

Virtualization: TCP/IP Performance Management in a Virtualized Environment

Orlando Share Session 9308

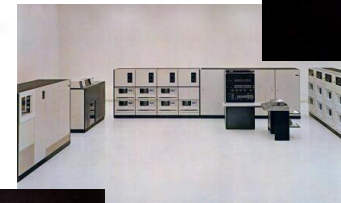
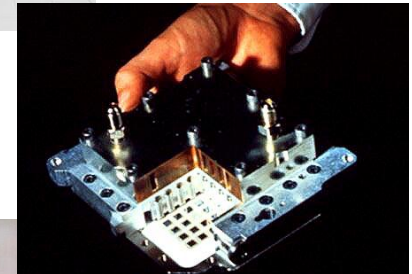
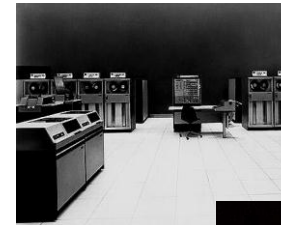


Laura Knapp
WW Business Consultant
Laurak@aesclever.com

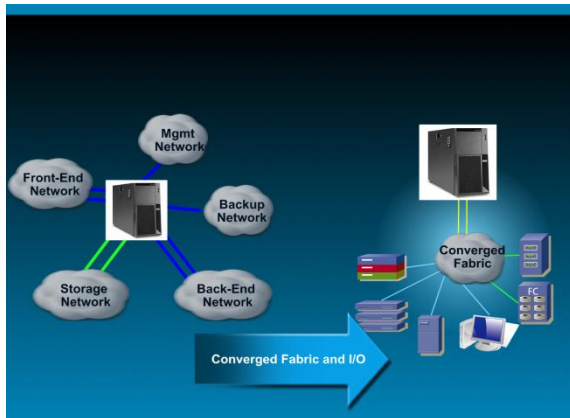
Background

Mainframe Specifics

Common Problems



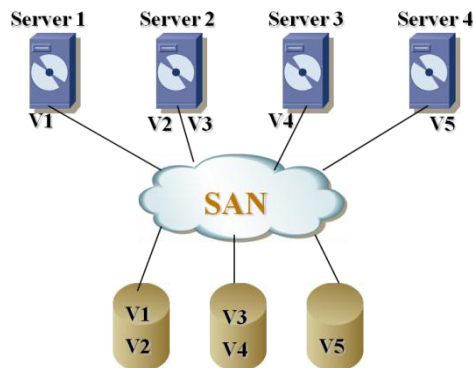
Right-Sizing IT Infrastructure



Consolidate...
entire farms of
.....servers ...
.....storage...
.....network....
.....etal



**...and dynamically
optimize to only
consume the
resources you
need!**



**...and dynamically
optimize to move
applications for high
availability and
performance!**

Always On, Optimized, Energy Efficient Datacenter

Dynamic Resource Scheduling

- > Balance workloads
- > Right-size hardware
- > Optimize real time

High Availability

- > Restart immediately when H/W or OS fail
- > Protect all apps

On-demand Capacity

- > Scale without disruption
- > Reconfigure on the fly
- > Save time

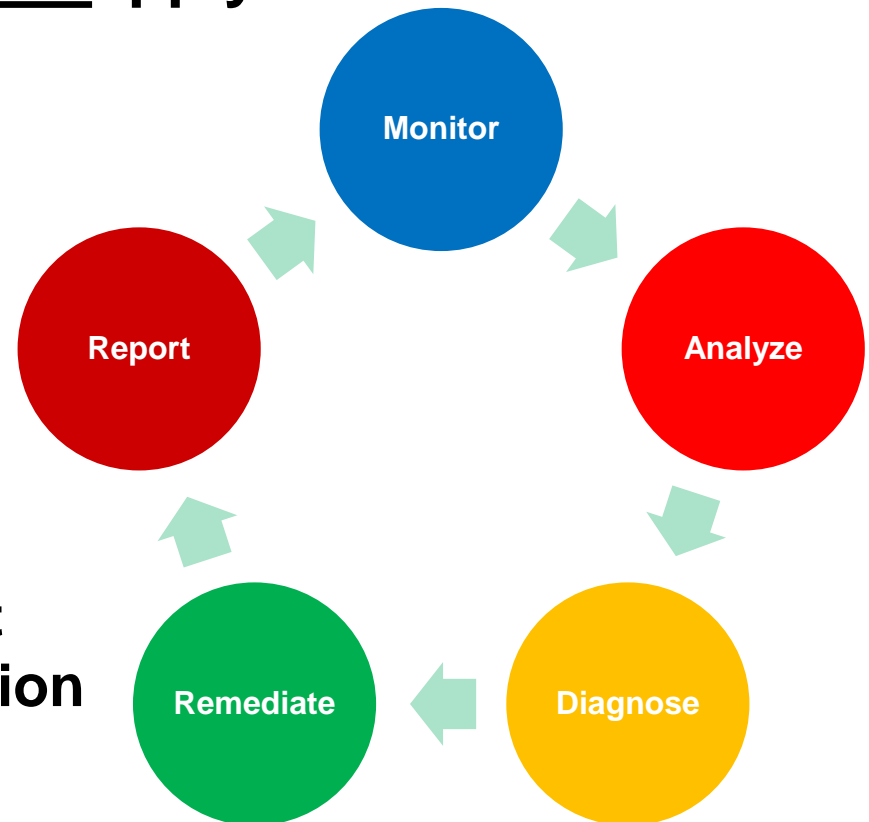


Managing Virtualized Data Center

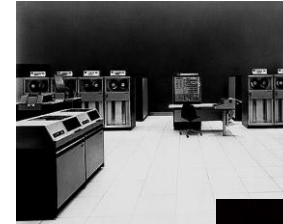
- **Fundamentals of management apply FCAPS**

- **Fault**
- **Configuration**
- **Availability**
- **Performance**
- **Security**

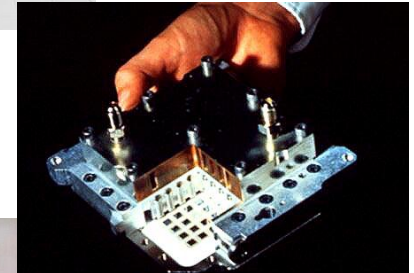
- **Leading to**
 - **Service Level Achievement**
 - **Optimum Resource Utilization**
 - **Highly available systems**
 - **High performing systems**



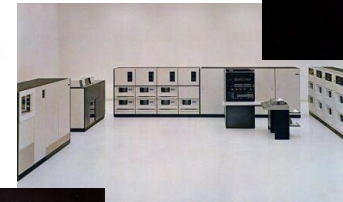
Background



Mainframe Specifics



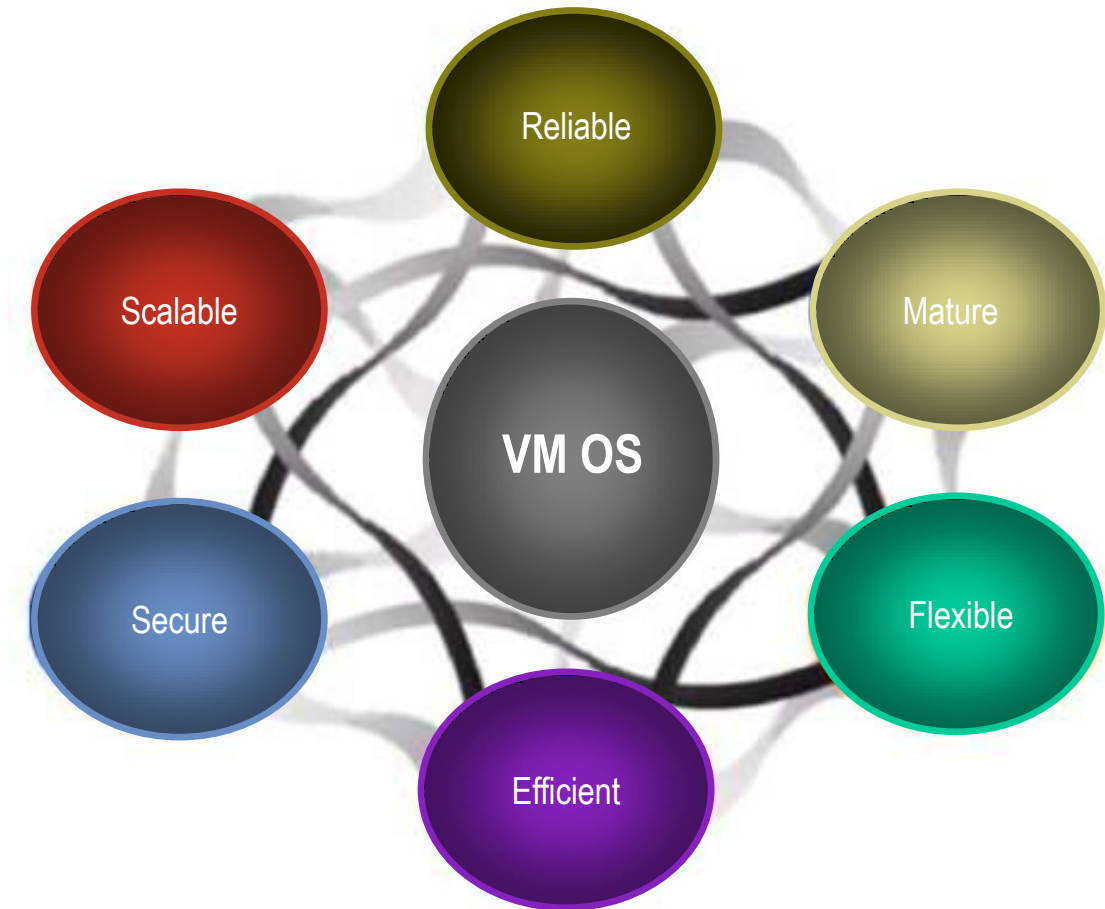
Common Problems



Best Practices



z/VM



Use of Mainframe as VM Grows

The momentum continues:

Shipped IFL engine volumes
increased 62% from 3Q07 to
3Q09

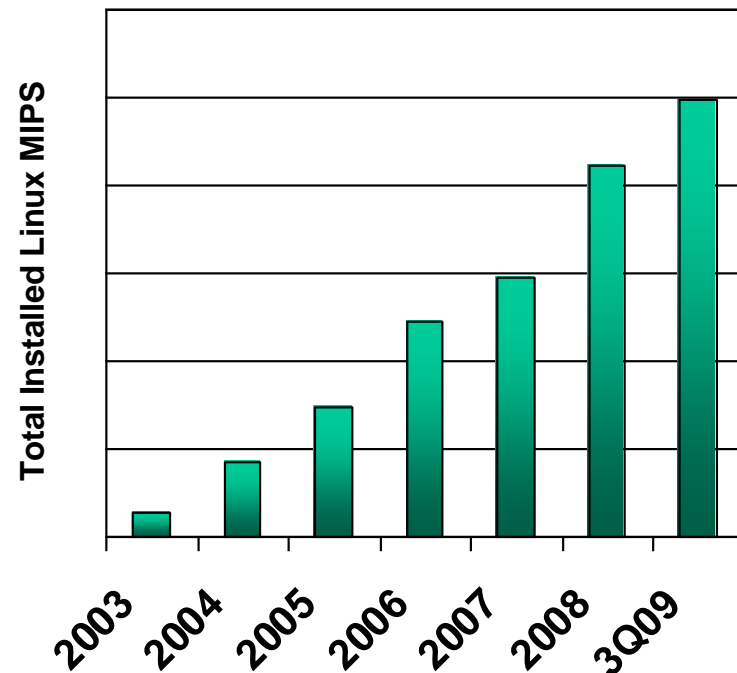
Shipped IFL MIPS increased 100%
from 3Q07 to 3Q09

**Linux is 16% of the System z
customer install base (MIPS)**

**70% of the top 100 System z
clients are running Linux on
the mainframe**

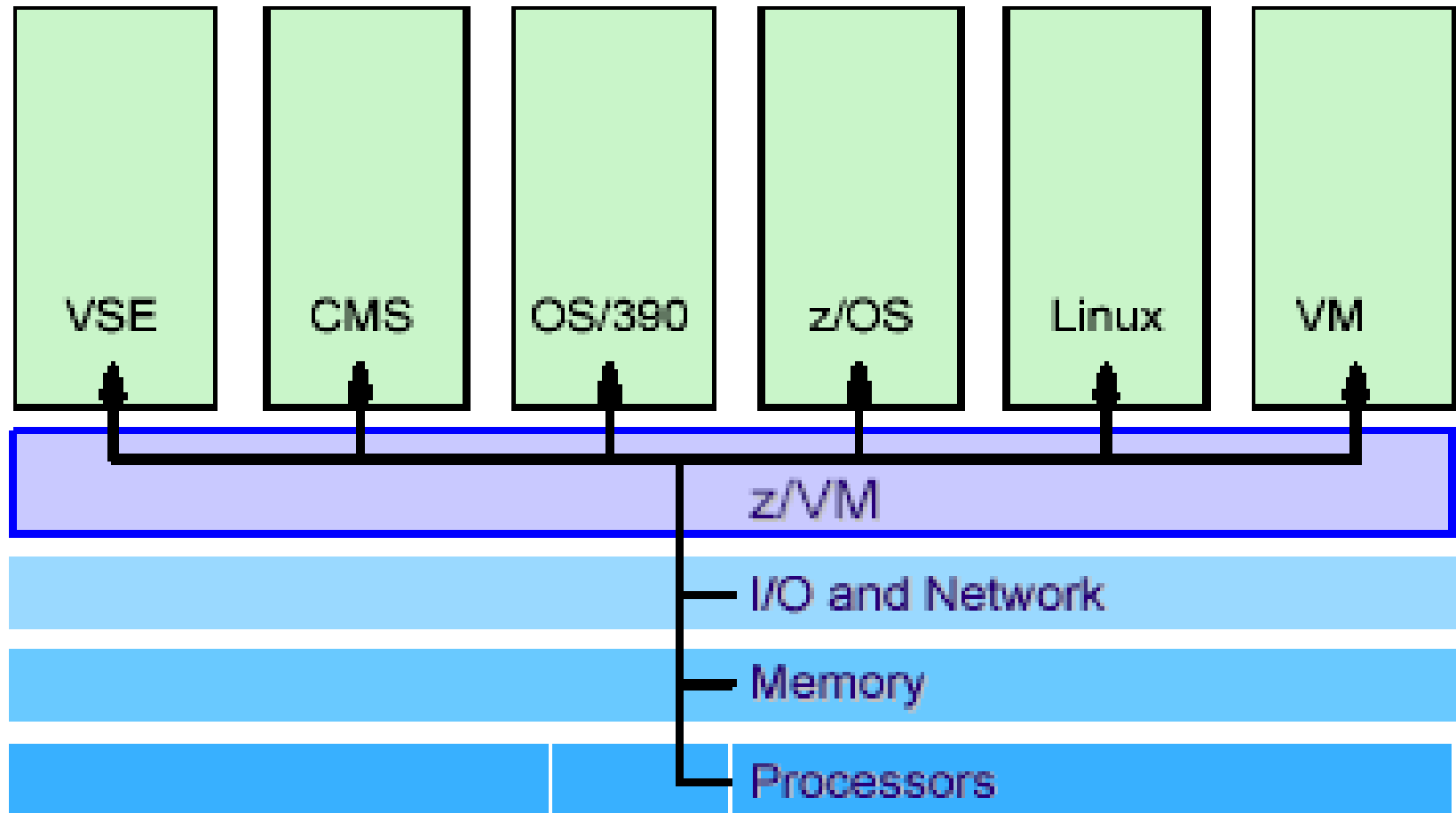
**>3,000 applications available
for Linux on System z**

Installed Linux MIPS

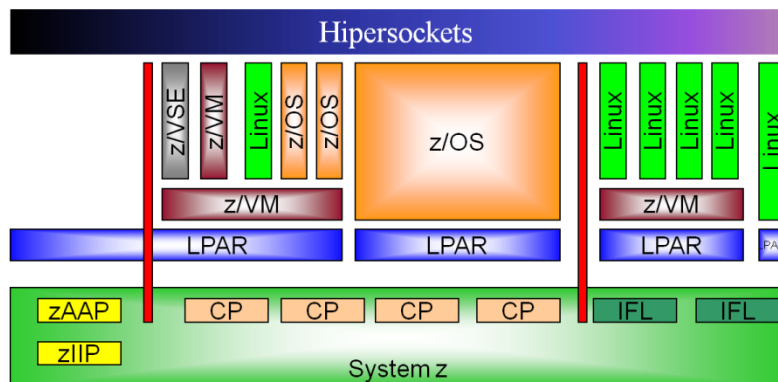


Source: Reed Mullen-IBM

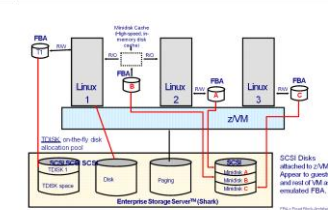
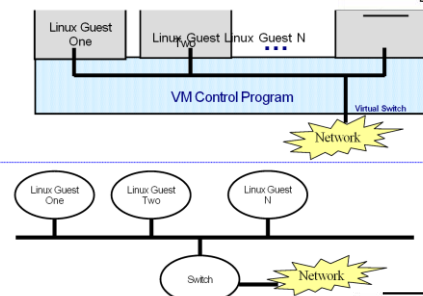
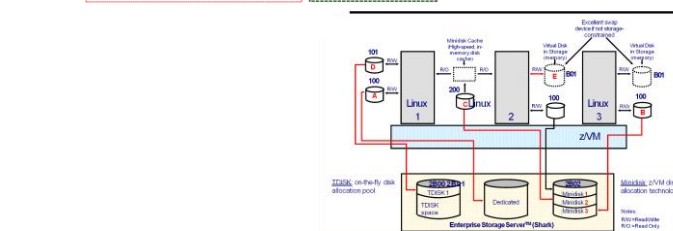
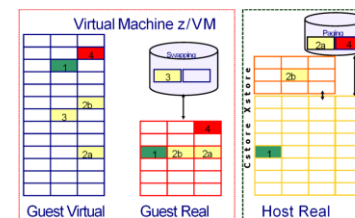
z/VM Structure



Advanced Virtualization on System z

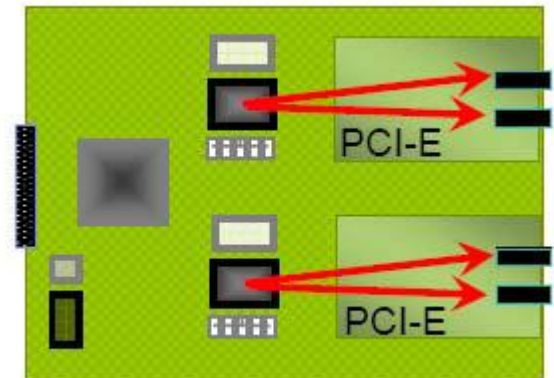


- **MVS (Multiple Virtual Storage)**
- **VM (Virtual Machine)**
- **LPAR (Logical Partition)**
- **Load Balancing**
- **VIPA (Virtual IP Addressing)**
- **HiperSockets**
- **Enterprise Extender (Virtual SNA)**
- **Linux for z/Series**
- **VLAN's (Virtual LAN)**
- **VSwitch (Virtual Switch)**



OSA Express3 Highlights

- New hardware data router bypasses firmware for packet construction, inspection, routing, etc
- New microprocessor (660 MHz versus 500/448 MHz)
- New PCI bus (PCI Express)
- New LC Duplex SM connectors for 10 Gbe feature
 - Dual density adapters Up to four ports per feature, two ports per CHPID
- Up to 45% improvement in latency over OSA-Express2
- 4x improvement over OSA-Express2 for 10g Ethernet feature (line speed)



Background

Mainframe Specifics

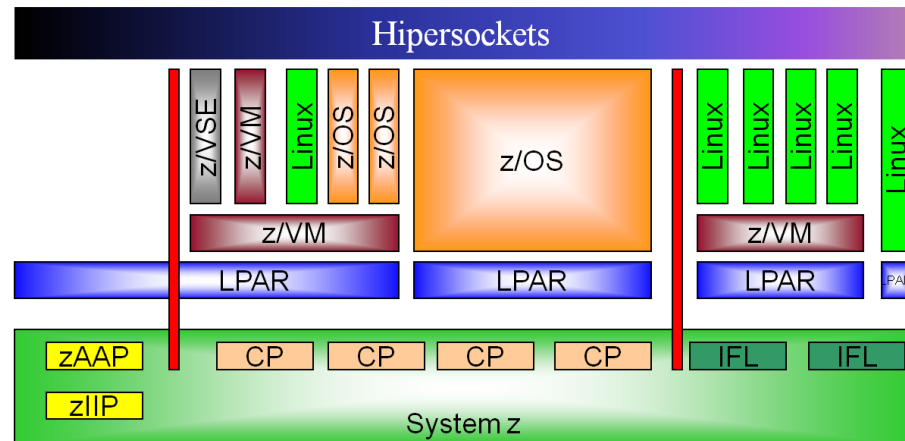
Common Problems



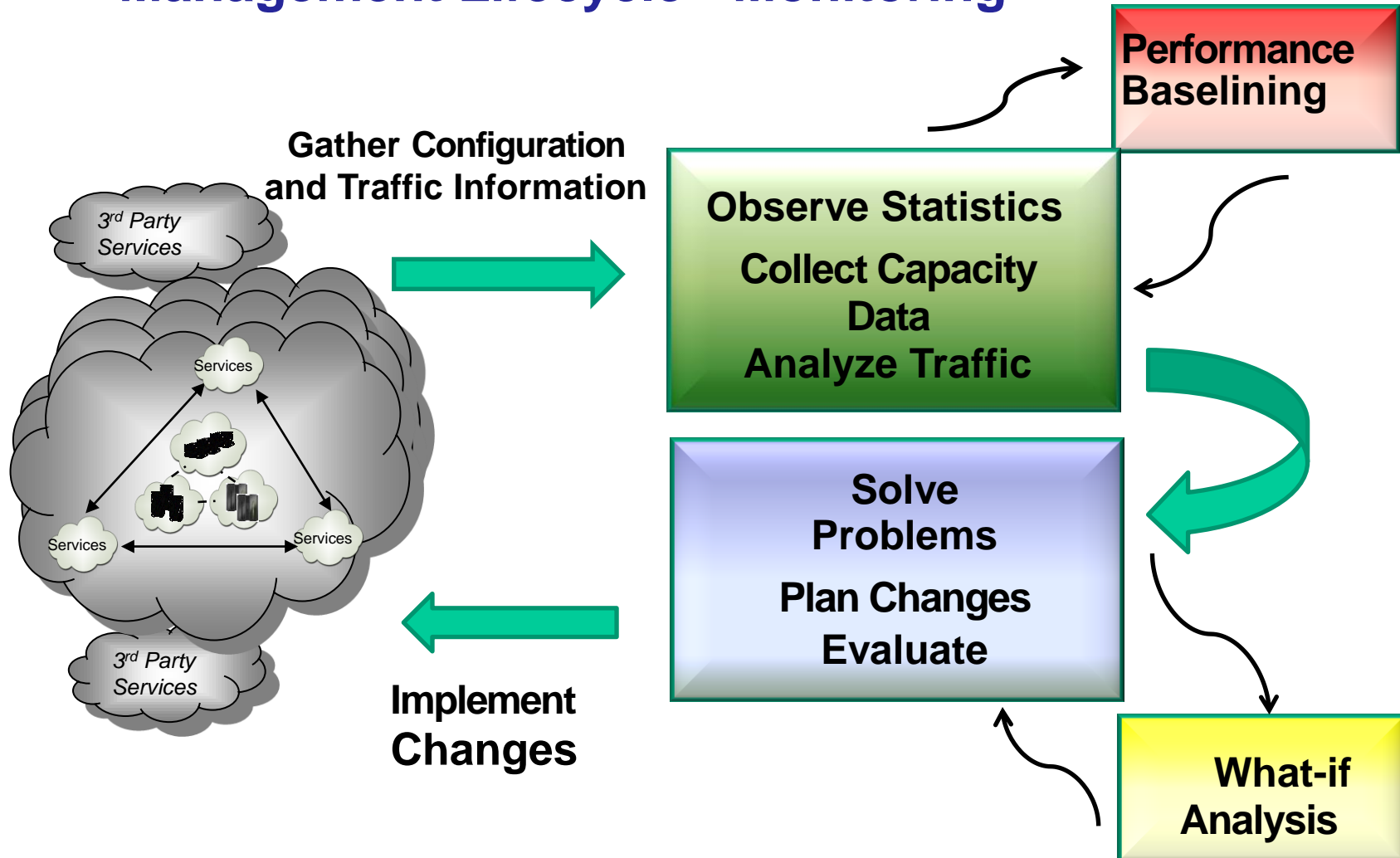
Approaches

Top Down or bottom Up doesn't matter
Consistency does

- Applications
- Middleware
- Guest OS
 - VM
- Network



Management Lifecycle - Monitoring



Linux: OSA LAN Timer or Blocking Timer

OSA inbound blocking function

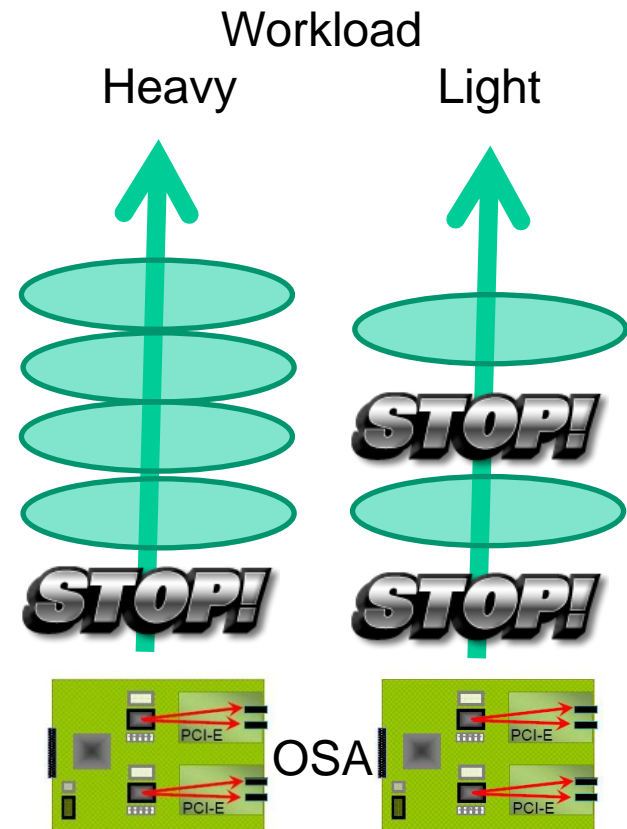
- Determines how long OSA will hold packets
- Indirectly affects
 - Frequency of host interrupt
 - Payload per interrupt

Linux has 3 potential values for OSA2

- For frames under 1536: Time between 2 incoming packets
- For Jumbo frames: Duration of inter-packet gap
- Total duration that OSA holds a single inbound buffer
- Default mode is NO LAN idle which is a good compromise for both transactional and streaming workloads

Linux behaves differently with OSAExpress3

- Using the default for OSA2 results in short latency but high CPU utilization



Scenario 1 –High CPU Utilization after move to OSA3

Situation

A system with an even mix of transactional and streaming workloads had a hardware upgrade and was now running with an OSA3 adapter. The Linux CPU became excessively high for no clearly visible reason.

Trouble Shooting

Historical data was viewed to ensure that the spike in CPU activity did occur when the OSA3 adapter was activated. In viewing the bytes in/out and other workload data no glaring inconsistencies were seen.

Solution

When the change was made the original OSA2 values for BLKT were used (inter=0, inter_jumbo=0, total=0). Due to the difference in OSA2 and OSA3 behavior these numbers were changed (inter=5, inter_jumbo=15, total=250). CPU utilization returned to normal

OSA2 default value on OSA3 results in shortest latency and highest CPU utilization

Best to use MTU size of 1492 for OSA3

Supported in
SLES10SP3+kernel update
SLES 11
RHEL 5.5

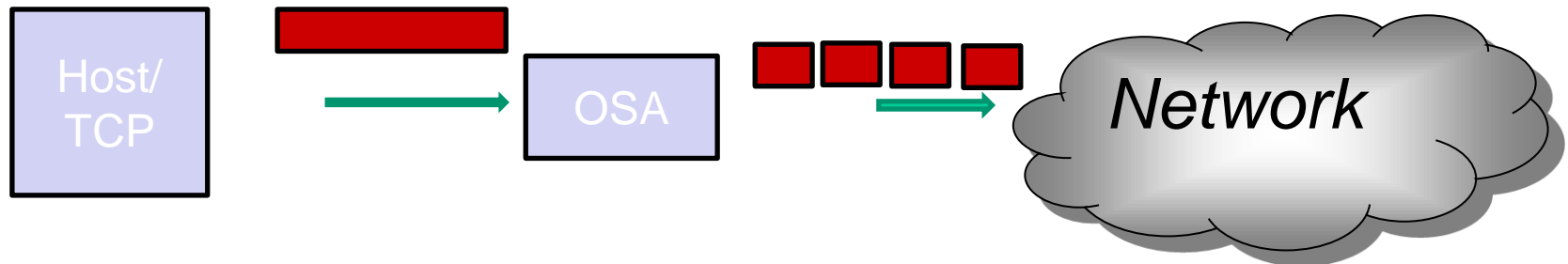
Red Hat:
/etc/sysconfig/network-scripts/ifcfg-eth0 add
OPTIONS="blkt/inter=5
blkt/inter_jumbo=15
blkt/total=250"

TCP Segmentation Offload (Large Send)

Segmentation consumes large amount of CPU

Allows most IPv4 TCP segmentation processing to be handled by OSA

Increases data transfer efficiency for IPv4 packets



Scenario 2 – 2 Tiered Database System

Situation

Client had a 2 tiered Database system and OSA 3 adapters. The front end database servers created many TCP/IP connections with high transactional volumes. The responses resulted in large TCP segments and the CPU utilization was unbearable.

Trouble Shooting

Look at the detailed TCP connections and transfer information. Use a packet trace tool. Is there a correlation between large segments sizes resulted in excessive CPU utilization. If so, go in and looking at the OSA adapter TCP Offload was not turned on.

Solution

TCP Segmentation was turned on for the OSA adapter.

On average anywhere from 25% to 45% CPU improvement was observed.

Use the large_send parameter

- no: no large send
- TSO: OSA adapter does segmentation
- EDDP: the qeth driver performs segmentation

TCP/IP will still do segmentation for:

- LPAR-LPAR packets
- IPSec encapsulated packets
-
- When Multipath is in effect (unless all interfaces support segmentation offload)

Scenario 3 – Mixed Servers on Virtualized Platform

Situation

Client had 12 servers on a single z/VM base. A single OSA adapter supported this z/VM base and several other LPARs. The workloads were mixed in nature and inconsistent in arrival. Some critical workloads when moved to the z/VM base were not providing the users with the desired response time.

Trouble Shooting

Using monitoring tools a detailed look was made of the overall IP stack and the flow of information. No inconsistencies were found. What was noted is that all traffic was treated equally. A look at the priority queue setting showed that everything was set to a queue setting of 3 (not important)

Solution

Due to the use of TOS bits for many of the workloads that required priority treatment the queue assignment was changed to 'prio_queuing_tos'

Workloads requiring special handling were now being provided this by the IP stack.

Service Type	Queue
Low Latency	0
High Throughput	1
High Reliability	2
Not Important	3

Impacts outgoing packets

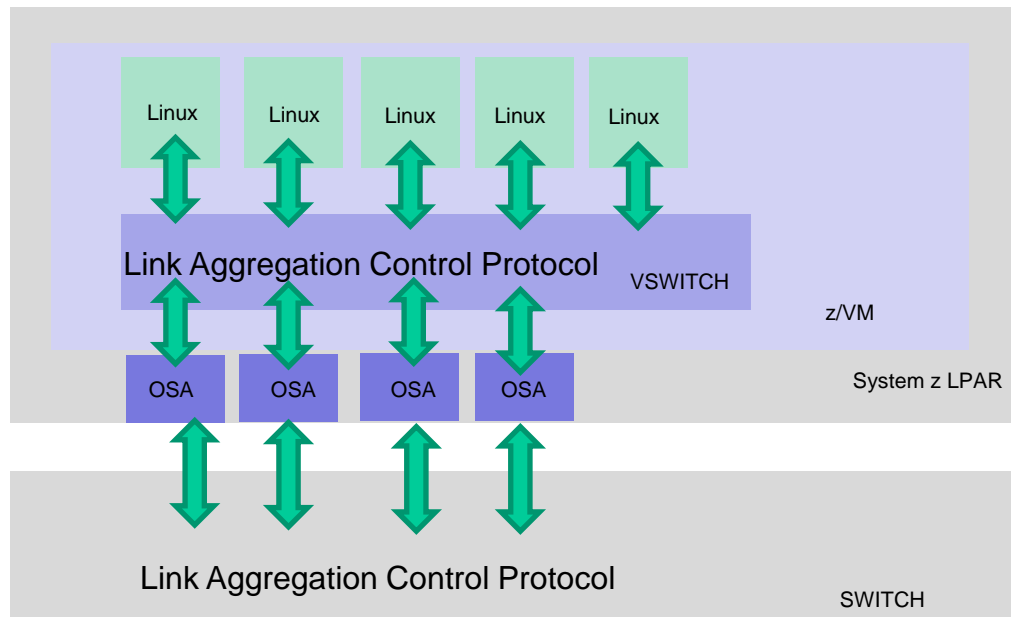
Extremely important in high traffic conditions

4 queues are available

z/VM Virtual Switch Link Aggregation

First introduced with z/VM 5.3

Aggregated links includes a physical switch supporting IEEE 802.3ad



Scenario 4 – Connection Lost after VSWITCH OSA Takeover

Situation

Client had a VSWITCH configured with OSA-P and backup OSA-B. When the cable connected to OSA-p was removed the communication was lost even though switching to OSA-B occurred.

Trouble Shooting

The following data was used in this analysis and was gathered using standard monitoring, tracing, and IP tools. First look at the status of the devices OSA-P and OSA-B. The monitor used showed us that there was a mismatch between speed of OSA-B and the switch that it was connecting to.

Solution

Match the speed of the switch and the OSA-B

Other items investigated included
ARP cache of z/VM TCP/IP
z/VM routing of Linux guest
information
Trace information for the
VSWITCH

Scenario 5 – Linux Hipersocket Performance Slow

Situation

A client had a very successful beta with Linux on system z. As they added additional workloads onto the Linux systems overall network performance declined. Hipersockets was used and the expectation was that performance should have been better.

Trouble Shooting

Using a Linux TCP/IP Monitor check the overall flow of information through both the IP and TCP layers. Verify that listeners are available for the applications. View alerts and determine if any would suggest the problem being seen. Check the buffer count. In this system the buffer count had never been raised and was still set at 16.

Solution

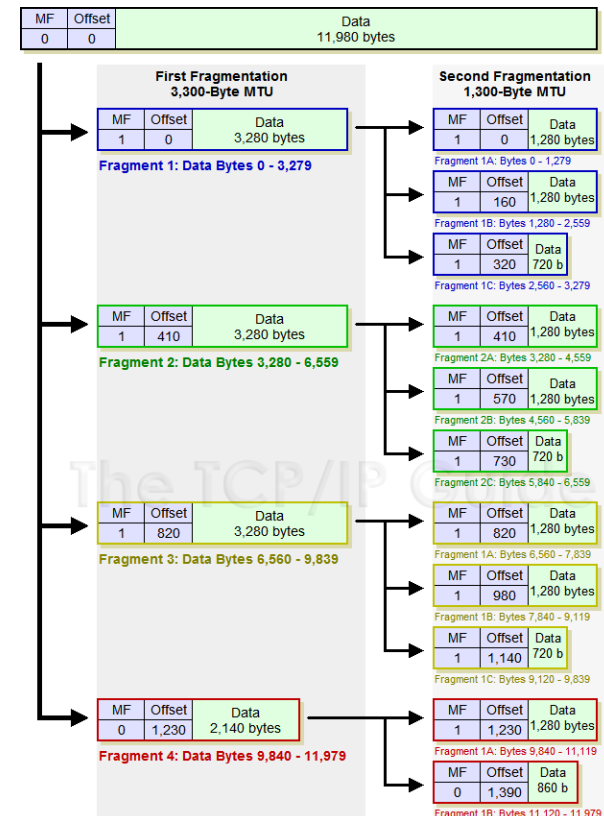
Tests by IBM have shown that using the default of 16 limits throughput as the number of parallel sessions increases with hipersockets. The buffers were increased to 50 with acceptable results

As you increase the buffer additional memory will be used

```
SUSE SLES11: in  
/etc/udev/rules.d/51-qeth-  
0.0.f200.rules add  
ACTION=="add",  
SUBSYSTEM=="ccwgroup",  
KERNEL=="0.0.f200",  
ATTR{buffer_count}="128"
```

MTU Size

- Optimizing MTU size can provide optimum performance improvements
- Set the maximum size supported by all hops between the source and destination
- Traceroute can provide details on the MTU size but some router administrators block traceroute
- If your application sends
- frames ≤ 1400 bytes use an MTU size of 1492
- Jumbo frames use an MTU size of 8992
- TCP uses MTU size for window size calculation
- For VSWITCH an MTU of 8992 is recommended



Scenario 6– Excessive Fragmentation

Situation

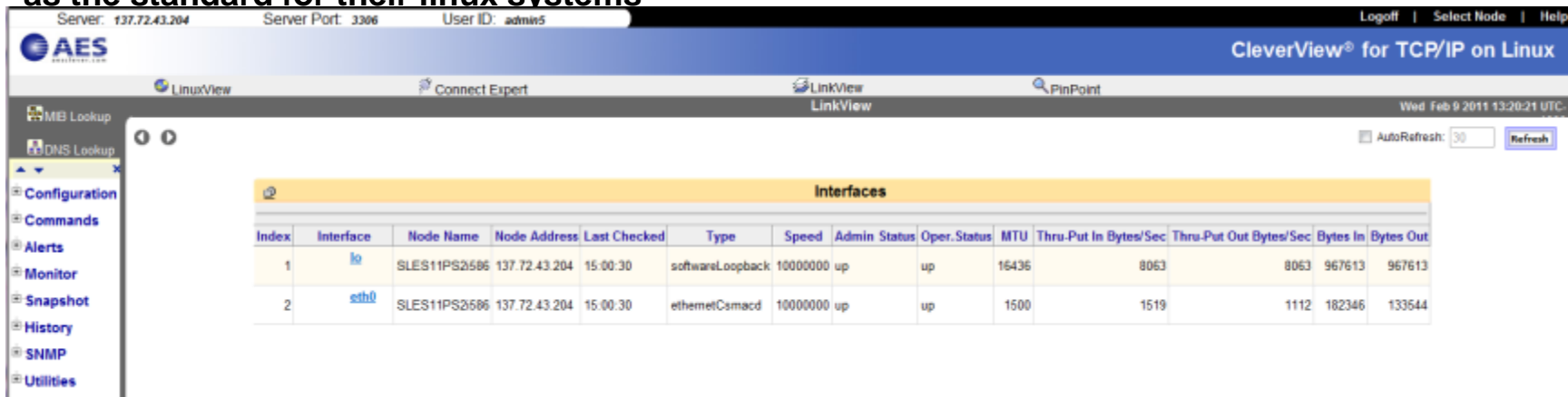
A client had a Linux on system environment and they were about ready to grow the production use of Linux. While they did not have any major problems they new of they asked for an overall health check.

Trouble Shooting

Using a Linux TCP/IP Monitor check the overall flow of information through both the IP and TCP layers. Look at the MTU settings on your links and the fragmentation on the IP stack. While there was not significant fragmentation, the MTU size was set at 1500.

Solution

In order to prevent future fragmentation issues we reset the MTU size to 1492 and defined that as the standard for their linux systems



The screenshot shows the AES CleverView for TCP/IP on Linux interface. The top bar displays the server information: Server: 137.72.43.204, Server Port: 3306, User ID: admin5. The main content area shows a table of network interfaces.

Index	Interface	Node Name	Node Address	Last Checked	Type	Speed	Admin Status	Oper. Status	MTU	Thru-Put In Bytes/Sec	Thru-Put Out Bytes/Sec	Bytes In	Bytes Out
1	lo	SLES11PS2686	137.72.43.204	15:00:30	softwareLoopback	10000000	up	up	16436	8063	8063	967613	967613
2	eth0	SLES11PS2686	137.72.43.204	15:00:30	ethernetCsmacd	10000000	up	up	1500	1519	1112	182346	133544

A few items to exploit in System z Kernel

Kernel 2.6.35

Offload outbound checksumming

- Move calculation of checksum for non-TSO packets from the driver to the OSA card
- NAPI support for QDIO and QETH (NAPI is Linux networking API)

Kernel 2.6.34

Hipersockets Network Traffic Analyzer

- Trace Hipersockets network traffic
- Layer 2 and Layer 3 supported

Kernel 2.6.33

OSA ADIO Data Connection Isolation

- Isolate data traffic between Linux guests on a VSWITCH
- Linux Guest communication needs to go through an external router
- Improves multi-tier security

AES Sessions at Share

Aug 8, 2011: 1:30-2:30 9288: [Keeping Your Network at Peak Performance as You Virtualize the Data Center](#)

Aug 10, 2011: 8:00-9:00 9266: [IPv6 Basics](#)

Aug 10, 2011: 4:30-5:30 9270: [Managing an IPv6 Network](#)

Aug 11, 2011: 3:00-4:00 9273: [CSI Maui: Forensics in The Case of the Attacked Browsers](#)

Aug 11, 2011: 11:00-12:00 9277: [Implementing IPv6 on Windows and Linux Desktop](#)

Aug 11, 2011: 1:30-2:30 9290: [Network Problem Diagnosis with OSA Examples](#)

Aug 12, 2011: 8:00-9:00 9308: [TCP/IP Performance Management in a Virtualized Environment](#)

Vielen
Dank

QUESTIONS?

Köszönettel

Obrigado!

Bedankt

Gracias

ขอบคุณ

شكراً

धन्यवाद

THANK YOU

Merci *Díky*

תודה

Ευχαριστώ

Hvala

Teşekkürler

laurak@aesclever.com

www.aesclever.com

650-617-2400