IBM Systems & Technology Group

z/VM Performance Update 2011

SHARE 117 - Orlando - Session 9452

IBM z/VM Customer Focus and Care
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.

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
- **z196 Performance**
- **Few thoughts on futures**
- **Thanks to the z/VM Performance Evaluation Team:**
  - Dean DiTommaso, Bill Guzior, Steve Jones, Virg Meredith, Patty Rando, Dave Spencer, Joe Tingley, Xenia Tkatschow, Brian Wade

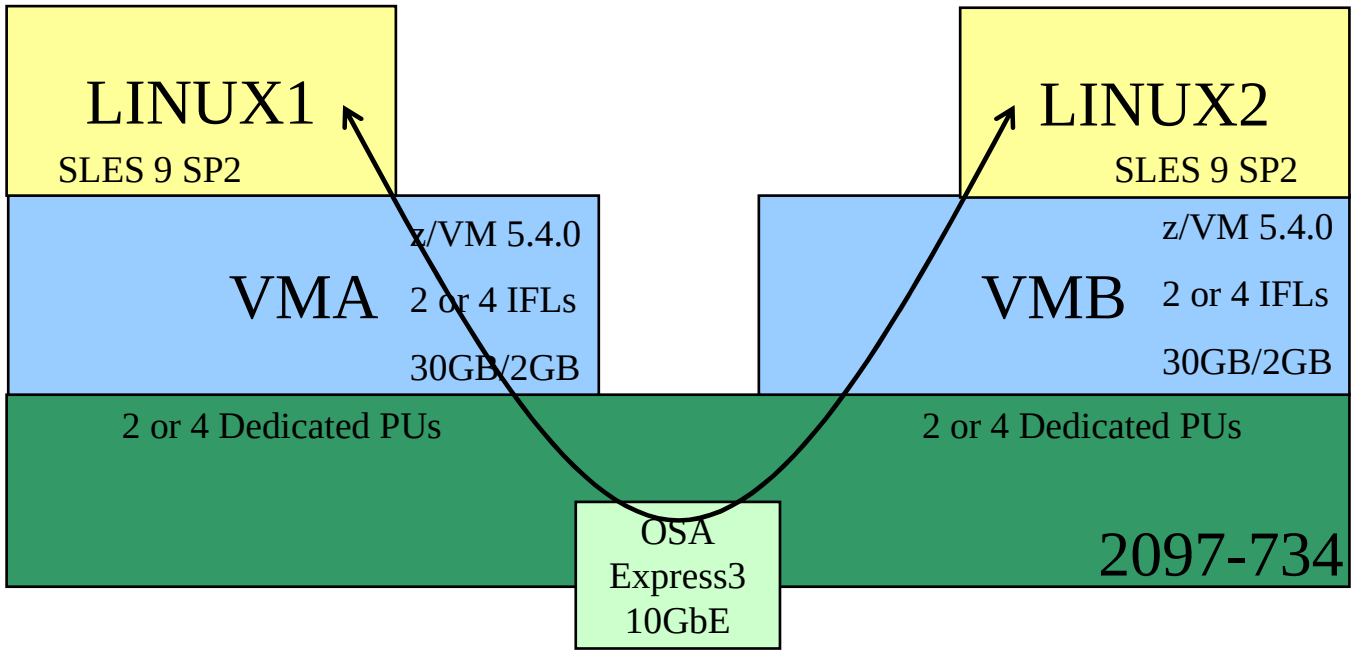
z/VM 6.1 Performance

- **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.
- **Service may start to show up that is z/VM 6.1.0 but not z/VM 5.4.0**

Network Performance Revisited

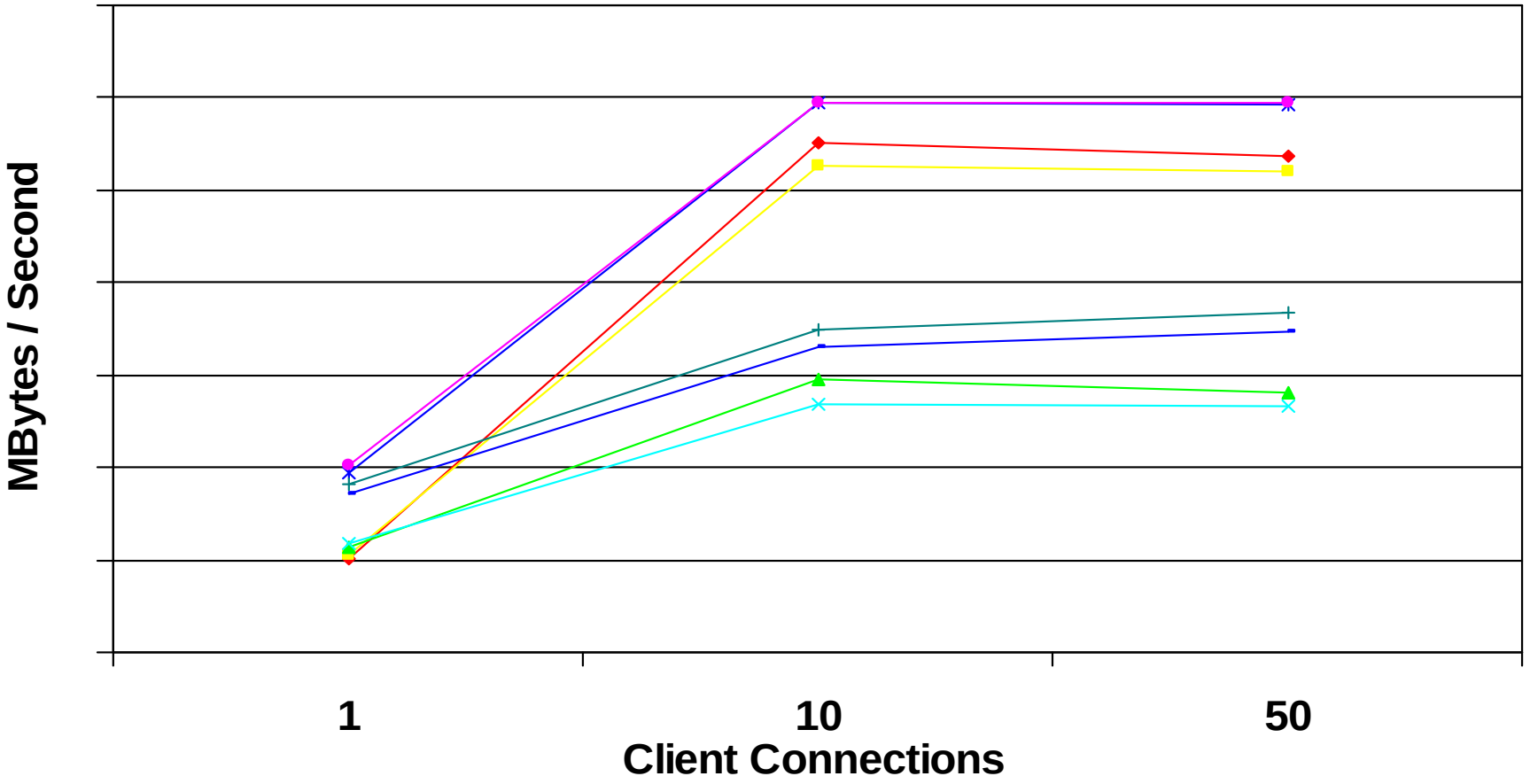
- **Measurement Environment and Workload Description**
- **Measurement Results**
  - Single Connection vs. Multiple Connections
  - MTU Size Comparisons
  - Dedicated OSA vs. VSwitch
- **Quantifying Throughput**
- **Hardware Performance Measurements**
- **Conclusions**

Measurement Configuration



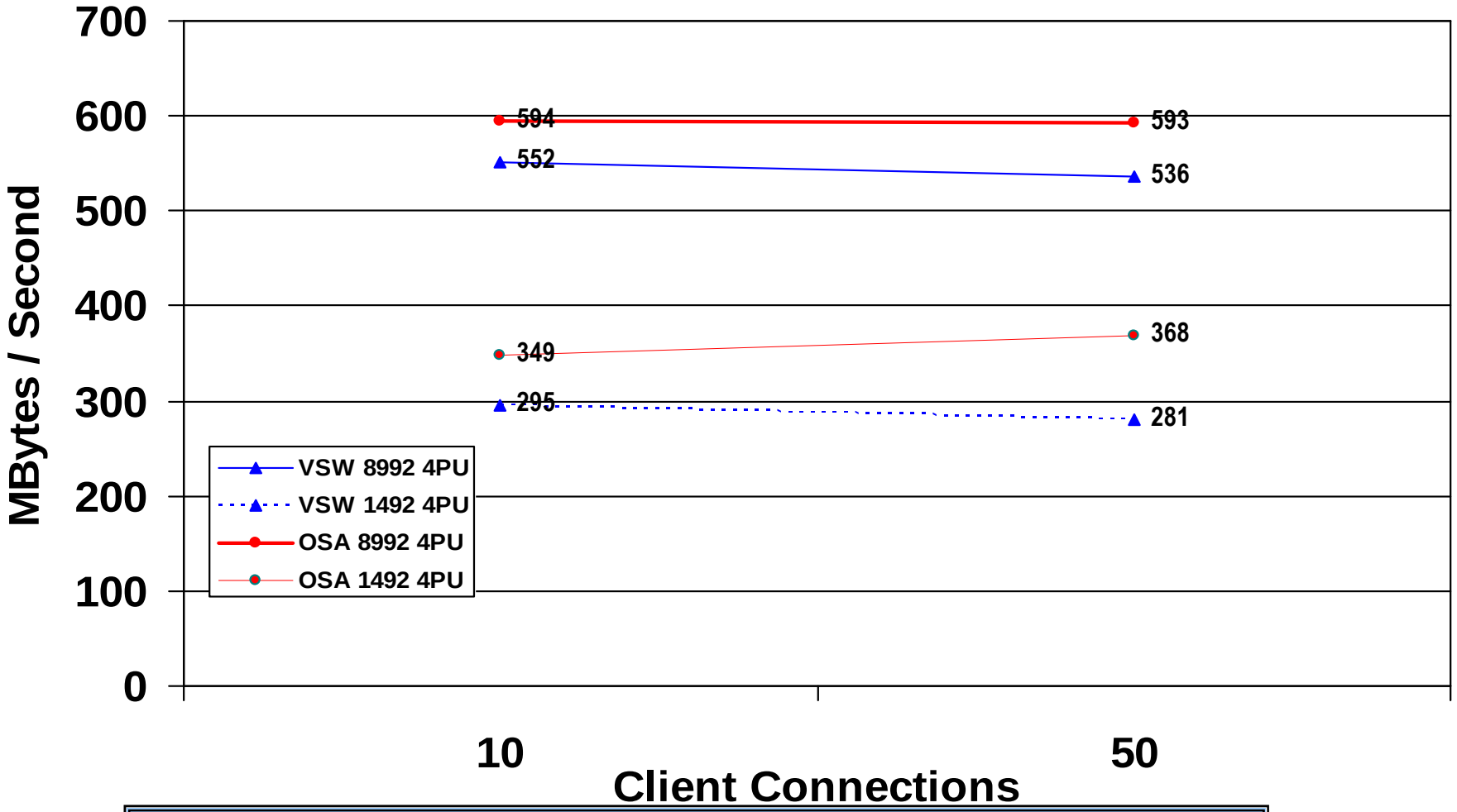
- Application Workload Modeler (AWM) used as the driver.
- Streaming workload – client sends 20 Bytes, receives 20 MBytes.
  - Throughput reported based on AWM data sent.
- Separate Ports on same OSA Express 3 card

Impact of Number of Connections



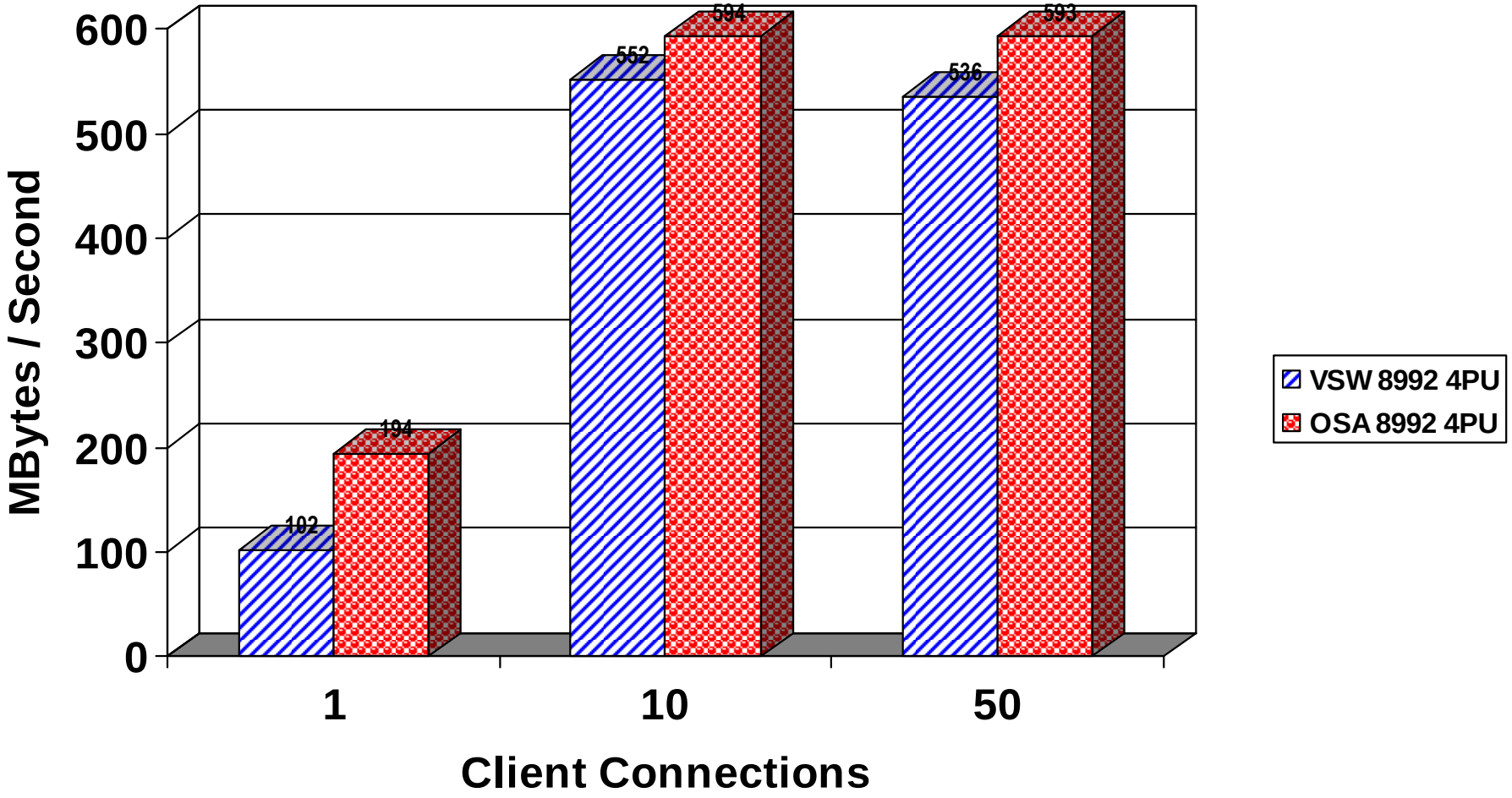
Need to be careful of single thread benchmark numbers. System z and z/VM optimize for large scale environments. Above, all measurements in these experiments are shown. You clearly see the difference between a single connection and multiple connection.

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

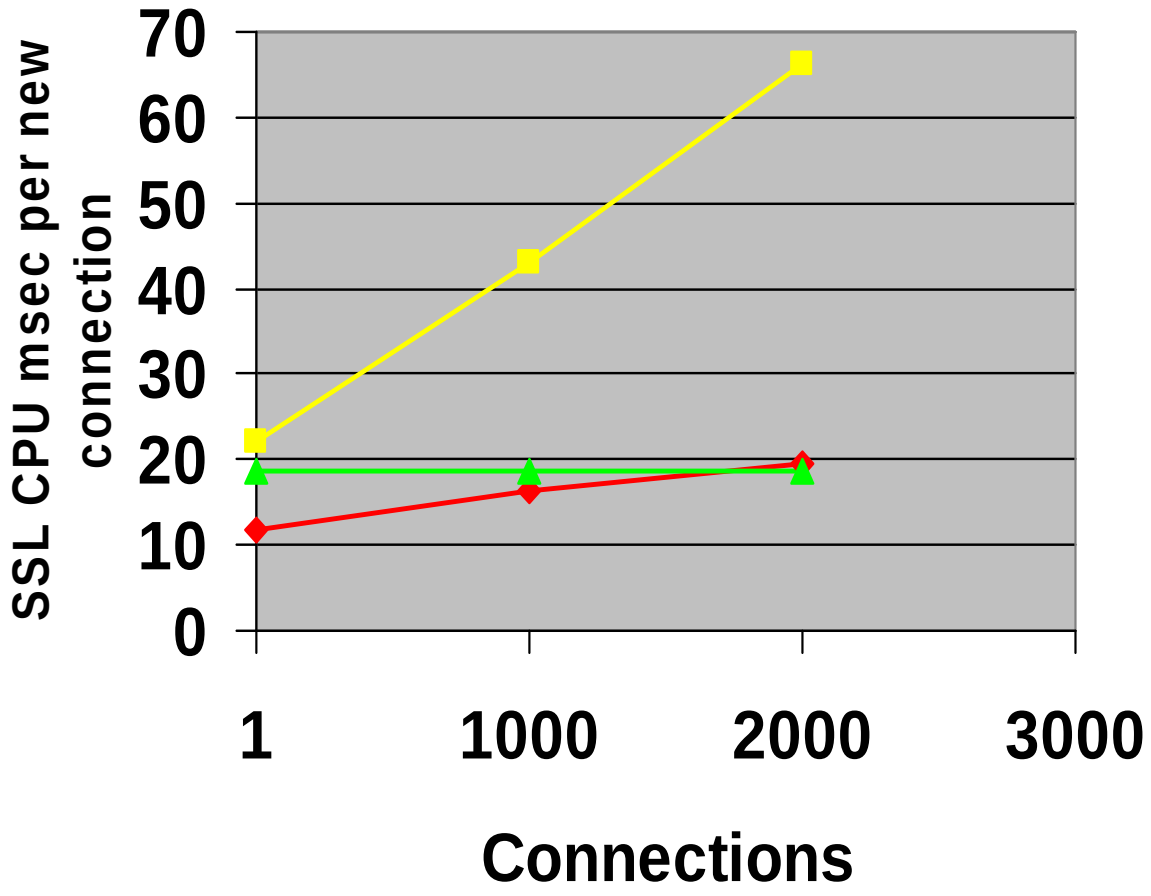
SSL Performance

- 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, closed in August 2010. PTFs for z/VM 5.4 and 6.1
- Because of significant changes in configuration for enhanced SSL, there is additional documentation
  - User memos
  - Red Alert material http://www.vm.ibm.com/service/redalert/
- http://www.vm.ibm.com/related/tcpip/vmsslinf.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 used 10 servers
  - 200 clients on each.
- Altitude of green line is a function of the configured maximum in the server.
- Cache Server not used for these measurements.

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 Significantly (Functional problems with high workload on old level)	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 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

REORDMON Example

Userid	Num. of Reorders	Average Rsdnt(MB)	Average Ref'd(MB)	Reorder Times	
LINUX002	2	18352	13356	13:29:05	14:15:05
LINUX001	1	22444	6966	13:44:05	
LINUX005	1	14275	5374	13:56:05	
LINUX003	2	21408	13660	13:43:05	14:10:05
LINUX007	1	12238	5961	13:51:05	
LINUX006	1	9686	4359	13:31:05	
LINUX004	1	21410	11886	14:18:05	

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”, moving to APAR.
- **Apply APAR VM64774 if necessary:**
  - SET and QUERY commands, system wide settings
  - Corrects problem in “patch” solution that inhibits paging of PGMBKs (Page Tables) for virtual machines where Reorder is set off.
  - z/VM 5.4.0 PTF UM33167 RSU 1003
  - z/VM 6.1.0 PTF UM33169 RSU 1003

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
- **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 closed and available on z/VM 5.3.0, 5.4.0, and 6.1.0**
  - R530 UM33028 Sept 2010
  - R540 UM33029 Sept 2010 RSU 1101
  - R610 UM33046 Sept 2010 RSU 1003

VM64527 MCW002 Abends from Memory Imbalance

- **Closed and available on z/VM 5.3, 5.4, and 6.1**
  - R530 UM32878 November 2009 RSU 1001
  - R540 UM32879 November 2009 RSU 1001
  - R610 UM32880 November 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

- **Closed and available on z/VM 5.4 and 6.1**
  - R540 UM33119 July 2010 RSU 1003
  - R610 UM33120 July 2010 RSU 1003
- **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**
- **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 went out for this.**
  - http://www.vm.ibm.com/service/redalert/
- **VM64862**
  - Abstract: HCPHRMDP MAY GET WRONG PTIL LOCK TO INVALIDATE STE
  - Dealt with PTIL lock and which VMDBKs were being used in various processing
- **z/VM 5.4.0 and z/VM 6.1.0 affected**
  - Problems include abends in HCPHRM
- **VM64965 closed and available on z/VM 5.4.0 and 6.1.0**
  - R540 UM33346 April 2011 Future RSU Candidate
  - R610 UM33347 April 2011 Future RSU Candidate

VM64887 – Erratic System Performance

- **VM64887 closed and available on z/VM 5.4.0 and 6.1.0**
  - R540 UM33213 October 2010 Future RSU Candidate
  - R610 UM33214 October 2010 Future RSU Candidate
- **The problem scenario:**
  - PLDV (Processor Local Dispatch Vector) overflow is later ignored for a period of time, resulting in virtual CPUs not being dispatch.
  - Appear as erratic system performance or ‘hang’ like scenarios.
  - Need to have enough work to overflow PLDV, >14 **active** virtual CPUs per logical processor.

VM64887 – Erratic System Performance

- **Snapshot of PLDV level**
  - CP INDICATE QUEUES EXP | COUNT LINES | CONS
    - Divide this by number of logical processors z/VM owns
- **Performance Toolkit PROCLOG FCX144**
  - "Mean when Non0" look for close to 14
  - Not conclusive, as after the overflow work can be removed from PLDV

			<-------- PLDV -------->				
	C		Pct	Mean	<-VMDBK-->		To
Interval	P		Em-	when	Mast	Stoln	Mast
End Time	U	Type	pty	Non0	only	/s	/s
11:37:00	0	IFL	37	9	0	286.0	3.2
11:37:00	1	IFL	43	9	....	374.2	.0
11:37:00	2	IFL	45	8	....	402.0	.0
11:37:00	3	IFL	43	8	....	337.5	.0

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.

VM64927 Additional Use of Diagnose x'9C' by CP

- **Excessive use of Diagnose x'44' can lead to higher LPAR Management Time**
- **Increases the number of spins attempted before issuing a diagnose**
- **Additional use of Diagnose x'9C' and SIGP Sense Running Status**
- **Additional Monitor Data for CP use of these diagnoses**
- **Closed and available for z/VM 6.1.0**
  - R610 UM33297 February 2011 RSU 1101

VM64795 Enhanced Contiguous Frame Coalescing

- **Environments existed where the pattern of requesting and returning contiguous frames resulted in a shortage of contiguous frames, leading to very high system processor requirements.**
- **Changes made to increase probability of having sufficient contiguous frames on available list.**
- **Closed and Available on z/VM 5.4.0 and z/VM 6.1.0**
  - R540 UM33244 November 2010 Future RSU Candidate
  - R610 UM33245 November 2010 RSU 1101

z/VM and CPUPLUGD Interactions

- **If using CPUPLUGD for Memory or CMM1 with guests >2GB, check with your Linux distribution for a fix to Linux use of Diagnose x'10' for pages greater than 2GB.**
  - Without fix, Linux never does page release on the pages above 2GB. Linux won't use those pages, but CP still lugs them around.
- **If using CPUPLUGD for CPU in paging environment, check with your Linux distribution for a change to Linux to workaround z/VM problem with PFAULT processing.**
  - Without fix, there is possibility that when a virtual CPU is stopped, threads waiting for completion of PFAULT (asynchronous page fault) may get lost and never run.

LPAR Improvement to Lower Overhead on z10

- **The required LPAR MCL with this change for driver 79F is N24404.008.**
  - **You would have to be at bundle 37a or later to have this.**
- **This is a pure performance improvement that was obtained by moving the LPAR lock that is getting serious contention to its own private cache line.**
- **Problem appears as high LPAR overhead.**
- **Benefit proportional to:**
  - **Larger physical n-way (>32 way)**
  - **Larger number of LPARs (>6)**
  - **Larger logical to physical processor ratio**
- **Problems does not apply to the z196.**

z196 Performance Expectations for z/VM

- **The glossy message appears to be “up to 60%” performance improvement for z/VM Linux environments.**
- **Official LSPR z/VM scaling ITR ratios of z196 to z10 are in the range 1.38 to 1.55**
- **Internal z/VM Lab scaling ITR ratios show similar range**
- **Overall better MP curves**
- **Factors:**
  - Larger N-ways tend to have better scaling ratios
  - Larger footprints that benefit from improved hardware cache have better scaling ratios
- **z9 to z10 comparison also had a wide range of ITR ratios**
  - Workloads that were lower in the range of ITRR for z9 to z10 comparison will tend to be higher in the range of ITRR for z10 to z196.

Preparing for z196 or z114

- **z/VM 5.4.0 and z/VM 6.1.0 with appropriate support**
- **Check buckets for very latest news and updates**
- **Cheat Sheet located on VM Home Page**
  - http://www.vm.ibm.com/service/vmreqze.html
- **No soup for you if you don't have the appropriate service**

- **Collect performance data** before **migrating to z196 or z114**
  - Raw z/VM Monwrite data
    - http://www.vm.ibm.com/perf/tips/collect.html
  - Application and/or information on load

LSPR Changes for z/VM & Linux Space

- **More current levels of various components**
  - Update SLES 9 to SLES 10
  - Update DB2 8.1 to 9.5
  - Update WebSphere 6.02 to 7.01
  - Update z/VM 5.2 to z/VM 5.4
- **Application Workload changed from Trade6 to Daytrader**
- **Higher n-way measured: 32-way LPARs**
- **Exploring workloads with resource over-commitment (virtual to real)**

Other LSPR Changes

- **z196 LSPR introduces new 'view' of performance**
  - Traditionally have used 'application' specific workloads for LSPR (CICS, IMS, etc.)
  - Move to workloads that map to how a workload interacts with the hardware, an evolving process.
  - Exploiting CPU Measurement Facility, introduced with z10
    - Data available in z/OS via SMF 113 records.
    - Known requirement for z/VM to support the CPU MF
- **Three new workloads that represent the Relative Nest Intensity (RNI):**
  - LOW (relative nest intensity)
    - Workload curve representing light use of the memory hierarchy
    - Similar to past high Nway scaling workload primitives
  - AVERAGE (relative nest intensity)
    - Workload curve expected to represent the majority of customer workloads
    - Similar to the past LoIO-mix curve
  - HIGH (relative nest intensity)
    - Workload curve representing heavy use of the memory hierarchy
    - Similar to the past DI-mix curve

CPU Measurement Facility

- **z/VM Support for Host Counters coming soon**
- **Main intent is for IBM use**
  - Customers may be directed to enable and collect…
    - for problem solving.
    - for help with capacity planning.
    - for IBM insight into future processor designs.

CPU MF – 2011 Update and WSC Experiences

Session 9999 with John Burg

Wednesday 3:00 Oceanic 3

Summary

- **The Adventure Continues**
- **New improvements and fixes coming out in the service stream.**
  - Most will be z/VM 5.4.0 and z/VM 6.1.0 but not all.
- **See http://www.vm.ibm.com/perf/**