



IBM Systems & Technology Group

z/VM Performance Update 2010

Session 7908

IBM z/VM Performance Evaluation
Bill Bitner bitnerb@us.ibm.com

Trademarks

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml: AS/400, DBE, e-business logo, ESCO, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/30, VM/ESA, VSE/ESA, Websphere, xSeries, z/OS, zSeries, z/VM

The following are trademarks or registered trademarks of other companies

Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation
Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
Linux is a registered trademark of Linus Torvalds
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.
SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.
Intel is a registered trademark of Intel Corporation
* 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.

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

Any proposed use of claims in this presentation outside of the United States must be reviewed by local IBM country counsel prior to such use.

The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

Permission is hereby granted to SHARE to publish an exact copy of this paper in the SHARE proceedings. IBM retains the title to the copyright in this paper, as well as the copyright in all underlying works. IBM retains the right to make derivative works and to republish and distribute this paper to whomever it chooses in any way it chooses.

Agenda

- **z/VM 6.1.0**
- **Revisit Network Performance**
- **SSL Performance**
- **Discuss some current z/VM performance questions and concerns**
- **Discuss key service related to performance**
 - Closed
 - Expected to close this year
- **Few thoughts on futures**
- **Thanks to the rest of the z/VM Performance Evaluation Team:**
 - Dean DiTommaso, Bill Guzior, Steve Jones, Virg Meredith, Patty Rando, Dave Spencer, Joe Tingley, Xenia Tkatschow, Brian Wade

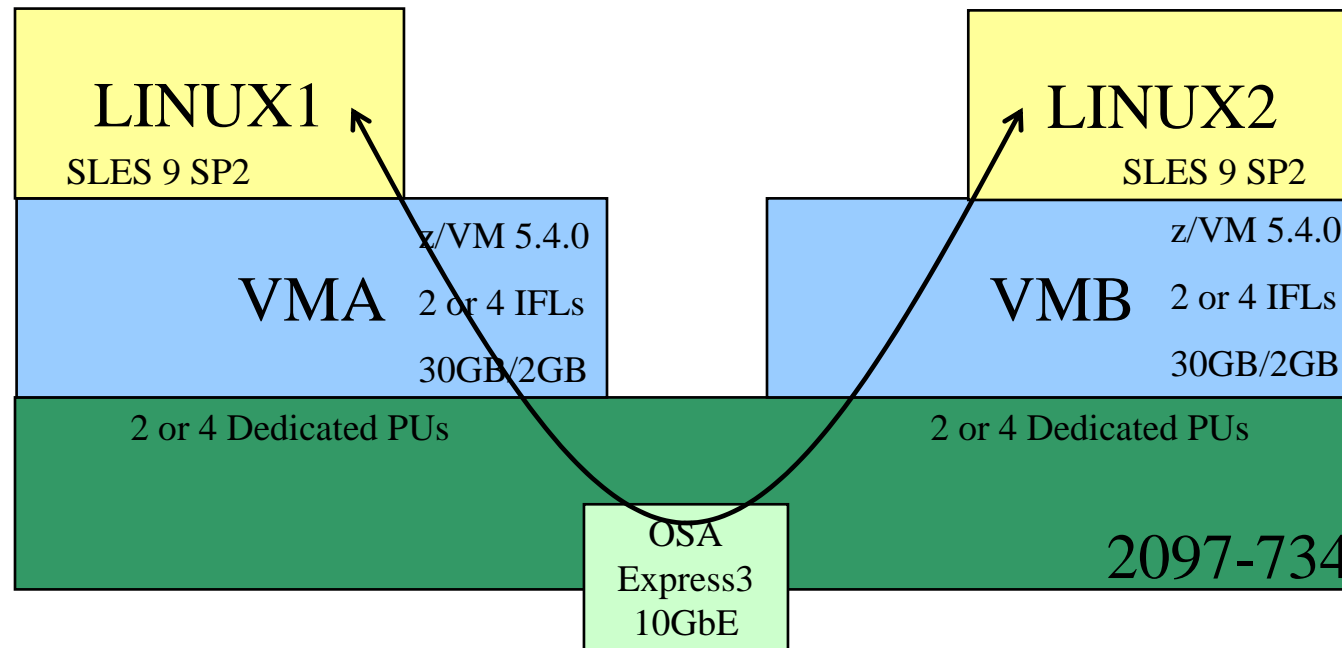
z/VM 6.1 Performance

- **One significant performance change: Guest LAN and VSwitch guest to guest improvement.**
- **Exploitation of instructions introduced in z10 that help avoid processor cache misses.**
- **Decreases processor time proportional to data movement intensity.**
- **Pure guest to guest data streaming showed up to 4% reduction in total processor time.**

Network Performance Revisited

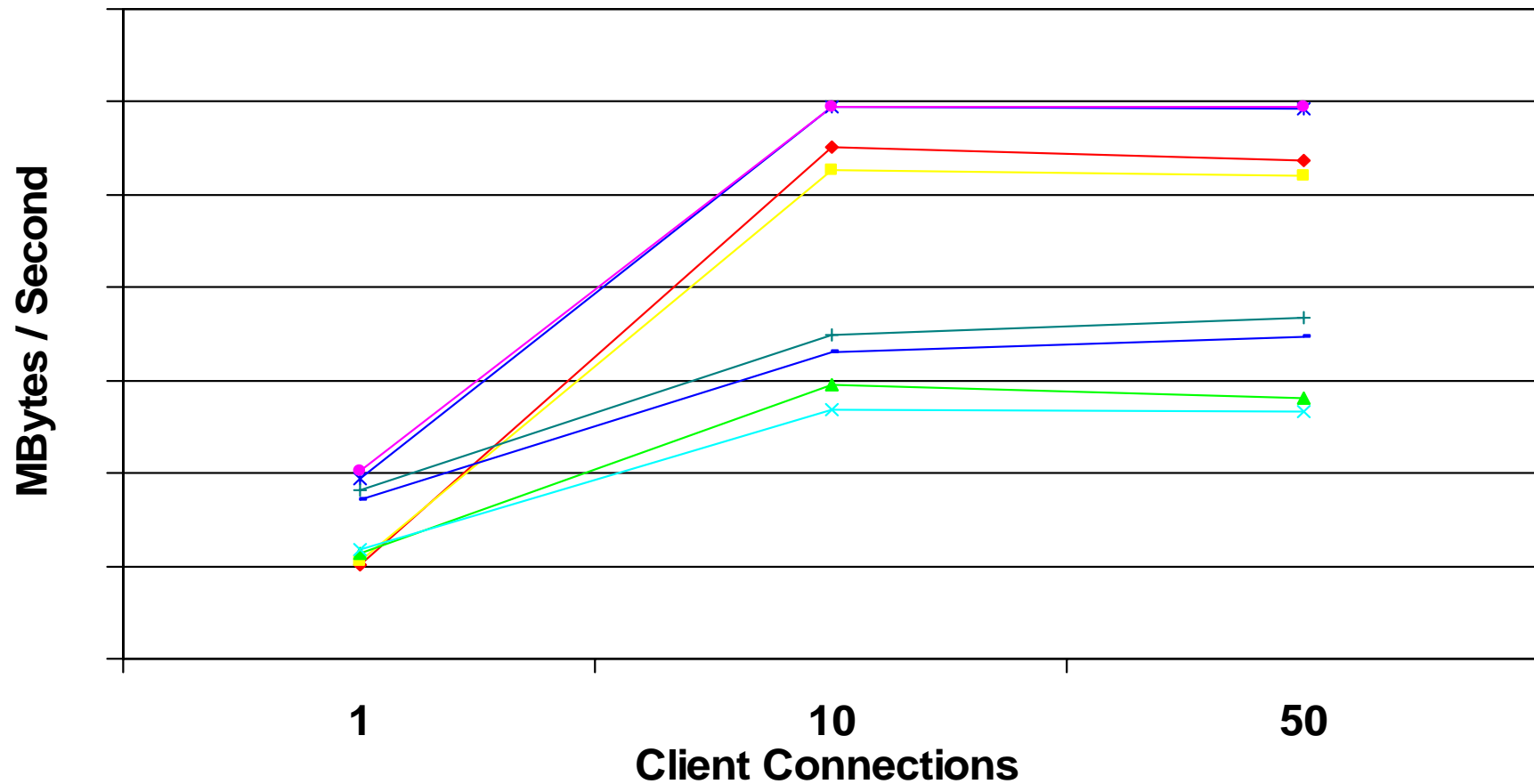
- **Measurement Environment and Workload Description**
- **Measurement Results**
 - Single Connection vs. Multiple Connections
 - MTU Size Comparisons
 - Dedicated OSA vs. VSwitch
- **Quantifying Throughput**
- **Hardware Performance Measurements**
- **Conclusions**

Measurement Configuration



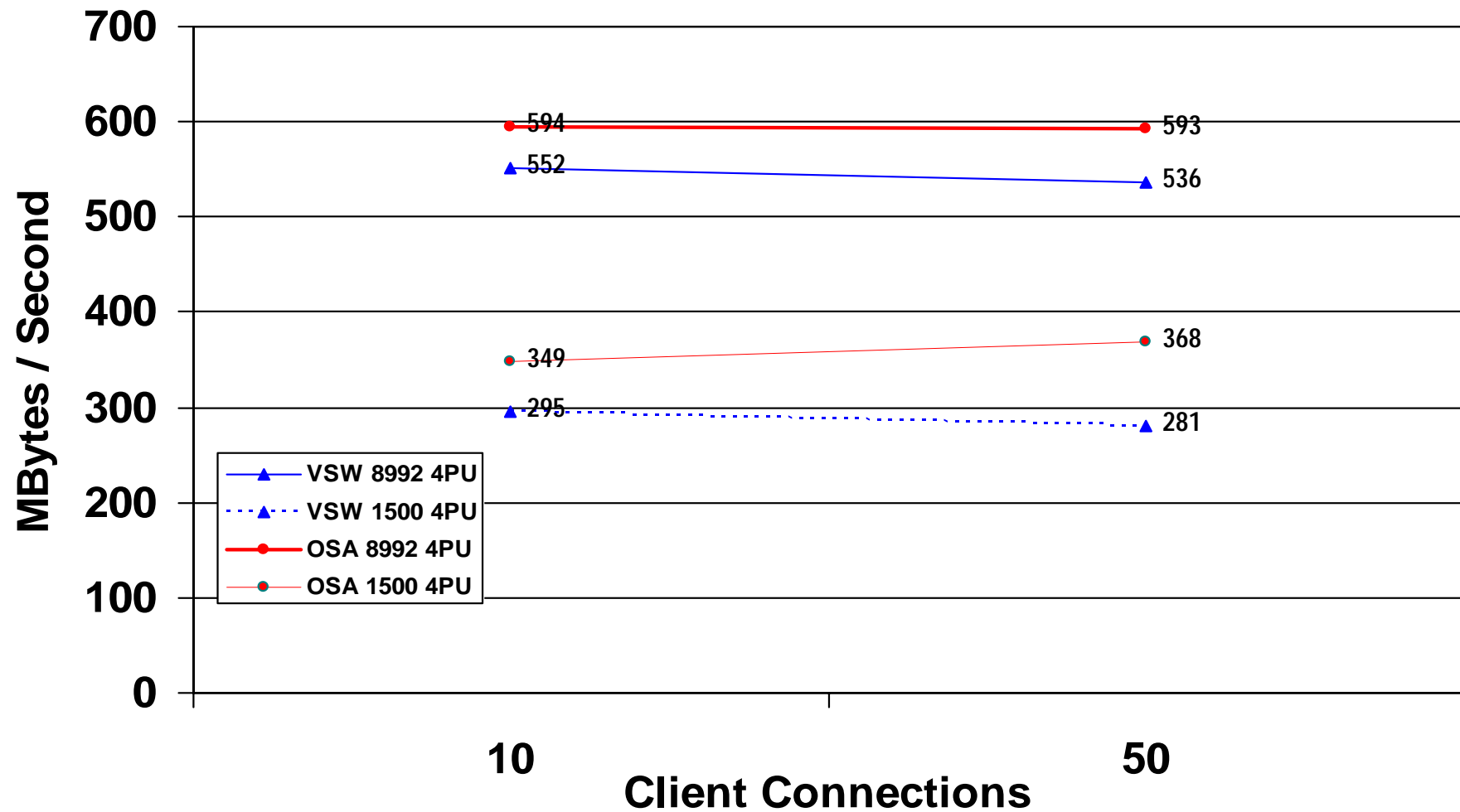
- Application Workload Modeler (AWM) used as the driver.
- Streaming workload – client sends 20 Bytes, receives 20 MBytes.
 - Throughput reported based on AWM data sent.
- Separate Ports on same OSA Express 3 card

Impact of Number of Connections



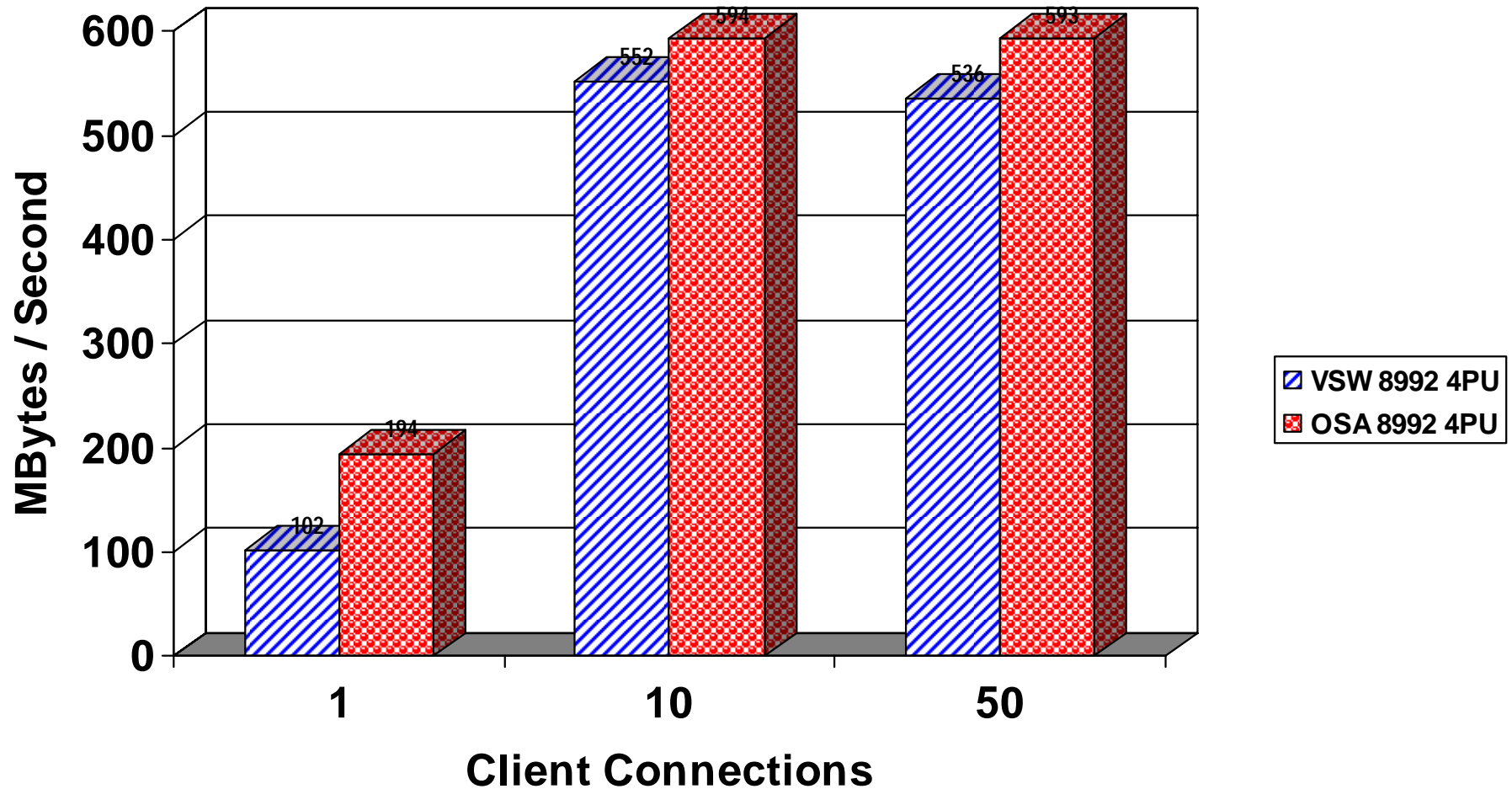
Need to be careful of single thread benchmark numbers. System z and z/VM optimize for large scale environments.

Impact of MTU Size



Using jumbo frames increases throughput between 61 to 91%.

Dedicated OSA vs. Virtual Switch



Except for single connection, OSA throughput is 6 to 7% higher

Quantifying Throughput

- **All measurements shown here were based on pure application data throughput.**
- **Other views or benchmarks may include additional bytes:**
 - Headers
 - Filler Space in packets
- **Example with MTU 8992:**
 - AWM reports 552.6 MBytes/Second
 - VSwitch reports 557.4 MBytes/Second (~1% additional)
- **Example with MTU 1492:**
 - AWM reports 269.3 MBytes/Second
 - VSwitch reports 327.2 MBytes/Second (~20% additional)
- **Workloads will show different ratios, as the data to header ratios differ. For this streaming workload, ratios are lower.**

System z HW OSA Performance Measurements

- **OSA-Express 3 Performance Report – November 2008**
- **Used AWM with z/OS as well as a ‘hand loop’ program that avoids all operating system overhead.**
- **Determined streaming type workloads with Jumbo frames deliver**
 - Mixed Direction: ~1110MB / Second
 - One Direction: ~660MB / Second
- **1 Byte Latency**
 - 66 microseconds
 - Roughly 40% improvement over OSA-Express 2

Network Conclusions

- **Both Dedicated OSA and Virtual Switch can provide throughput approaching 600MB/Second for application data being streamed in a single direction.**
- **Using MTU of 8992 is key**
- **Benchmark Considerations**
 - Single connections
 - Application data vs. Total data
 - Mixed Direction traffic vs. One Direction traffic

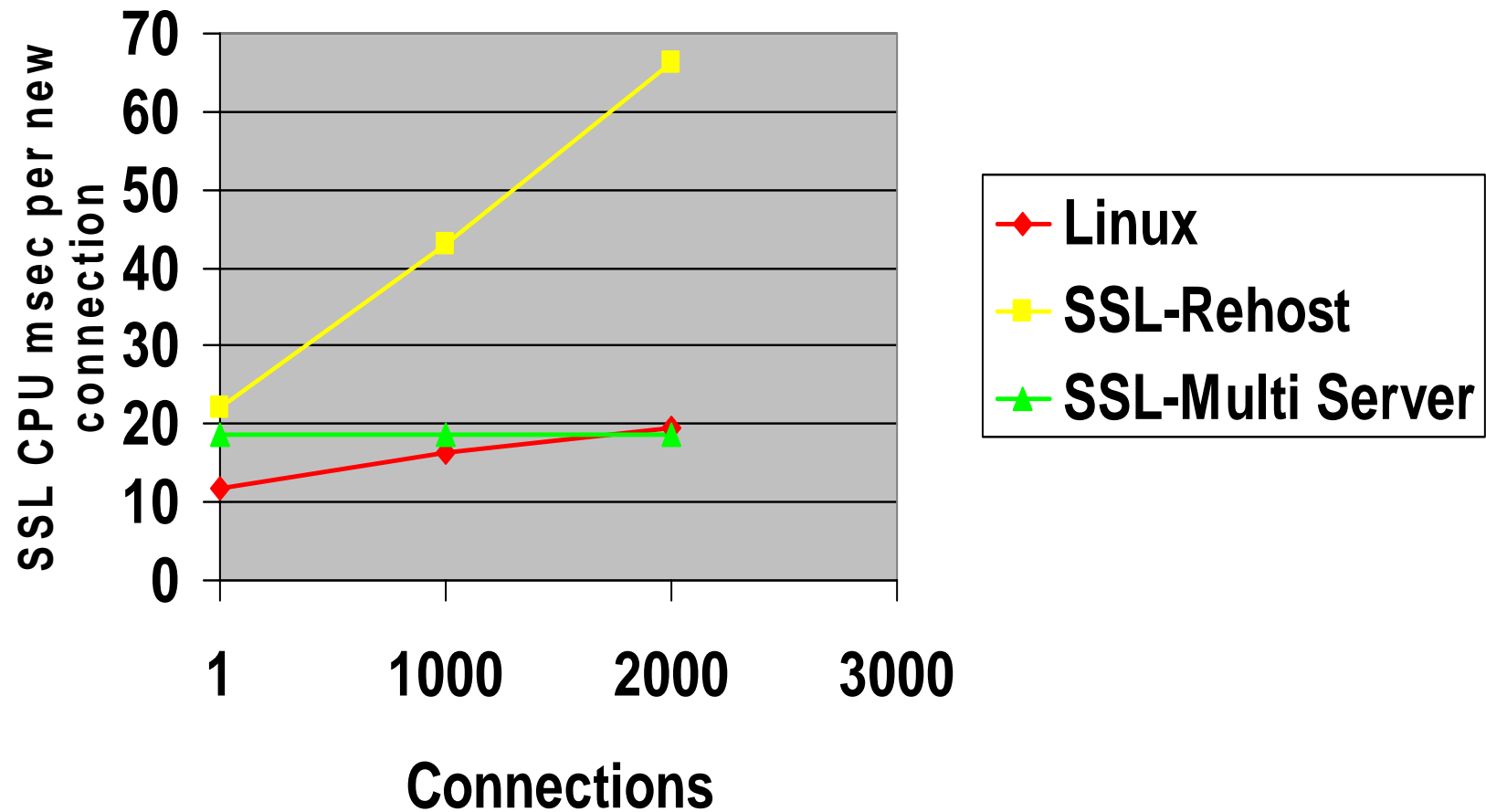
SSL Performance

- In z/VM 5.4, the z/VM SSL Server moved from Linux based server to a CMS based server with APAR PK65850.
 - Performance concerns compared to Linux based server.
- A group of related APARs to address performance, targeted to close August 2010. PTFs for z/VM 5.4 and 6.1
 - PK75662 (stack), PK97437 (packaging), PK97438 (SSL), VM64313 (CMS), VM64740 (CMS), PM06244 (System SSL) - SSL Performance Enhancements
- Because of significant changes in configuration for enhanced SSL, there will be additional documentation
 - User memos
 - Red Alert material

SSL Enhancement Objectives

- **Increase scalability (Support multiple SSL servers per TCP/IP stack)**
- **Increase the number of supported connections while maintaining the CPU cost of a connection stable**

2000 Connection Rampup



Results For Various TCP/IP Interfaces

Interface	Percentage Improvement (CPU/tx)	Comments
FTP	Degraded by 38%	The 'Select' code imported from z/OS is very inefficient. z/OS rewrote their 'Select' code for performance concerns. We did not have the capacity or bandwidth to rewrite the 'Select' code
Telnet	Improved by 8%	A slight improvement but again, the z/OS 'Select' code held us back from obtaining better performance results
SMTP	Improved Infinitely	The SMTP environment in the SSL-Rehost environment was not functioning. This problem was fixed in the current level of SSL.

Reorder Processing - Background

- ***Page reorder* is the process in z/VM of managing user frame owned lists as input to demand scan processing.**
 - It includes resetting the HW reference bit.
 - Serializes the virtual machine (all virtual processors).
 - In all releases of z/VM
- **It is done periodically on a virtual machine basis.**
- **The cost of reorder is proportional to the number of resident frames for the virtual machine.**
 - Roughly 130 ms/GB resident
 - Delays of ~1 second for guest having 8 GB resident
 - This can vary for different reasons +/- 40%

Reorder Processing - Diagnosing

■ Performance Toolkit

- Check resident page fields (“R<2GB” & “R>2GB”) on FCX113 UPAGE report
 - Remember, Reorder works against the resident pages, not total virtual machine size.
- Check Console Function Mode Wait (“%CFW”) on FCX114 USTAT report
 - A virtual machine may be brought through console function mode to serialize Reorder. There are other ways to serialize for Reorder and there are other reasons that for CFW, so this is not conclusive.

■ REORDMON

- Available from Bill Bitner now and the VM Download Page <http://www.vm.ibm.com/download/packages/>
- Works against raw MONWRITE data for all monitored virtual machines
- Works in real time for a specific virtual machine
- Provides how often Reorder processing occurs in each monitor interval

Reorder Processing - Mitigations

- **Try to keep the virtual machine as small as possible.**
- **Virtual machines with multiple applications may need to be split into multiple virtual machines with fewer applications.**
- **See <http://www.vm.ibm.com/perf/tips/reorder.html> for more details.**
 - Several customers running with a “patch”.
- **Known requirement at IBM to bring relief in this area: APAR VM64774**
 - Provide System Config settings. SET and QUERY commands, system wide settings
 - Corrects problem in “patch” solution that inhibits paging of PGMBKs for virtual machines where Reorder is set off.

VMDUMP Processing Concern

- **VMDUMP is a very helpful command for problem determination.**
- **Some weaknesses:**
 - Does not scale well, can take up to 40 minutes per GB.
 - It is not interruptible
 - APAR VM64548 is open to address this.
- **Linux provides a disk dump utility which is much faster relative to VMDUMP.**
 - It is disruptive
 - Does not include segments outside the normal virtual machine.
- **See <http://www.vm.ibm.com/perf/tips/vmdump.html>**

VM64271 SET SHARE ABSOLUTE LIMITHARD

- **Customers reported both underlimiting and overlimiting**
- **Problematic configurations:**
 - Sum of absolute shares > 100%
 - Guest with low relative minimum and larger absolute maximum
 - LIMITHARD used and system not very busy
- **Status:**
 - VM64721 Closed and Available for z/VM 5.3, 5.4, and 6.1
 - R530 UM32851 October 2009 RSU 1001
 - R540 UM32852 October 2009 RSU 1001
 - R610 UM32853 October 2009 RSU 1001
 - Introduces new SET SRM LIMITHARD options:
 - DEADLINE = current behavior and default
 - CONSUMPTION = new approach. Will become the default in a future release.
 - Only applies to ABSOLUTE

Excess Share Distribution: Background

- **Shares are relative to other users that want to run.**
- **Example:**
 - Four virtual machines that want to run on real 1-way:
 - LINUX01 Relative 100 = 17%
 - LINUX02 Relative 100 = 17%
 - LINUX03 Relative 200 = 33%
 - LINUX04 Relative 200 = 33%
 - Total Shares = 600
 - What happens if LINUX04 only wants to use 3%?

Excess Share Distribution Problem

Userid	Share	Normalize	Uses	Should Get	Problem Scenario
LINUX01	100	17%	17%	24.5%	17%
LINUX02	100	17%	17%	24.5%	17%
LINUX03	200	33%	33%	47%	63%
LINUX04	200	33%	3%	3%	3%

Excess Share Distribution Problem Status

- **IBM is aware, has recreated the problem, and is working on correcting.**
- **No APAR currently open. No customer has open problem report.**
- **There was a previous problem like this that was changed by major code changes in VM/ESA 1.2.2, June 1994.**
 - <http://www.vm.ibm.com/perf/reports/vmesa/vm122prf.pdf> describes the changes
- **Unclear when the problem was re-introduced.**

MDC and FlashCopy Interaction

- **FlashCopy requests require z/VM to flush MDC for the entire minidisk.**
- **MDC Flush processing is very expensive even when there is no data in MDC to flush**
 - System Time becomes very high.
- **z/OS DFSMS and other utilities can make extensive use of FlashCopy for functions such as defragmentation**
- **Mitigations**
 - Turn off MDC for minidisks that are FlashCopy targets

VM64767: VARY ON PROCESSOR Hangs

- ***VARY ON PROCESSOR* n might sometimes never complete**
 - Mishandling of VARY lock in save area reclaim
- **Other work requiring the VARY lock can pile up behind this indefinite postponement**
- **Eventually the system can hang**
- **VM64767 is open**

CMM Futures

- **CMM 2 (aka CMMA, MEMASSIST)**
 - Linux support limited to SLES 10
 - Off by default at the Linux Level
 - Check “[cmma=on](#)” option with “[cat /proc/cmdline](#)” to see if in use.
- **CMM 2 Lite**
 - Form of CMMA that only uses the “Stable” and “Unused” states (isolated to architecture-specific code).
 - Direction of future distributions
- **CMM 1**
 - Can be used via VMRM support
 - Originally thought to be more of a tactical solution with CMM 2 being the strategic solution
 - Service to improve: VM64439
 - Expect more investigation in this area in future.
- **For more performance information, see:**
 - <http://www.vm.ibm.com/perf/reports/zvm/html/530cmm.html>

VM64527 MCW002 Abends from Memory Imbalance

- **z/VM 5.3, 5.4, and 6.1**
 - R530 UM32878 Nov 2009 RSU 1001
 - R540 UM32879 Nov 2009 RSU 1001
 - R610 UM32880 Nov 2009 RSU 1001
- **Imbalance in storage management free storage pools when using dedicated FCP or OSA devices may lead to host abend.**
- **Very large dumps because memory has been consumed by FOB blocks**

VM64850 Avoids Problem with VSwitch Failover

- **z/VM 5.4 and 6.1**
 - R540 UM33119 July 2010 Future RSU
 - R610 UM33120 July 2010 Future RSU
- **The problem scenario:**
 - After a fail-over to a backup OSA adapter or
 - Adding an additional port to a LinkAG port group
 - When multiple LPARs, VSWITCHes and OSA devices are involved.
- **The VSwitch erroneously starts using only a single 64K buffer.**
 - Normally, it is 128 x 8MB buffers.

VM64715 Page Release Serialization

- **z/VM 5.4 and 6.1 – Still Open, Target Fall 2010**
- **The problem scenario:**
 - Page release serialization changes from z/VM 5.2 and service resulted in the Page Table Invalidation Lock (PTIL) exclusive in cases that result in poor performance.
 - Worse in environments with significant segment creation/deletion, such as large DB2 for VM & VSE data space exploitation scenarios
- **The fix:**
 - Change various PTIL exclusive locks to PTIL shared
 - Restructure code appropriately

Monitor and Performance Toolkit

- **Enhancements in monitor for various service items 3Q2010**
 - VM64818: new fields to help determine which function introduced in service is available.
- **Support in Performance Toolkit shipping in service 3Q2010**
 - VM64819: 64 internal fixes and enhancements
 - VM64820: New function in conjunction with z196, Scheduler Changes, etc.
 - VM64821: New function in conjunction with STP support.

Future Performance – Some Thoughts

- **z196**

- “IBM zEnterprise 196 Hardware Overview”
 - Tuesday 1:30 Room 302
- “To MIPS or Not to MIPS, That is the CP Question!”
 - Thursday 11:00 Room 303
- “IBM zEnterprise Unified Resource Manager Overview”
 - Wednesday 11:00 Room 305

- **z/VM SSI and Guest Relocation**

- “z/VM Single System Image and Guest Mobility Preview”
 - Tuesday 1:30 Room 208
- “z/VM System Limits”
 - Wednesday 1:30 Room 305

Summary

- **New improvements and fixes coming out in the service stream.**
- **The Adventure Continues**
- **See <http://www.vm.ibm.com/perf/>**