IBM Systems & Technology Group

z/VM Performance Update for z/VM V6.2 VM02

Alan Altmark, Alan_Altmark@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.

Acknowledgements – Your z/VM Performance Team

- **Dean DiTommaso**
- **Bill Guzior**
- **Steve Jones**
- **Virg Meredith**
- **Patty Rando**
- **Dave Spencer**
- **Susan Timashenka – Dept Manager**
- **Xenia Tkatschow**
- **Brian Wade**

Agenda

- **z/VM V6.2 thoughts**
  - LGR and SSI
    - Performance notes
    - Management and monitoring thoughts
  - Various other line items
  - Monitor record changes
  - Performance-related service
- **Other thoughts**
  - z114 at a glance
  - Continued evolution of z/VM LSPR

z/VM V6.2 Highlights – A Performance View

- **Regression performance**
- **SSI and LGR considerations**
- **Memory management improvements**
- **MONDCSS and SAMPLE CONFIG increases**
- **STORBUF changes**
- **z/CMS and implications**
- **CPU Measurement Facility exploitation**
- **Monitor records**
- **z/VM Performance Toolkit changes**

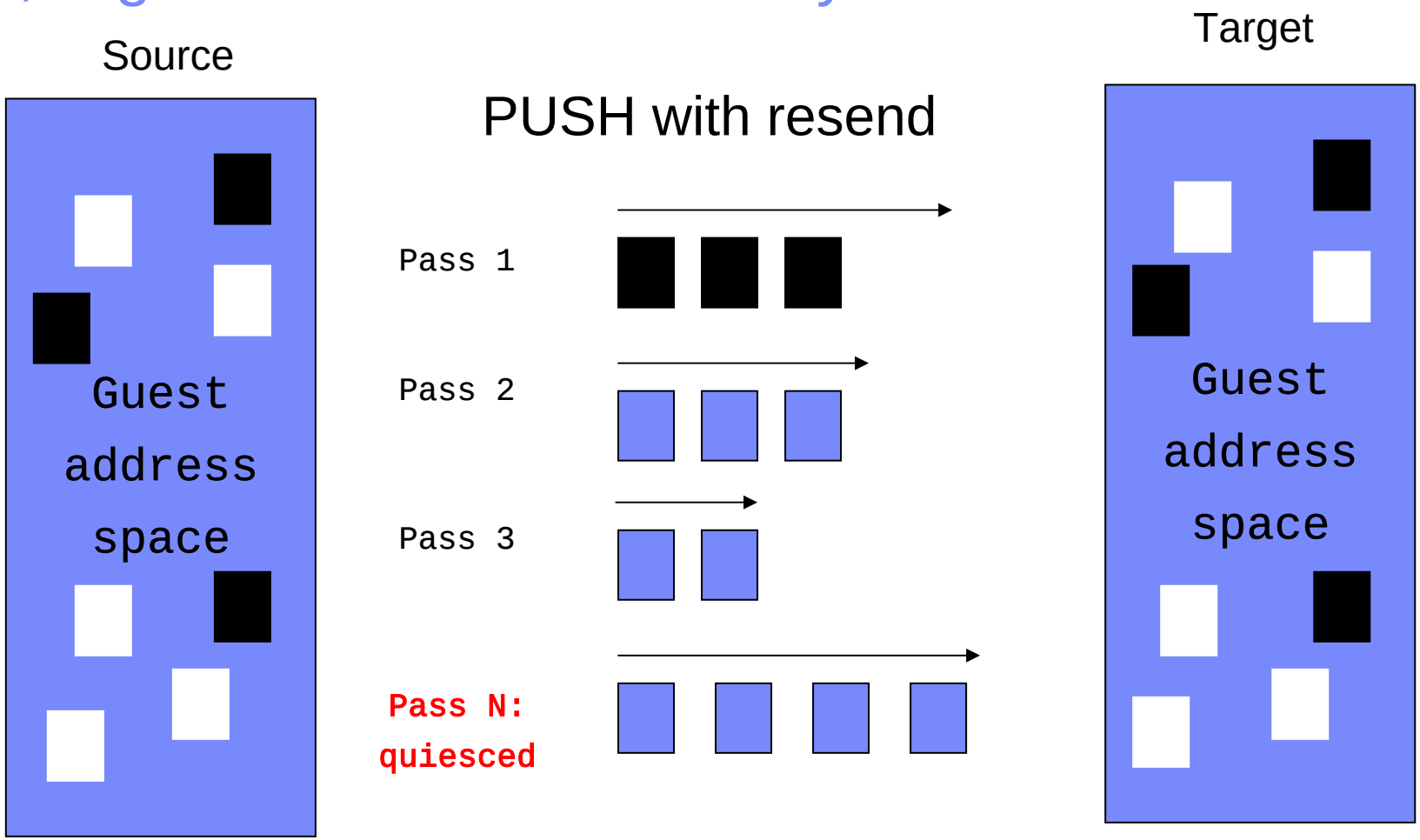
z/VM V6.2 Regression Performance

- **Ran our usual library of workloads**
  - CMS interactive, various Apache configurations
- **Results are within usual 5% regression criteria**
- **Some workloads will see improvements:**
  - Overprovisioned for logical PUs compared to utilization
  - Storage-constrained with heavy contention for <2 GB real storage
  - High virtual CPU to logical CPU overcommit with virtual CPUs often in a ready-to-run state

IBM Systems & Technology Group

SSI and LGR Thoughts

LGR, High-Level View of Memory Move



Technique: iterative push with change detection

Live Guest Relocation – Key Performance Metrics

- **Quiesce Time (QT)**
  - **Elapsed time that the guest is stopped (stunned) so z/VM can move the guest's last set of storage pages – probably the frequently-changed ones**
  - **To tolerate relocation, the guest and its applications must tolerate the quiesce time**
  - **VMRELOCATE can be invoked with a specified maximum quiesce time**
    - **If the quiesce would run past the maximum, z/VM cancels the relocation**
- **Relocation Time (RT)**
  - **Elapsed time from when the VMRELOCATE command is issued to when the guest is successfully restarted on the destination system.**
  - **Elapsed time must fit within the customer's window of time for planned outages for system maintenance, etc.**
- **Bottom line: there are some scenarios where LGR is not feasible as a result of the requirements for relocation time and quiesce time**

LGR: Factors Affecting QT and RT

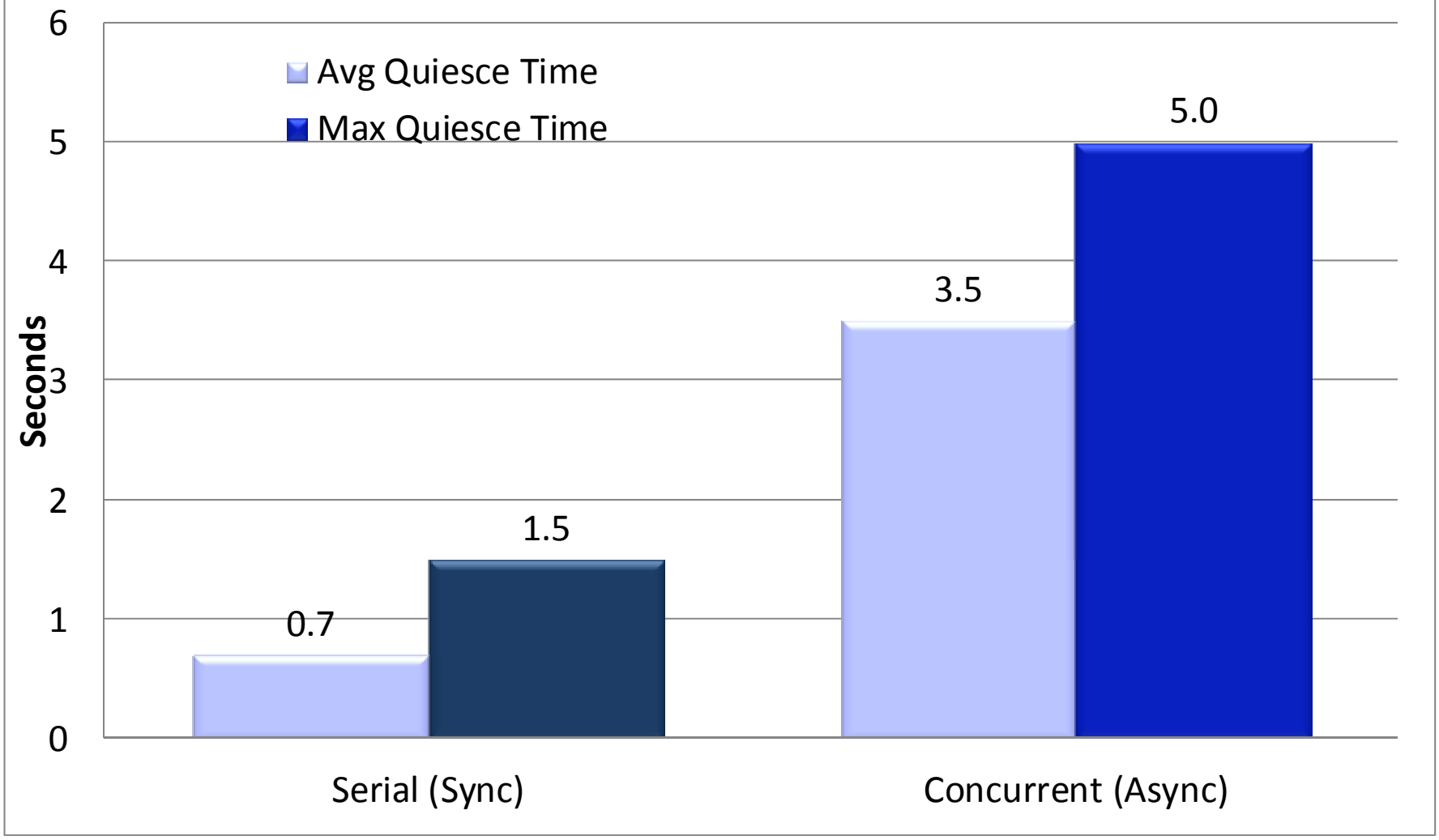
- **Size of the guest**
  - Amount of memory to move, time required to walk its DAT tables
- **How broadly or frequently the guest changes its pages**
  - It's an iterative memory push from source to destination
- **Time needed to relocate the guest's I/O configuration**
  - I/O device count, I/Os to quiesce, OSA recovery on target side
- **Capacity of the ISFC logical link**
  - Number of chpids, their speeds, number of RDEVs
- **Storage constraints on source and target systems**
- **Performance of paging subsystem**
- **Other work the systems are doing**
- **Other relocations happening concurrently with the one of interest**
  - There shouldn't be any!
- **LGR throttling of relocation to protect the z/VM system**

LGR: Serial vs. Concurrent Relocations

- **By default, the VMRELOCATE command operates synchronously.**
  - There is a command option to run it asynchronously (a la SPXTAPE)
- **You could achieve concurrent relocations, though:**
  - Use the asynchronous version of VMRELOCATE
  - Run VMRELOCATE commands in multiple users concurrently
- **The best practice, though, is to run only one relocation at a time.**
- **QT and individual RT improves substantially when relocations are done serially**
  - … and total RT elongates only slightly

Individual Virtual Machine Quiesce Time

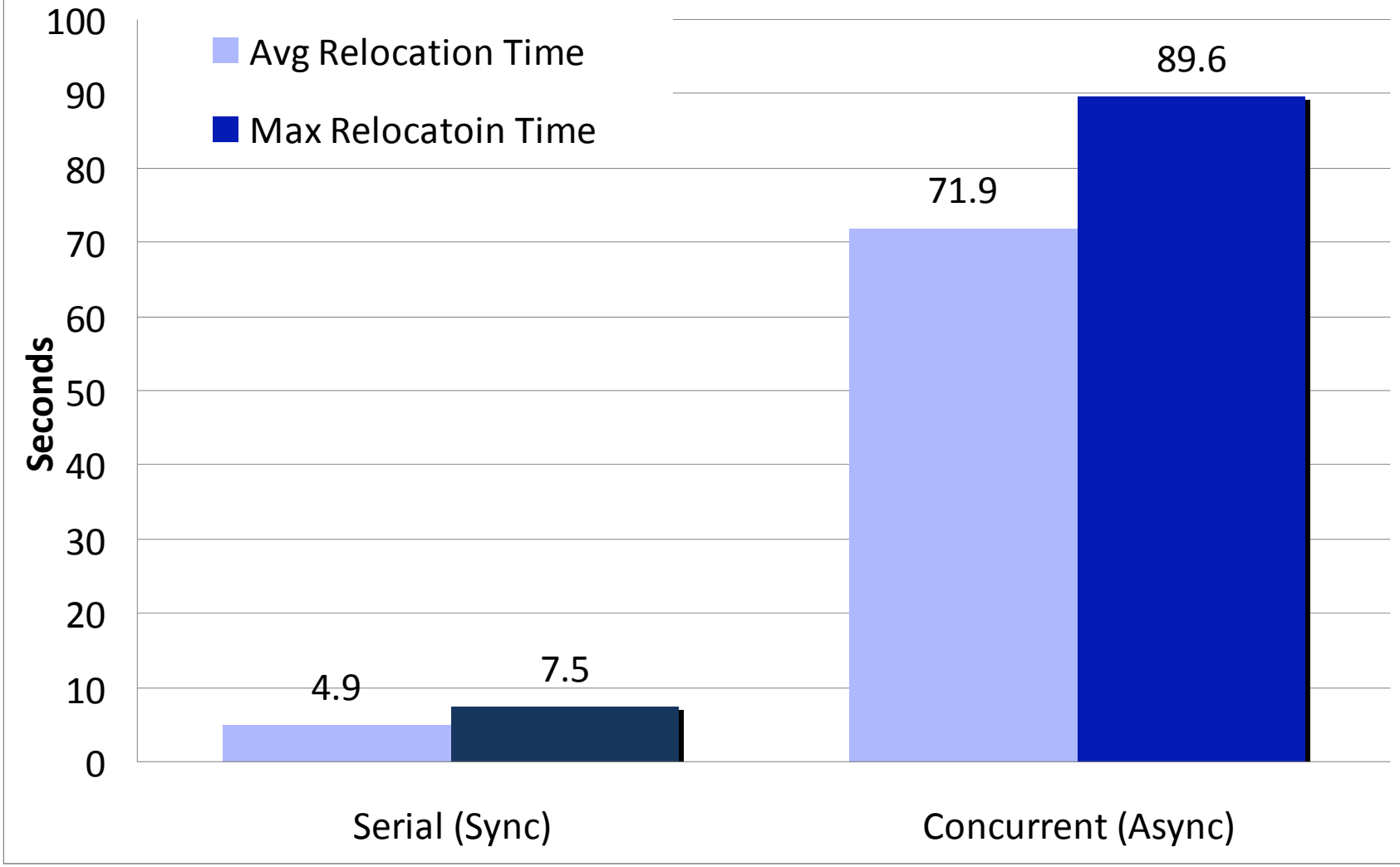
LGRSPEC Workload - 25 Virtual Machines



Preliminary results. z/VM 6.2 Performance Report is due later this year.

Individual Virtual Machine Relocation Time

LGRSPEC Workload - 25 Virtual Machines



Preliminary results. z/VM 6.2 Performance Report is due later this year.

ISFC Channel Impact on Quiesce Time

- For an 8 GB guest changing 6 GB of its pages
  - **20 seconds (4 Gb chpid)**
  - **14 seconds (2 Gb, 4 Gb chpids)**
  - **9 seconds (2 Gb, 4 Gb, 4 Gb chpids)**
  - **6 seconds (2 Gb, 4 Gb, 4 Gb, 8 Gb chpids)**
- For a 40 GB idle guest
  - **< 5 seconds (2 Gb, 4 Gb, 4 Gb, 8 Gb chpids)**

Notes:

1. CHPIDs (CTCs) are all FICON
2. Maximum efficiency is achieved with 4 to 5 RDEVs per CHPID
3. Best practice is to use CHPIDs of all the same speed… more on this later

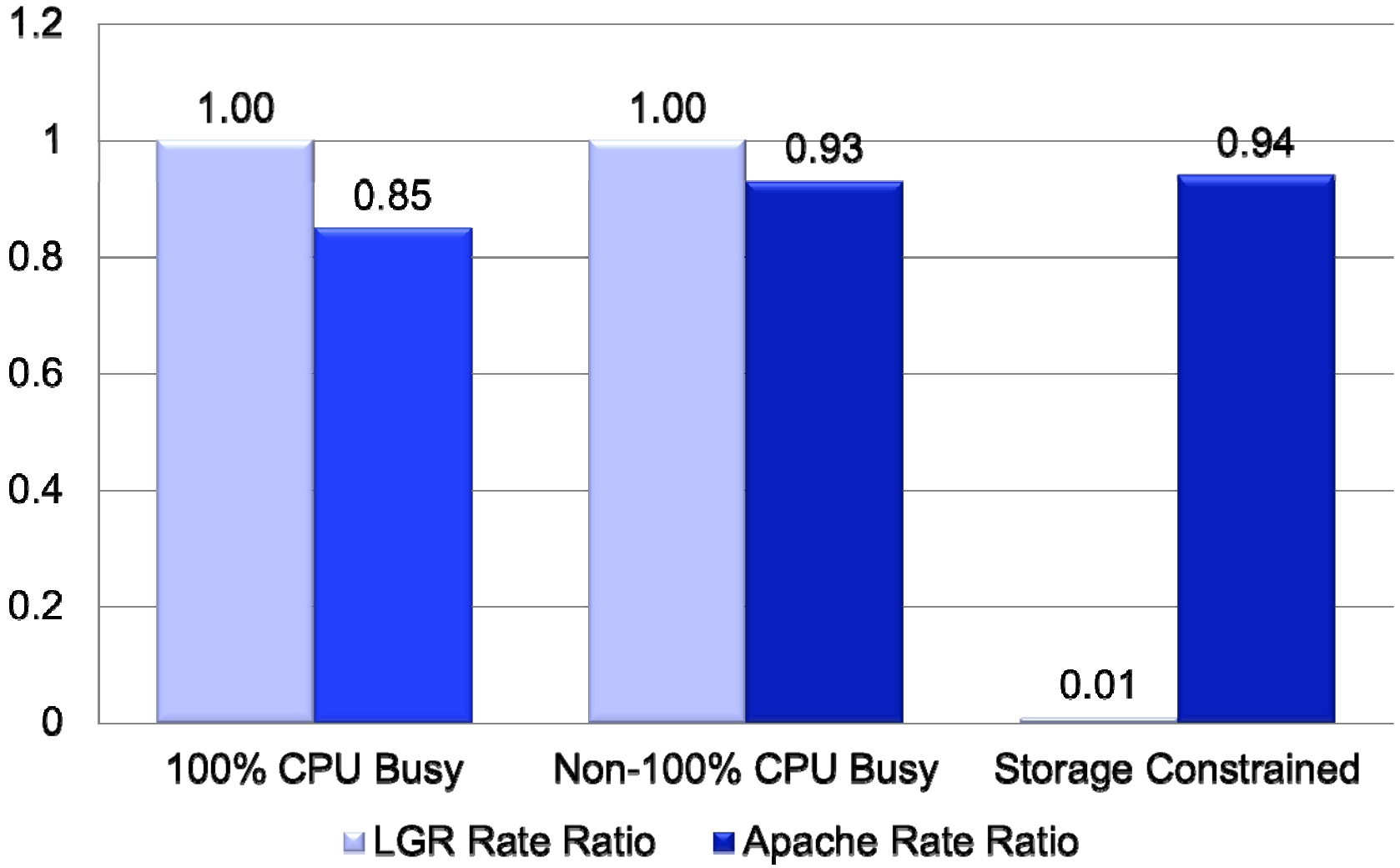
Preliminary results. z/VM 6.2 Performance Report is due later this year.

LGR: CPU and Memory Use Habits

- **CPU: generally LGR gets what it needs**
  - Taken “off the top” compared to your workload
- **Memory: CP tries really hard not to interfere**
  - End-to-end throttling, ISFC buffer limits, …
  - Socket memory-move throttling – triggered by memory consumption
  - ISFC logical link throttling – triggered by ISFC running out of queued traffic buffers
  - Considers effect on paging, memory use for specific relocations, …

Effect of LGR on Existing Workload

LGR Bounce and Apache Web Serving Workloads



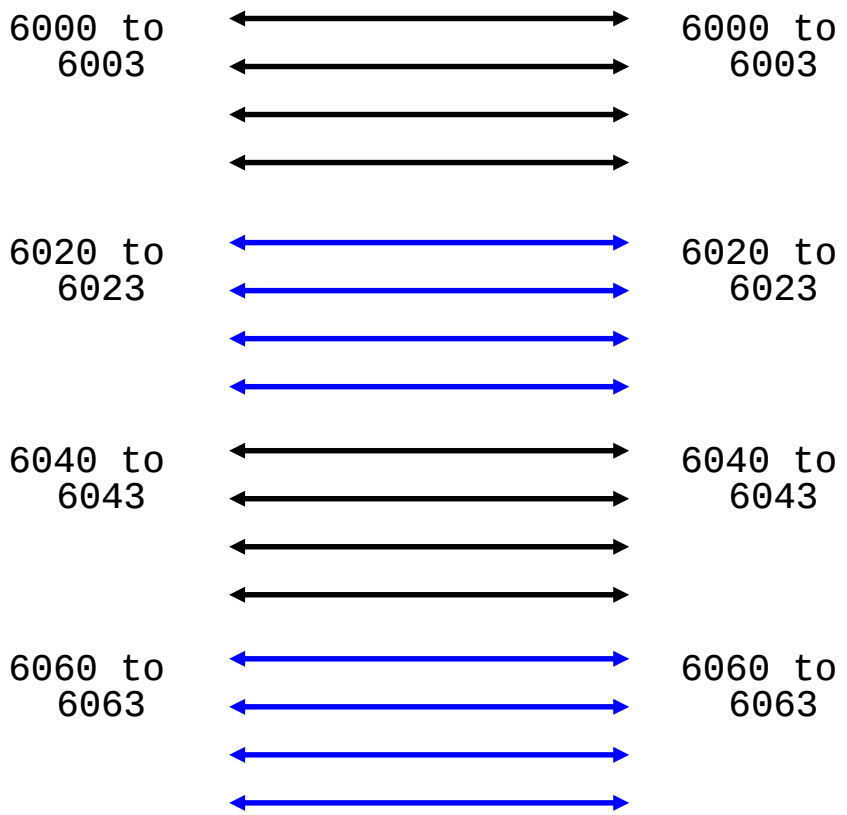
Preliminary results. z/VM 6.2 Performance Report is due later this year.

LGR: Keep These in Mind…

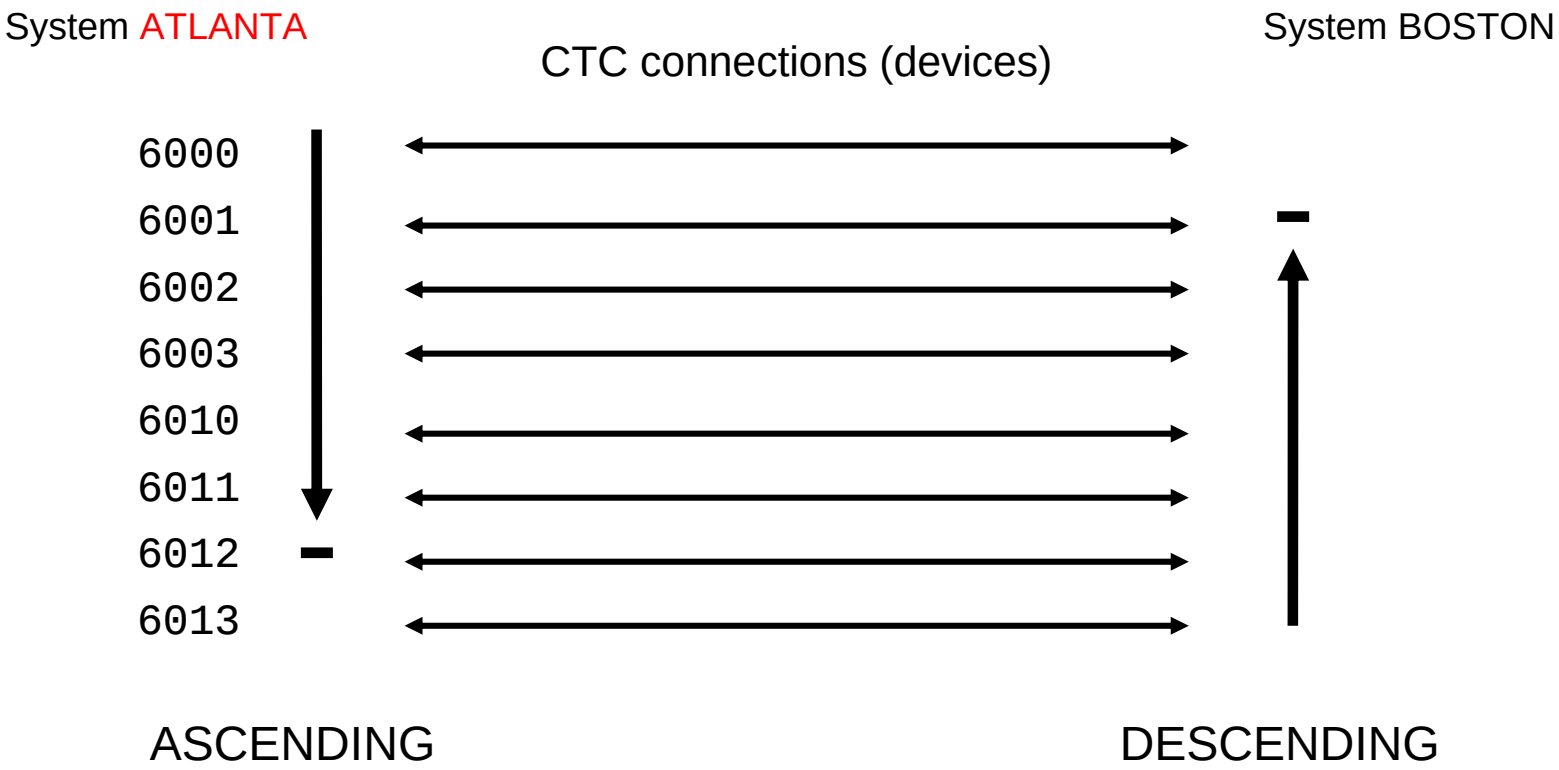
- **Charge back: can your procedures handle guests that suddenly disappear and then reappear somewhere else?**
- **VMRM: if VMRM-A tweaks the guest and then the guest moves to system B, what happens? And then what happens when the guest comes back?**
  - Best practice is to exclude relocating guests from VMRM-managed groups.
- **Second-level schedulers: do you have them? Can they handle guest motion?**

SSI: ISFC Logical Link Configuration Best Practices

- **Use four FICON chpids of all same speed**
- **Use four CTC devices per chpid**
- **Use same RDEV numbers on both ends**
- **More esoteric configurations are certainly possible**



SSI: ISFC Logical Link Write Scheduling, under the covers



Moral: put the fast chpids in the **middle** of ATLANTA's RDEV range.

Selection of where to start in selecting write path is alphabetical.

SSI: Contrived Workload Illustrates ISFC Traffic Scheduling

From H001569C PERFKIT B M-HL P-50 R-12

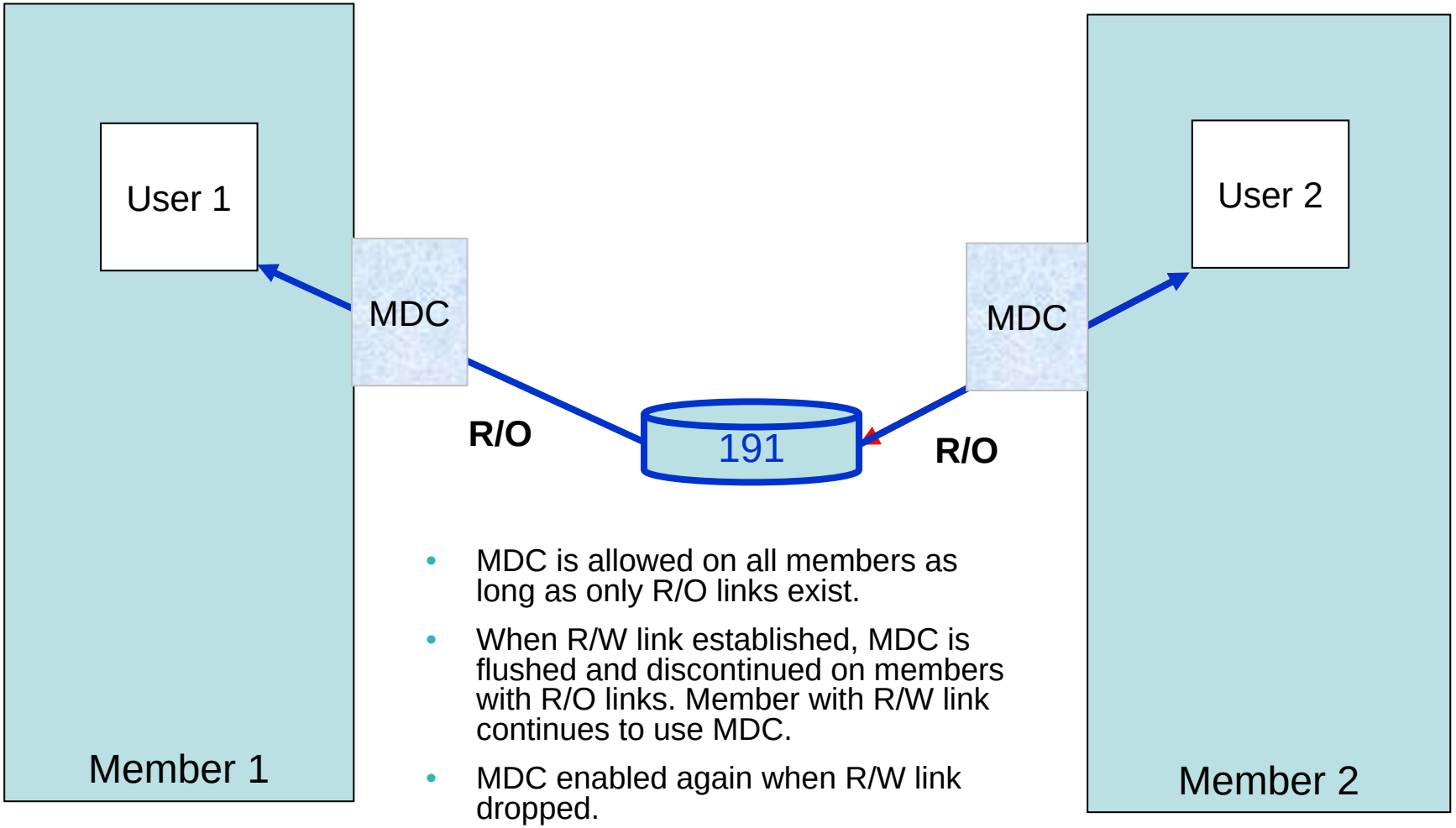
<-- Device Descr. -->			Mdisk	Pa-	<-Rate/s->		<------- Time (msec) ------->				Req.	<Percent>	
Addr	Type	Label/ID	Links	ths	I/O	Avoid	Pend	Disc	Conn	Serv	Resp	CUWt	Qued	Busy	READ
6000	CTCA		...	1	61.8	...	.5	1.7	13.7	15.9	15.9	.0	.0	98	..
6001	CTCA		...	1	61.7	...	.5	1.7	13.7	15.9	15.9	.0	.0	98	..
6002	CTCA		...	1	61.6	...	.5	1.7	13.7	15.9	15.9	.0	.0	98	..
6003	CTCA		...	1	61.6	...	.5	1.7	13.7	15.9	15.9	.0	.0	98	..
6020	CTCA		...	1	61.3	...	.5	1.8	13.7	16.0	16.0	.0	.0	98	..
6021	CTCA		...	1	61.5	...	.5	1.7	13.7	15.9	15.9	.0	.0	98	..
6022	CTCA		...	1	61.3	...	.5	1.7	13.8	16.0	16.0	.0	.0	98	..
6023	CTCA		...	1	61.4	...	.5	1.8	13.7	16.0	16.0	.0	.0	98	..
6040	CTCA		...	1	173	...	.4	1.9	3.2	5.5	5.5	.0	.0	95	..
6041	CTCA		...	1	173	...	.4	1.8	3.2	5.4	5.4	.0	.0	94	..
6042	CTCA		...	1	.9	...	.3	.3	1.0	1.6	1.6	.0	.0	0	..
6043	CTCA		...	1	525	...	.2	.1	.8	1.1	1.1	.0	.0	58	..

Run H001569C talking over link GDLBOFVM, config HL, P=50, R=12

____ISO-UTC_____	_TXPENDCT_	_WCol/sec_	_WMB/sec__	_WMsg/sec_	_WPkg/sec_	_WByt/pkg_	_WMsg/pkg_
2011-09-27 02:59:50.402251	32.0	0.0	663.2	5876.9	838.2	829648.3	7.0
2011-09-27 03:00:50.399438	26.0	0.0	662.0	5867.9	836.9	829443.2	7.0
2011-09-27 03:01:50.411664	20.0	0.0	662.1	5869.2	837.0	829460.3	7.0
2011-09-27 03:02:50.397239	18.0	0.0	661.2	5860.3	835.9	829386.5	7.0

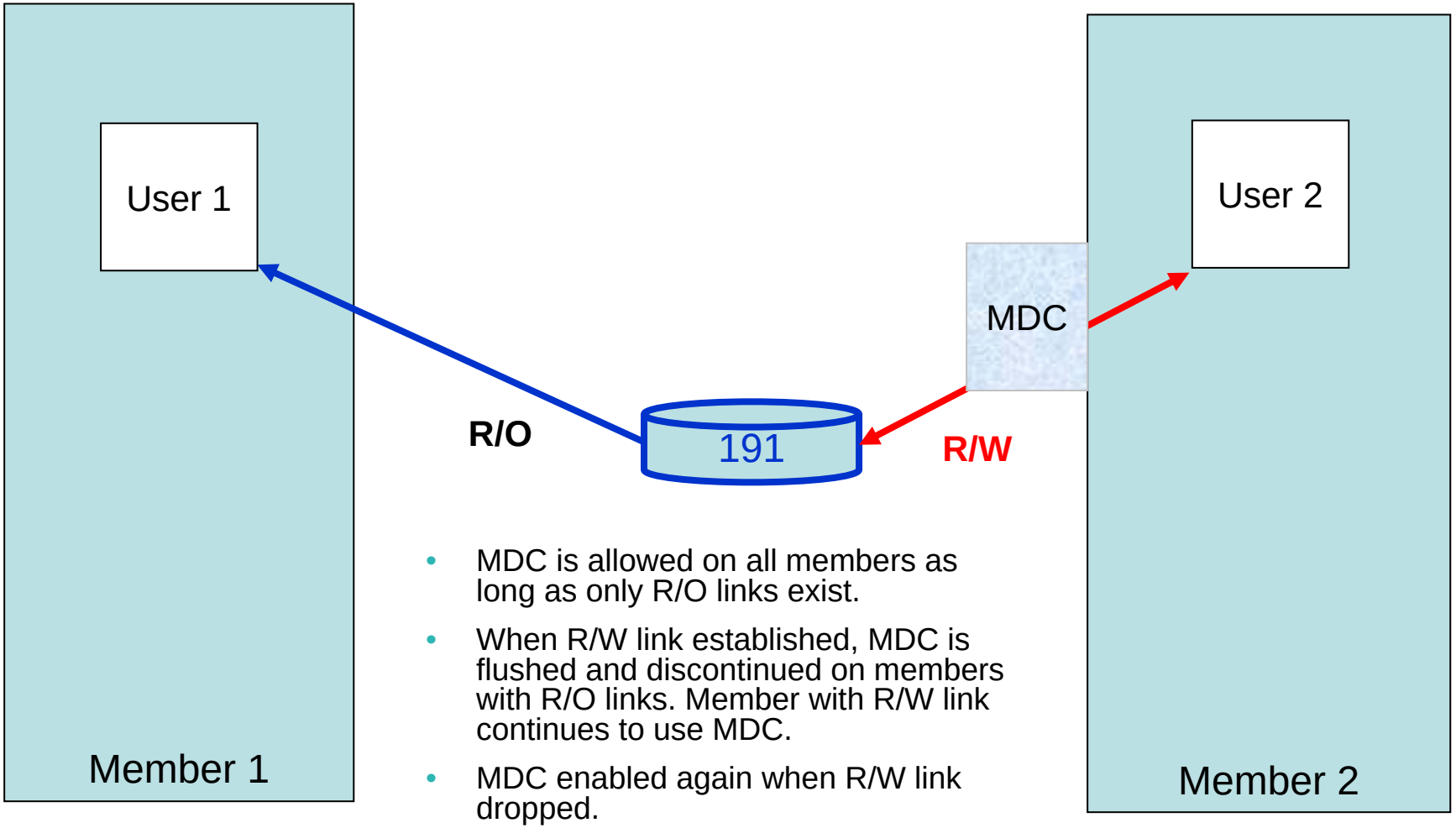
6000-6003 2 Gb/sec; 6020-6023 2 Gb/sec; 6040-6043 4 Gb/sec

SSI: Automatic MDC Management



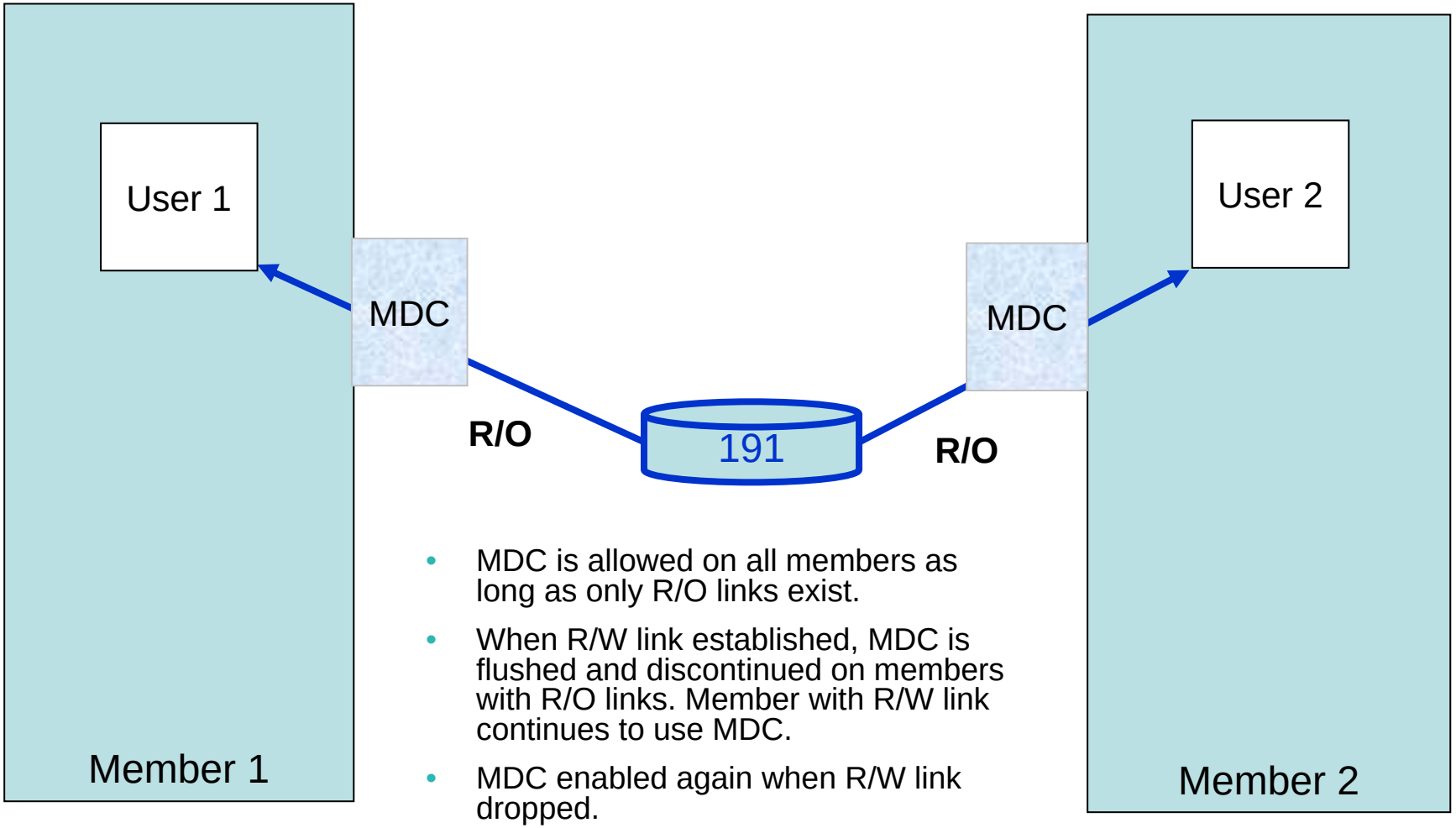
- MDC is allowed on all members as long as only R/O links exist.
- When R/W link established, MDC is flushed and discontinued on members with R/O links. Member with R/W link continues to use MDC.
- MDC enabled again when R/W link dropped.

SSI: Automatic MDC Management



- MDC is allowed on all members as long as only R/O links exist.
- When R/W link established, MDC is flushed and discontinued on members with R/O links. Member with R/W link continues to use MDC.
- MDC enabled again when R/W link dropped.

SSI: Automatic MDC Management



- MDC is allowed on all members as long as only R/O links exist.
- When R/W link established, MDC is flushed and discontinued on members with R/O links. Member with R/W link continues to use MDC.
- MDC enabled again when R/W link dropped.

SSI: Performance Toolkit, Considerations

- **Performance Toolkit continues to run separately on each member of the cluster**
  - There continues to be a unique z/VM monitor data stream for each member.
  - There will be a PERFSVM virtual machine on each member
- **Configuration and usage**
  - Configure so that you will log onto or connect to a different PERFSVM on each system.
  - Configure Performance Toolkit to use the Remote Performance Monitoring Facility, which allows local and remote performance monitoring from a single screen.
- **In general, Performance Toolkit does not produce "cluster view" reports**
  - DASD device-busy view, for example

SSI: Performance Toolkit, New Reports

- **New Reports for SSI**
  - SSICONF: SSI configuration
  - SSISCHLG: SSI state change synchronization activity log
  - SSISMILG: SSI state/mode information log
- **New ISFC reports related to SSI**
  - ISFECONF: ISFC end point configuration
  - ISFEACT: ISFC end point activity
  - ISFLCONF: ISFC logical link configuration
  - ISFLACT: ISFC logical link activity
  - ISFLALOG: ISFC logical link activity log

SSI: MONWRITE Considerations

- **IBM often asks you to run MONWRITE**
  - PMR diagnosis, for example
- **You should be running MONWRITE anyway**
- **You should now be running MONWRITE on every member of the cluster**
- **Make sure it's easy to go find the MONWRITE data for all members for a specified time interval**

SSI: Dump and PMR Considerations

- **To solve your PMR,**
- **… IBM might need concurrently-taken dumps.**

- **Just be prepared:**
  - Know how to take a SNAPDUMP. Practice.
  - Know the effect of SNAPDUMP on your workload.
  - Know how to take a restart dump.

SSI: Capacity Planning

- **Great flexibility in managing multiple LPARs**
  - Previously, if you split work across LPARs and had an imbalance, it was more difficult to rebalance
  - With SSI, virtual machines can run anywhere in the cluster without a lot of additional work
- **Greater responsibility in planning, at two levels**
  - Individual members
    - Need to ensure sufficient capacity and resources for the workload on each member
    - Track growth in requirements to limits of the member
  - Cluster-wide
    - Track growth in requirements of overall cluster to the limits of that cluster
    - Need to ensure sufficient white space for planned outages where LGR will be used to move workload out of a given member.

The “Getting Started With Linux” book has been updated with SSI and LGR planning tips.

SSI & LGR: Planning White Space

- **Need white space for planned outages where you move work off of a given member .**
- **How will work move off the member?**
  - Use existing HA solutions to redirect work to existing servers on other members or elsewhere in enterprise.
  - Use LGR to move to another member.
  - Log off and then logon to another member.
  - Shutdown non-critical virtual machine for duration of unplanned outage.
- **To where do you move the virtual machines?**
  - To a single member or multiple members?
  - To a member on same CEC or another CEC?
  - To a member held in reserve (such as a DR LPAR)?
  - It's not just one z/VM image anymore

SSI & LGR: Planning White Space

- **CPU**
  - Shared logical processors?
  - Adjust LPAR weight settings?
  - Vary on additional engines?
- **I/O**
  - Ensure sufficient resources at all levels:
    - Channel, switch, control unit, device
  - Shared channels?
- **Memory white space is not as easy to manage**
  - Ensure sufficient paging space and concurrency or data rate capability
  - Increase real memory over commitment?
  - Temporarily decrease size of some virtual machines?
  - Use Dynamic Memory Upgrade?
    - No downgrade available

IBM Systems & Technology Group

z/VM 6.2 – More Than Just SSI and LGR

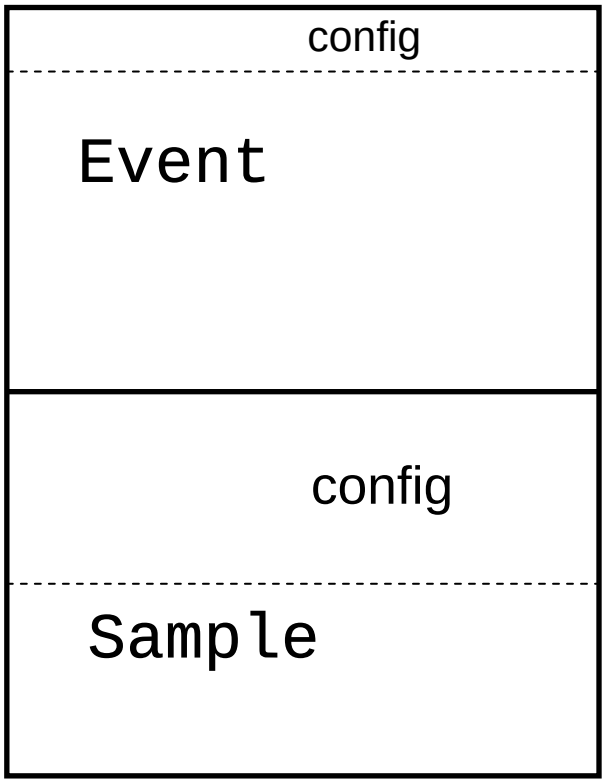
Memory Management: Needle-in-Haystack Searches

- **Searching for a below-2-GB frame in lists dominated by above-2-GB frames**
  - In months of study we identified about 10 of these searches
  - Development prototype that shut off all unnecessary use of <2GB storage gave us tremendous results
- **z/VM now does not allocate pageable buffers <2GB if:**
  - Dynamically, usable >2GB to usable <2GB is beyond a certain threshold
  - Statically, if the partition is beyond a certain size, for the life of the IPL
- **Result: no more needle searches**
- **Practically speaking, systems with 128 GB or more of real memory use below-2-GB memory only when it is architecturally required.**

MONDCSS and SAMPLE CONFIG Changes

- **The old defaults are too small for most systems nowadays**
- **So we have changed the default layout**
- **MONDCSS is 64 MB now (16384 pages)**
  - Half (32 MB) for EVENT
  - Half (32 MB) for SAMPLE
    - Half (16 MB) for SAMPLE CONFIG
- **As before, empty pages are not instantiated**
- **Remember, config pages evaporate after a short time**
- **MONWRITE 191 disk also increased to 300 cylinders.**
- **If you use your own MONDCSS, the new default SAMPLE CONFIG size may be too large, requiring you to set it manually or to change your MONDCSS.**

MONDCSS – 16384 pages



Default STORBUF Changes

- **Many parties were noticing that the old defaults of 125 105 95 were not appropriate for Linux workloads**
- **We considered several different proposals**
  - From IBM ATS
  - From vendors
  - From Redbooks
  - From customer data
- **After careful consideration by “top people” we came to 300 250 200 as new defaults**
- **If you already override defaults, the only impact would be if you also use SET SRM STORBUF INITIAL at some point.**
- **For CMS-intensive workloads, the old defaults might be more appropriate, and you should validate the settings for these workloads when you migrate to z/VM 6.2**

z/CMS

- **Prior to z/VM V6.2, z/CMS was supplied as a sample.**
- **z/VM V6.2 supports z/CMS as an optional alternative to the standard CMS that runs in ESA and XC mode virtual machines and 31-bit addressing.**
- **z/CMS can run in a z/Architecture guest**
  - Allows programs to use z/Architecture instructions, including 64-bit addressing
- **Standard CMS function does not exploit memory above 2GB**
- **Remember that z/Architecture is not XC**
  - No VM Data Spaces
  - No SFS DIRCONTROL directories in a data space
  - No DB2 for VM and VSE data space use
- **The standard, usual, XC-mode CMS is still there**

CPU Measurement Facility Counters

- **CPU MF counters are a System z hardware facility that characterizes the performance of the CPU and nest**
  - Instructions, cycles, cache misses, and other processor related information
- **Available on z10 EC/BC, z196, and z114**
- **The CPU MF counter values:**
  - Help IBM to understand how your workload stresses a CEC for future design
  - Help IBM to map your workload into the LSPR curves for better sizing results
  - Help IBM better understand your system when there is a processor performance related problem.
- **z/VM 6.2, 6.1, and 5.4 can all collect the CPU MF counters**
  - z/VM 5.4 and 6.1: VM64961, UM33440 (5.4), UM33442 (6.1)
- **We want volunteers to send us MONWRITE data!**
  - Your contributions will help us to understand customer workloads!

CPU MF Counters and CP Monitor, Details

- **Counter sample record is in the Processor domain**
- **MONITOR SAMPLE command manipulates counter collection**
- **QUERY MONITOR reveals whether counter collection is on**
- **z/VM writes the collected counters into the Monitor data stream**
  - Domain 5 Record 13: MRPRCMFC, Processor domain, sample record
- **The D5 R13 records land in your MONWRITE data**

IBM Wants Your CPU MF Counter Data

- **Your data will help IBM to build a library of customer workloads**
- **Collect an hour's worth of MONWRITE data…**
  - From a peak period,
  - With CPU MF counters enabled,
  - With one-minute sample intervals
- **Contact Richard Lewis at rflewis@us.ibm.com**
- **Richard will send you instructions on how to transmit the data to IBM**
- **No deliverable will be returned to you**
- **We will be ever grateful for your contribution**

Monitor Records – Highlights – New and Almost-New

- **In domain 1 (monitor), ISFC and SSI config records**
- **In domain 1, system topology record (PU-book-chip)**
- **In domain 4 (user), LGR start and LGR end**
- **In domain 5 (processor), CPU MF and system topology**
- **In domain 6 (I/O), minidisk MDC setting change event**
- **New domain 9 – ISFC performance records**
- **New domain 11 – SSI performance records**

- **Other changes to report on LGR, mostly in user domain**

z/VM 6.2: Service Integrated in Base of z/VM 6.2

- **VM64774 SET/QUERY REORDER command**
- **All of the SSL scaling fixes**
- **VM64721 LIMITHARD now works**
  - SET SRM LIMITHARD CONSUMPTION is default now
- **VM64767/64876 VARY PROCESSOR causes hangs**
- **VM64850 VSWITCH failover buffer mixup**
- **VM64795 Enhanced Contiguous Frame Handling**
- **VM64927 Spin Lock Manager Improvement**
- **VM64887 Erratic System Performance (PLDV overflow)**
- **VM64756 Long CPEBK Chains, Master-only work, and SYSTEMMP**

IBM Systems & Technology Group

z/VM Performance: Other Thoughts

z114 Performance

- **We ran workloads to help evaluate z114 compared to z196**
- **Equal N-way: about 0.65 of a z196**
- **Remember, it's a smaller machine than z196**
  - Only 10 engines, not 80
  - Only 248 GB, not 3072 GB
- **For more information:**

  http://www-03.ibm.com/systems/z/hardware/zenterprise/z114_specs.html

Evolution of z/VM LSPR Workload

- **From memory-rich to memory-constrained**
- **From 16-way to 32-way**
- **From equally-active to unequally-active**
- **From workload-indexed to RNI-indexed**
  - We do want your CPU MF counter data
- **Our goal is a lab setup that represents z/VM customers' environments**

Summary

z/VM Performance Update: Summary

- **z/VM v6.2: SSI and LGR, plus more**
  - Loose clustering for guest mobility
  - Recognition of systems becoming larger
    - Memory management improvements
    - Better defaults: MONDCSS, SAMPLE CONFIG, STORBUF
  - CPU MF counters: help us help you
  - Lots of good service rolled into the base
  - See http:/www.vm.ibm.com/perf/ for more details
- **LSPR: we keep trying to improve**
- **The adventure continues**

Contact Info:

Alan Altmark
Senior Managing IT Consultant
IBM STG Lab Services
Endicott, NY
Alan_Altmark@us.ibm.com

IBM Systems & Technology Group

Backup Charts – Just keep these in the deck in case somebody asks, “Remind me about that again, please?”

z/VM 6.2: Open Switches and Things Worth Watching

- **Reorder: /perf/tips/reorder.html**
- **VMDUMP: /perf/tips/vmdump.html**
- **Overconfiguring logical and virtual PUs**
- **VM64715 Page Release Serialization Impact**
- **Excess share distribution**
- **Low-utilization guests and memory management**
- **MDC and FlashCopy interaction**

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

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**
- **Available April 2011:**
  - 5.4: UM33346
  - 6.1: UM33347

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**

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?

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).

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

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

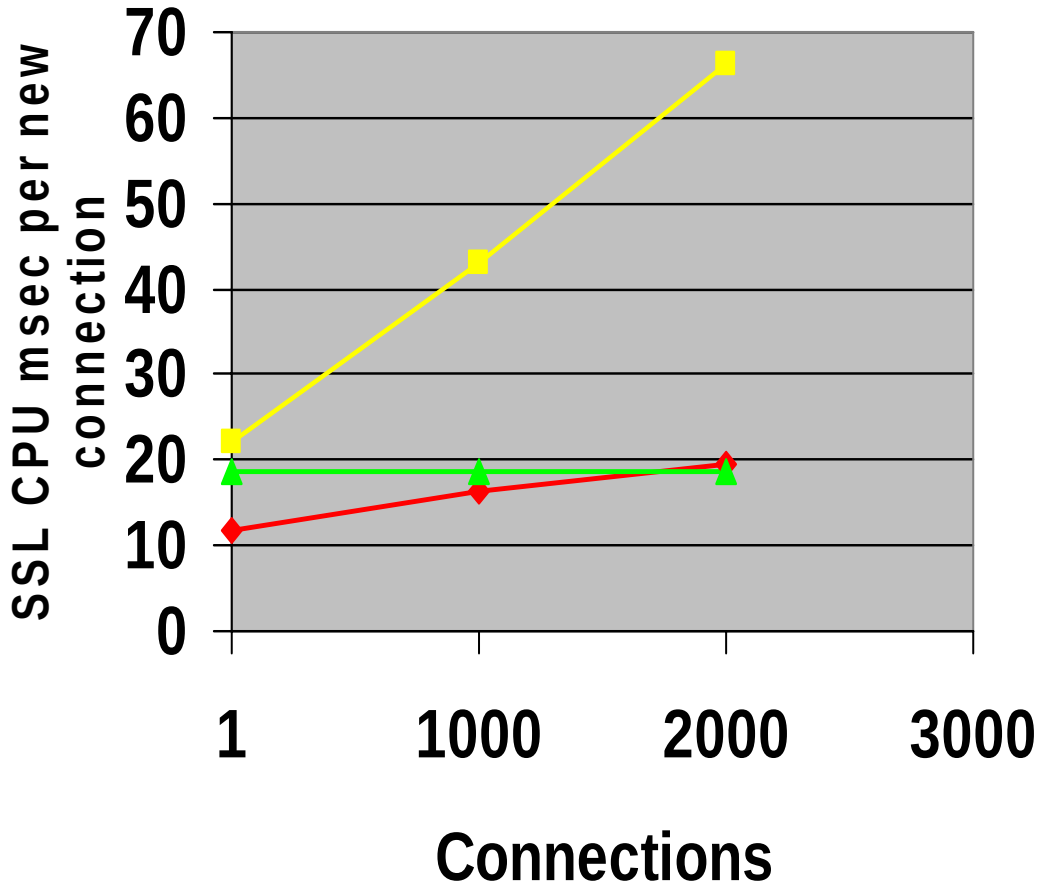
SSL Performance

- **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.

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.

IBM Systems & Technology Group

The End