V26

Link Aggregation with the z/VM Virtual Switch

Tracy Adams

IBM System z Expo

September 17-21, 2007
San Antonio, TX

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

APPN*
CICS*
DB2*
DFSMSMVS
DFSMS/VM*
DirMaint
Distributed Relational Database Architecture*
DRDA*
e-business logo*
ECKD
Enterprise Storage Server*
Enterprise Systems Architecure/390*
ESCON*
FICON*
GDDM*

* Registered trademarks of IBM Corporation

GDPS*
Geographically Dispersed Parallel Sysplex
HiperSockets
HyperSwap
IBM*
IBM eServer
IBM logo*
IBMlink
Language Environment*
MQSeries*
Multiprise*
On demand business logo
OS/390*
Parallel Sysplex*
Performance Toolkit for VM
POWER5

POWERPC*
PR/SM
Processor Resource/Systems Manager
QMF
RACF*
Resource Link
RMF
RS/6000*
S/390*
S/390 Parallel Enterprise Server
System 370
System 390*
System z9
Tivoli*
Tivoli Storage Manager
TotalStorage*

Virtual Image Facility
Virtualization Engine
VisualAge*
VM/ESA*
VSE/ESA
VTAM*
WebSphere*
z/Architecture
z/OS*
z/VM*
z/VSE
zSeries*
zSeries Entry License Charge

The following are trademarks or registered trademarks of other companies.

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
Linux is a trademark of Linus Torvalds in the united States and other countries..
UNIX is a registered trademark of The Open Group in the United States and other countries.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation in the United States and other countries.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

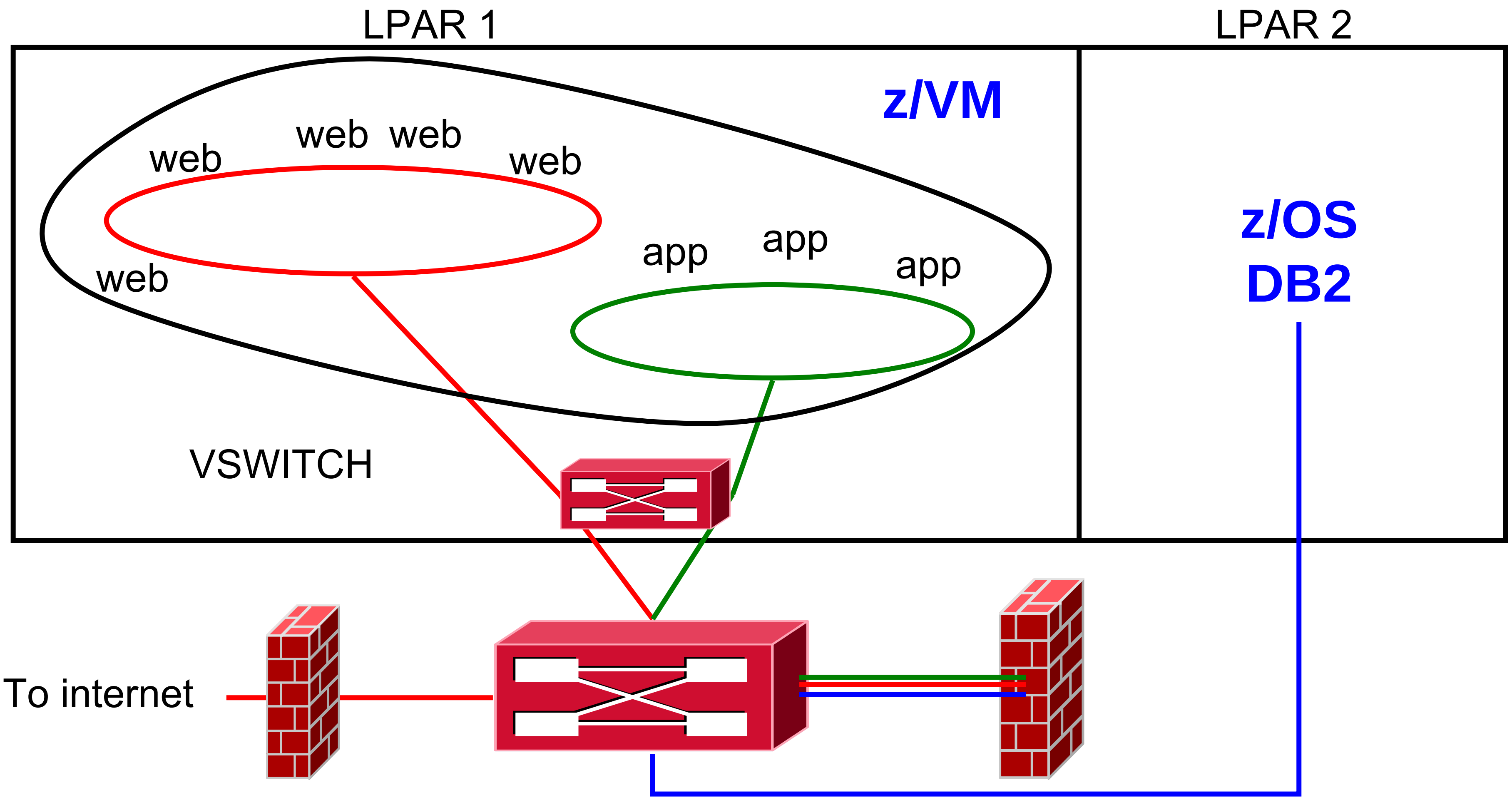
Session Objectives

At the end of this session you will understand the following:

- Virtual Switch Technology
- Concept of Link Aggregation
- Software and Hardware Requirements
- Journey to the World of Link Aggregation
- Benefits

Virtual Switch Overview

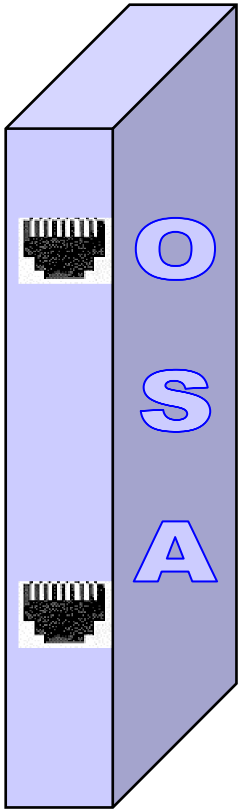
Network with VSWITCH



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

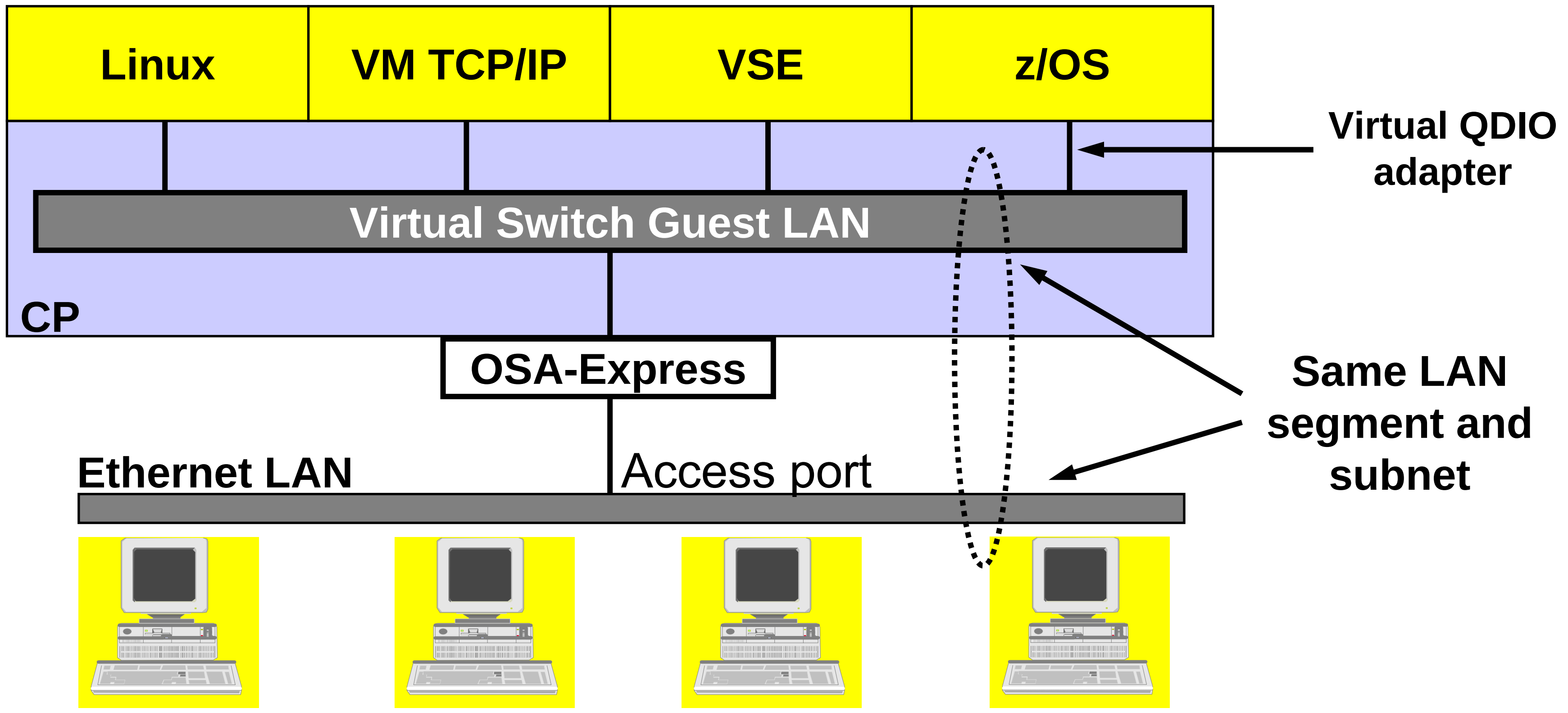
What's a 'switch' anyway?

© Cisco Corp

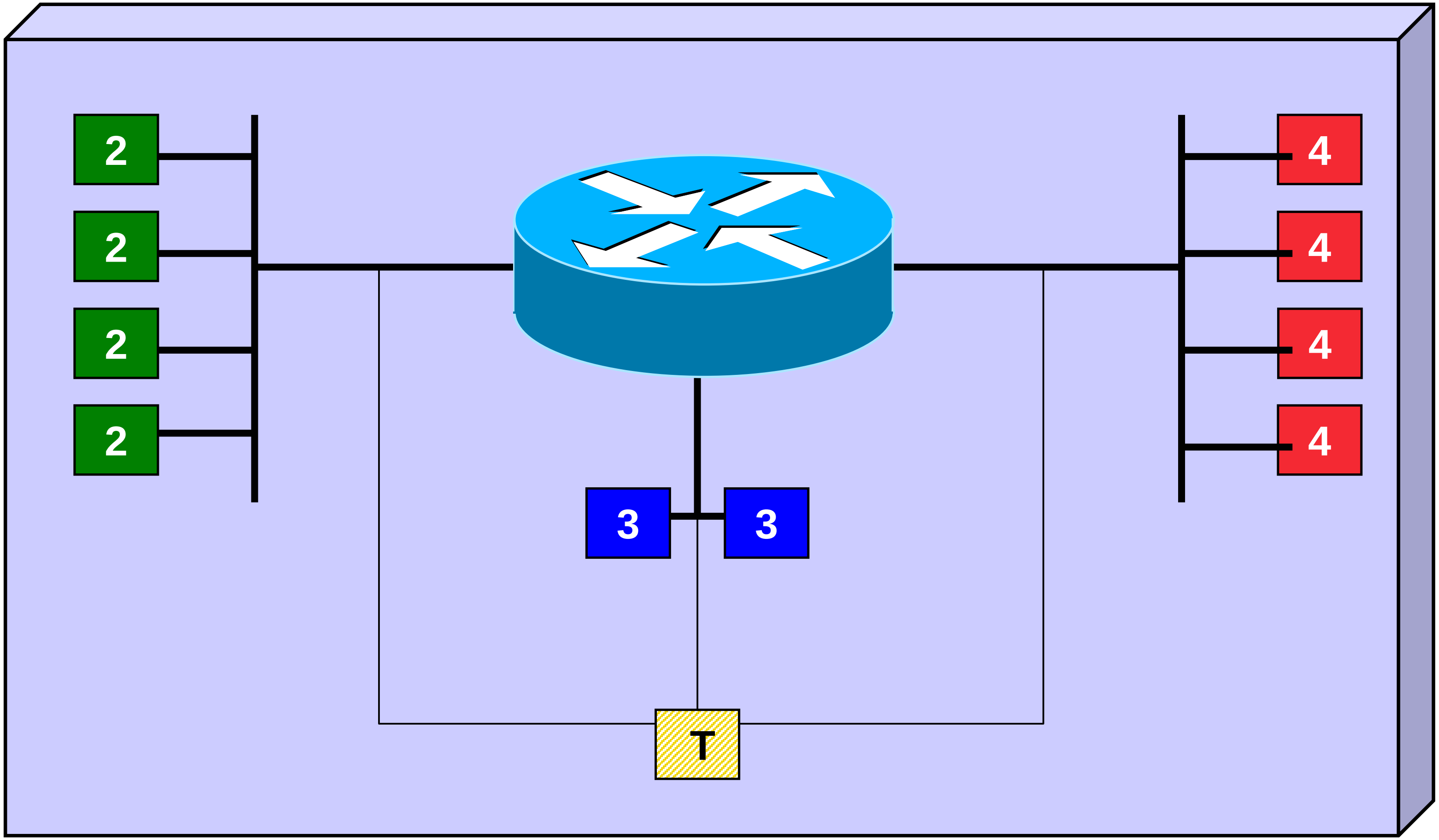


- A box that creates a LAN
- It can be remotely configured
  - E.g. Turn ports on and off
- Similar to a home router

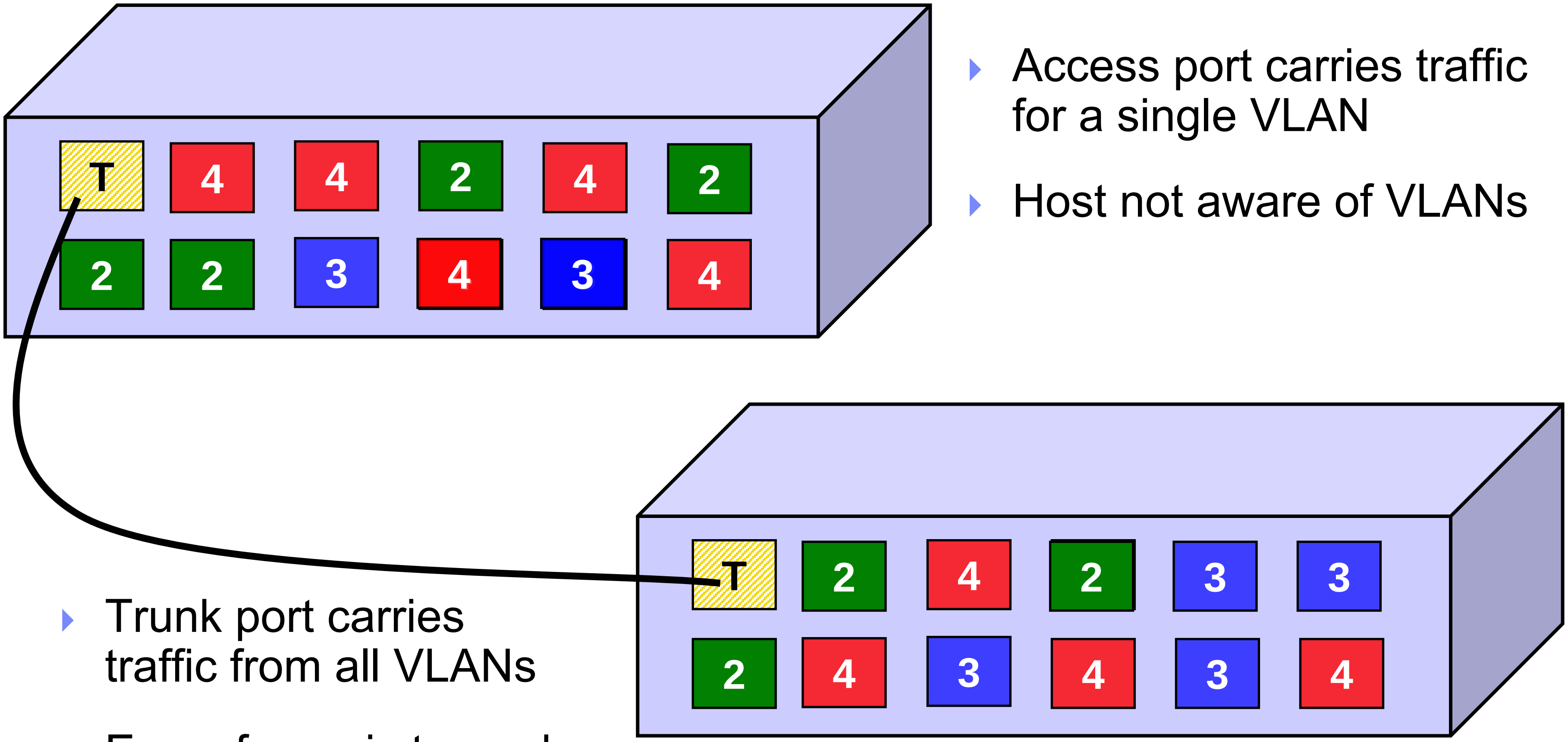
z/VM Virtual Switch – VLAN unaware



A VLAN-aware switch: An inside look



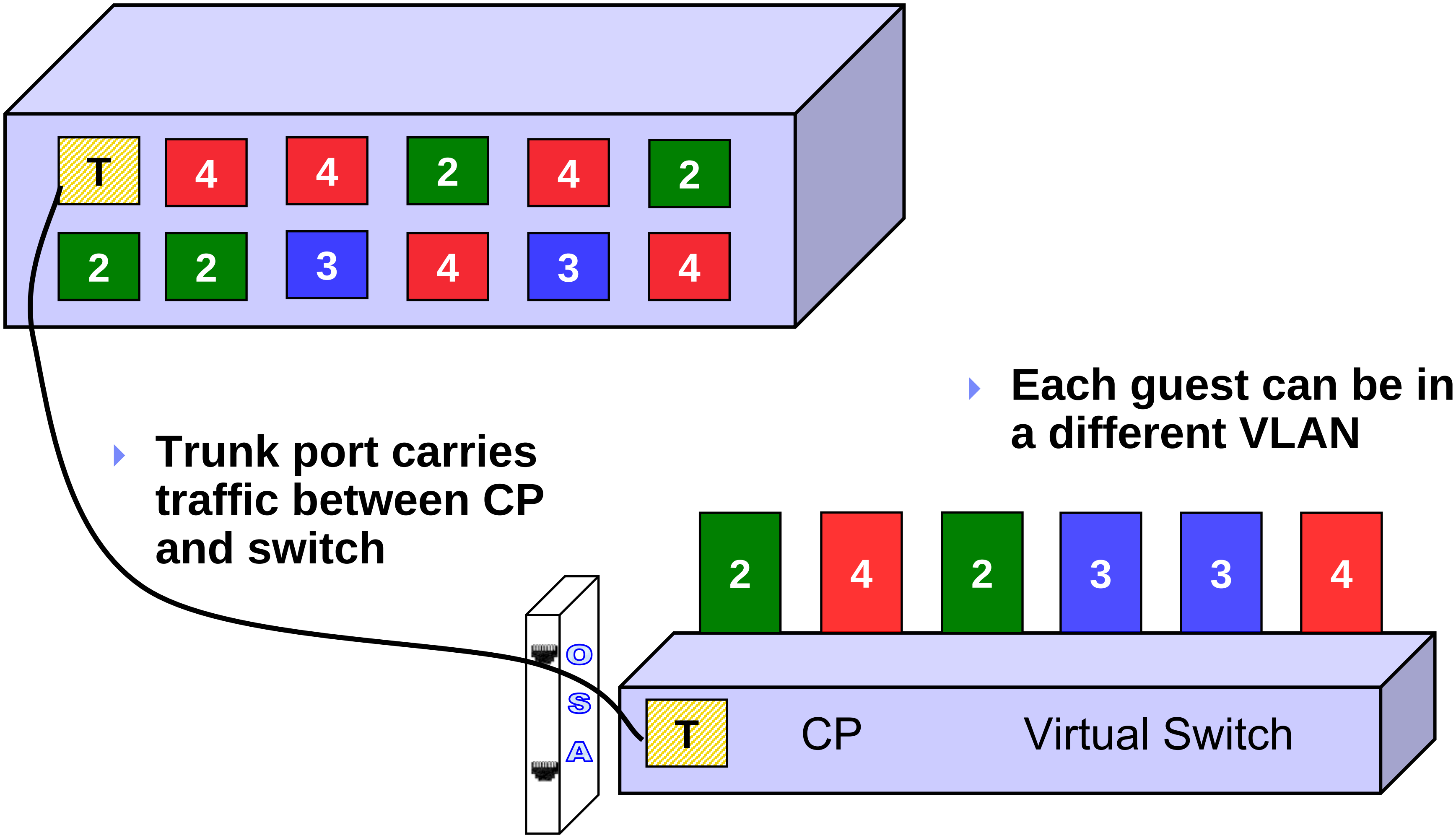
Trunk Port vs. Access Port



- 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

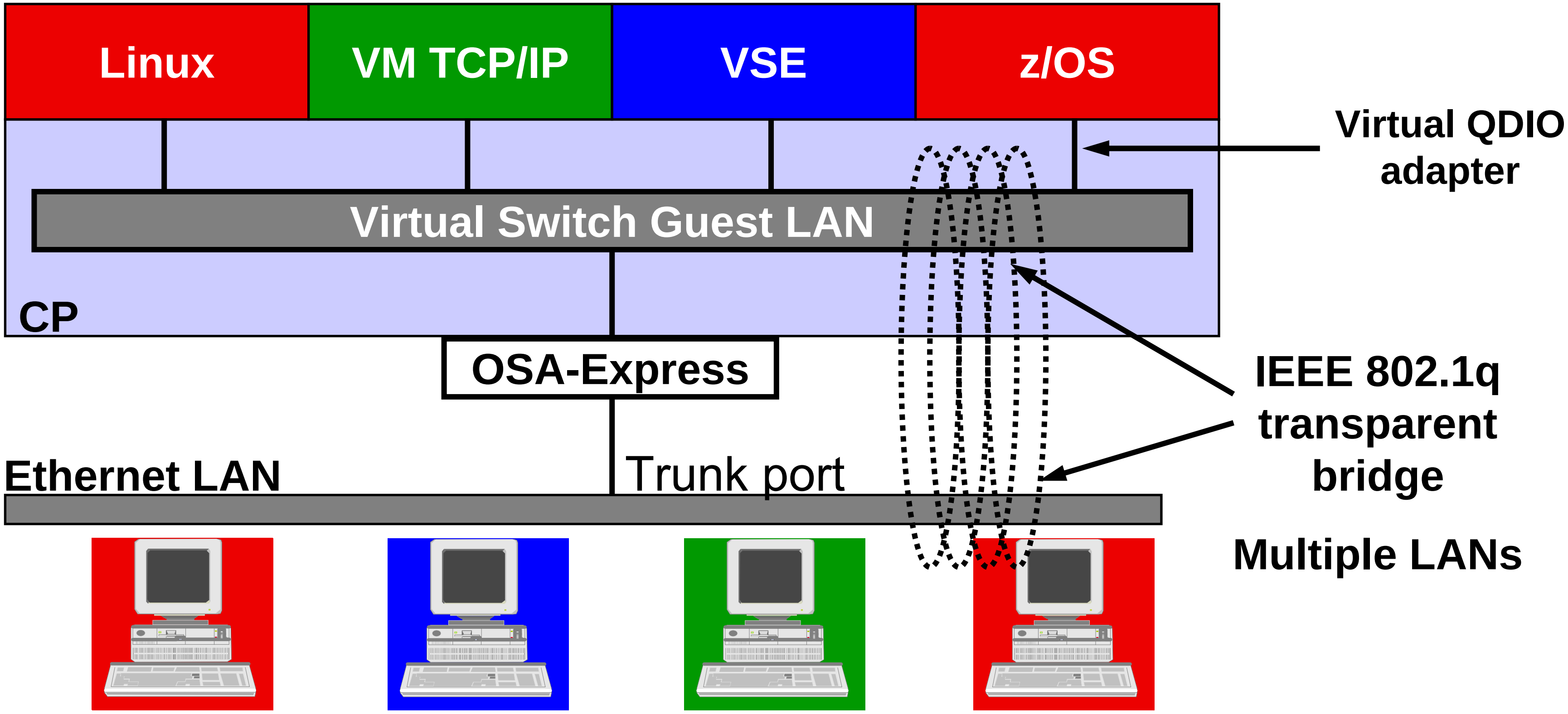
Physical Switch to Virtual Switch



- Trunk port carries traffic between CP and switch

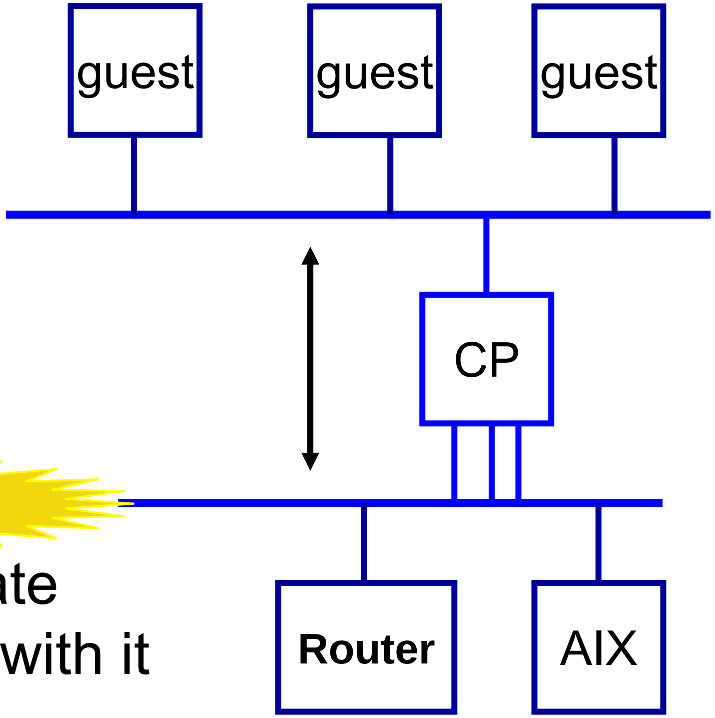
- Each guest can be in a different VLAN

z/VM Virtual Switch – VLAN aware



z/VM Virtual Switch

- A special-purpose Guest LAN
  - Ethernet IPv4 and IPv6
  - Built-in IEEE 802.1q bridge to outside network
  - IEEE VLAN capable



Per 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 controller virtual machines (minimal VM TCP/IP stack servers)
  - Controller not involved in data transfer
  - Do not ATTACH or DEDICATE
  - Use 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
```

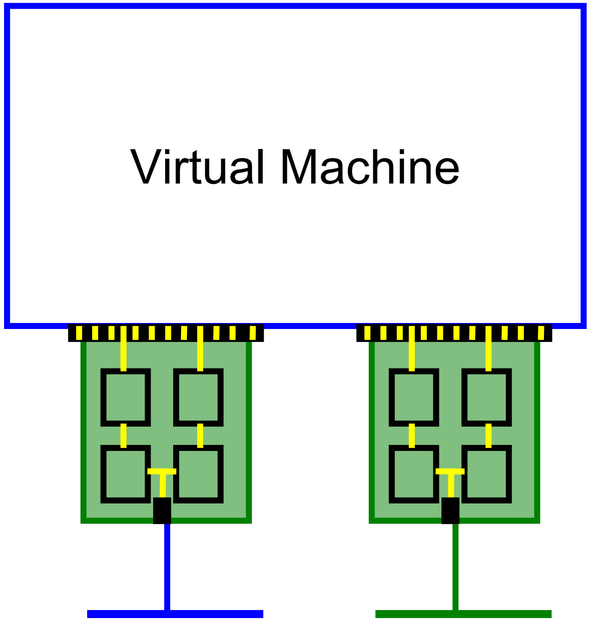
z/VM 5.2

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

Virtual Network Interface Card

Virtual Network Interface Card (NIC)

- A simulated network adapter
  - OSA-Express QDIO
  - HiperSockets
  - Must match Guest LAN or VSWITCH transport 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



Guest LAN or virtual switch

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
```

Link Aggregation

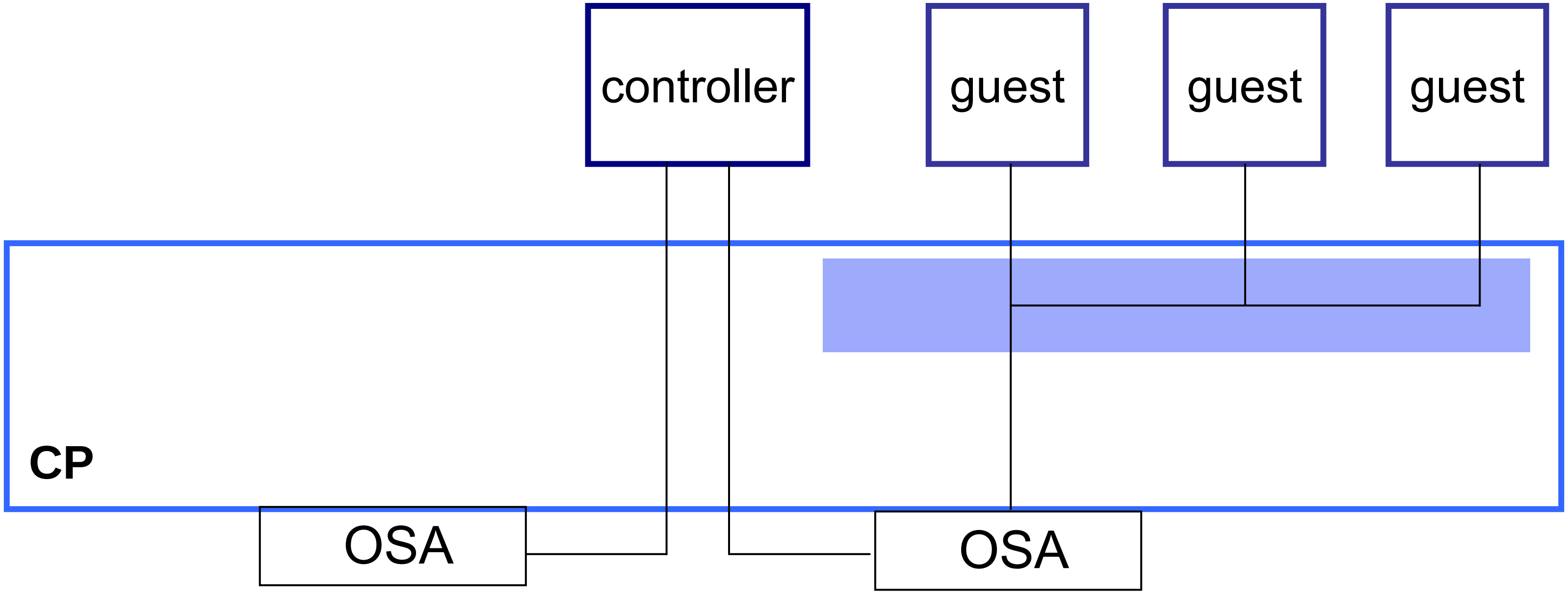
VSWITCH LinkAG Motivation

“But why aren’t you using my back up OSA card?”

VSWITCH Traditional Setup

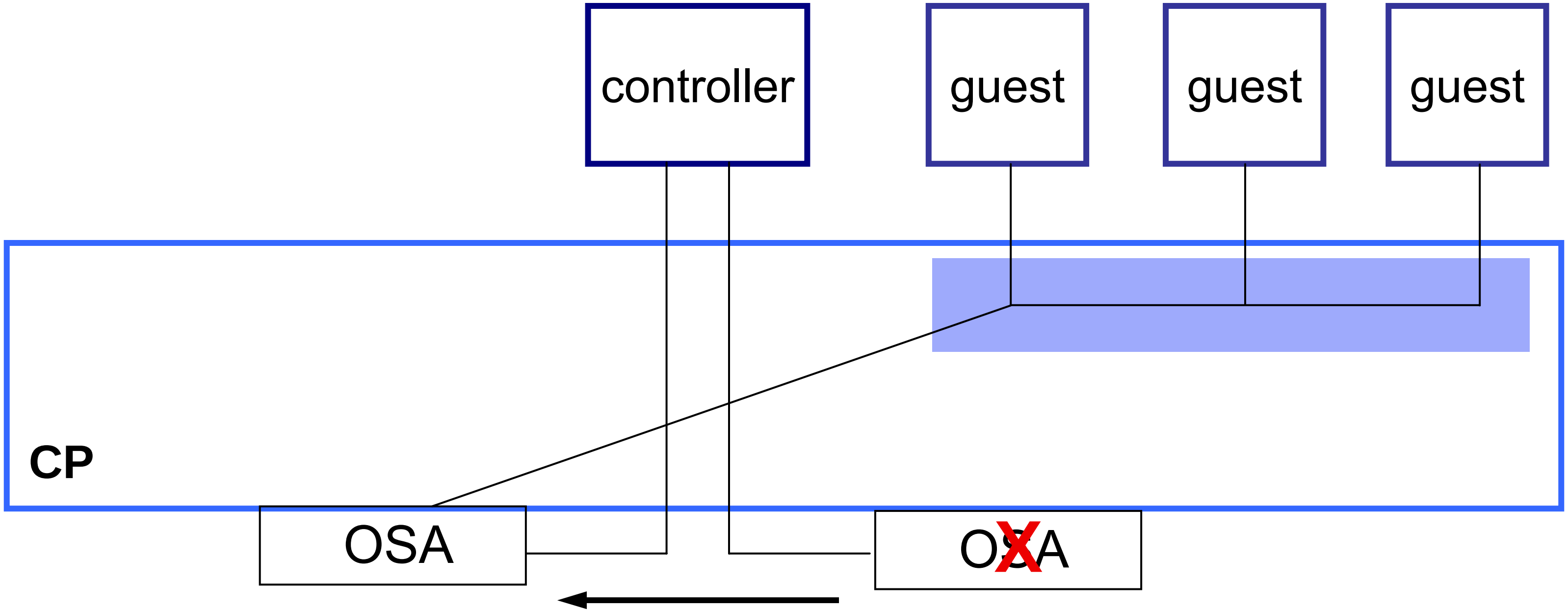
- Define VSWITCH with 3 RDEVS
- Use one OSA for data transfer
- Keep 2 OSA's as back up devices
- Failover to a back up OSA causes a brief network outage
  - Has been improved from release to release but customers always want more

OSA Failover



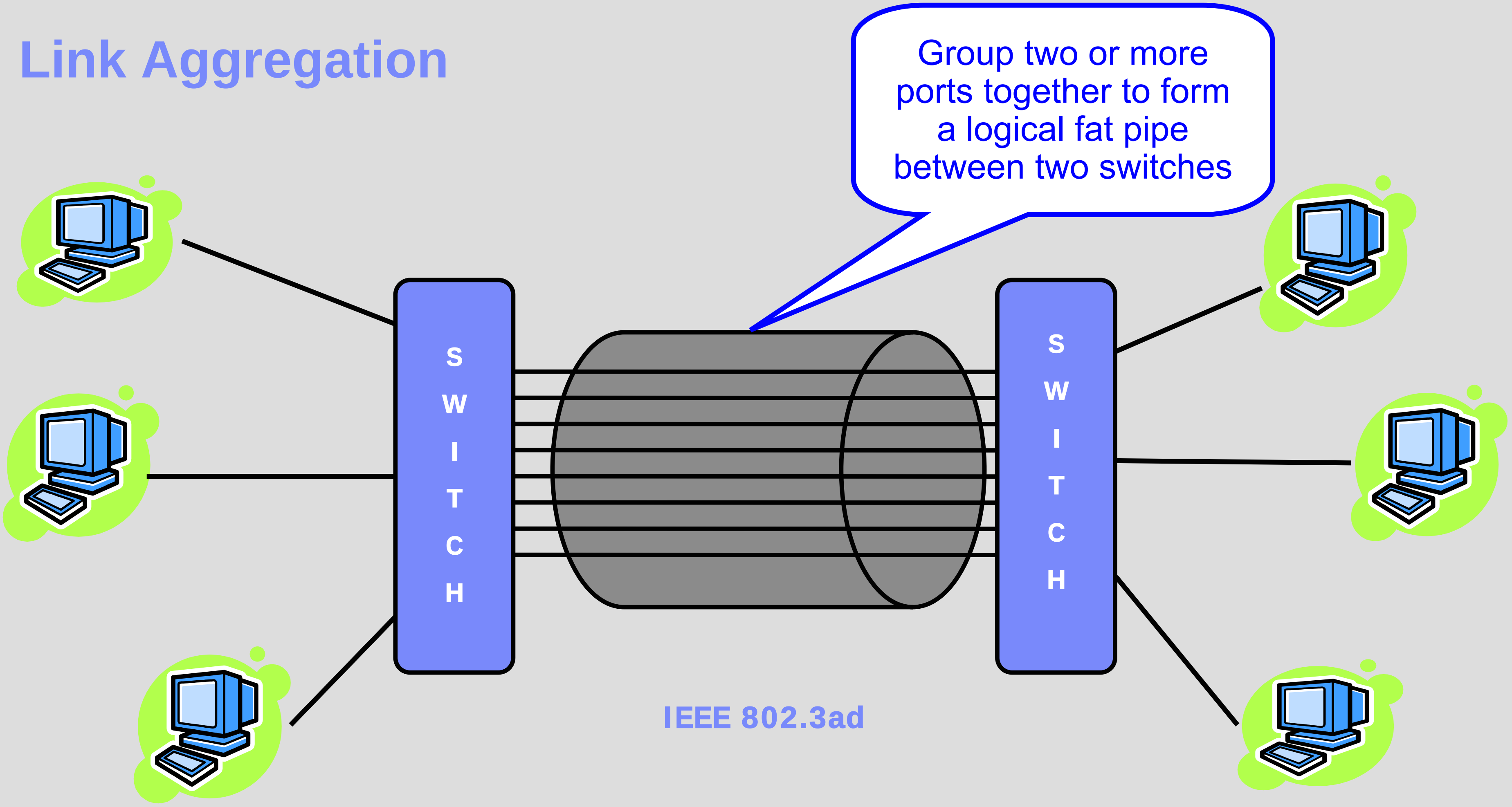
- **Up to 3 OSAs per VSWITCH**
- **Automatic failover**

OSA Failover



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

Link Aggregation


Group two or more ports together to form a logical fat pipe between two switches

IEEE 802.3ad

Cascading Switches

VSWITCH LinkAG Specifications

- Group multiple active QDIO VSWITCH real OSA connections as a single logical group (No support for aggregation of virtual NICs)
  - Up to 8 OSA ports (within a group or as backup devices)
  - Synchronized conversations over the same OSA link
  - Only one aggregate group per VSWITCH
- 802.3ad compliance for layer 2 **ETHERNET VSWITCH** only
- MAC level implementation which makes it totally transparent to all connected NICs or protocols

VSWITCH LinkAG Specifications

- Port group management
  - Dynamic (LACP ACTIVE)
  - Static (LACP INACTIVE)
- Near seamless failover
  - Port failover to another port within the group
  - Group failover to a single backup port (existing failover support)
- Minimal link selection overhead
- Ability to distribute single guest port traffic across multiple OSA connections.
- External controls using existing commands and a new SET PORT Command

Hardware Requirements

- Dedicated OSA Express2 Ports
  - Same type of NIC card (10, 100,1000 and 10000 mbps)
  - Point to point connection to the same switch
  - Support of IEEE 802.3ad by both switches
  - Full duplex mode (send and receive paths)
  - VLANs considerations
    - All member OSA ports within the group must be trunk links to provide the virtual LAN connectivity in which to flow tagged traffic
    - Aggregated link should be viewed as one logical trunk link containing all the VLANs required by the LAN segment

New OSA Express2 Hardware Feature

Exclusive Port Mode

Single QDIO Connection

The ability to establish an exclusive QDIO connection on an OSA port . Once the connection is established, the port can no longer be shared within this or any other LPAR. Any attempt to establish another connection on the port will be prevented as long as the exclusive QDIO connection is active.

Automatic Port Disablement / Enablement

When an exclusive QDIO connection leaves the “QDIO Active” state, the OSA port will be automatically disabled until the next QDIO connection is established. By disabling the OSA port, the connected switch port is notified the link is no longer operational. This provides a signal to the partner switch to route future traffic to another port within the group.

Simple Virtual Switch LAN Segment (VSWITCH)

Create a simulated Layer 2 or Layer 3 switch device

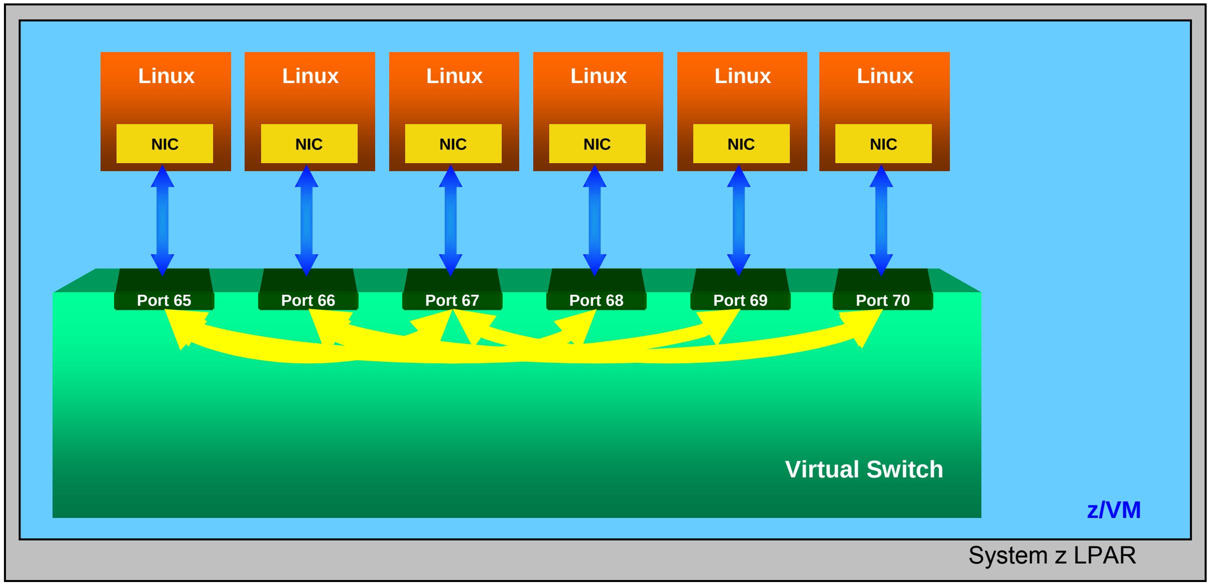
Virtual machine access control and VLAN authorization

Create ports and connect NIC to virtual switch (LAN Segment)

Provides full MAC address management (generation and assignment)

Forwards traffic between Guest Ports by either IP or MAC address

1-n VSWITCHs per z/VM image



Example

Create VSWITCH from PRIVCLASS B User ID

```
DEF VSWITCH VSWITCH1 ETHERNET
SET VSWITCH VSWITCH1 GRANT {user ID}
```

From Linux Virtual Machines

```
DEF NIC 100 TYPE QDIO
COUPLE 100 SYSTEM VSWITCH1
```

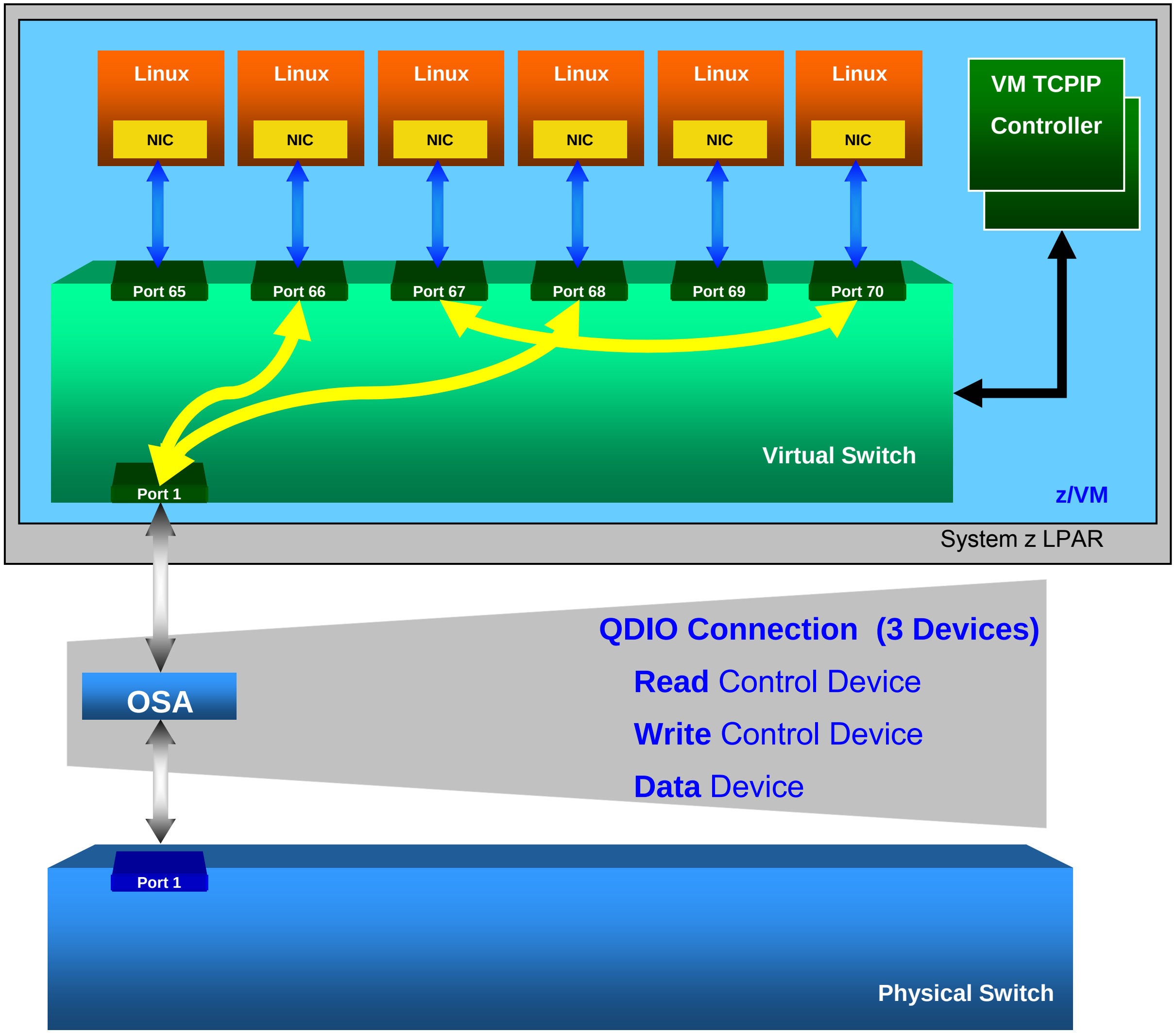
Cascading a Virtual to a Physical Switch

Start VM TCPIP Controllers

XAUTOLOG DTCVSW1
XAUTOLOG DTCVSW2

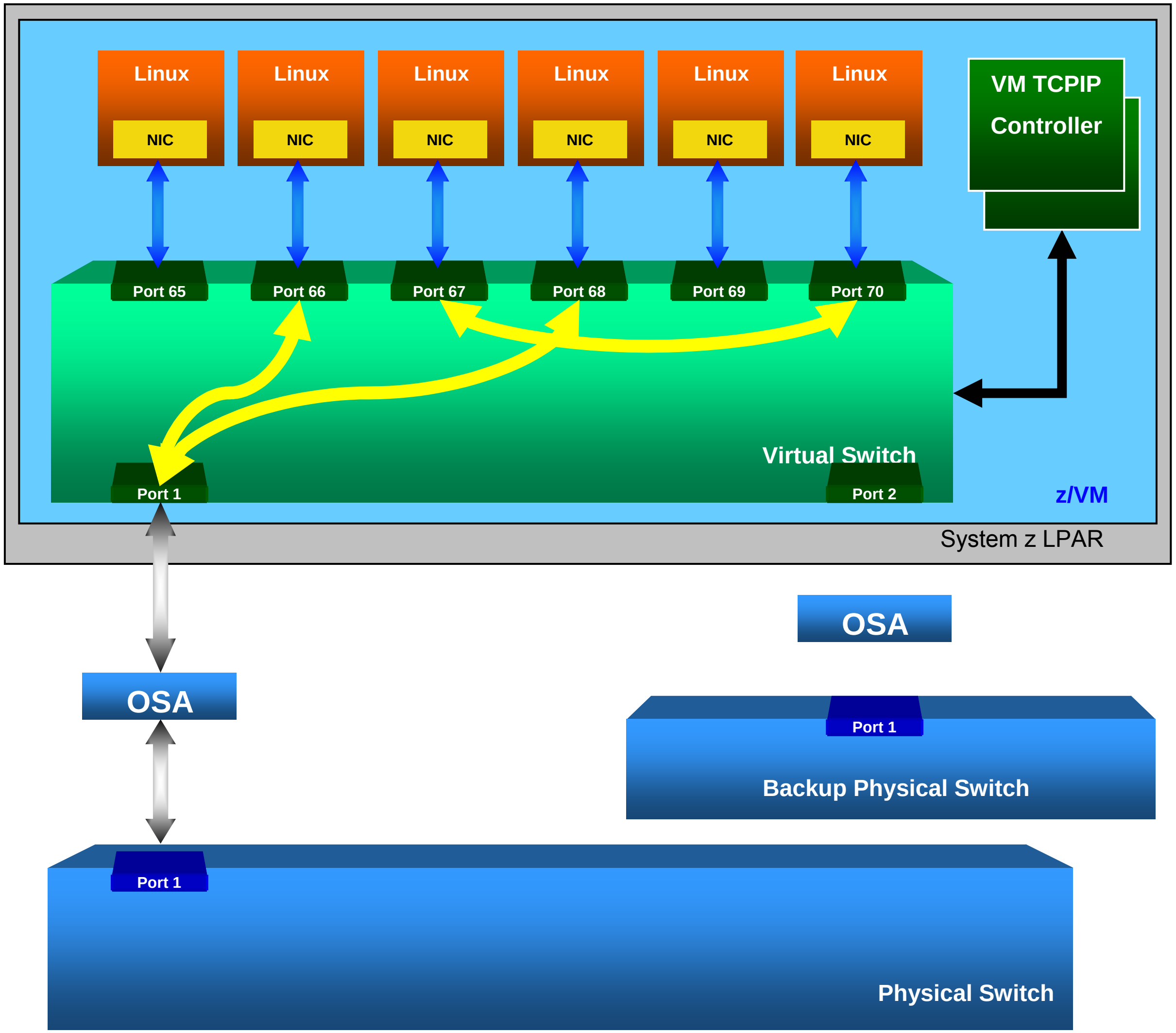
Connect the Real Switch

SET VSWITCH VSWITCH1 RDEV 100



Adding a Failover Device

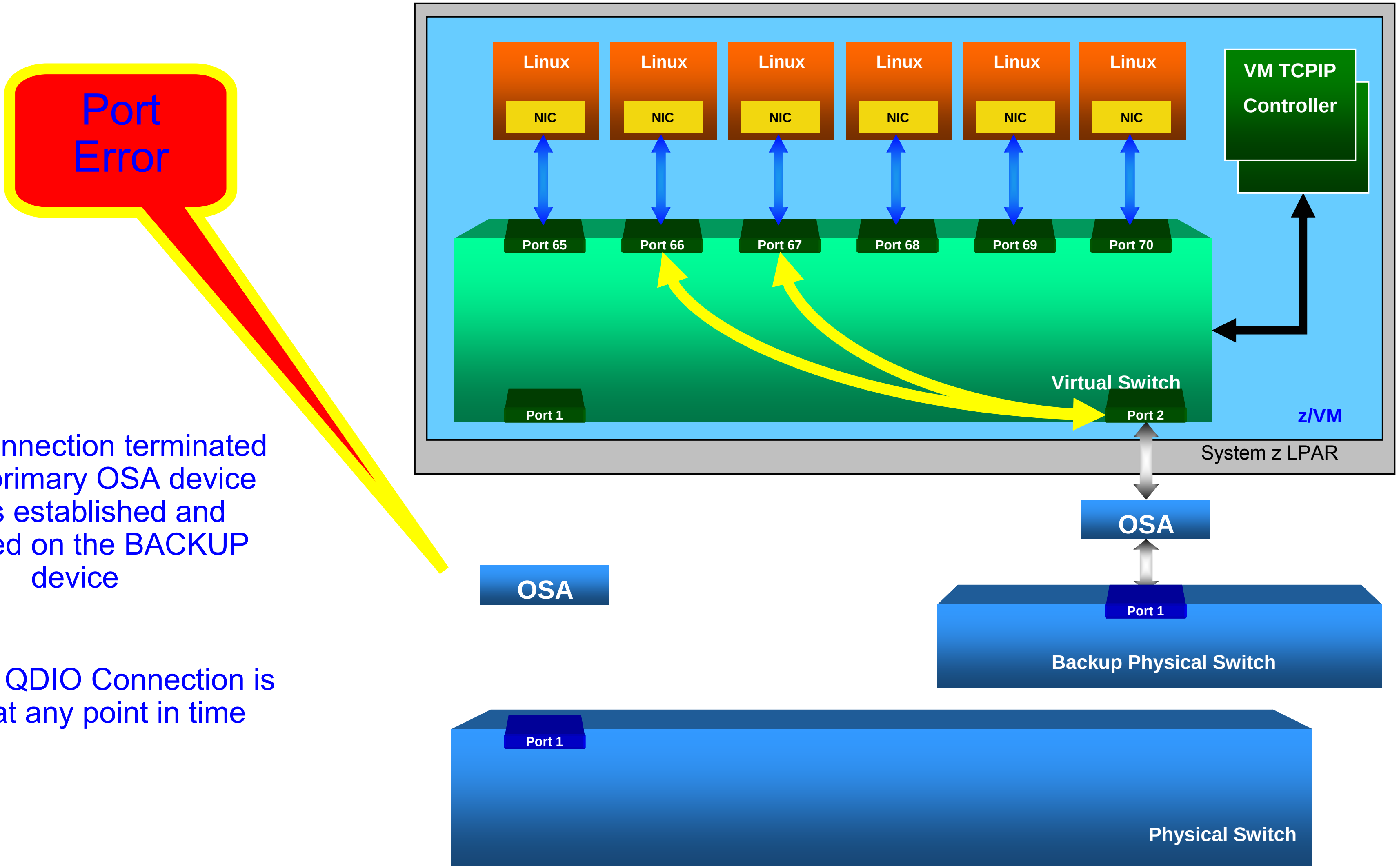
Issue the SET VSWITCH command and include the new RDEV



Example

```
SET VSWITCH VSWITCH1 RDEV 100 500
SET VSWITCH VSWITCH1 CONNECT
```

Port Failover



QDIO connection terminated on the primary OSA device and is established and activated on the BACKUP device

Only one QDIO Connection is active at any point in time

Defining Port Groups

Two step process to create a LinkAG port configuration

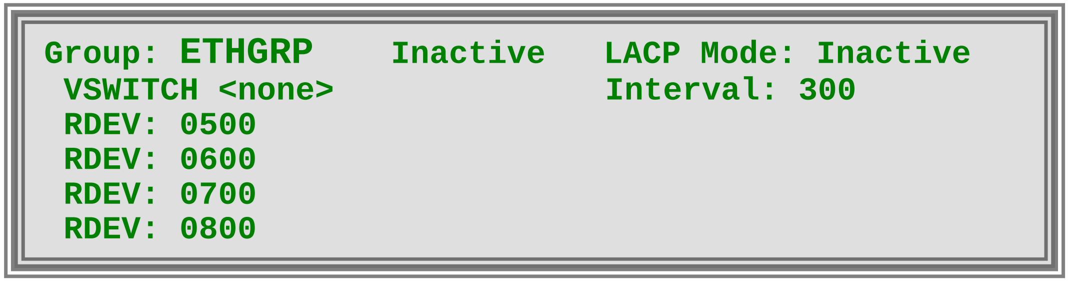
1. Create a port group using new SET PORT CP Command
2. Associate a port group with an ETHERNET type VSWITCH

Create a Port Group

```
SET PORT GROUP ETHGRP JOIN 500 600 700 800
SET PORT GROUP ETHGRP LACP INACTIVE
```

Display INACTIVE Port Groups

```
Q PORT GROUP INACTIVE
```



```
Group: ETHGRP     Inactive    LACP Mode: Inactive
 VSWITCH <none>               Interval: 300
 RDEV: 0500
 RDEV: 0600
 RDEV: 0700
 RDEV: 0800
```

Display ACTIVE Port Groups

```
Q PORT GROUP
```

```
HCPSWP2837E No active groups found.
```

SET or MODIFY PORT GROUP

Use the SET or MODIFY PORT command to define or change the OSA Express2 devices that make up a link aggregation group and to set the attributes of a link aggregation group.

```
 Privilege Class: B                        +-------+
                                           |       |(1)
                                           v       |
>>---SET-PORT-GROup groupname -+- JOIn --+---rdev --------+---><
                               +- LEAve -+                |
                               +- DELete -----------------+
                               +- LACP -+- ACTive -----+--+
                               |        +- INActive ---+  |
                               +- INTerval--+- nnnn -+----+
                                            +- OFF --+
                                                (2)
```

Note: (1) You can specify a maximum of 8 real device numbers
(2) Operands that may be specified while the group is ACTIVE

QUERY PORT GROUP CP Command

Use the QUERY PORT command to display information about link aggregation groups or devices that have been defined for virtual switches on the system.

```
Privilege Class: B
```

```
                         +-ALL--ACTive-------+
>--Query--PORT--+-GROup--+-------------------+-+--+---------+-----><
                |        |      +-ACTive---+ | |  +-DETails-+
                |        +-ALL--+----------+-| |
                |        |      +-INActive-+ | |
                |        +-groupname---------+ |
                '-+------------+---------------+
                  +-RDEV--rdev-+
```

Display Routing Table

Query PORT GROup *name* DETails

```
Group: ETHGRP           Active            LACP Mode: Active
 VSWITCH SYSTEM SWITCH1  Interval: 300
 GROUP Information:
   PORT Information - Total Frames per Interval:
     Device  Status     Previous
      0510   Active     11            7
      0520   Active     11            7
   ROUTING Information - Frame Distribution per Interval:
     MAC     Device     Previous      Current
      0      0510       0             0
      1      0520       0             0
      2      0510       0             0
      3      0520       0             0
      4      0510       0             0
      5      0520       0             0
      6      0510       0             0
      7      0520       0             0
```

LACP INACTIVE LinkAG Group

Associate a port group with an ETHERNET type VSWITCH

Disconnect the Physical Switch

SET VSWITCH VSWITCH1 DISCON

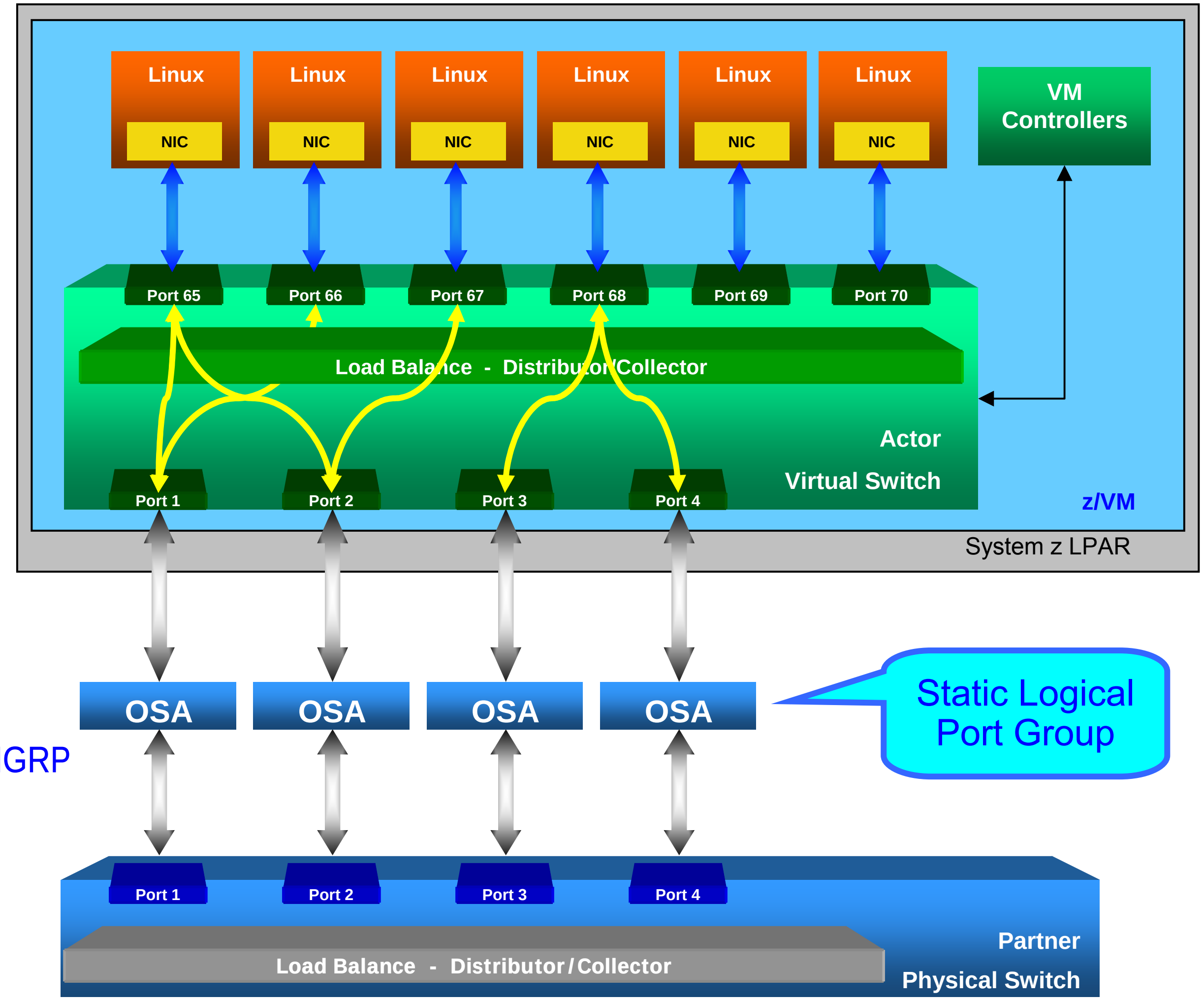
Setup Partner Switch for a LACP INACTIVE port

Associate the Port Group

SET VSWITCH VSWITCH1 GROUP ETHGRP

Connect the Port Group

SET VSWITCH VSWITCH1 CONNECT



LACP ACTIVE LinkAG Group

Create a Dynamically Managed LinkAG Port Group

Disconnect the Physical Switch

SET VSWITCH VSWITCH1 DISCON

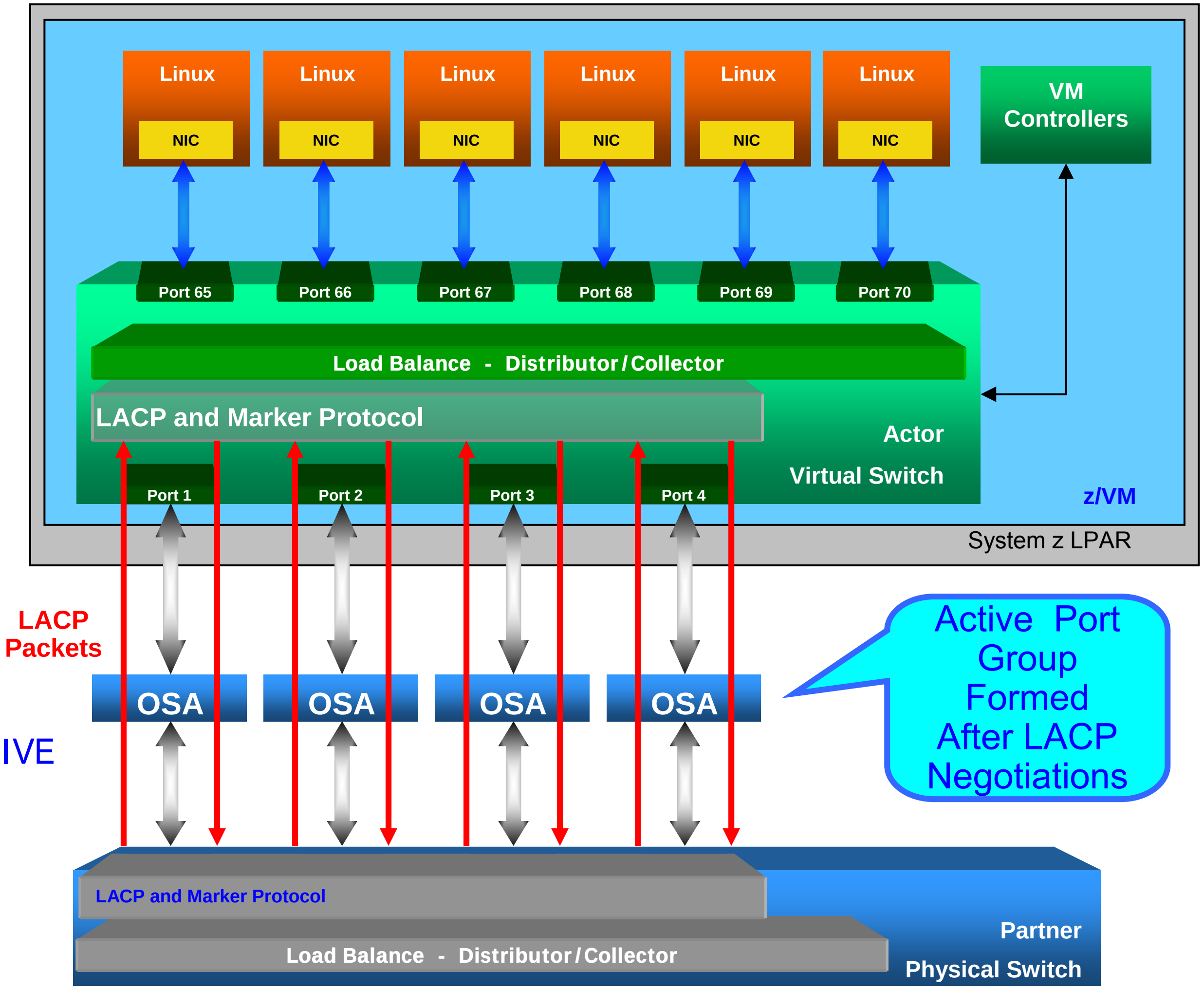
Setup Partner Switch for a LACP ACTIVE port

Make Port Group LACP ACTIVE

SET PORT GROUP ETHGRP LACP ACTIVE

Connect the Port Group

SET VSWITCH VSWITCH1 CONNECT



Switch Failover to Traditional Backup Device

LinkAG group can be setup to failover to a single port on another switch

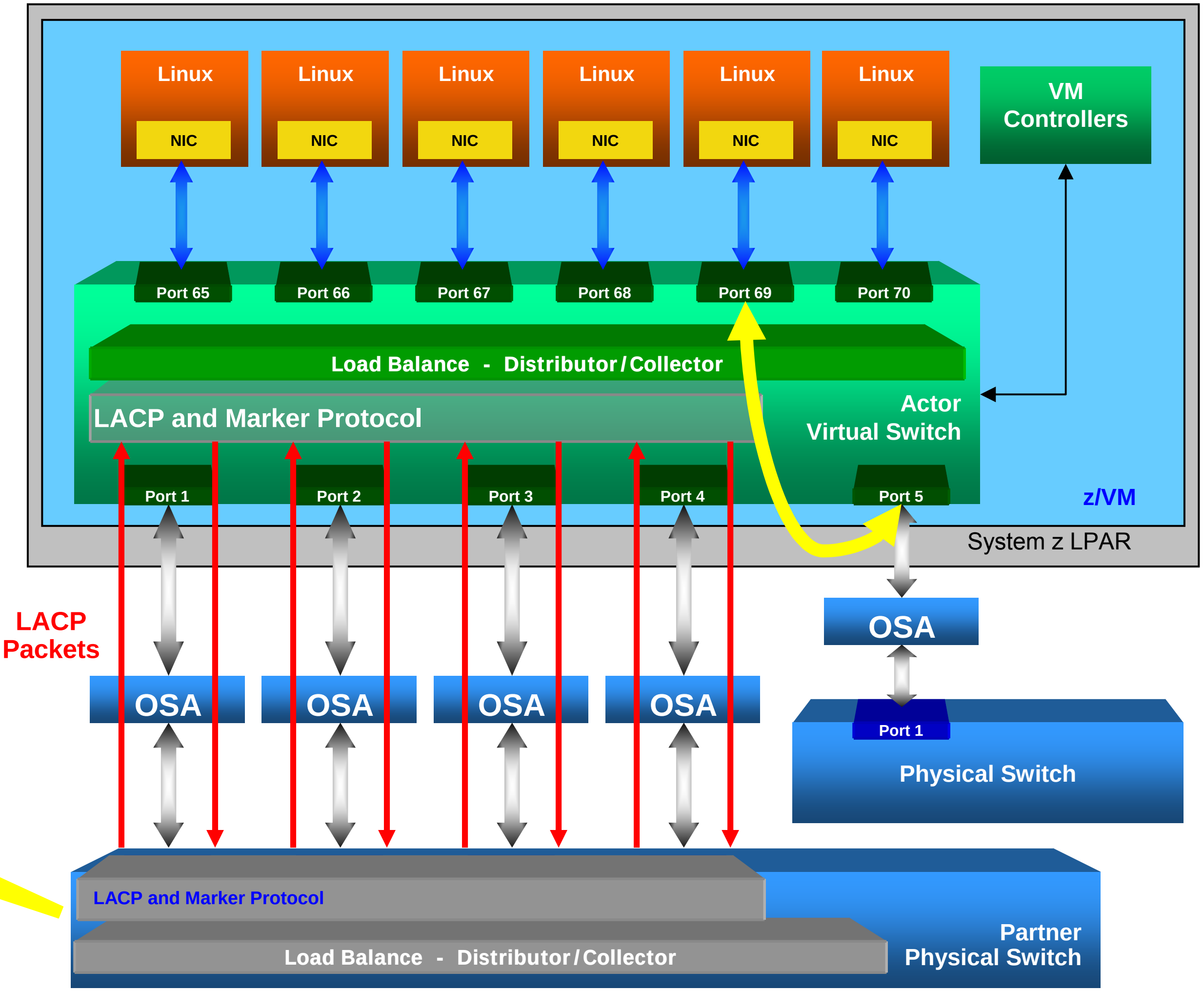
Select another physical switch on the same LAN segment

Add the BACKUP device

SET VSWITCH VSWITCH1 RDEV 100

Switch Error

VM automatically establishes and activates the QDIO connection on the BACKUP device



Advantages of a LACP ACTIVE Port Group (Recommended)

- Ports can be added or removed dynamically within the LinkAG group
  - Changes made on one switch are automatically made on the other switch
  - Immediate packet rerouting
- Fast near seamless failover to another port within the group
- Adding or removing capacity is not disruptive
- LACP Protocol provides a heartbeat mechanism
- Marker Protocol allows greater flexibility to dynamically move work from one port to another within the group
- Automatic fail-back from the backup device to a port group

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!