IBM Systems & Technology Group

z/VM & CMM

Bill Bitner
VM Performance Evaluation
bitnerb@us.ibm.com

4/14/2008

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.

Acknowldgements

- **Thanks to following for material borrowed and insight**
  - Chris Casey
  - Bill Holder
  - Virg Meredith
  - Damian Osisek
  - Martin Schwidefsky
  - Xenia Tkatschow
  - Don Wilton

z/VM & CMM

- **High level challenge**
- **z/VM Memory Management Concepts**
- **Approaches to improve the situation**
  - Asynchronous Page Fault
  - VMRM use of Cooperative Memory Management
    - Aka: CMM 1, VMRM-CMM
  - Collaborative Memory Management Assist
    - Aka: CMM 2, CMMA, MEMASSIST
- **Uses for each**

Linux as a pageable guest: challenges

- **Linux is optimized for physical machine**
  - Uses all available memory (disk cache expands to consume)
- **“Double paging” by both hypervisor and Linux**
  - Each employs “least recently used” algorithm
  - 2 LRUs may conflict, degrade performance
    - If host under greater memory pressure, host may have paged out data which guest will shortly decide to page out itself; must then reread from host disk to write to guest disk.
    - Guest will optimize to “guest physical” memory size, rather than ideal footprint within host

z/VM Concepts – Available Lists

- **Series of list of frames available to be used**
- **Avoids having to take time to search for a usable frame when a request from other subsystems or pagefault for guest occurs**
- **Depending on release, lists for: <2GB vs. >2GB, single frames vs. contiguous frames.**
- **Dynamic Thresholds**
  - Low threshold – kick off processing to get more available frames
  - High threshold – number of frames we want to have to stop process kicked off by low threshold
- **Most severe impact is when Available List goes empty**

Performance Toolkit FCX254 AVAILLOG

```
<---- Thresholds -----> <---------- Page Frames ---------->
<---Low---> <--High---> <Available> <Obtains/s> <Returns/s>
 <2GB  >2GB  <2GB  >2GB   <2GB  >2GB  <2GB  >2GB  <2GB  >2GB
    0   356   920  1271   175k   904  16.0  9070   1.4  9070

       <-Times->  <----------- Replenishment -----------> Perct
       <-Empty->  <--Scan1--> <--Scan2--> <-Em-Scan-> Scan Emerg
       <2GB >2GB  Compl Pages Compl Pages Compl Pages Fail  Scan
          0    5   6472 5431k    30 33864  6101 3985k  443    50
```


z/VM Concepts – Demand Scan

- **Process where we try to replenish available list**
- **Series of passes through various lists of 'owned' frames.**
  - Pass 1, Pass 2, Emergency Pass
  - Passes oriented to 'steal' pages that will hurt the users and system the least
  - Example of information considered:
    - Long term dormant, short term dormant, active virtual machine
    - Referenced or not referenced
    - Shared or private
    - …

Performance Toolkit FCX259 DEMNDLOG

	<--------- Demand Scan Pass 1 ---------->								<------- Demand Scan Pass 2 -------->								<------- Emergency Scan ------->							
	<-- Ended After -->				<--- Page Frames --->				<-Ended After->				<--- Page Frames --->				<-Ended After->				<-Page Frames-->			
Interval End Time	Lng Drm	Drm ant	NSS Shr	Eli Lst	Dsp Lst	Long Dorm	Dor- mant	NSS Shr	Eli Lst	Dsp Lst	Lng Drm	Drm ant	Eli Lst	Dsp Lst	Long Dorm	Dor- mant	NSS Shr	Eli Lst	Dsp Lst	Drm ant	NSS Shr	Eli Lst	Dsp Lst	Dor- mant	NSS Shr	Eli Lst	Dsp Lst	Scan Failed
>>Mean>>	0	3	0	0	181	1886	9129	0	0	1M	0	13	0	0	0	21k	0	0	0	3	9	0	1k	3331	7k	0	2M	3346
09:58:36	0	0	0	0	17	0	543	0	0	81k	0	0	0	0	0	960	0	0	0	0	1	0	46	29	261	0	99k	155
09:59:06	0	0	0	0	3	0	22	0	0	42k	0	0	0	0	0	1162	0	0	0	0	0	0	49	161	229	0	77k	88
09:59:36	0	0	0	0	3	1	401	0	0	38k	0	0	0	0	0	709	0	0	0	1	0	0	47	39	117	0	62k	87
10:00:06	0	1	0	0	1	2	517	0	0	20k	0	0	0	0	0	538	0	0	0	0	0	0	27	84	197	0	37k	46
10:00:36	0	0	0	0	4	0	859	0	0	54k	0	2	0	0	0	417	0	0	0	0	1	0	35	17	789	0	0M	140

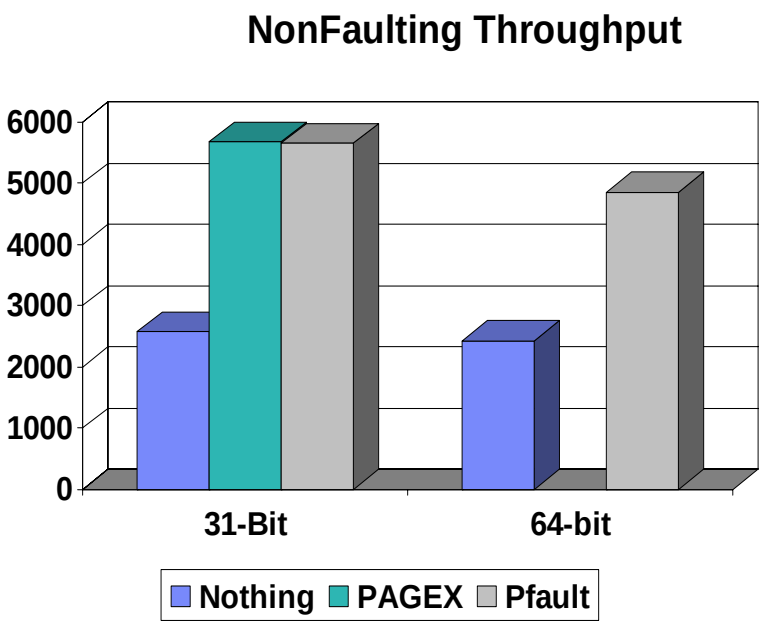
z/VM Concepts – Reorder & Page Release

- **Reorder**
  - Processing of reordering frame ‘owned’ lists
  - Reference bit processing
- **Page Release**
  - Ability for a guest to tell CP that it no longer needs a guest real page.
  - CP no longer needs to back the page
  - Two methods:
    - Diagnose x’10’ – Page Release – Used by Linux
    - Diagnose x’214’ – Pending Page Release – Used by CMS

Asynchronous Page Fault Processing

- **Ordinarily, page faults serialize the virtual machine. This can be a throughput and response time problem for guest systems**
- **PAGEX was implemented for VSE**
  - Limited to 31-bit
- **Enhancements designed for Linux**
- **PFAULT macro**
  - Accepts 64-bit inputs
  - Provides 64-bit PSW masks
- **Diagnose x'258'**

Page Fault Tests



CMM 1: Overview

- **Customer identifies guests to participate**
- **VM Resource Manager (VMRM) tracks system memory utilization / demand, computes target "resident footprint" for each guest**
- **VMRM sends messages to guest to adjust footprint**
- **Guest device driver receives messages, uses "ballooning" to comply**
  - Driver requests specified number of pages from kernel (*kmalloc()*)
  - *kmalloc()* will trigger reclaim of guest real memory
    - Capitalizes on existing guest logic to cast out the least valuable pages
  - Driver instructs VM to release page contents (Diagnose x'10')

CMM 1 Verbs

- **Sent to Linux guests through VM "SMSG"**
- **Message formats:**
  - CMM SHRINK *npages* – resize "permanent" balloon to *npages* (absolute count; may be higher or lower than previous SHRINK)
  - CMM RELEASE *npages* – inflate "temporary" balloon by an additional *npages* pages
  - CMM REUSE *npages nseconds* – sets deflation rate for temporary balloon
    - Pages redeployed gradually: total of *npages* pages every *nseconds* seconds, until temporary balloon deflated
    - Setting persists until next REUSE message (through multiple RELEASE cycles)

Enabling guests for VMRM-CMM

- **In Linux guests, indicate that VMRMSVM is eligible to send CMM1 commands.**
- **In VMRM, supply list of virtual-machine names to VMRM**
  - New VMRM CONFIG statement
    - NOTIFY MEMORY *userid1* [*userid2… useridx*]
  - Not related to other statements (WORKLOAD, GOAL, ..)

VMRM-CMM Suggested Levels

- **Suggested levels, please follow normal service research**
- **z/VM**
  - z/VM 5.2.0 + VM64085
  - z/VM 5.3.0
- **Linux**
  - SLES9 SP3
    - See http://www.vm.ibm.com/sysman/vmrm/vmrmcmm.html for additional patches.
  - RHEL5.1

CMM 2 - Objectives

- **Pass memory usage information from pageable guest to host**
  - Attributes per 4K-byte block of guest absolute storage
- **Make guest aware of page state changes by host**

Benefits

- **Host memory management efficiency**
  - More intelligent selection of page frames to be reclaimed
  - Reduced reclaim overhead: avoid page writes where possible
- **Guest memory management efficiency**
  - Avoid double-clearing of page on reuse
  - Option to favor host-resident pages on allocation requests
- **Reduce guest memory footprint**
- **Support greater memory overcommit ratios**

CMM 2 states

- **Cross-product of guest-specified and host states**
- **Guest-specified states (“block-usage states”)**
  - Stable (S)
  - Unused (U)
  - Volatile (V)
  - Potentially Volatile (P)
- **Host states (“block-content states”)**
  - Resident (r) – contained in host main memory
  - Preserved (p) – backed on host auxiliary storage
  - Logically zero (z) – not backed; will appear as zeros on reference

4 block-usage states

- **Stable: host must preserve page contents**
  - Block-content state may be resident, preserved, or logically zero
  - If logically zero, millicode assist (introduced with QDIO V=V Passthrough) can back with host frame without SIE exit
- **Unused: contents meaningless to guest**
  - Host may discard page contents (invalidating PTE to make non-resident)
  - Highest-priority page frames for host to reclaim
  - Guest not expected to reference pages in unused state

4 block-usage states …

- **Volatile: contents meaningful but guest can reconstruct**
  - Host may discard page and key contents if desired
  - (Volatile, resident) state means contents are intact
  - (Volatile, logically zero) means contents were discarded
    - Guest reference results in *block-volatility exception*
      - Guest response: reconstruct contents (*e.g.,* reread from disk)
    - Guest must change usage state back to stable before reusing page
    - Thereafter, next reference will be backed with a frame of zeros
      - Handled by millicode (HPMA)
- **Potentially Volatile**
  - Treated as volatile if guest change bit off, stable if on
  - Benefits of “volatile,” yet safe for files write-mapped into user space

Instruction: Extract and Set Storage Attributes

ESSA r1,r2,m3

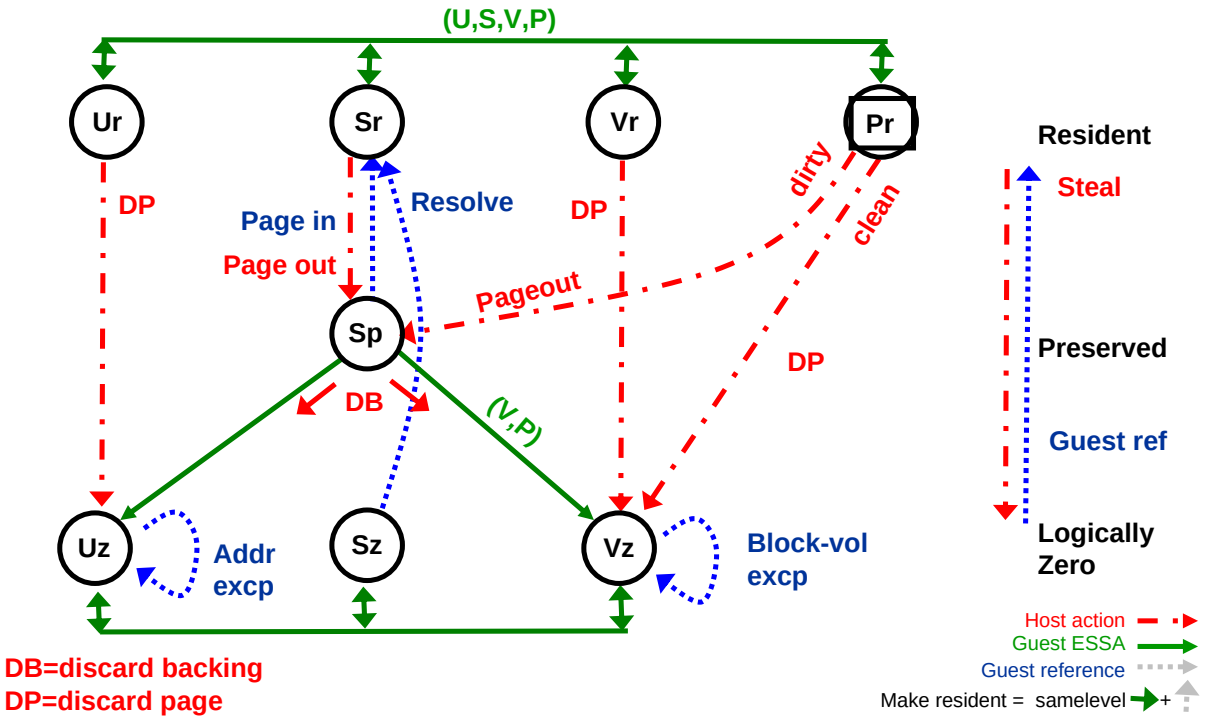
- **r1 (output): receives old block states**
  - 2- bit block-usage state (Stable, Unused, Potentially Volatile, Volatile)
  - 2-bit block-content state (resident, preserved, logically zero)
- **r2 (input): contains guest absolute address of target block**
- **m3 (immediate operand): specifies operation to be performed**

ESSA operations

All operations extract old state to r1

- **Extract (fetch states only)**
- **Set Stable**
- **Set Unused**
- **Set Volatile**
- **Set Potentially Volatile**
- **Set Stable and Make Resident**
  - If logically zero, invokes HPMA Resolve to bind to frame of zeros
- **Set Stable If Resident**
  - No state change if non-resident
  - Returned block-content state indicates whether change occurred
  - Useful atomic operation in preparation for Linux page write
  - Allows guest to favor reuse of still-resident pages, avoid growing footprint

Finite state machine for CMM 2



Enabling for CMMA

- **z/VM – enabled by default**
  - To disable, issue before guests boot

    CP SET MEMASSIST OFF
- **Linux – disabled by default**
  - To enable, add following in Linux parm file

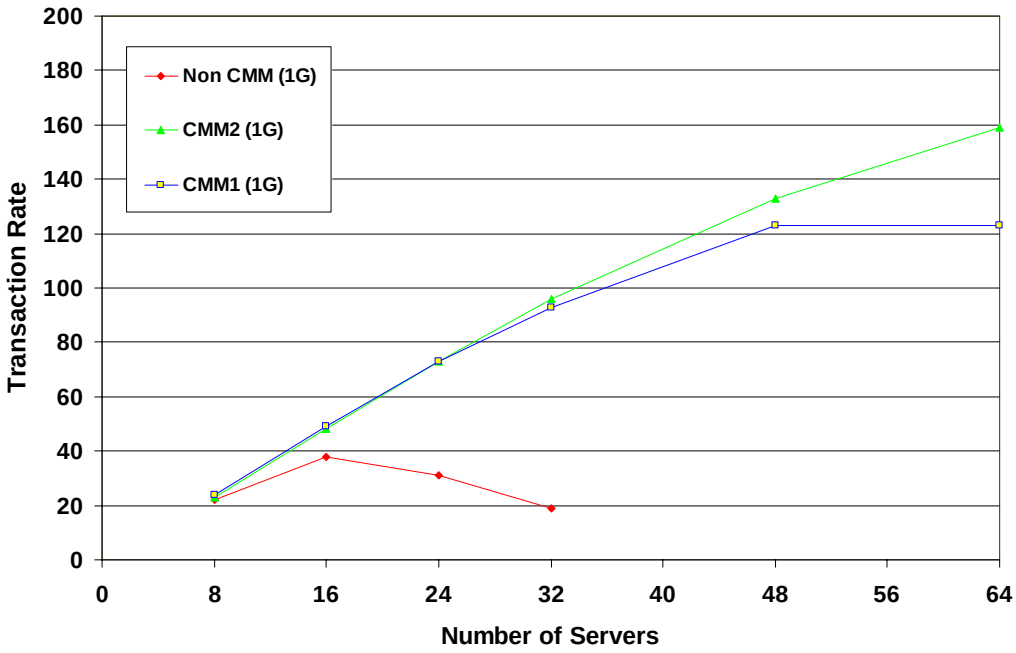
    cmma=on

CMMA Suggested Levels

- **Suggested levels, please follow normal service research**
- **z/VM**
  - z/VM 5.3.0 + APAR VM64265 + APAR VM64297
- **Linux**
  - SLES10 SP1 update kernel 2.6.16.53-0.18
- **Hardware**
  - z9 or z10 for ESSA instruction

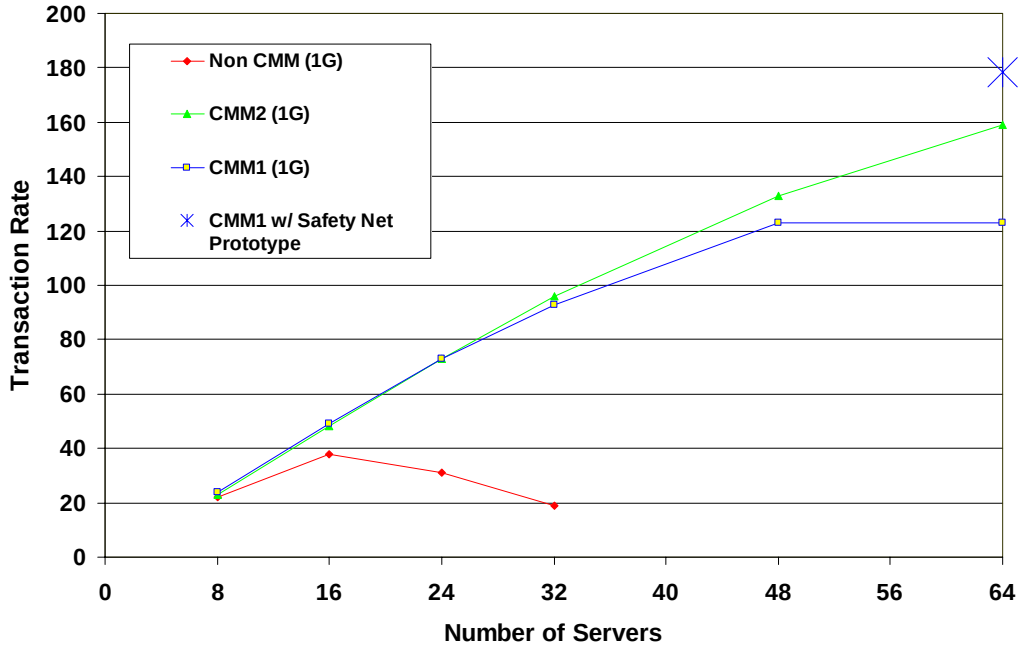
Transaction Rate vs. Number of Servers

for various Storage Management Products using Apache servers with a virtual storage size as shown in parenthesis in the legend; z9 6GB / 2GB



Transaction Rate vs. Number of Servers

for various Storage Management Products using Apache servers with a virtual storage size as shown in parenthesis in the legend



Other Uses and Approaches

- **Use VMRM-CMM for test virtual machines and leave production virtual machines off the Notify statement.**
- **Use of CP SET RESERVE for mission critical guests**
- **CMMA on older machines (pre z9)**
  - Simulation of ESSA instruction
  - Our workloads showed 26% increase in CP Processor time per transaction

More Information

- **z/VM Performance Report**
  - http://www.vm.ibm.com/perf/reports/zvm/html/530cmm.html
- **VMRM & CMM**
  - http://www.vm.ibm.com/sysman/vmrm/vmrmcmm.html
- **z/VM Large Memory – Linux on System z**
  - http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101151
- **Linux information on Out of Memory notifier patch**
  - http://www.ibm.com/developerworks/linux/linux390/linux-2.6.16-s390-09-october2005.html

Summary

- **System z continues to show leadership in virtualization**
  - Asynchronous Page Fault
  - VMRM-CMM
  - CMMA
- **Questions?**

