IBM Systems & Technology Group

z/VM Performance Update 2011

Revision 2011-07-22 (BKW)

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

Agenda

- **z196 availability**
- **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 whole z/VM Performance Evaluation team:**
  - Bill Bitner, Dean DiTommaso, Bill Guzior, Steve Jones, Virg Meredith, Patty Rando, Dave Spencer, Joe Tingley, Xenia Tkatschow, Brian Wade

z196 Availability

- **zEnterprise z196 began shipping Sep 10, 2010**
  - 96 engines (80 configurable), 5.2 GHz (1.18x z10)
  - 3 TB memory (2x z10), 192 MB cache per book (4x z10)
  - See http://www.ibm.com/systemz
- **z/VM requires service to run on a z196**
  - See http://www.vm.ibm.com/service/vmreqze.html
- **z/VM LSPR scaling ratios are in the range 1.38 to 1.55**
  - Larger N-ways have larger scaling ratios
  - The very workloads that were modest from z9 to z10 will do much better from z10 to z196, owing to emphasis in z196 on processor cache
  - MP rolloff curve is slightly more shallow than z9 or z10
    - 32-way ratio is 0.62 rather than 0.59 (z9) or 0.57 (z10)
- **Do your homework before swapping… get that MONWRITE data!**

LSPR Suite Changes for z/VM and Linux

- **More current levels of various components**
  - Updated from SLES 9 to SLES 10
  - Updated from DB2 8.1 to 9.5
  - Updated WebSphere from 6.02 to 7.01
  - Updated from z/VM 5.2 to z/VM 5.4
- **Application workload changed from Trade6 to Daytrader**
- **Measured up to a 32-way partition**
- **We are now tinkering with running storage-overcommitted workloads**
  - They stress the processor cache differently
  - They force the machine to run different instruction mixes

Other LSPR Changes

- **z196 LSPR introduces new view of how a workload stresses a CEC**
  - Old way: run specific application suites (IMS, etc.)
  - New way: try to measure the pressure the running workload exerts on the CEC, especially on the cache or "nest"
  - We are using CPU Measurement Facility counters for this (new in z10)
    - z/OS: SMF 113 records
    - z/VM: we are well aware of the exploitation requirement
- **"Nest intensity" (aka workload's cache habits) is key**
  - Low RNI: light use of memory hierarchy – high N-way scaling
  - Average RNI: centrist, similar to old LoIO
  - High RNI: very hard on the cache, similar to old DI-mix
- **We have a ways to go here**
  - Is RNI alone a sufficient predictor of how any given workload will scale?
  - Is there an additional metric that might be illuminating to collect?
  - How might we factor said additional metrics into what you read in LSPR?

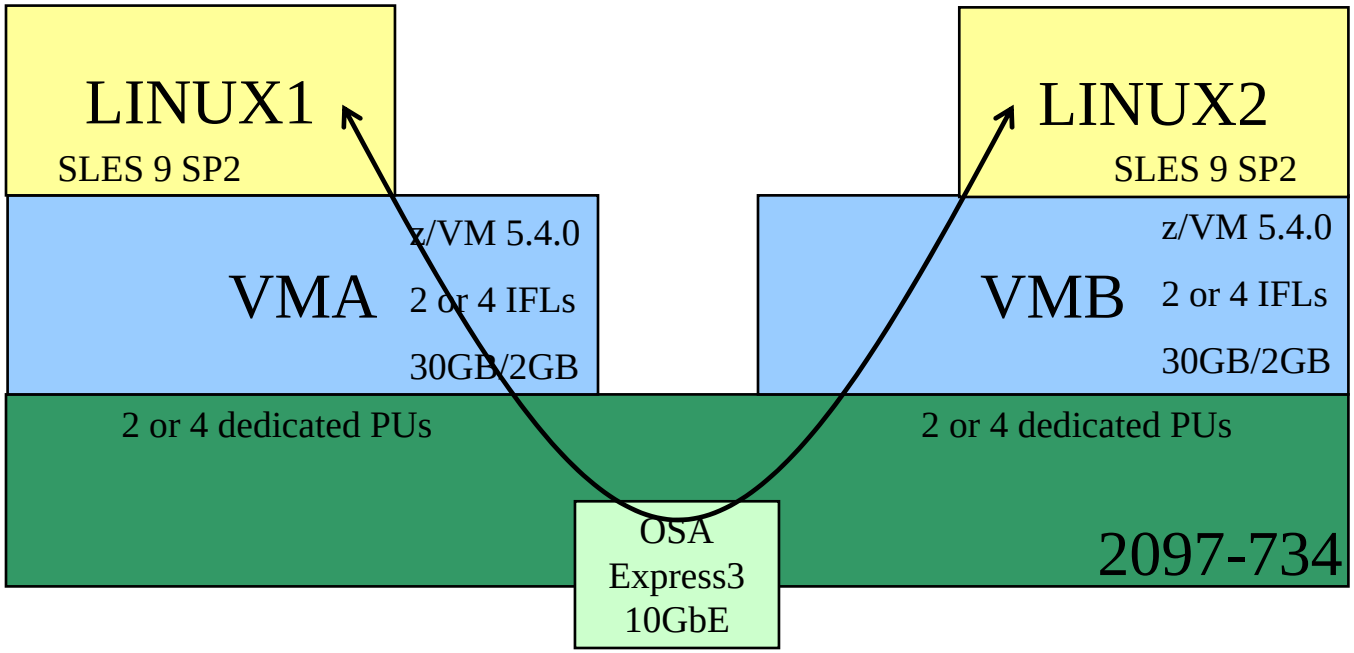
z/VM 6.1 Performance

- **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

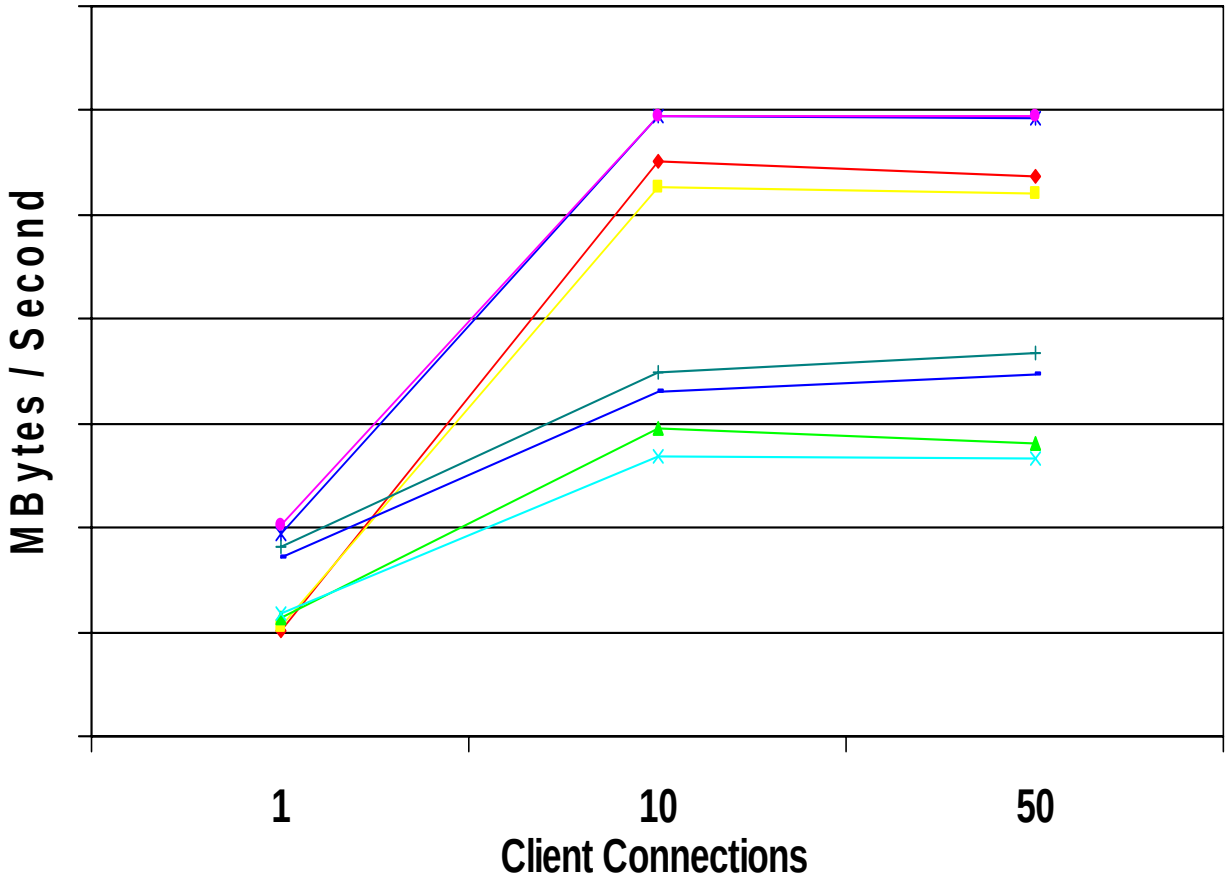
- **Measurement environment and workload description**
- **Measurement results**
  - Single connection vs. multiple connections
  - MTU size comparisons
  - Dedicated OSA vs. VSWITCH
- **Quantifying throughput**
- **Hardware performance measurements**
- **Conclusions**

Measurement Configuration



- Application Workload Modeler (AWM) used as the driver.
- Streaming workload: client sends 20 bytes, receives 20 MB.
  - Throughput reported based on AWM data sent.
- Separate ports on same OSA-Express3 card

Impact of Number of Connections



OSA, 8992, 2 or 4 PUs

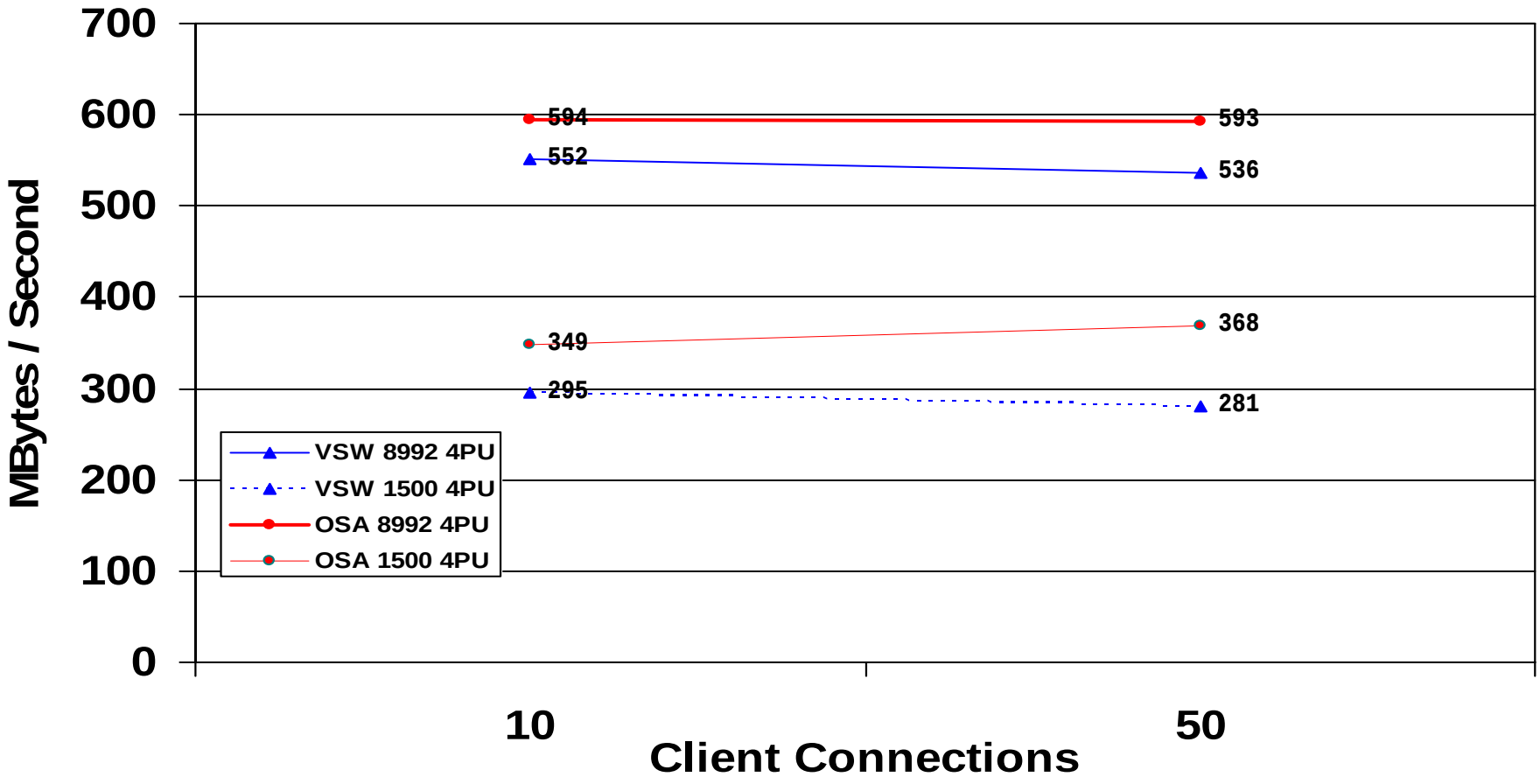
VSWITCH, 8992, 2 or 4 PUs

OSA, 1500, 2 or 4 PUs

VSWITCH, 1500, 2 or 4 PUs

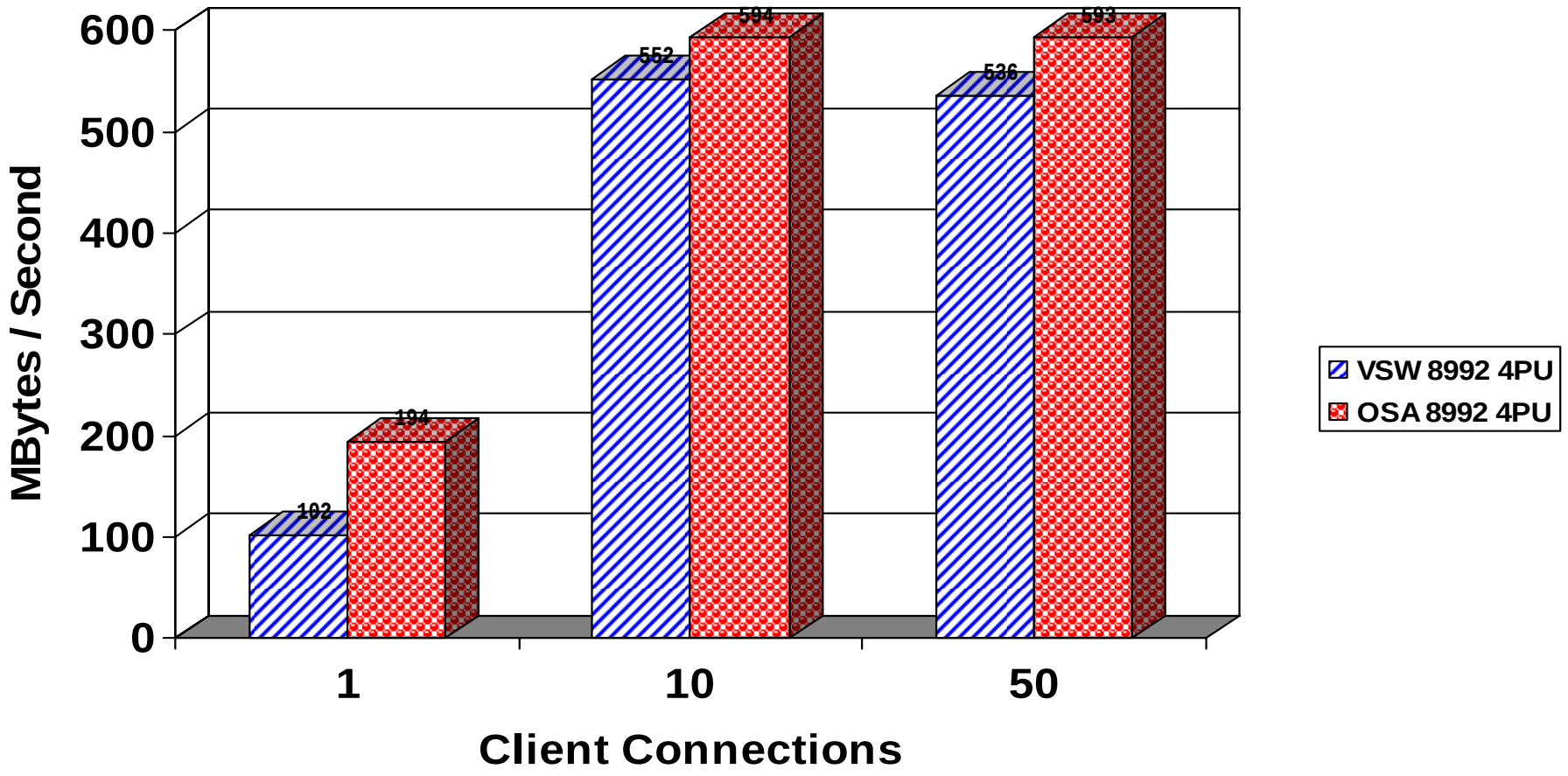
Need to be careful of single-thread benchmark numbers.
System z and z/VM optimize for large-scale environments.

Impact of MTU Size



Using jumbo frames increases throughput between 61% and 91%.

Dedicated OSA vs. Virtual Switch



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

Throughput of What, Exactly?

- **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-Express3 Performance Report – November 2008**
- **Used AWM with z/OS as well as a "hand loop" program that avoids all operating system overhead.**
- **Determined streaming workloads with jumbo frames deliver:**
  - Mixed direction: ~1110 MB / second
  - One direction: ~660 MB / second
- **1-byte latency**
  - 66 microseconds
  - Roughly 40% improvement over OSA-Express2

Network Conclusions

- **Both dedicated OSA and VSWITCH can provide throughput approaching 600 MB/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

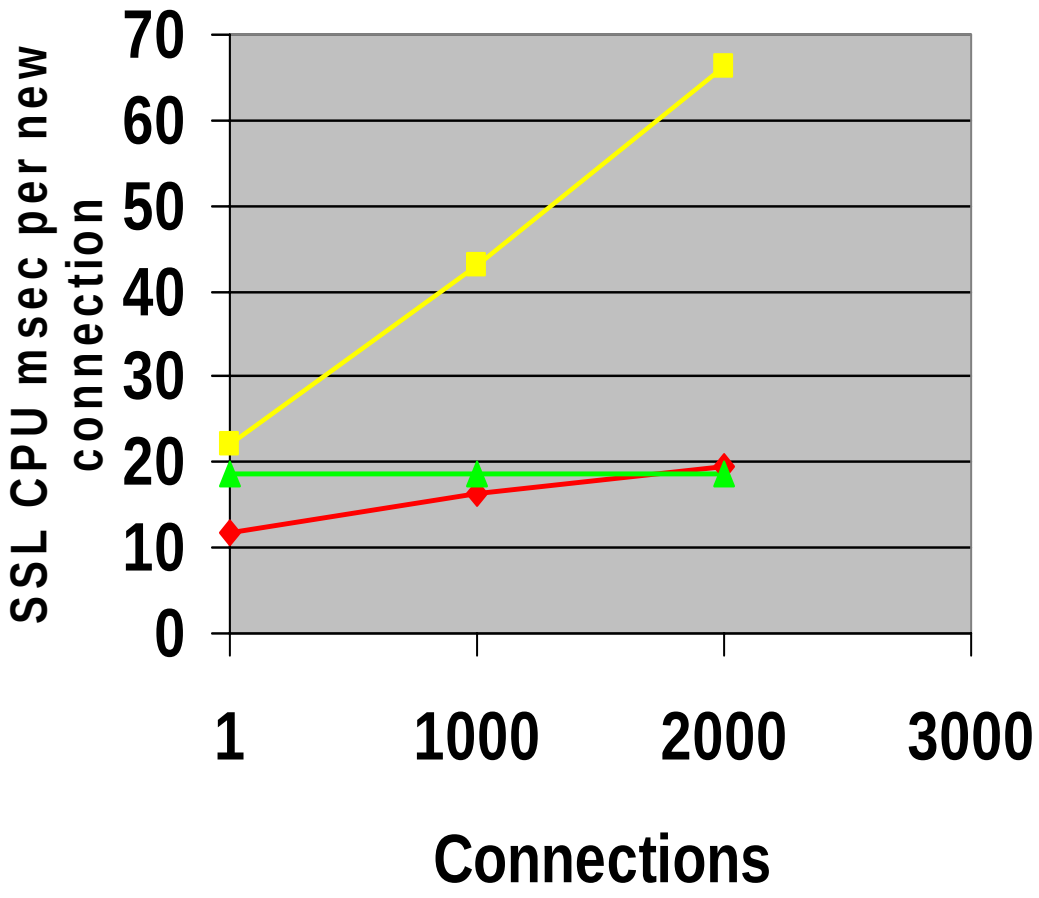
SSL Performance

- **In z/VM 5.4, the z/VM SSL server moved from being Linux-based to being CMS-based.**
  - APAR PK65850 shipped the support
- **Performance concerns compared to Linux-based server**
- **A group of related APARs to address performance**
  - All for z/VM 5.4 and 6.1
  - All now closed
  - PK75662 (stack)
  - PK97437 (packaging)
  - PK97438 (SSL)
  - VM64313 (CMS)
  - VM64740 (CMS)
  - PM06244 (SSL)
- **Because of significant changes in configuration for enhanced SSL, there is new documentation**
  - Overview: http://www.vm.ibm.com/related/tcpip/tcsslspe.html
  - Config: http://www.vm.ibm.com/related/tcpip/tcspepvs.html

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



SSL Multi was 10 servers with 200 clients on each.

Default configuration is 5 servers, 600 each.

System programmer can change config as needed.

Altitude of green line is a function of the configured maximum in the server.

Reorder Processing - Background

- ***Page reorder* is the process 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 on a z10
  - Delays of ~1 second for guest having 8 GB resident
  - This can vary for different reasons +/- 40%

Reorder Processing - Diagnosing

- **Performance Toolkit**
  - Check FCX113 UPAGE resident page fields R<2GB and R>2GB
  - Check FCX114 USTAT Console Function Mode wait %CFW
    - Reorders and CFW are somewhat correlated
- **REORDMON tool**
  - From Bill Bitner, on http://www.vm.ibm.com/download/packages/
  - Works against MONWRITE data or running system
  - Displays how often reorder happens

Reorder Processing - Mitigations

- **Keep guests as small as practical**
  - Perhaps split large guests with multiple applications each into several guests with one application each
- **Consider applying APAR VM64774**
  - Provides SET and QUERY commands with system-wide or per-user control
  - Corrects problem in earlier “patch” solution that inhibits paging of PGMBKs for virtual machines where reorder is set off.
  - z/VM 5.4 PTF UM33167 RSU 1003
  - z/VM 6.1 PTF UM33169 RSU 1003
- **See http://www.vm.ibm.com/perf/tips/reorder.html for more details.**

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**

VM64721 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.
    - Applies to only ABSOLUTE

Excess Share Distribution: Background

- **Shares are relative to other users that want to run.**
- **Example:**
  - Four compute-bound virtual machines on a 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 wants to use only 3%?

Excess Share Distribution Problem

User ID	Share	Normalize	Should Get	Problem Scenario
LINUX01	100	17%	24.5%	17%
LINUX02	100	17%	24.5%	17%
LINUX03	200	33%	48%	63%
LINUX04	200	33%	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 opened a 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

- **Sometimes, z/OS guests have minidisks**
- **Sometimes, z/OS guests do FlashCopy functions**
  - z/OS DFSMS and other utilities can make extensive use of FlashCopy for functions such as defragmentation
- **These two things do NOT play together well**
  - FlashCopy channel programs induce large numbers of MDC track invalidations
  - This can send z/VM storage management into a tizzy
  - Symptom is very high unexplained system time
- **Mitigations**
  - Turn off MDC for minidisks that are FlashCopy targets

VM64767: VARY PROCESSOR Hangs

- ***VARY PROCESSOR* command 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**
- **Order and apply the PTFs for these two APARs:**
  - VM64876, then
  - VM64767, which pre-reqs ‘876.
- **Fits z/VM 5.3, 5.4, and 6.1**

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 free storage pools when using dedicated FCP or OSA devices may lead to z/VM 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 64 KB buffer.**
  - Normally, it is 128 64 KB buffers (8 MB altogether).

VM64715 Page Release Serialization

- **z/VM 5.4 and 6.1 – still open, target 3Q 2011**
- **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

VM64965 – PE Correction for VM64862

- **Red alert: www.vm.ibm.com/service/redalert/**
- **VM64862**
  - HCPHRMDP may get wrong PTIL lock to invalidate STE
  - Locked wrong VMDBK's address space by mistake!
- **Affects z/VM 5.4 and 6.1**
- **Can cause abends in HCPHRM**
- **Watch for VM64965 (the correction) to close.**

VM64795 Enhanced Contiguous Frame Coalescing

- **Old way for coalescing free adjacent frames was exposed in certain scenarios**
- **Improved the coalesce function so as to help keep contiguous free frame lists populated**
- **Available now for z/VM 5.4 and 6.1**
  - 540 UM33244 November 2010 -- future RSU candidate
  - 610 UM33246 November 2010 -- future RSU candidate

Excessive PR/SM Overhead

- **CPU consumption falls into three categories**
  - Consumed by guests (FCX144 PROCLOG)
  - Consumed by z/VM Control Program (FCX144 PROCLOG)
  - Consumed by PR/SM hypervisor (FCX126 LPAR)
- **Some installations have seen the third category >100%**
  - Multiple engines burned up running PR/SM functions
  - Correlated with high CPU time in the z/VM Control Program
- **Usually due to poor configuration practices:**
  - Too many logical PUs compared to partitions' needs
  - Too many virtual PUs compared to guests' needs
- **Best practices:**
  - For each partition,
    - Configure just enough logical PUs to cover demand
    - Set LPAR weights appropriately
  - For each guest,
    - Configure just enough virtual PUs to cover demand
    - Set share appropriately
    - For Linux guests, consider cpuplugd to shut off unneeded virtual PUs

VM64927 z/VM Spin Lock Manager Improvement

- **When a z/VM logical PU senses lock contention, the logical PU tells PR/SM it wants to give up its physical PU**
  - So some other logical PU can run and thereby finish up and release the lock
- **Old way: z/VM just issues Diag x'44' to PR/SM**
  - Not a functionally rich interface – basically a dumb yield
- **New way: z/VM acts very differently**
  - Logical PU now knows which other logical PU is holding the lock it wants
  - SIGP Sense-Running to see if the holding logical PU is already running
    - If not already running, use Diag x'9C' to tell PR/SM to run the holder
    - If so, just spin
- **Behavior change is…**
  - z/VM stays out of PR/SM much better
  - When z/VM does in fact call PR/SM, z/VM tells PR/SM something genuinely useful
- **Savings for you is decreased PR/SM overhead**
  - "%Ovhd" in FCX126, first table
  - "%LPOVHD" and "%NCOVHD" in FCX126, second table
- **z/VM 6.1 UM33297 February 2011 -- and future RSU candidate**

More on Excessive PR/SM Overhead, z10

- **PR/SM itself was found in some workloads to be the cause of excessive PR/SM overhead**
- **Problem related to how PR/SM manages mutual exclusion (locking) in some situations**
  - Cache line getting dragged around
- **Benefits mostly seen in:**
  - High physical N-way (>32)
  - Larger numbers of partitions (>6)
  - Larger logical-to-physical ratios
- **MCL N24404.008, driver 79F, bundle 37a**

VM64887 Erratic System Performance

- **In systems with runnable VMDBKs >> logical PUs,**
  - … during reshuffle,
  - … PLDV overflow was not getting recorded.
- **Thus, after a logical PU cleared its PLDV,**
  - … it didn't know overflow had happened,
  - … so it didn't know to go check the dispatch list for work.
- **Thus, runnable VMDBKs would sit in the dispatch list,**
  - … forlorn and forgotten,
  - … until next reshuffle.
- **VM64887, UM33213 (5.4), UM33214 (6.1)**
  - Not on an RSU, but under consideration for a future one

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

- **z/VM Single System Image and Live Guest Relocation**
  - Start thinking how you would use it
  - Start planning for configuration whitespace
  - Start planning for what horizontal scaling might mean
  - Start planning for FICON capacity
    - ISFC will want much more FICON than it did previously

Summary

- **The adventure continues**
- **New improvements and fixes coming out in the service stream.**
- **See http://www.vm.ibm.com/perf/**

IBM Systems & Technology Group

Retired charts

Results For Various TCP/IP Services

Service	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 capacity available 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.