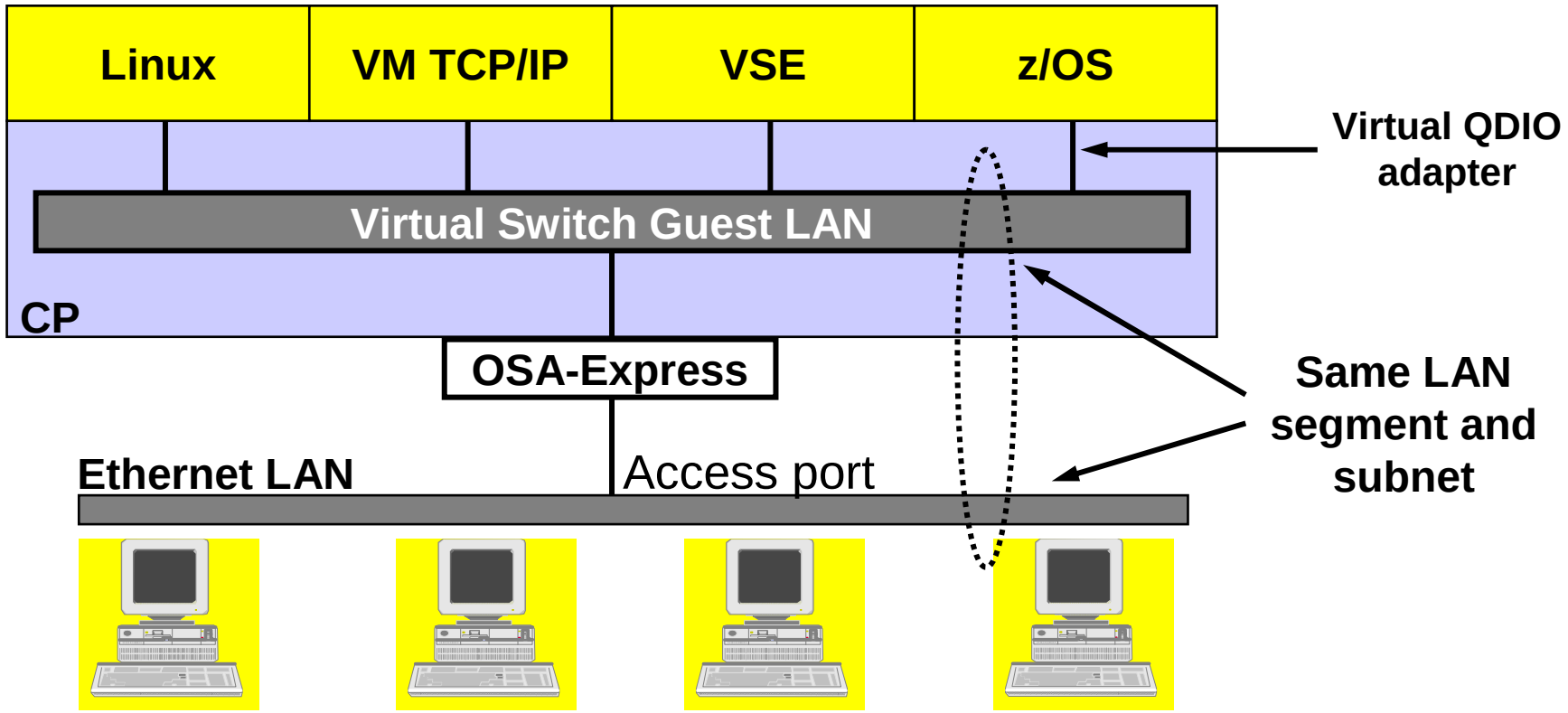
- The role of the VSWITCH is to provide external network connectivity through an OSA-Express device for a Guest LAN
- Using a VSWITCH reduces the CPU utilization cost and latency associated with providing external connectivity through a router virtual machine
- The switching logic resides in the z/VM Control Program (CP) which owns the OSA-Express connection and performs all data transfers between Guest LAN nodes and the OSA-Express
- In most configurations, primary router (PRIROUTER) OSA-Express connections are no longer required
  - To receive inbound packets for all guests behind real device
  - Permits further sharing of the OSA-Express among virtual switches

What's a 'switch' anyway?

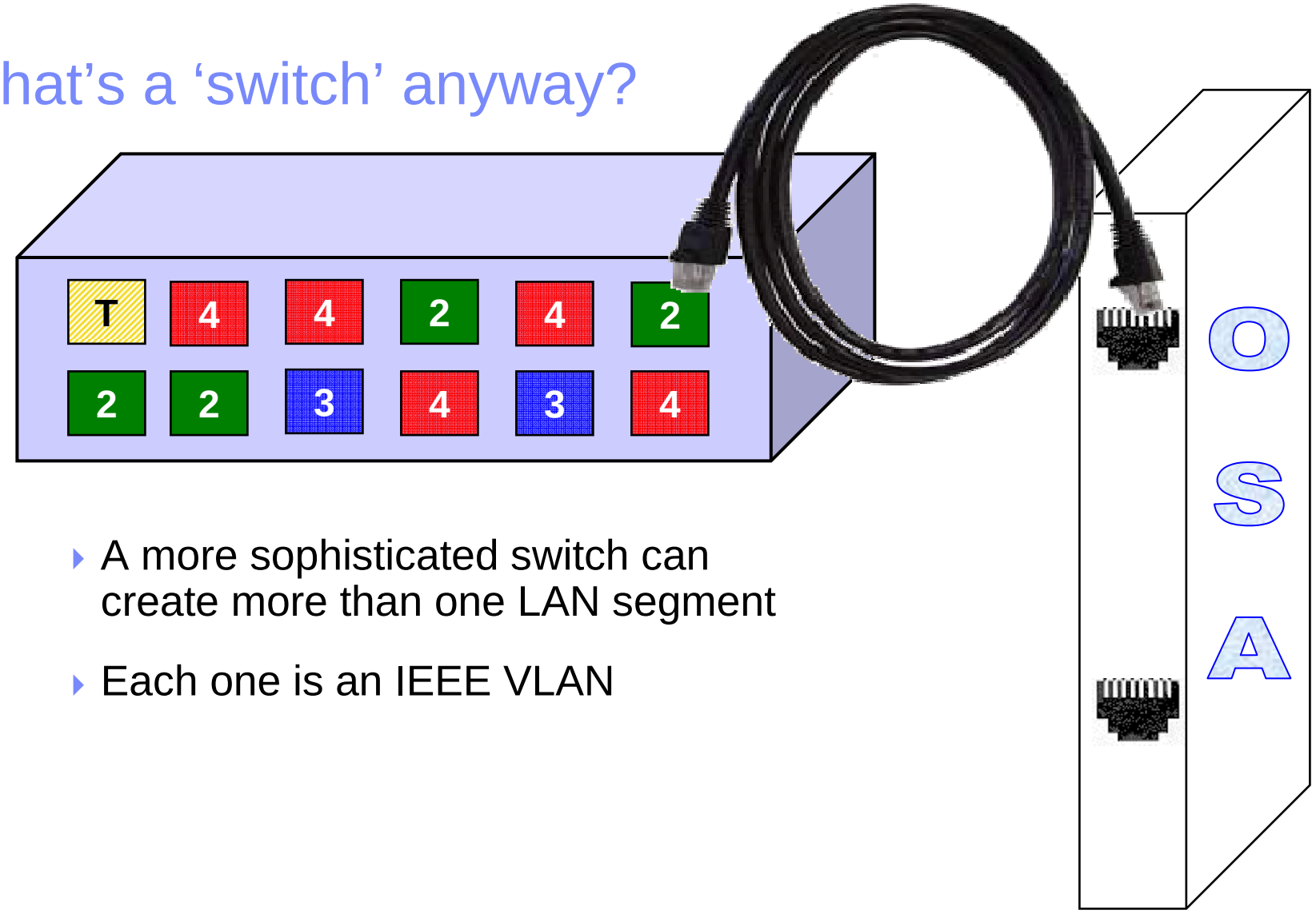


- A box that creates a LAN
- It can be remotely configured
  - E.g. Turn ports on and off
- Similar to a home router

z/VM Virtual Switch – VLAN unaware

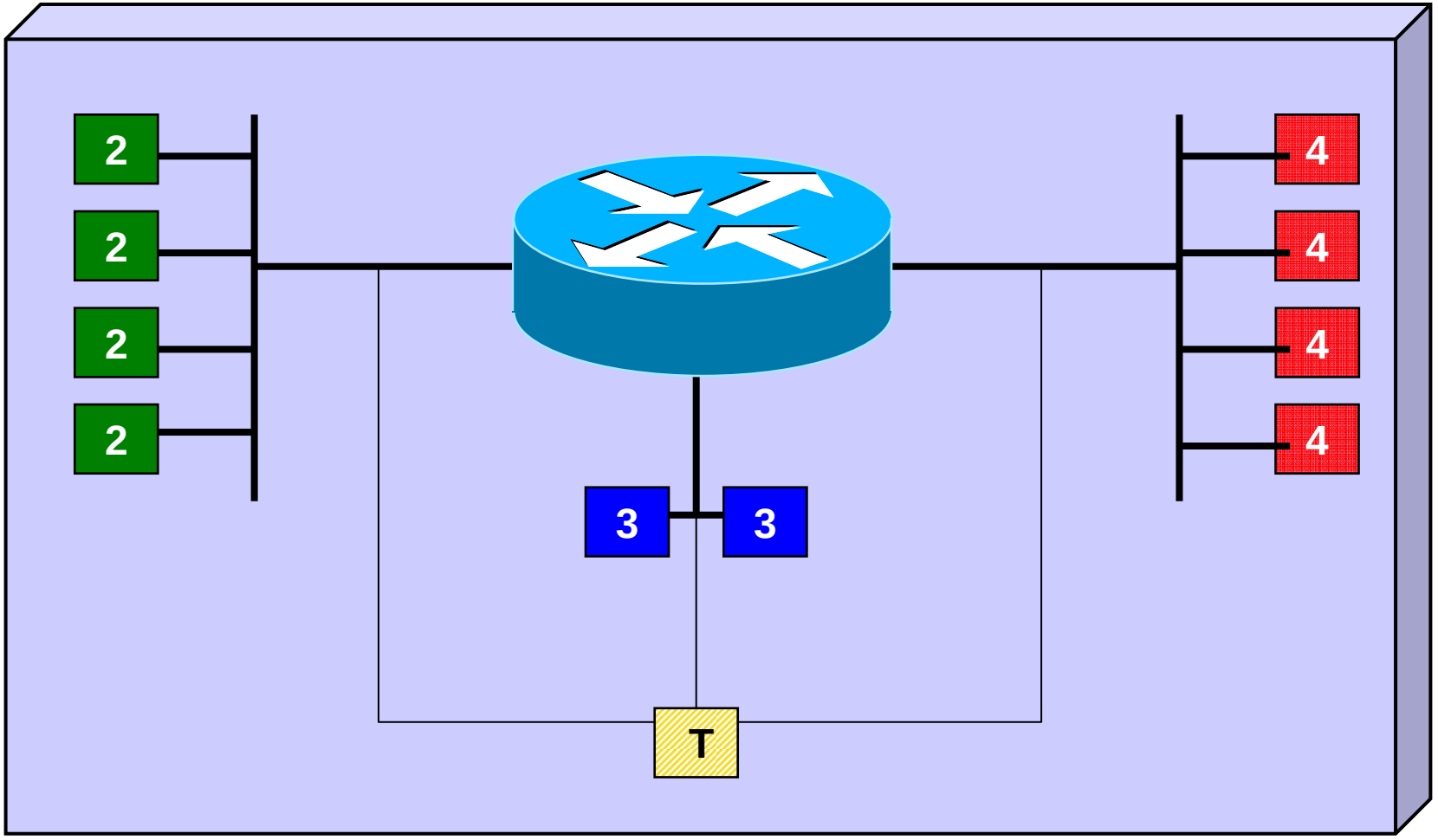


What's a 'switch' anyway?

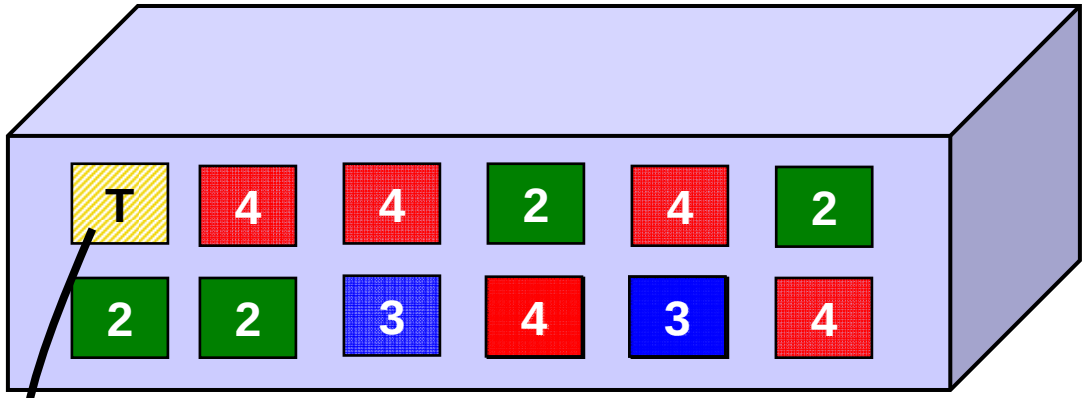


- A more sophisticated switch can create more than one LAN segment
- Each one is an IEEE VLAN

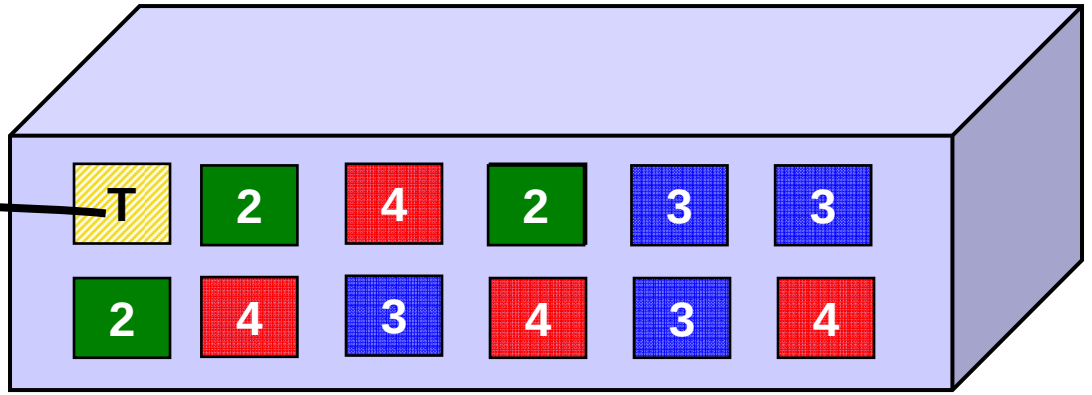
A VLAN-aware switch: An inside look



Trunk Port vs. Access Port

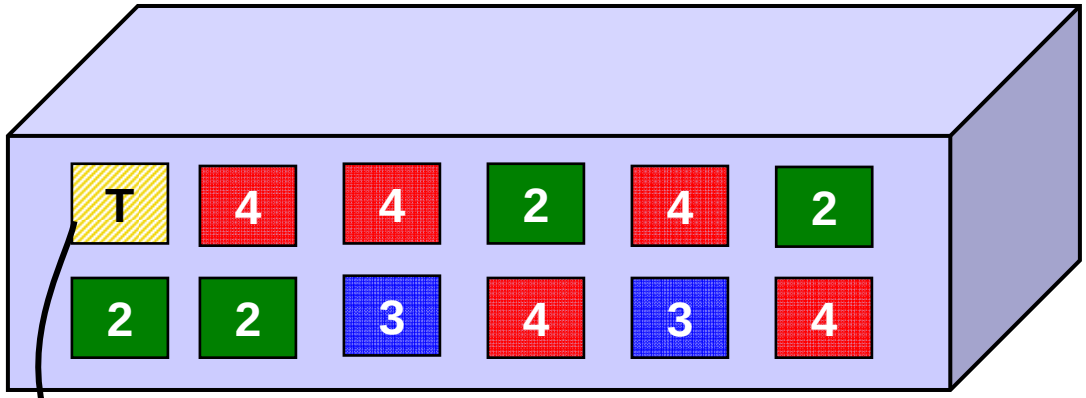


- Access port carries traffic for a single VLAN
- Host not aware of VLANs



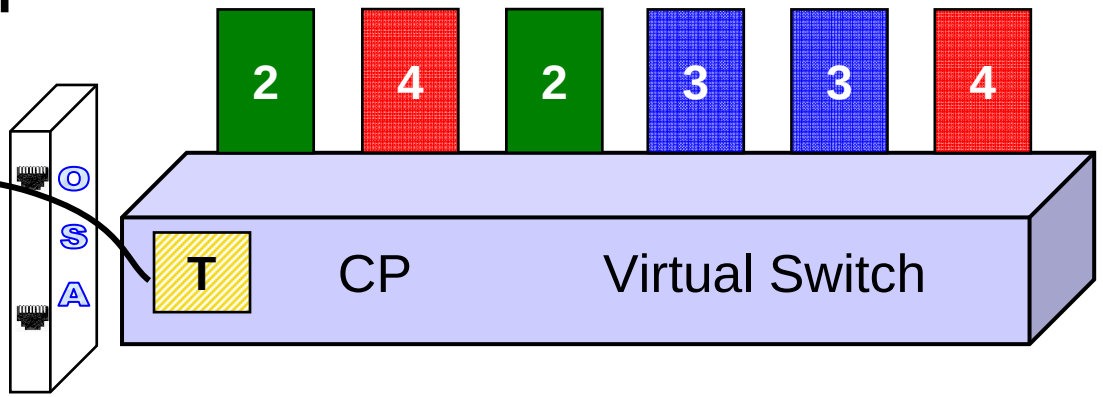
- Trunk port carries traffic from all VLANs
- Every frame is tagged with the VLAN id

Physical Switch to Virtual Switch

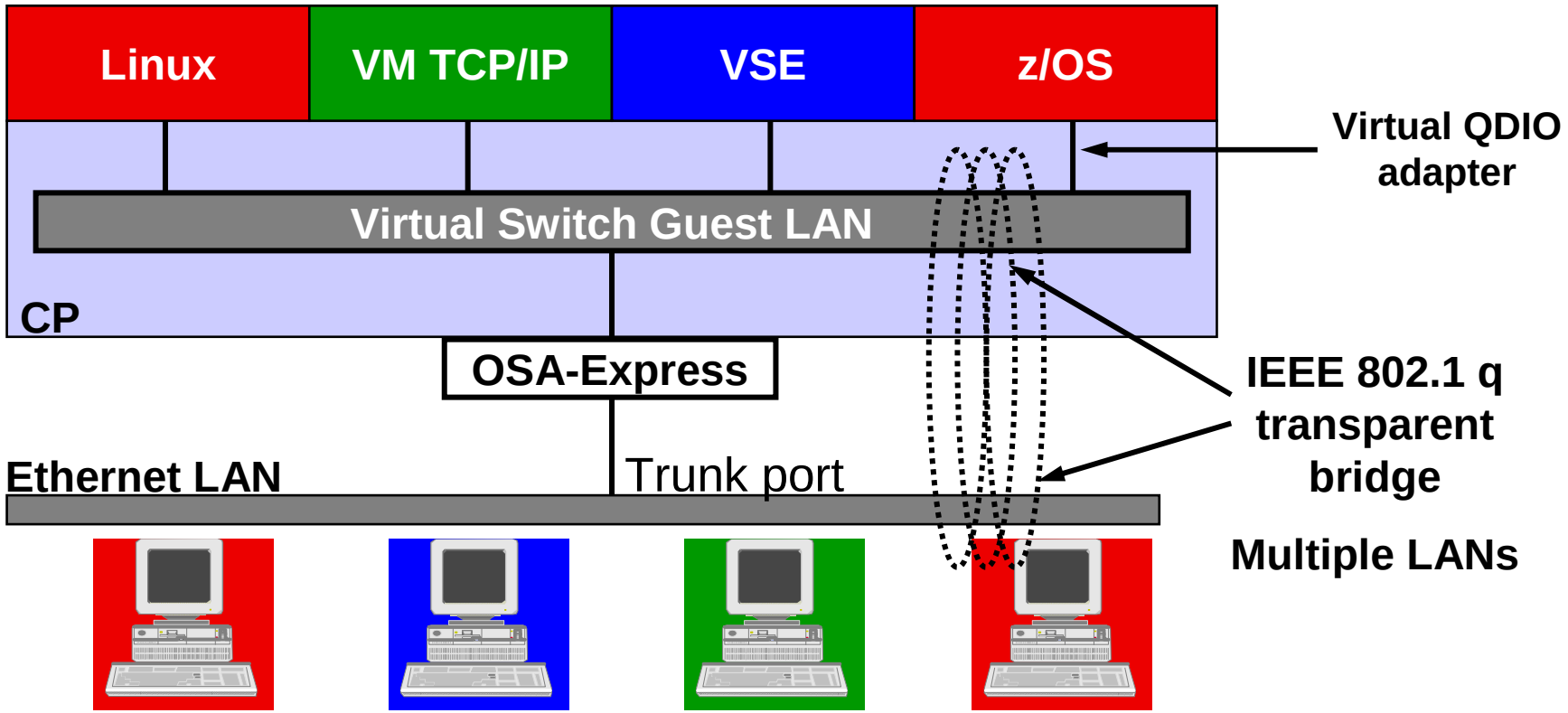


- **Trunk port carries traffic between CP and switch**

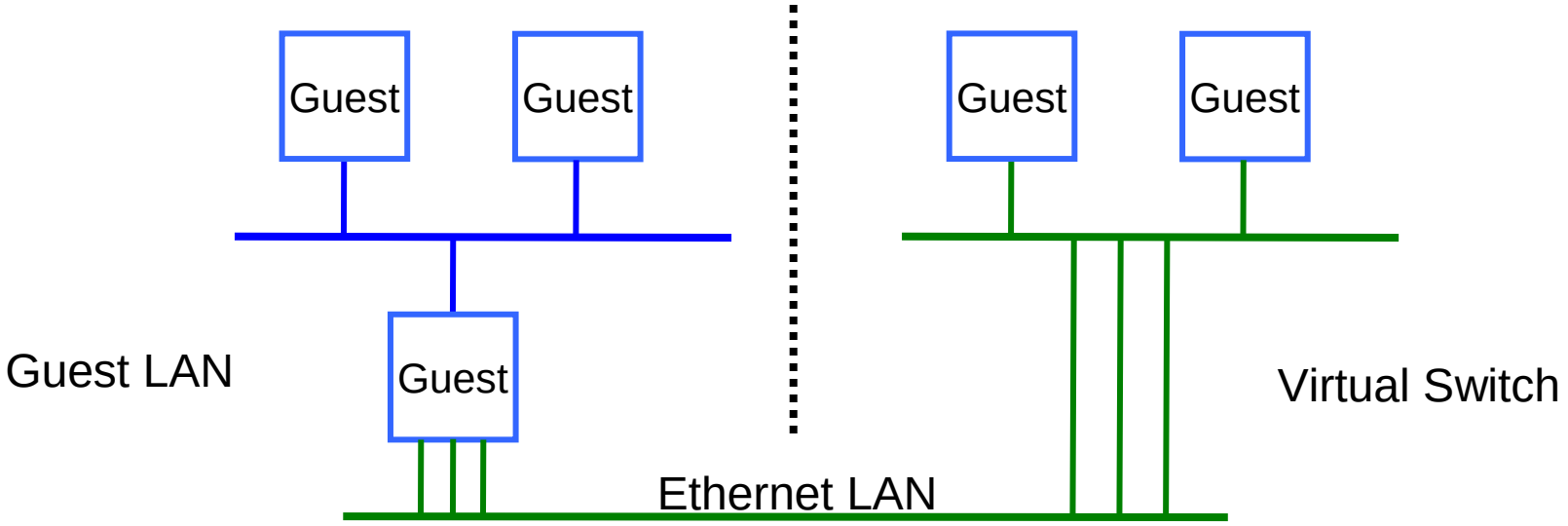
- **Each guest can be in a different VLAN**



z/VM Virtual Switch – VLAN aware



Guest LAN vs. Virtual Switch



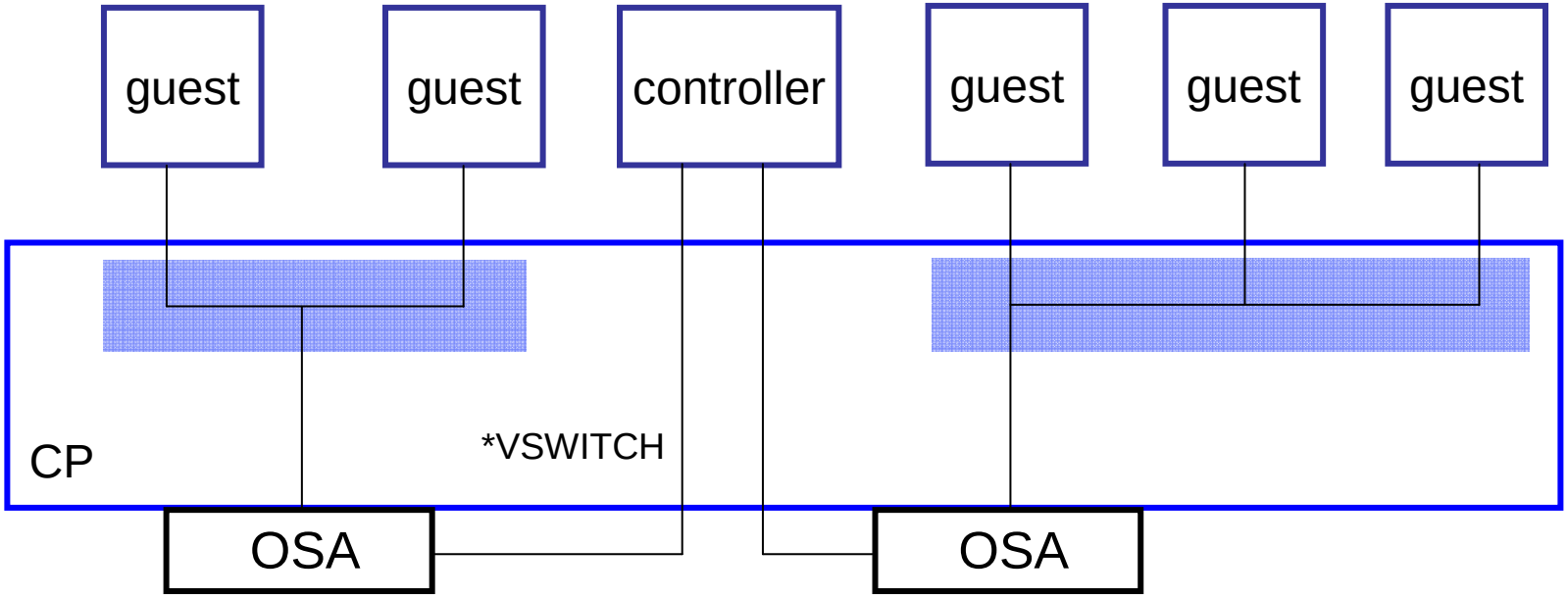
- Virtual router is required
- Different subnet
- External router awareness
- Guest-managed failover
- More data copies

- No virtual router
- Same subnet
- Transparent bridge
- CP-managed failover
- Fewer data copies

Advantages of the Virtual Switch

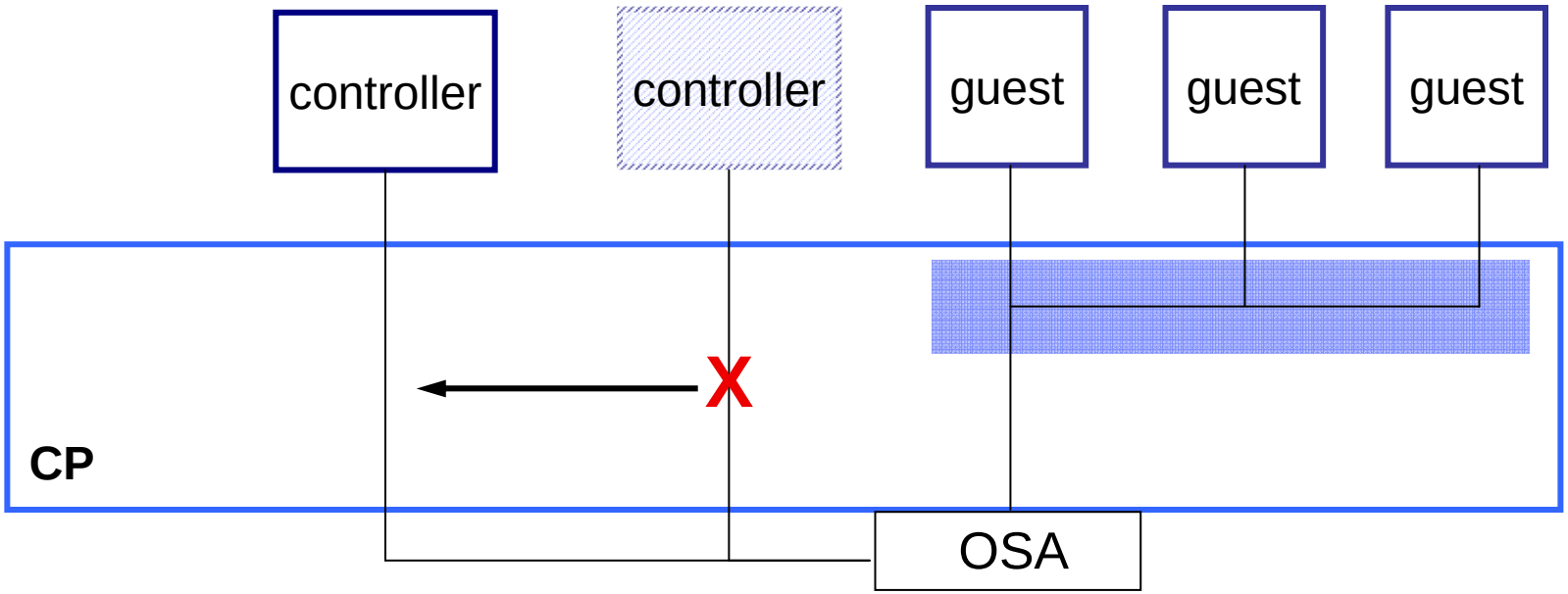
- Enables virtual QDIO connections to physical LAN segments without requiring a router
  - Reduces overhead associated with router virtual machines
- Virtual machines on the Guest LAN are in the same subnet as the physical LAN segment
- Reduces copying of the data being transported
- Provides centralized network configuration and control

VSWITCH Controller



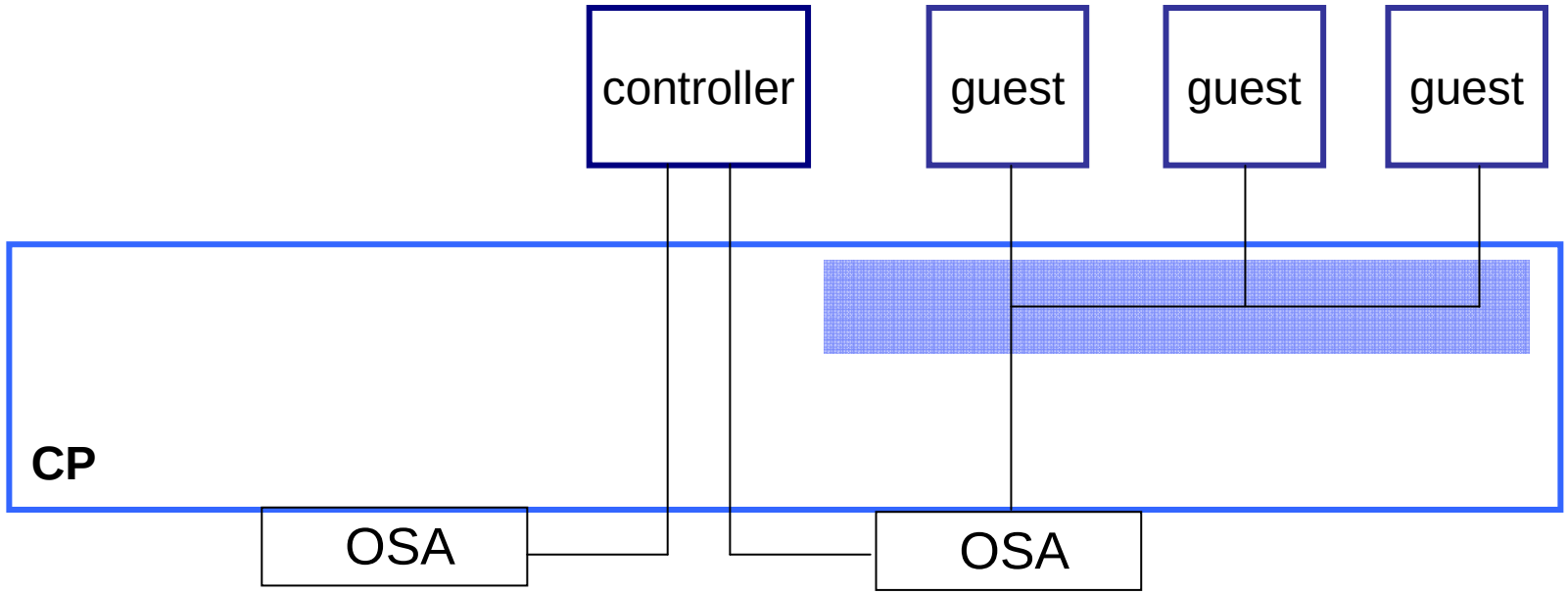
- **A controller is a VM TCP/IP stack, but it doesn't have to be your production stack. Create another one.**
- **Not involved in data transfer; only handles OSA housekeeping.**

VSWITCH Controller Failover



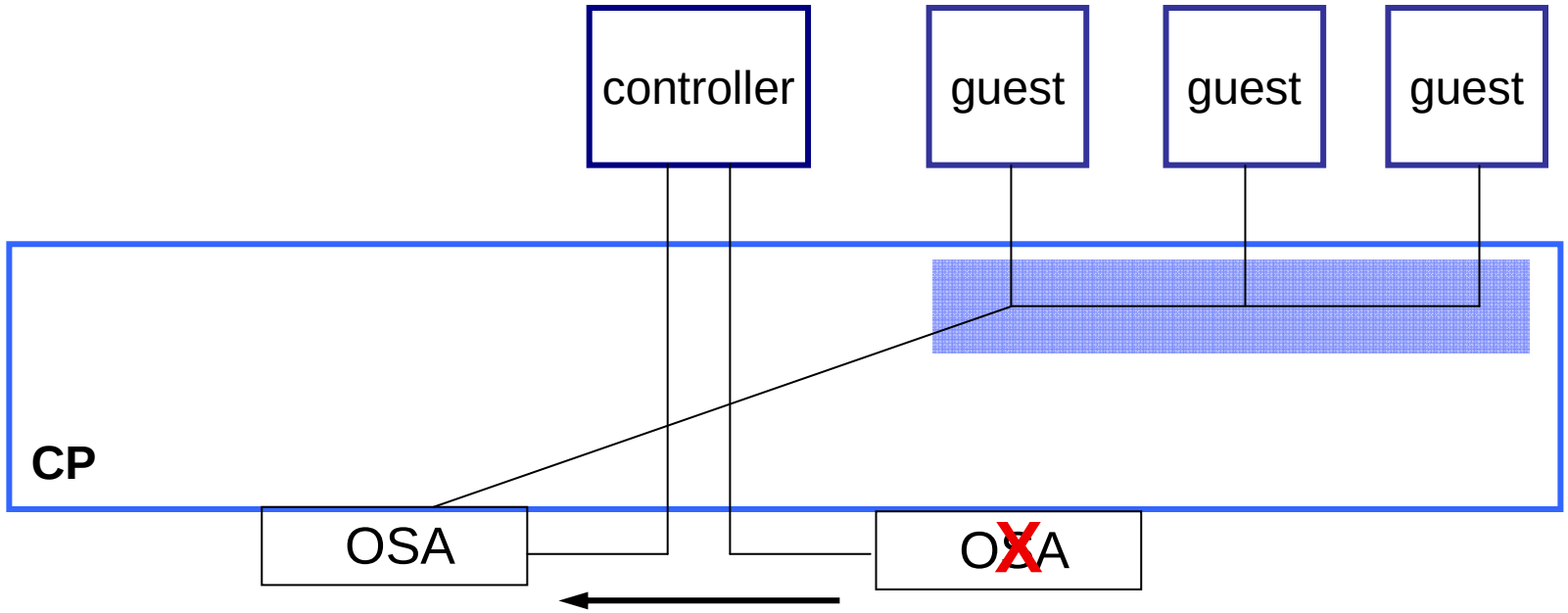
- **In case a controller fails or is forced off, CP will find another, if available.**
- **A VSWITCH can be limited to a specific controller, but is not recommended.**
- **If no controller, VSWITCH external connection is deactivated.**

OSA Failover



- **Up to 3 OSAs per VSWITCH**
- **Automatic failover**

OSA Failover



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

Controller Configuration

- An **IUCV *VSWITCH** statement must be included in the TCP/IP stack's CP directory entry
- You need **DIAG98** on the directory OPTION statement
- Standard directory entry for user TCPIP has both of these
- The VSWITCH CONTROLLER ON statement must be added to the TCP/IP configuration file (PROFILE TCPIP)
  - Do not code Gateway, Device, or Link statements for these devices
  - For standalone controller, only VSWITCH CONTROLLER ON is needed in the PROFILE TCPIP

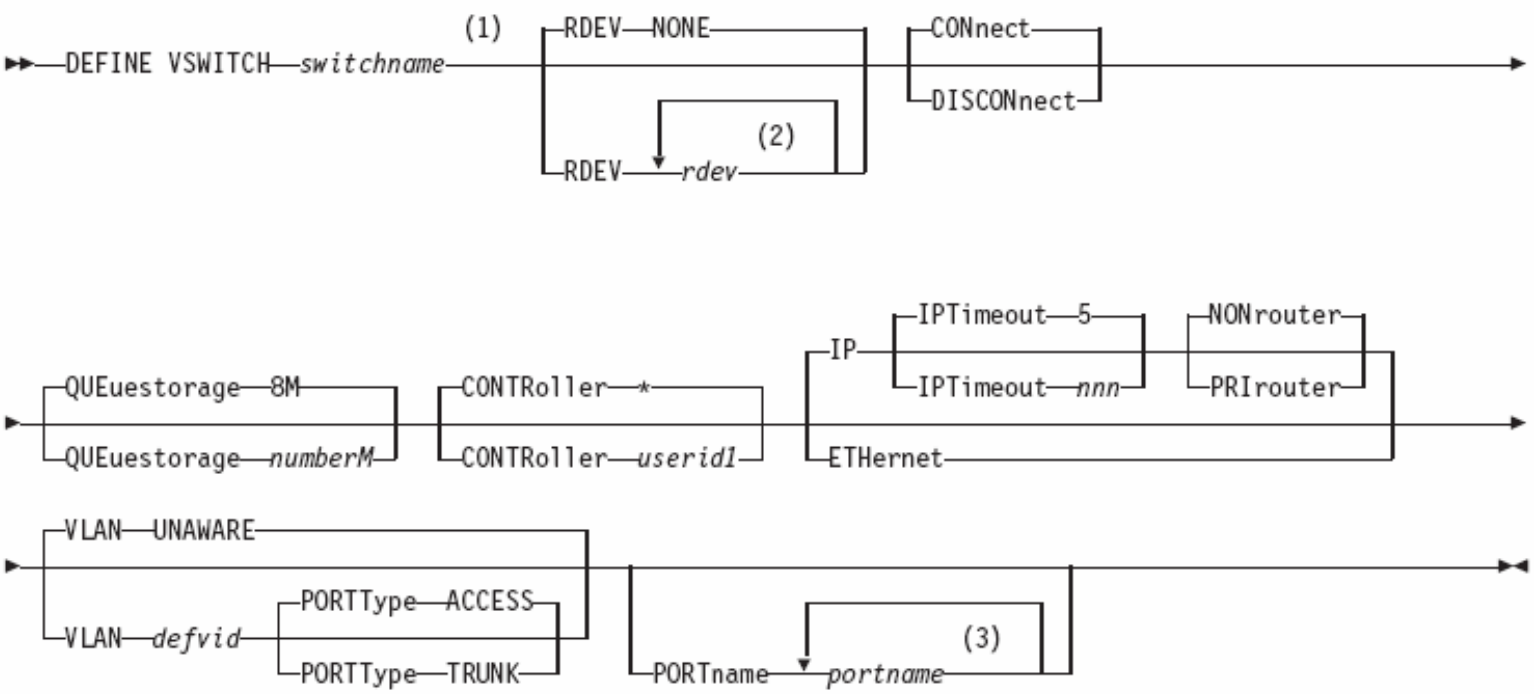
New in z/VM 5.2.0

- 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

Defining a VSWITCH is easy

- By CP command
  - DEFINE VSWITCH
  - SET VSWITCH
- In SYSTEM CONFIG
  - DEFINE VSWITCH
  - MODIFY VSWITCH
- To connect to the VSWITCH
  - DEFINE NIC or NICDEF in the user directory
  - COUPLE

DEFINE VSWITCH



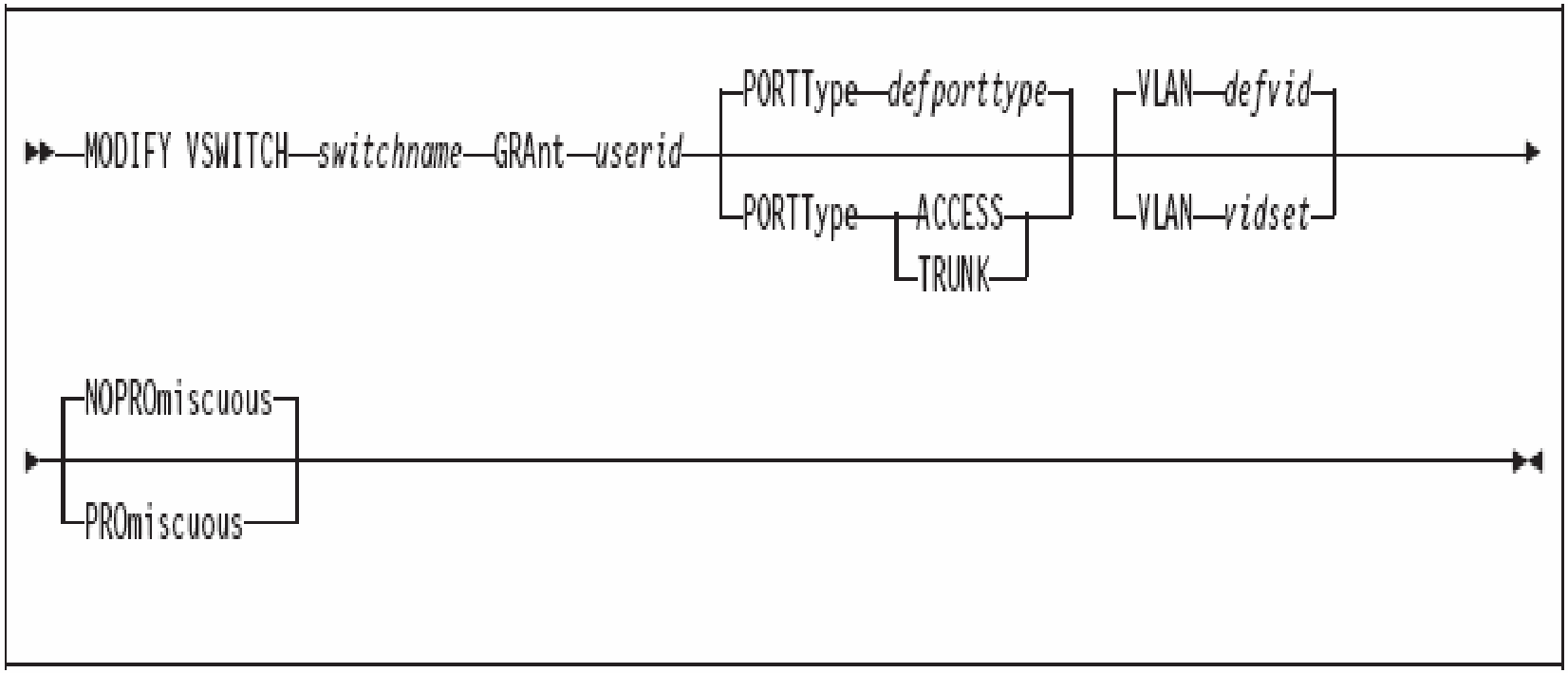
Notes:

1 You can specify the operands in any order, as long as *switchname* is the first operand specified, and *portname* is the last operand specified, if applicable.

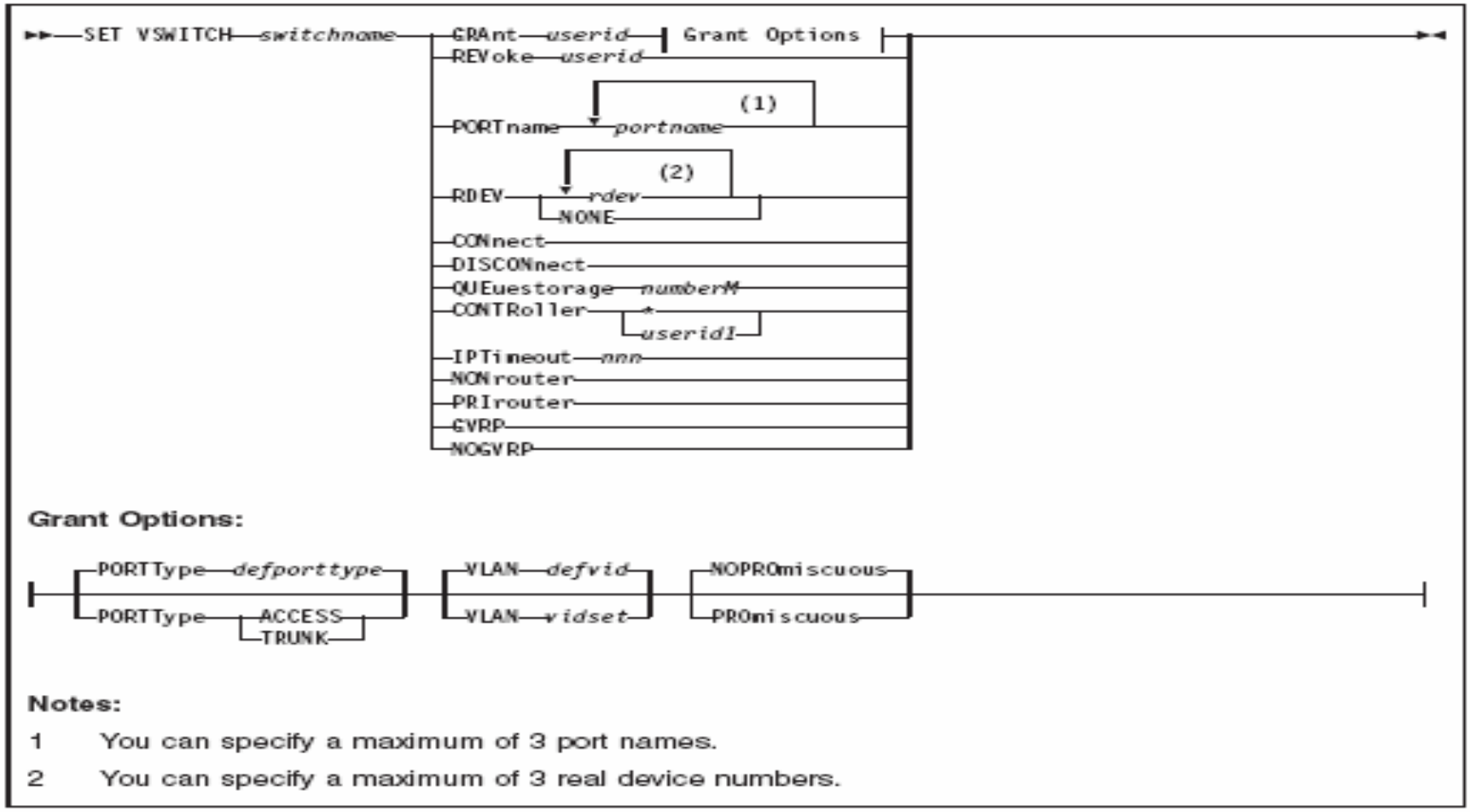
2 You can specify a maximum of 3 real device numbers.

3 You can specify a maximum of 3 port names.

MODIFY VSWITCH configuration statement



SET VSWITCH command

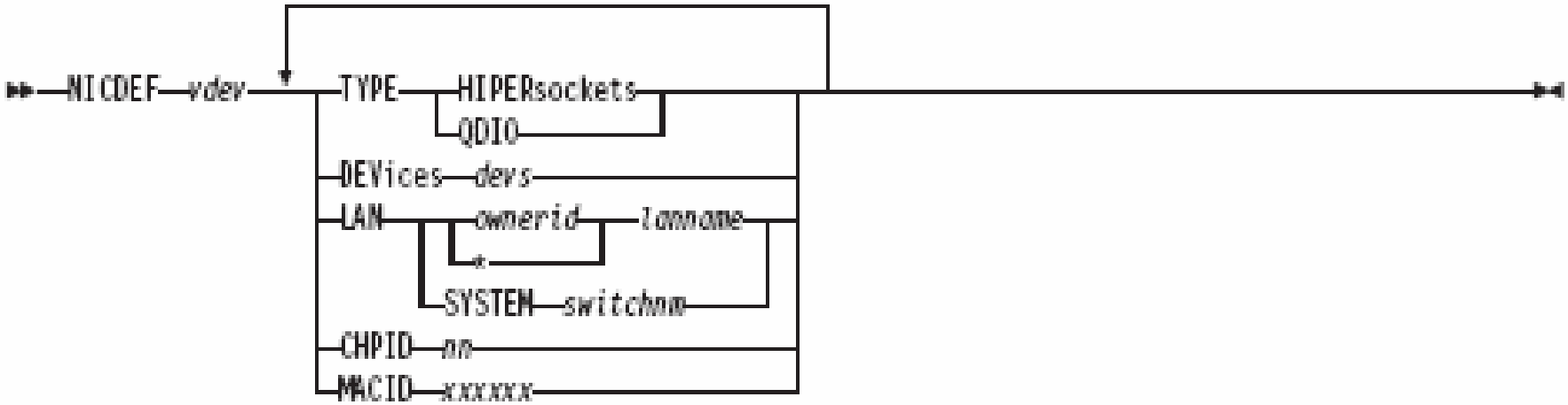


Notes:

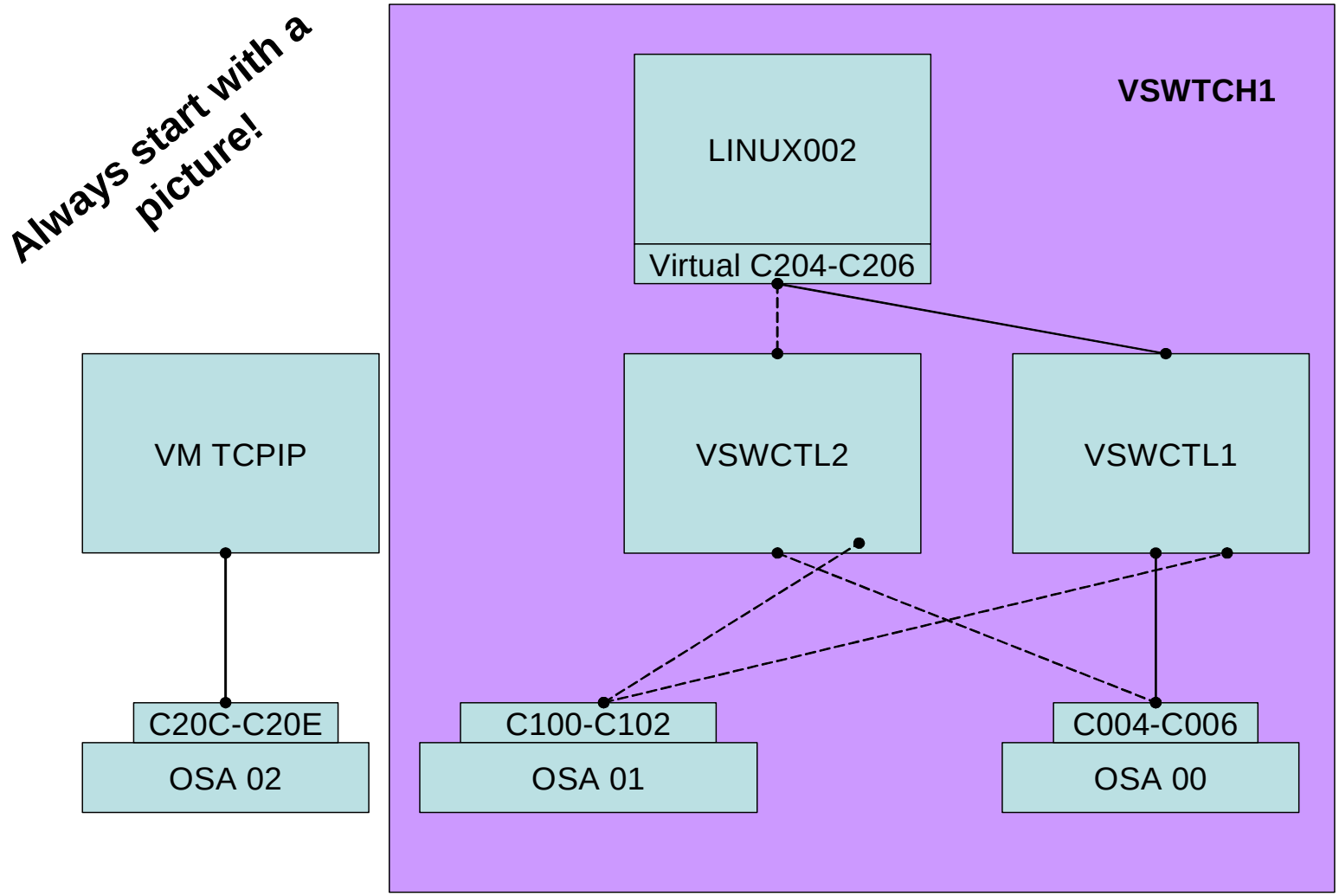
1 You can specify a maximum of 3 port names.

2 You can specify a maximum of 3 real device numbers.

NICDEF directory statement



Example VSWITCH Configuration



Example VSWITCH Configuration

- Exclusive use TCP/IP stack for both controllers
- Normal TCP/IP traffic through separate stack
- 3 OSA connections on different CHPIDs
- If you want the controlling z/VM TCP/IP stack to have connectivity through the virtual switch, define a NIC and couple to the virtual switch just like the other guests do

VSWITCH Configuration

- VSWITCH defined in SYSTEM CONFIG
- Controller setup
  - Controllers started by AUTOLOG1
  - Directory entries copied from TCPIP
  - SYSTEM DTCPARMS updated to include new Controllers
  - Used custom PROFILE TCPIP
- User setup
  - Used NICDEF directory statement
    - no need to issue COUPLE command
    - matched old physical OSA addresses
  - Did not make any changes to Linux!!!

Controller directory entry

```
USER VSWCTL1 XXXXXXXX 32M 128M ABG
 INCLUDE TCPCMSU
 OPTION QUICKDSP SVMSTAT MAXCONN 1024 DIAG98 APPLMON
 SHARE RELATIVE 3000
 IUCV ALLOW
 IUCV ANY PRIORITY
 IUCV *CCS PRIORITY MSGLIMIT 255
 IUCV *VSWITCH MSGLIMIT 65535
 LINK 4TCPIP40 491 491 RR
 LINK 4TCPIP40 492 492 RR
 LINK TCPMAINT 591 591 RR
 LINK TCPMAINT 592 592 RR
 LINK TCPMAINT 198 198 RR
 MDISK 191 3390 1706 005 440W02 MR
```

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 VSWCTL1
ADDRESS COMMAND CP XAUTOLOG VSWCTL2
```

SYSTEM DTCPARMS

```
:nick.TCPIP     :type.server
                :class.stack
                :attach.C20C-C20E

:NICK.VSWCTL1   :TYPE.SERVER
                :class.stack

:NICK.VSWCTL2   :TYPE.SERVER
                :class.stack
```

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
```

PROFILE TCPIP

```
 PROFILE  TCPIP    D1  V 86  Trunc=86 Size=1 Line=1 Col=1 Alt=0

====>
===== * * * Top of File * * *
===== VSWITCH CONTROLLER ON
===== * * * End of File * * *
```

Yes, there's only one line in it.

VSWITCH Failover

Initial state

```
q osa
OSA  C004 ATTACHED TO VSWCTL1  C004
OSA  C005 ATTACHED TO VSWCTL1  C005
OSA  C006 ATTACHED TO VSWCTL1  C006
OSA  C100 ATTACHED TO VSWCTL1  C100
OSA  C101 ATTACHED TO VSWCTL1  C101
OSA  C102 ATTACHED TO VSWCTL1  C102
OSA  C20C ATTACHED TO TCPIP    C20C
OSA  C20D ATTACHED TO TCPIP    C20D
OSA  C20E ATTACHED TO TCPIP    C20E

q controller
Controller VSWCTL1   Available: YES   VDEV Range: *          Level 510
  Capability: IP ETHERNET VLAN_ARP
   SYSTEM VSWTCH1    Primary          Controller: *          VDEV: C004
   SYSTEM VSWTCH1    Backup           Controller: *          VDEV: C100
Controller VSWCTL2   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: VSWCTL1  VDEV:  C004
  Portname: UNASSIGNED RDEV: C100 Controller: VSWCTL1  VDEV:  C100 BACKUP
```

Simulate failures

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

FORCE VSWCTL1

```
force vswctl1
USER DSC   LOGOFF AS  VSWCTL1  USERS = 15    FORCED BY MAINT

HCPSWU2843E The path was severed for TCP/IP Controller VSWCTL1.
HCPSWU2843E It was managing device C004 for VSWITCH SYSTEM VSWTCH1.

HCPSWU2830I VSWITCH SYSTEM VSWTCH1 status is in error recovery.
HCPSWU2830I VSWCTL2 is new VSWITCH controller.

HCPSWU2830I VSWITCH SYSTEM VSWTCH1 status is ready.
HCPSWU2830I VSWCTL2 is VSWITCH controller.

q controller
Controller VSWCTL2   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 VSWCTL2 is VSWITCH controller.

HCPSWU2830I VSWITCH SYSTEM VSWTCH1 status is in error recovery.
HCPSWU2830I VSWCTL2 is new VSWITCH controller.

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

HCPSWU2830I VSWITCH SYSTEM VSWTCH1 status is ready.
HCPSWU2830I VSWCTL2 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: VSWCTL2  Error: No RDEV
  Portname: UNASSIGNED RDEV: C100 Controller: VSWCTL2  VDEV:  C100
```

Summary

- IT WORKS!
- Very easy to setup - TRY IT!
- Need more information?
  - z/VM Connectivity
    - Part 2: Planning Virtual Networks
  - z/VM TCP/IP Planning and Customization
  - z/VM CP Planning and Administration
  - Getting Started with Linuz on System Z

Questions?

Contact Information

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

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

Thanks for Listening!