Anatomy of a z Penguin

A Customer Experience

Helping A Colony Thrive Under Extreme Conditions

Rick Barlow

Richard.Barlow@nationwide.com

August 14, 2008

Session 9213

Overview and Disclaimer

Disclaimer:

The content of this presentation is for information only and is not intended to be an endorsement by Nationwide Insurance. Each site is responsible for their own use of the concepts and examples presented.

Overview:

With a few exceptions, this is an overview! Where possible there are technical details you may be able to use. As you frequently hear, when anyone asks for recommendations, "IT DEPENDS"!

The information in this session is based on my experiences as a long-time VM-er adding virtual Linux.

Interactive is good! Please ask questions. We'll all get the most out of this session that way.

Topics

- Our Environment
- Simple “Logical” TCO - Why Virtualize Hardware?
- Virtual Networking
- High Availability
- Disaster Recovery Enablement
- Performance
- Conclusions

Our Environment

– Then - two z990 installed in 2005, each with:

- § Development box
  - o 5 IFL engines on development box
    Grew to 8 and then to 16
  - o 64GB memory
    Grew to 120GB
  - o 5 z/VM LPARs (sandbox LPAR for system programmer test)
- § Production box
  - o 3 IFLs
    Grew to 7 and then to 15
  - o 56GB memory
    Grew to 112GB
  - o 4 z/VM LPARs

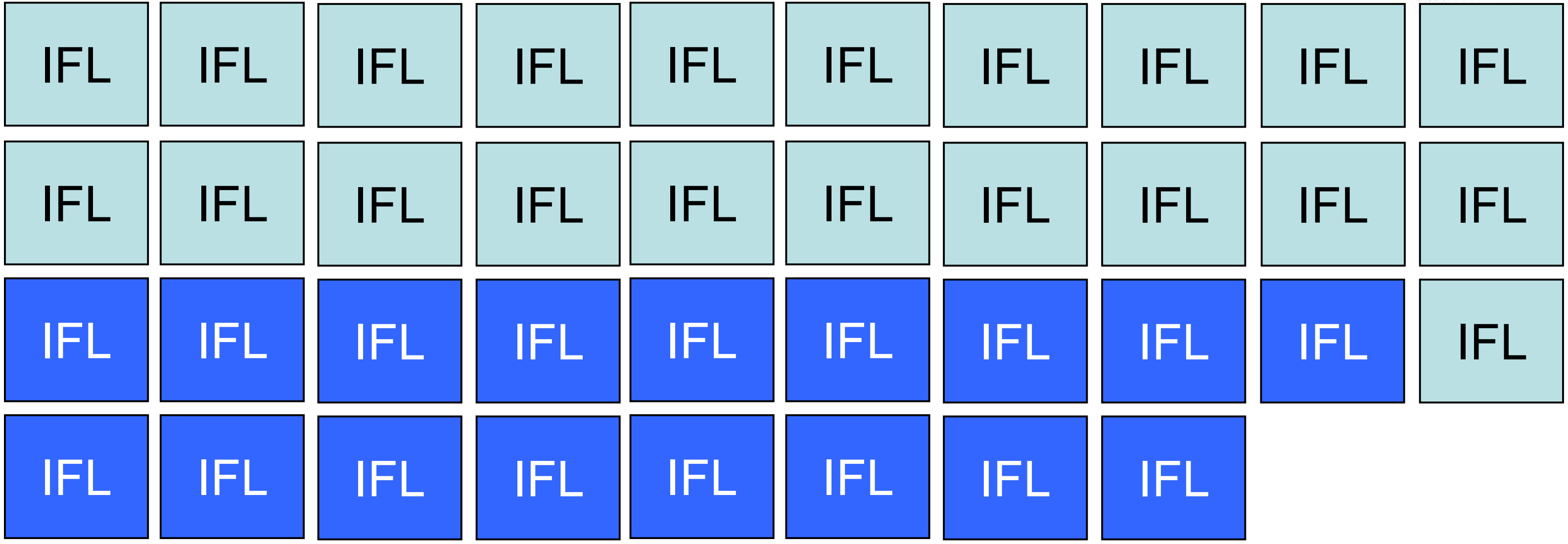
Environment

– Upgrade to two z9 in 4Q2006, each with:

- § Development box
  - o 8 IFL engines on development box
    Grew through several upgrades - now 17
  - o 128GB memory
    Grew through several upgrades - now 352GB
  - o 5 z/VM LPARs (sandbox LPAR for system programmer test)
- § Production box
  - o 7 IFLs
    Grew through several upgrades - now 18
  - o 128GB memory
    Grew through several upgrades - now 240GB
  - o 4 z/VM LPARs
- § Growing FAST!

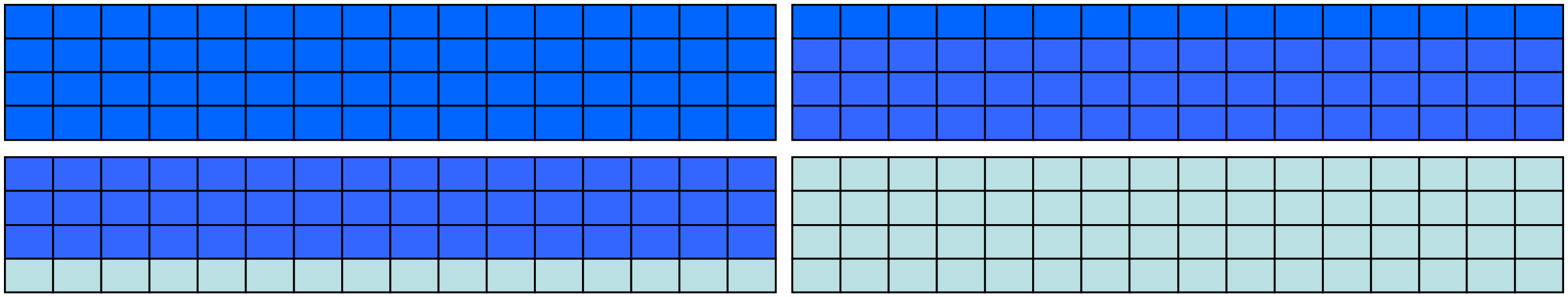
IBM z9 Platform (Test / Dev)

Processors
17 IFLs
Max 38
(45%)

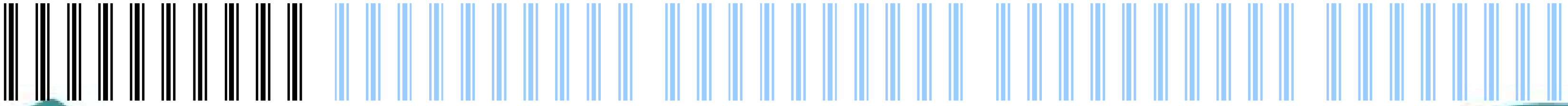

IFL IFL IFL IFL IFL IFL IFL IFL IFL IFL
IFL IFL IFL IFL IFL IFL IFL IFL IFL IFL
IFL IFL IFL IFL IFL IFL IFL IFL IFL IFL
IFL IFL IFL IFL IFL IFL IFL IFL

Max MIPS on the z9 is >13,000. System z Linux (dev/test + prod) has already out-MIPd the z/OS & z/VM traditional z environments combined!

Memory
352GB
Max 512GB
(69%)



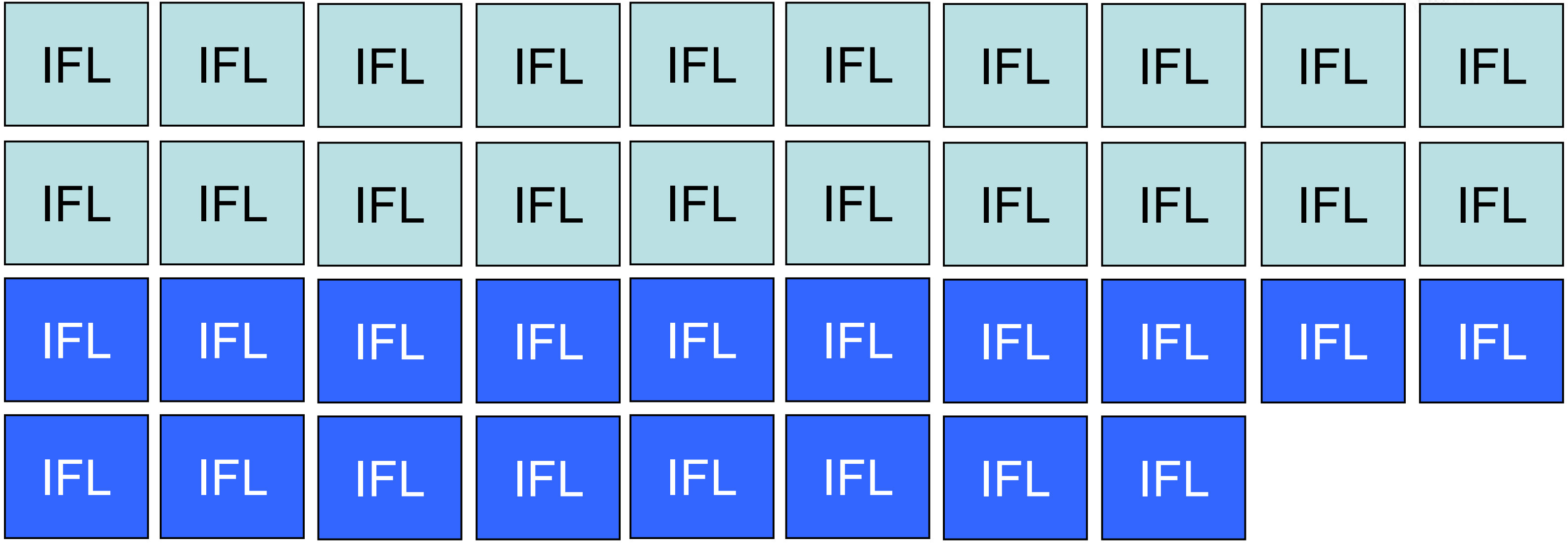
Network
10 OSA Cards
Max 48
(20%)



This information is for sharing only and not an endorsement by Nationwide Insurance

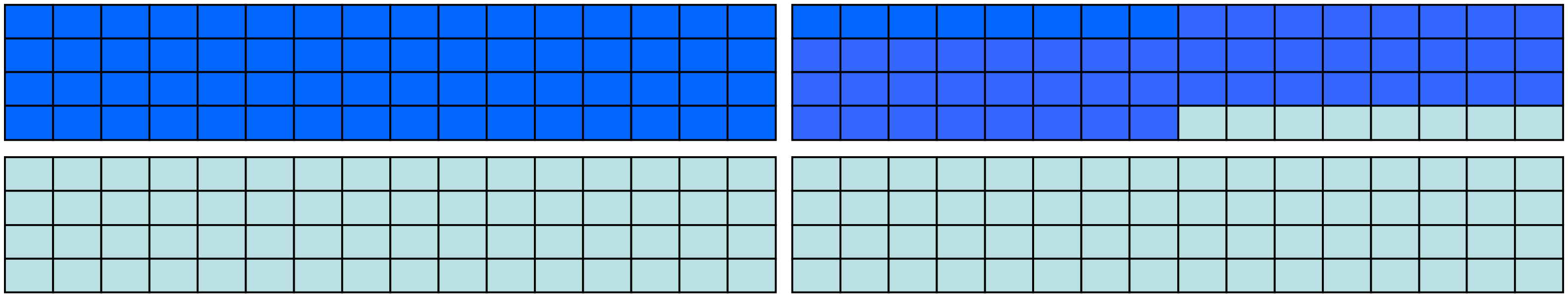
IBM z9 Platform (Prod)

Processors
18 IFLs
Max 38
(47%)

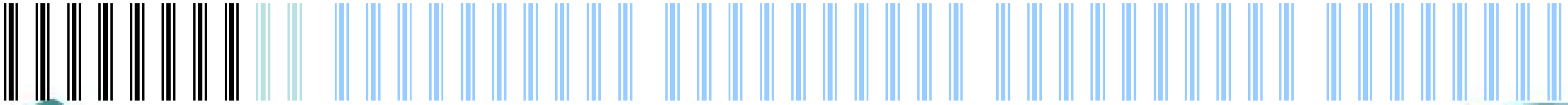


IFL IFL IFL IFL IFL IFL IFL IFL IFL IFL
IFL IFL IFL IFL IFL IFL IFL IFL IFL IFL
IFL IFL IFL IFL IFL IFL IFL IFL IFL IFL
IFL IFL IFL IFL IFL IFL IFL IFL

Memory
240GB
Max 512GB
(49%)



Network
6 OSA Cards
Max 48
(13%)

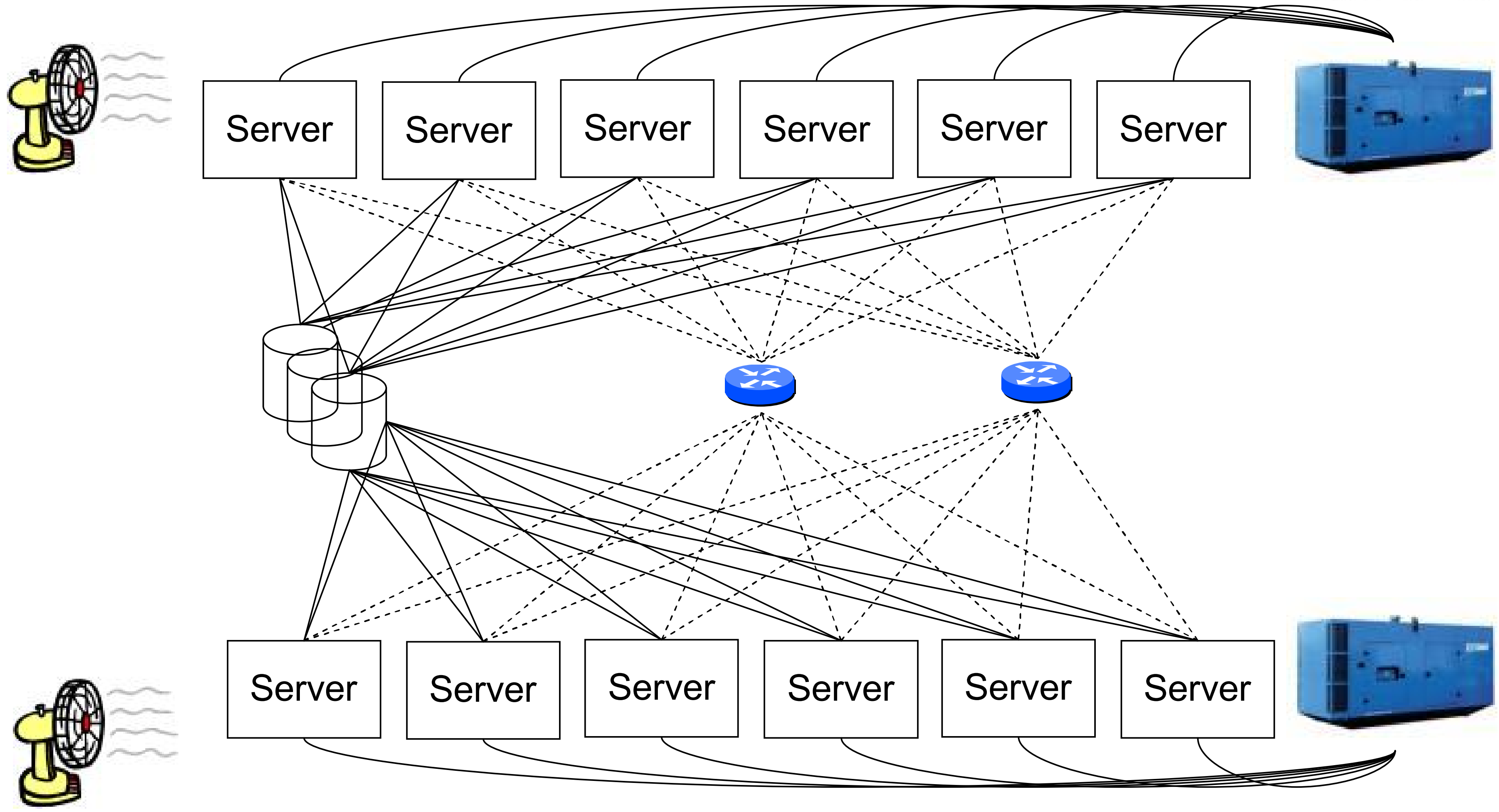


Simple “Logical” TCO
Why Hardware Virtualization?

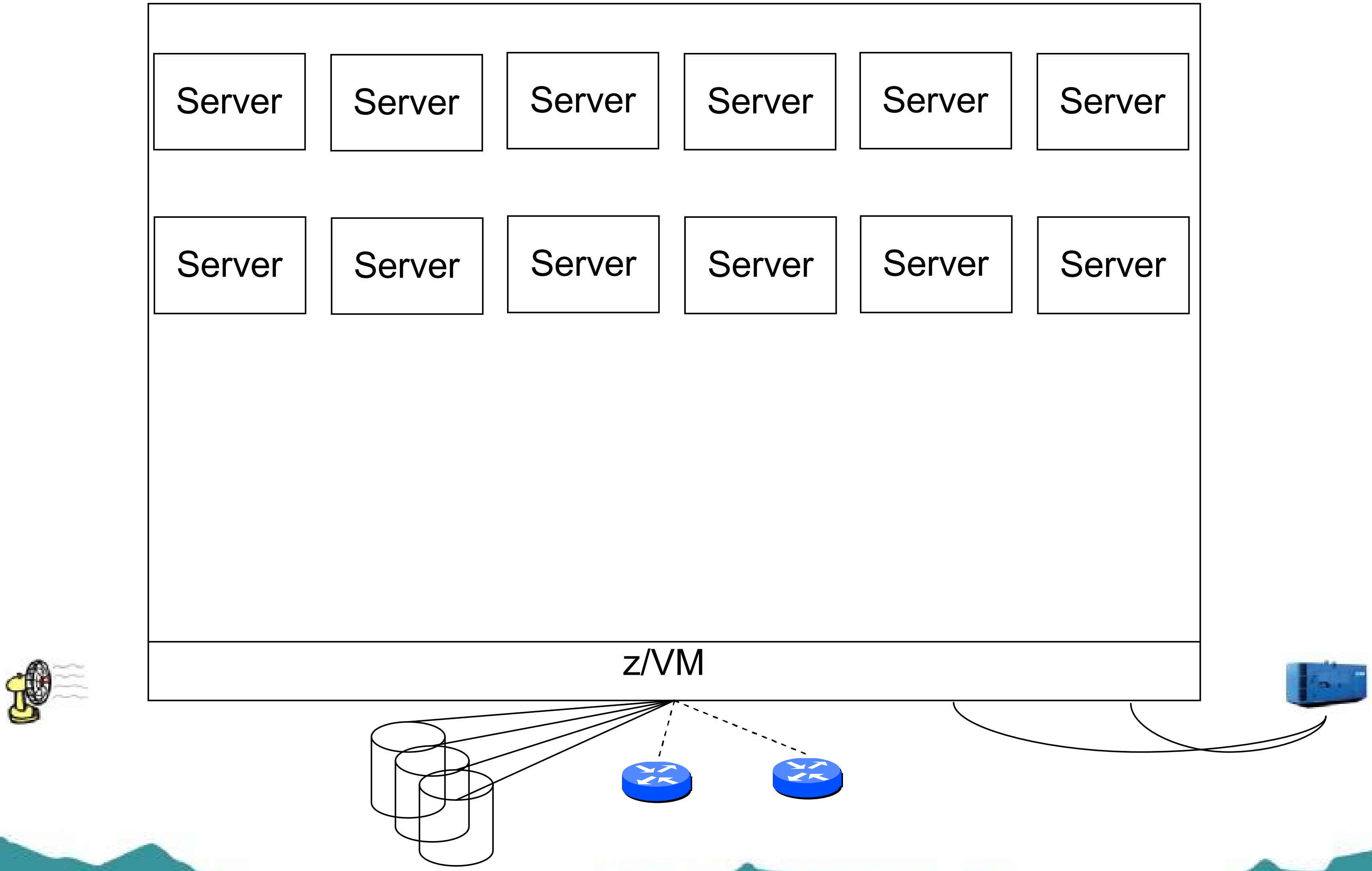
- Reduce complexity
  - Physical servers
  - Network connections
  - Disk connections
- Reduce facility resources
  - Floor space
  - Power consumption
  - Cooling
- Opportunities
  - Shared disk
  - Shared memory
  - Reduced total capacity because of sharing

This information is for sharing only and not an endorsement by Nationwide Insurance

Distributed Server Model



Virtual Server Model



Why Hardware Virtualization?

- 19” Server Rack with 1U x86 (e.g. HP Proliant DL360 G4)
  - Physical width 20” per rack
  - 40 servers per rack
  - Max servers in 3 racks is 120
  - 585W * 40 * 3 = 70.2kw
  - 2kBTU/hr * 40 * 3 = 240kBTU/hr

Note: The numbers on this and the following charts are approximate and are based on information located on vendor sites on the Internet.

Why Hardware Virtualization?

- z9 EC – Model 2094-S38
  - Physical width 60”; about 3 19” racks
  - 332 servers on test/development using 42% of IFLs and 69% of memory
    151 servers on production using 47% of IFLs and 49% of memory
  - Max servers per z9 is:
    - 480 for test/development (332/69% of memory currently used)
    - 310 for production (151/49% of memory currently used)
  - 18.3kw
  - 62.2kBTU/hr
    - z9 live power consumption display show 35kBTU on test/dev box

Why Hardware Virtualization?

- 19” Server Rack with 1U x86 (e.g. HP Proliant DL360 G4)
  - Software (25% web server; 55% application server; 20% database server)
    Assumption: 2/3 of servers are at least dual processor
    - 120 licenses for OS
    - (30 web server licenses) – depending on the tool
    - 66 WAS licenses
    - 50 DB2 or Oracle licenses (no DB servers with only one processor)
    - ? Miscellaneous software licenses (e.g. MQ)

Why Hardware Virtualization?

- z9 EC – Model 2094-S38
  - Software (25% web server; 55% application server; 20% database server)
    - 38 licenses for OS
    - (30 web server licenses) – depending on the tool
    - 38 WAS licenses
    - 38 DB2 or Oracle licenses
    - ? Miscellaneous software licenses (e.g. MQ)

Why Hardware Virtualization?

- 19” Server Rack with 1U x86 (e.g. HP Proliant DL360 G4)
  - Cabling (25% web server; 55% application server; 20% database server)
    - 91-182 SAN connections (91 servers with SAN connection; many dual path)
      Assumption: All DB and half Application servers use SAN disk
    - 120+ Ethernet connections (1 per server minimum; most have >1)
    - Serial connections for console most likely

Why Hardware Virtualization?

- z9 EC – Model 2094-S38
  - Cabling (25% web server; 55% application server; 20% database server)
    - 8 4Gb FCP SAN connections
    - 10 Ethernet connections

Virtual Networking

- Overcoming Terminology
  - VLAN, VLAN, Guest LAN
    - § VLAN – native, hardware, management – the one the routers, switches and OSAs use
    - § VLAN – logical – the ones used to separate / isolate servers
    - § Guest LAN – a VM emulation of a network
  - Switches, Routers , VSWITCH
    - § Switch – a device that acts as a connector to create a network
    - § Router – a device that forwards data packets between computer networks
    - § VSWITCH – a logical extension of the physical network inside the System z

Virtual Networking

- System z Hardware
  - Open System Adapter (OSA) Express 2
    - § Gigabit adapter with a smart network controller
    - § System z LPAR microcode allows:
      - o Sharing of the same OSA across LPARs
      - o Multiple Read / Write / Data groups to be attached to virtual server or defined as a VSWITCH
    - § Gigabit Ethernet
      - o Fiber
    - § 1000BaseT
      - o Copper Cat6
      - o Can be configured as Integrated Console Controller (ICC)

Virtual Networking

- System z Hardware with z/VM
  - Virtual Switch (VSWITCH)
    - Combination of System z microcode and z/VM CP code to create an extension of a network switch
    - Layer 3
      - Forwarding based on IP address
      - Sufficient for most implementations
      - Defined as "IP"
      - Common MAC included for all guests
      - z/VM 4.4.0 or higher

Virtual Networking

- System z Hardware with z/VM
  - Virtual Switch (VSWITCH)
    - Layer 2
      - Forwarding based on MAC address
      - Allows non-IP protocols like NETBIOS or IPX
      - Defined as "Ethernet"
      - New on z990 with OSA Express 2 and z/VM 5.1.0
      - Recommended by IBM
      - Unique MAC for each virtual server
        - Local MAC addressing must be administered
      - z/VM TCPIP cannot connect to a Layer 2 VSWITCH

Virtual Networking

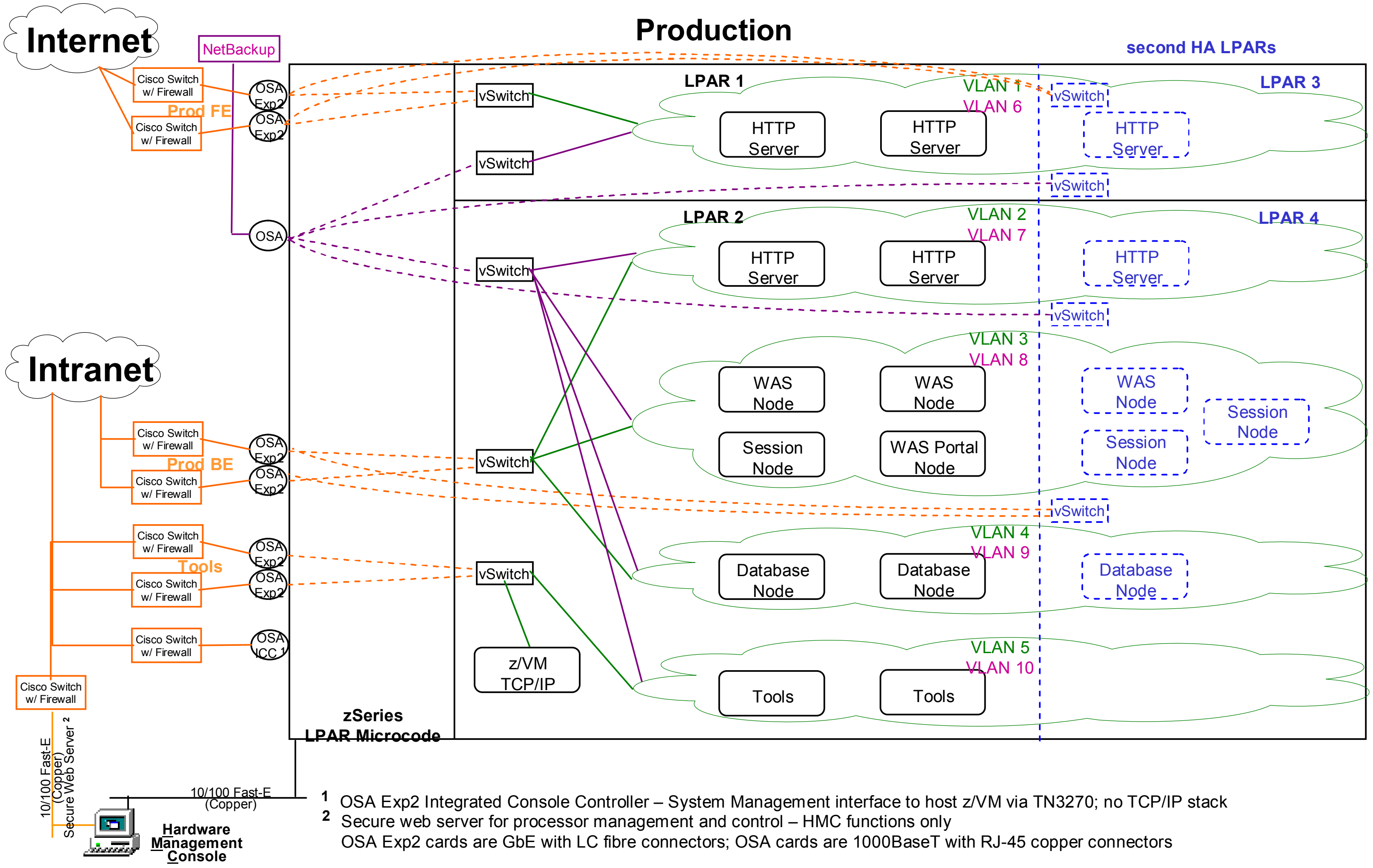
- z/VM
  - Guest LAN
    - § Use to create isolated LAN within a z/VM LPAR
      - o Can be owned by SYSTEM or a virtual machine
      - o Can be restricted to authorized users or open to anyone
    - § HIPERSOCKET – emulate System z HIPERSOCKET hardware
    - § QDIO – emulate gigabit ethernet
    - § Define in CP SYSTEM CONFIG or by CP command (syntax is the same)

```
DEFINE LAN GLAN1 OWNERID SYSTEM TYPE HIPERS MAXCONN INFINITE
DEFINE LAN GLAN2 OWNERID SYSTEM TYPE QDIO MAXCONN INFINITE
```

Our Network

- Our OSA / VSWITCH configuration
  - Production has 3 cards / 6 ports OSA Express 2 Gigabit Ethernet cards
  - Test/Development has 6 cards / 12 ports OSA Express 2 Gigabit Ethernet cards
  - 2 OSA Express 1000BaseT (2 ports each; 2 defined as ICC)
  - 6 different network zones; 12 VSWITCHes defined
    - § 2 VSWITCHes on each pair of OSA ports for redundancy and load distribution
      - o Paired OSA ports are on separate cards for redundancy
    - § Each pair of ports is in a specific network zone
      - o Each OSA port in a pair is connected to a different physical switch

Our Network



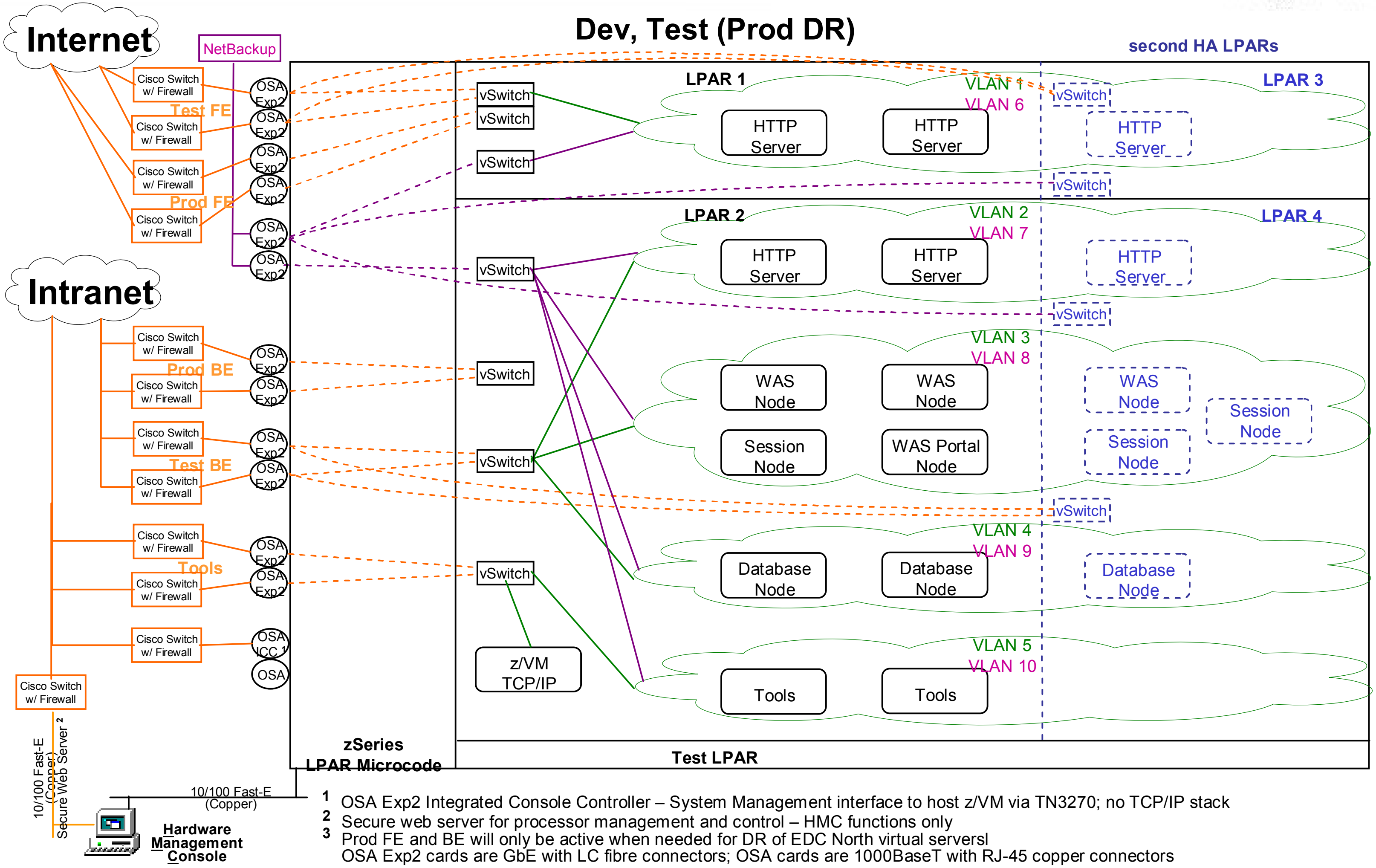
[1] OSA Exp2 Integrated Console Controller – System Management interface to host z/VM via TN3270; no TCP/IP stack

[2] Secure web server for processor management and control – HMC functions only

OSA Exp2 cards are GbE with LC fibre connectors; OSA cards are 1000BaseT with RJ-45 copper connectors

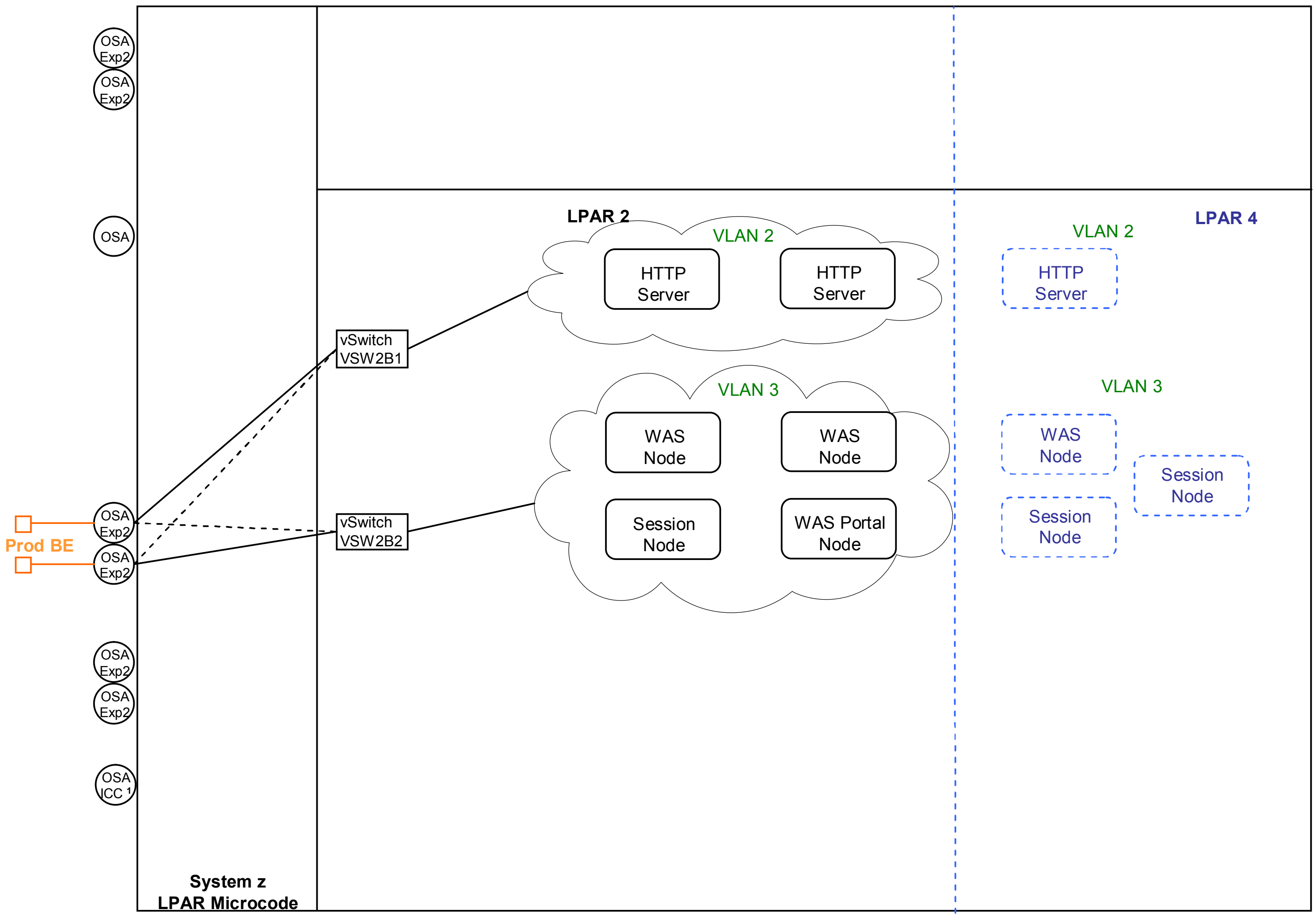
This information is for sharing only and not an endorsement by Nationwide Insurance

Our Network

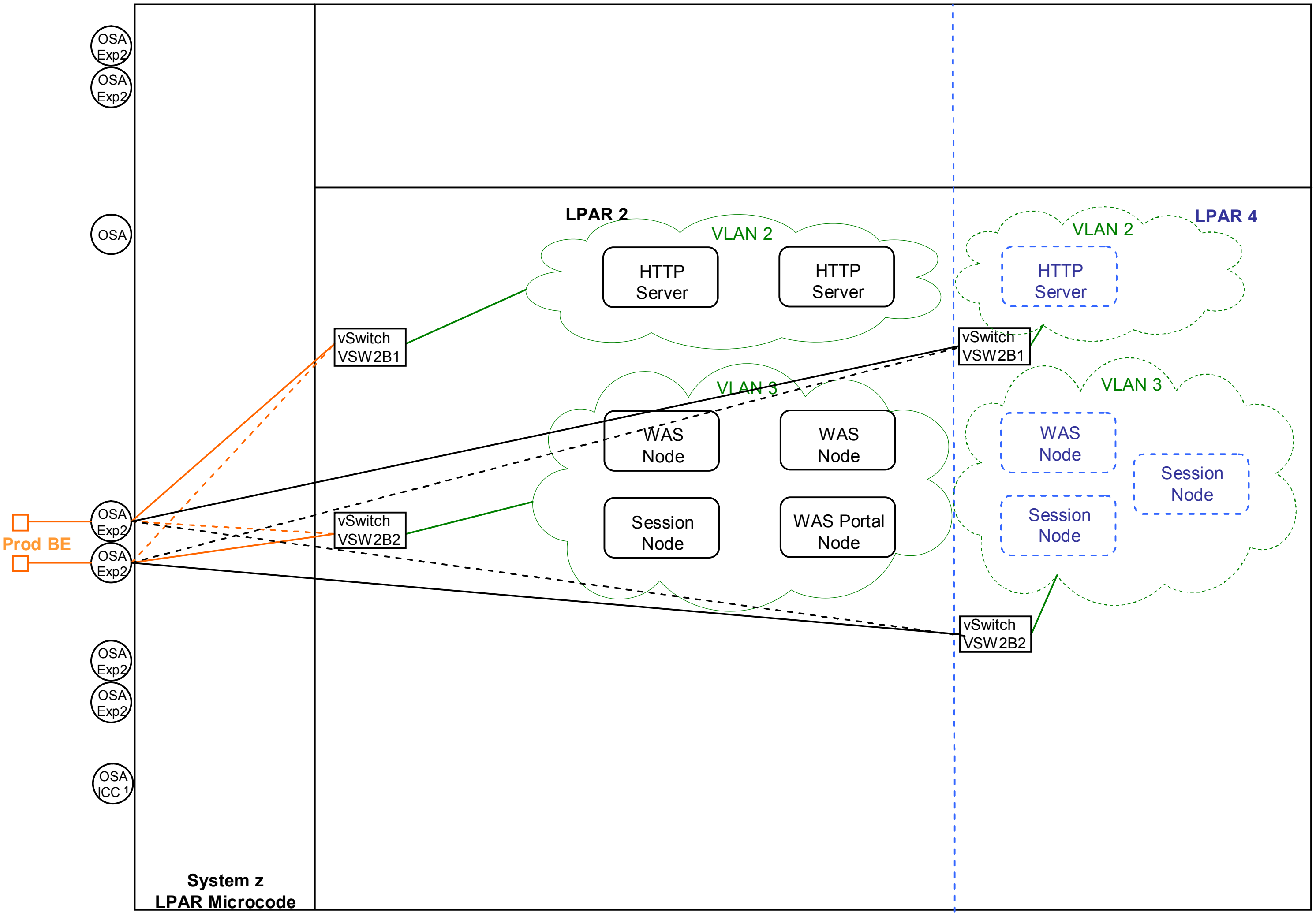


1 OSA Exp2 Integrated Console Controller – System Management interface to host z/VM via TN3270; no TCP/IP stack
2 Secure web server for processor management and control – HMC functions only
3 Prod FE and BE will only be active when needed for DR of EDC North virtual serversl
OSA Exp2 cards are GbE with LC fibre connectors; OSA cards are 1000BaseT with RJ-45 copper connectors

Network


OSA Exp2
OSA Exp2
OSA
LPAR 2
VLAN 2
HTTP Server
HTTP Server
LPAR 4
VLAN 2
HTTP Server
vSwitch VSW2B1
VLAN 3
WAS Node
WAS Node
VLAN 3
WAS Node
Session Node
Session Node
Session Node
WAS Portal Node
Prod BE
OSA Exp2
OSA Exp2
vSwitch VSW2B2
OSA Exp2
OSA Exp2
OSA ICC 1
System z LPAR Microcode

Network


OSA Exp2
OSA Exp2
OSA
Prod BE
OSA Exp2
OSA Exp2
OSA Exp2
OSA Exp2
OSA ICC 1
System z LPAR Microcode
LPAR 2
VLAN 2
HTTP Server
HTTP Server
vSwitch VSW2B1
VLAN 3
WAS Node
WAS Node
Session Node
WAS Portal Node
vSwitch VSW2B2
LPAR 4
VLAN 2
HTTP Server
vSwitch VSW2B1
VLAN 3
WAS Node
Session Node
Session Node
vSwitch VSW2B2

VSWITCH Detail

- Defining VSWITCH
  - In SYSTEM CONFIG or via CP command by authorized user
    (same syntax in both places)
    - Example of a pair of VSWITCHes:

```
CP DEFINE VSWITCH VSW2B1 RDEV C100 C204 CONTROLLER * IP VLAN 4094
CP DEFINE VSWITCH VSW2B2 RDEV C200 C104 CONTROLLER * IP VLAN 4094
```

    - VLAN on the VSWITCH is the default VLAN used by the hardware switches.

VSWITCH Detail

- Authorizing virtual servers to use VSWITCH
  - SYSTEM CONFIG format
    - § Example 1: 2 virtual servers in same zone on opposite VSWITCHes

```
MODIFY VSWITCH VSW2B1 GRANT LINSERV1 VLAN 1001
MODIFY VSWITCH VSW2B2 GRANT LINSERV2 VLAN 1001
```

    - § Example 2: 1 virtual server on 2 VSWITCHes in different zones

```
MODIFY VSWITCH VSW2B1 GRANT LINSERV1 VLAN 1001
MODIFY VSWITCH VSW2F1 GRANT LINSERV1 VLAN 2001
```

  - CP command format
    - § Example 1: 2 virtual servers in same zone on opposite VSWITCHes

```
CP SET VSWITCH VSW2B1 GRANT LINSERV1 VLAN 1001
CP SET VSWITCH VSW2B2 GRANT LINSERV2 VLAN 1001
```

    - § Example 2: 1 virtual server on 2 VSWITCHes in different zones

```
CP SET VSWITCH VSW2B1 GRANT LINSERV1 VLAN 1001
CP SET VSWITCH VSW2F1 GRANT LINSERV1 VLAN 2001
```

VSWITCH Detail

- Defining Guest NIC
  - CP DIRECTORY format

```
NICDEF 5708 TYPE QDIO DEVICES 3 LAN SYSTEM TOOL2
NICDEF 1E00 TYPE QDIO DEVICES 3 LAN SYSTEM
NETBKUP1
```

  - CP command format

```
CP DEFINE NIC 5708 TYPE QDIO DEVICES 3
CP COUPLE 5708 TO SYSTEM TOOL2
CP DEFINE NIC 1E00 TYPE QDIO DEVICES 3
CP COUPLE 1E00 TO SYSTEM NETBKUP1
```

VSWITCH Detail – Linux definitions

- Hardware configuration script

```
cat /etc/sysconfig/hardware/hwcfg-qeth-bus-ccw-0.0.5708
#!/bin/sh
#
# hwcfg-qeth-bus-ccw-0.0.5708
#
# Hardware configuration for a qeth device at 0.0.5708
# Automatically generated by netsetup
#
STARTMODE="auto"
MODULE="qeth"
MODULE_OPTIONS=""
MODULE_UNLOAD="yes"
# Scripts to be called for the various events.
SCRIPTUP="hwup-ccw"
SCRIPTUP_ccw="hwup-ccw"
SCRIPTUP_ccwgroup="hwup-qeth"
SCRIPTDOWN="hwdown-ccw"
# CCW_CHAN_IDS sets the channel IDs for this device
# The first ID will be used as the group ID
CCW_CHAN_IDS="0.0.5708 0.0.5709 0.0.570a"
# CCW_CHAN_NUM set the number of channels for this device
# Always 3 for an qeth device
CCW_CHAN_NUM=3
# CCW_CHAN_MODE sets the port name for an OSA-Express device
CCW_CHAN_MODE="suselin7"
```

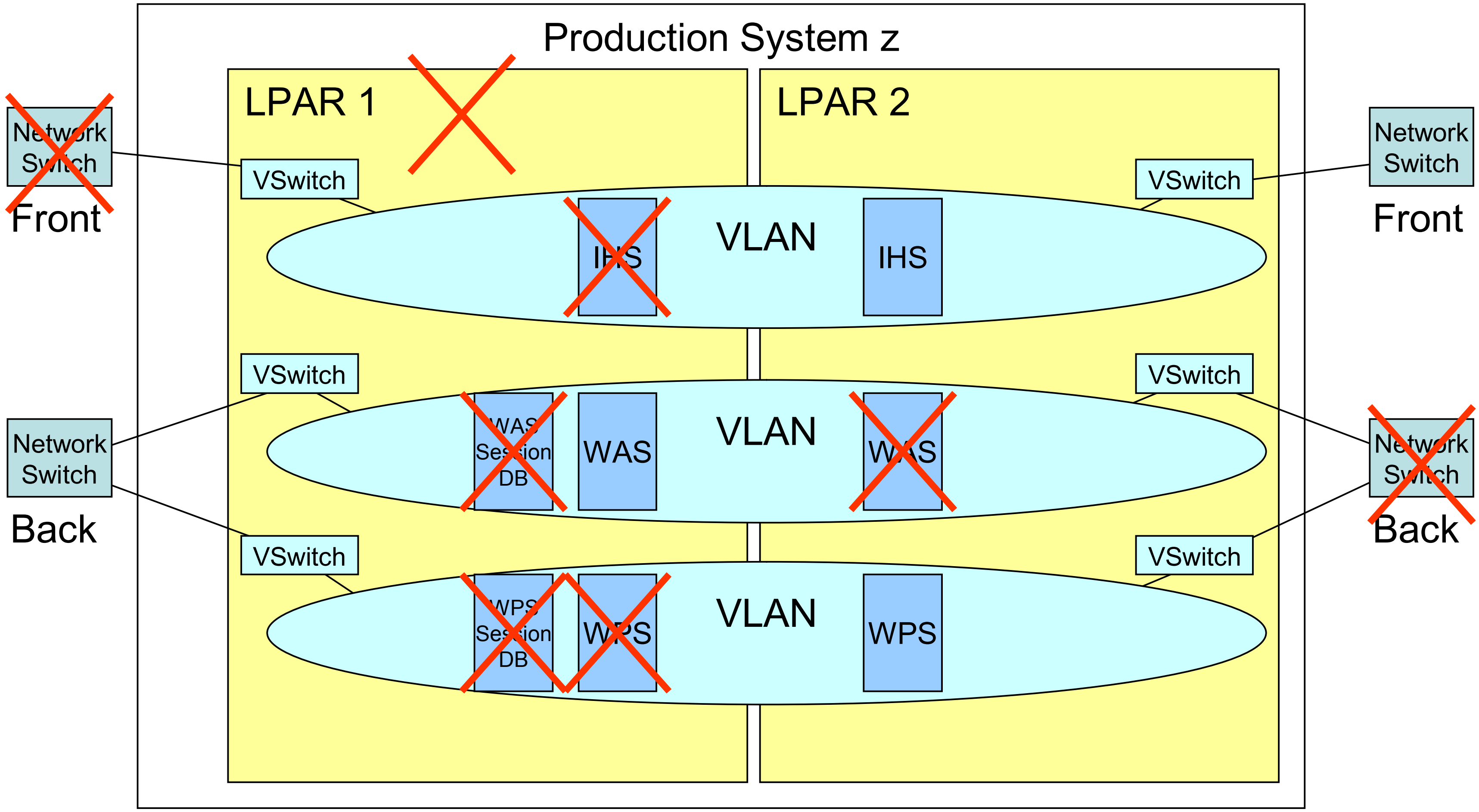
VSWITCH Detail – Linux definitions

- Confirmation

```
ifconfig
eth0      Link encap:Ethernet  HWaddr 02:00:00:00:00:05
          inet addr:10.1.1.1  Bcast:10.1.1.1  Mask:255.255.255.0
          inet6 addr: fe80::200:0:100:5/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:6 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:0 (0.0 b)  TX bytes:652 (652.0 b)
eth1      Link encap:Ethernet  HWaddr 02:00:00:00:00:04
          inet addr:10.2.1.1  Bcast:10.2.1.1  Mask:255.255.255.0
          inet6 addr: fe80::200:0:100:4/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:150122 errors:0 dropped:0 overruns:0 frame:0
          TX packets:66742 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:32348101 (30.8 Mb)  TX bytes:17319537 (16.5 Mb)
lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:16436  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:0 (0.0 b)  TX bytes:0 (0.0 b)
```

This information is for sharing only and not an endorsement by Nationwide Insurance

High Availability
Failure Scenarios Tested



This information is for sharing only and not an endorsement by Nationwide Insurance

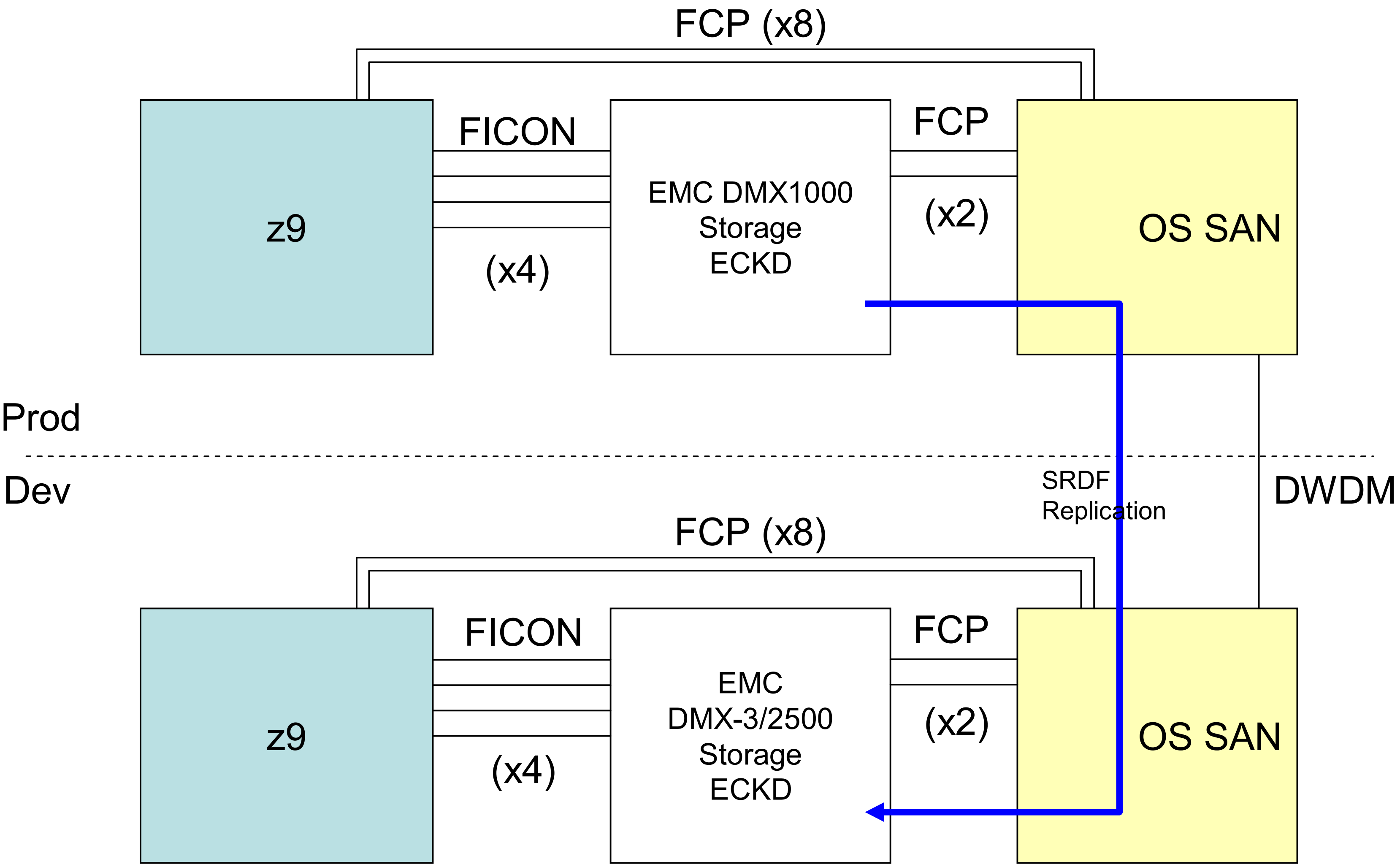
High Availability Clustering

- Scenarios tested
  - Loss of clustered web server
  - Loss of network switch
  - Loss of clustered application server
  - Loss of entire z/VM LPAR
- Current Limitations
  - Single System z
    Increased availability if LPARs are spread across CPCs

Disaster Recovery

- Included with High Availability offering
  - Disk replication between sites
  - Complete server definition (VM Directory) at second site
  - Physical network connections in place
  - Standby network definitions
  - Automated script for network personality at second site
    - Script on virtual server "asks" where it is running and sets network parameters
    - External DNS swap process must be performed
  - If primary site is unavailable, virtual servers are booted at second site

Disaster Recovery - Disk



This information is for sharing only and not an endorsement by Nationwide Insurance

Performance Measurement Tools

Performance 001 (way less than 101)

- Basic metrics to watch – z/VM
  - CPU utilization
    - While System z runs fine at 100%, Linux workload is much more demanding than traditional mainframe workloads.
      - Significant impact of memory over-commit
      - May need to keep peak periods at 85-90%.
  - Memory
    - Many Linux guests have huge working set sizes and many don't go idle
    - Keep memory over-commit less than 2:1 (ratio of combined working set sizes to real memory available)
  - Paging
    - z/VM has no problem with high page rates
      - Keep Expanded Storage for high-speed page buffer
    - Guests may not be tolerant
    - Allocate enough VM page space for twice the total of the working set of expected guests

Performance 001

- Basic metrics to watch – Linux Guests
  - Don't wake guests to ask
    - Choose performance tools that understand that Linux is running on z/VM
  - Pick one tool
    - Multiple monitoring tools adds a lot of overhead
    - ½% CPU per server adds up fast when there are 100s of servers
  - CPU measured inside guest is not very meaningful
    - Improved with CPU accounting enhancement at later kernel levels
  - Avoid TOP – significant overhead
    - Use vmstat or nmon

Performance 001

- Basic metrics to watch – Linux Guests
  - Memory
    - Don't over provision. Large virtual storage sizes drive up z/VM paging.
    - Use a swap hierarchy with z/VM VDISK as the highest priority swap space. It is not a problem for Linux to do some swapping.
    - Show all snapshot of memory/swap: free or cat /proc/meminfo
    - Avoid multiple caching
      - DB2: Use directio=yes to prevent it from doing its own I/O caching and rely on Linux
    - Default Linux memory management may not be optimal
      - Kernel parm: vm.swapiness=60
        Default may be too high – causes memory to be consumed
        Lower values cause Linux to reuse memory allocations more often to reduce memory demand
  - Paging
    - Prevent Linux from paging. z/VM paging is much more efficient.
    - Show Linux pagein/pageout: cat /proc/vmstat | grep pgpg

Performance 001

- Basic metrics to watch – Linux Guests
  - Look at guest CPU demand from z/VM
  - Watch for excessive paging on behalf of a guest.
    - May indicate inefficient memory usage or excessive virtual storage allocation
  - Watch for guests with poor I/O response
    - System z handles high I/O rates fine but bottlenecks can occur
  - Watch for % of active time that guests spend in various queues
    - Run
    - CPU queue
    - Page queue
    - etc

Performance 001

- Linux Guests internal performance
  - Tools to analyze guests functions vary greatly
    - Some have a lot of tools – WAS
    - Some may have little to offer
  - Application developers debugging skills may be limited
    - Accustomed to working with excessive capacity
    - Not accustomed to shared environment

Performance 001

- Ideas that may help
  - Utilize Cryptographic hardware
    - Dramatically improves SSL calls for secure web pages
  - Minimize external network hops
    - Use virtual firewall solutions
    - Staying inside the System z hardware operates at memory speeds
  - Turn off NTP (or only run occasionally)
  - Minimize or stagger cron scheduling

Performance Future Options

- Shared Read-Only disks
  - Requires separation of code from configurations
    - Special mount points or symbolic links
  - Testing complete; roll out in progress
    - Session 9216 has more details
- Cooperative Memory Management
  - Some testing done; working on automated control; implementation expected this year
- Execute In Place (xipfs)
- DCSS – shared code in z/VM storage

Conclusions

- Linux virtualization on System z does:
  - Reduce complexity
  - Improve provisioning time
    - No hardware acquisition
    - No physical installation to perform
  - Reduce environmental demand
    - Less cooling
    - Less power
    - Less floor space

Conclusions

- Things are changing rapidly
- Prepare for availability and continuity
- Performance is an iterative learning process
  - As much an art as a science
- Be careful what you ask for because you may get it!

References

- Other sessions this week:
  - Thu 11:00 9216 Extreme Filesystem Sharing Experiences with Linux using a Read-Only Root FS at Nationwide
  - Thu 04:30 9112 z/VM TCP/IP Stack Configuration

- Documentation
  - REDP3719 **Linux on IBM eServer zSeries and S/390: VSWITCH and VLAN Features of z/VM 4.4**

Contact Information

"And I thought we were busy ***before*** Linux showed up!"

Nationwide®

Rick Barlow

Senior z/VM Systems Programmer

Phone: (614) 249-5213

Internet: Richard.Barlow@nationwide.com

This information is for sharing only and not an endorsement by Nationwide Insurance