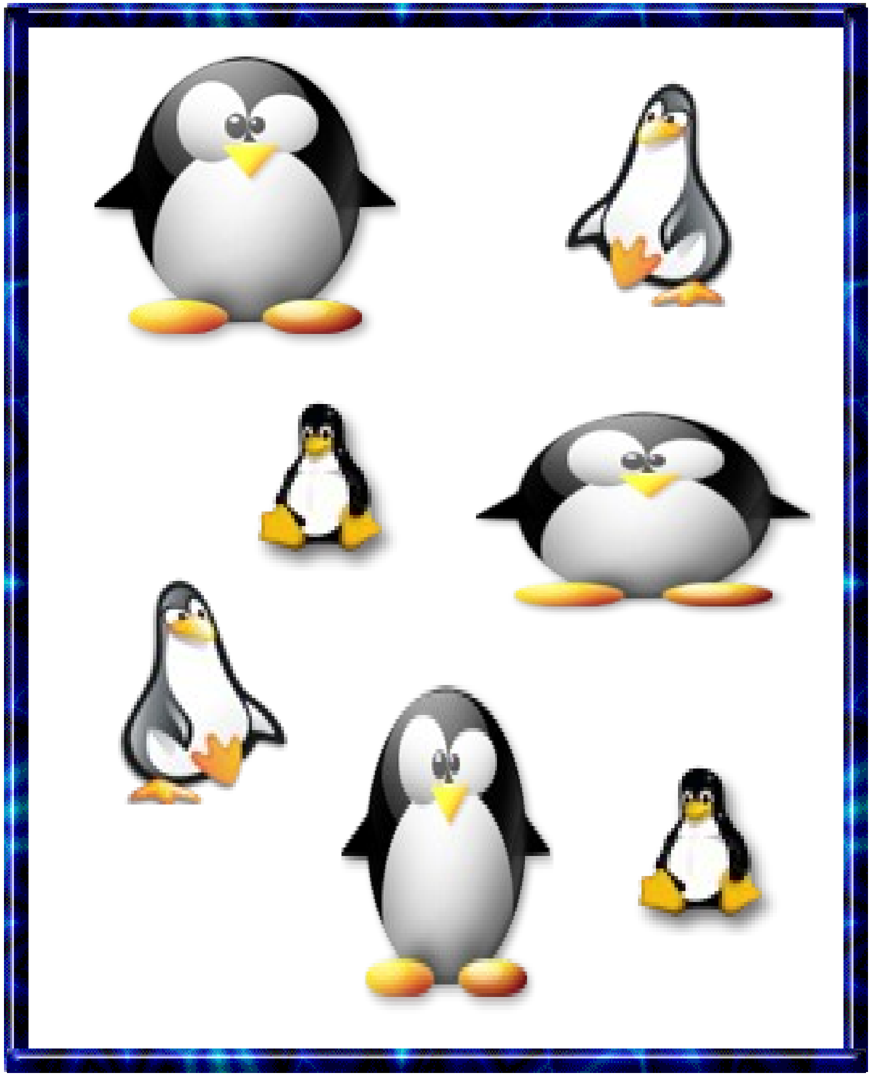
Speicheroptimierung mit z/VM und Linux auf System z durch Collaborative Memory Management 2

Martin Schwidefsky (schwidefsky@de.ibm.com)
Linux on System z Development
IBM Lab Boeblingen, Germany

Session V13, GSE Frühjahr 2008, Bonn

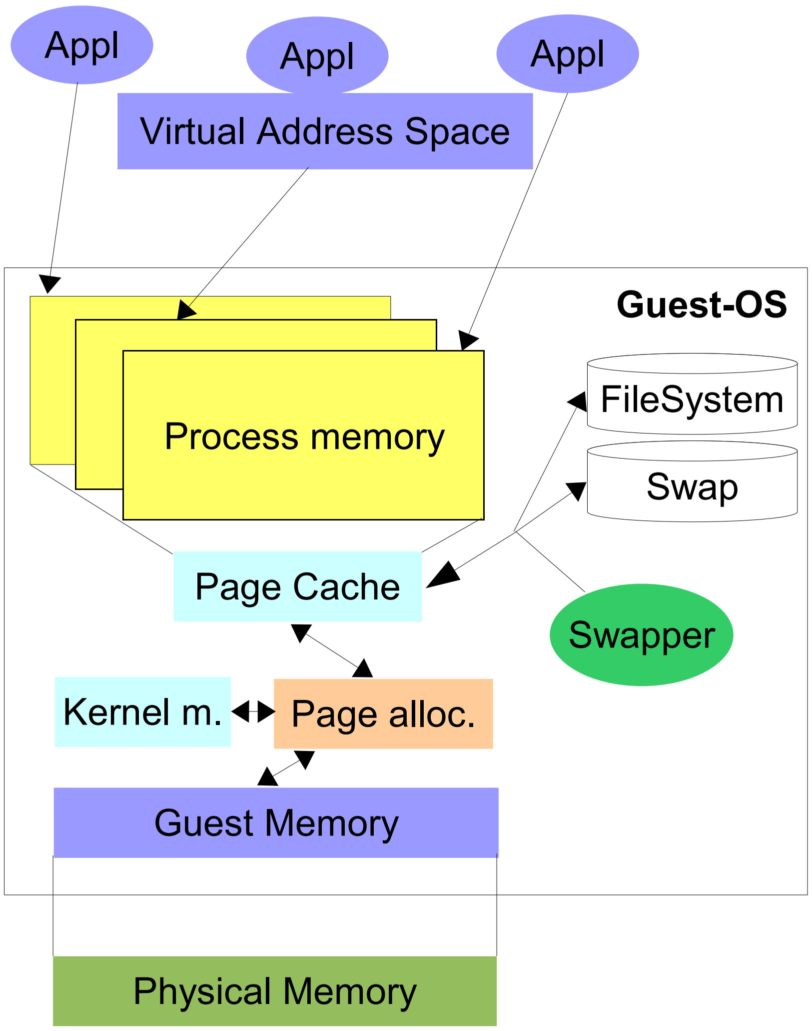


Agenda

- Guest vs. host memory
- Linux as a guest
- Page hints
- State machine
- Guest/host communication
- Overhead
- Performance results

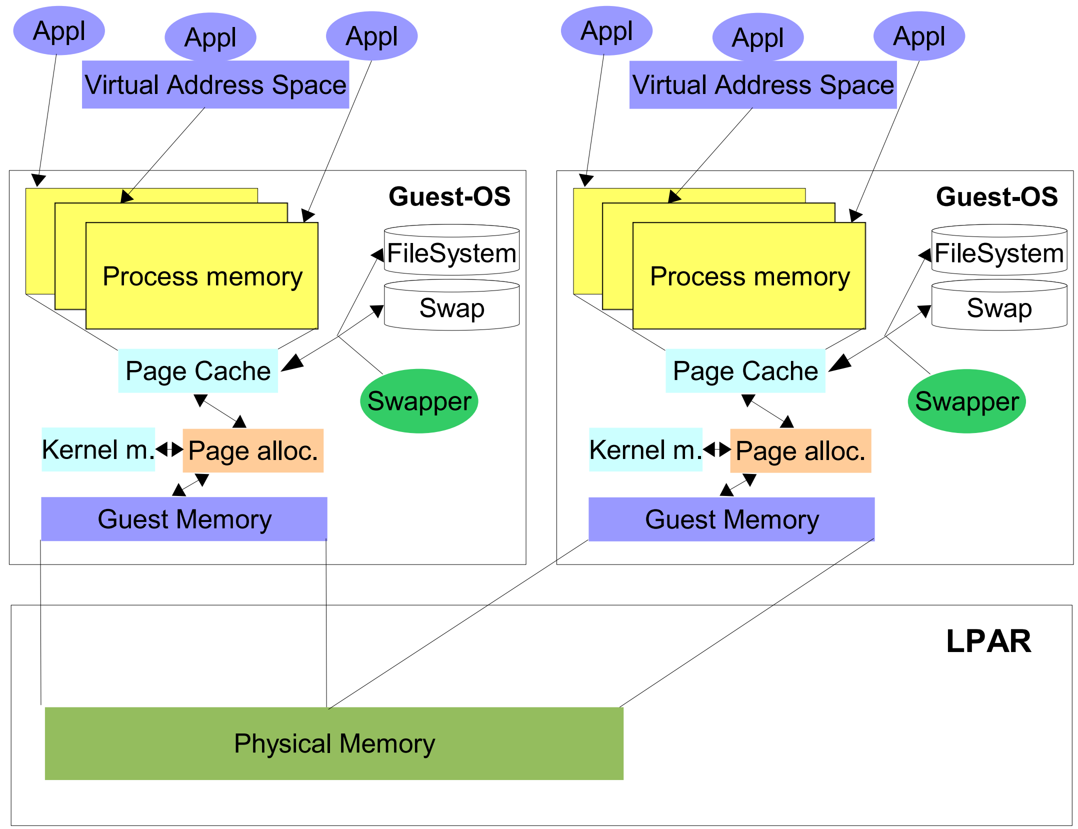
OS Virtual Memory Management (VMM)

- Multiple virtual address spaces
- Good VMM only holds the most commonly access data in real memory, rest is stored/ retrieved from secondary storage
- VMM must create the illusion of a single level storage
  - Hardware provides protection and VA-PA address translation and exceptions
  - Management granularity is based on pages



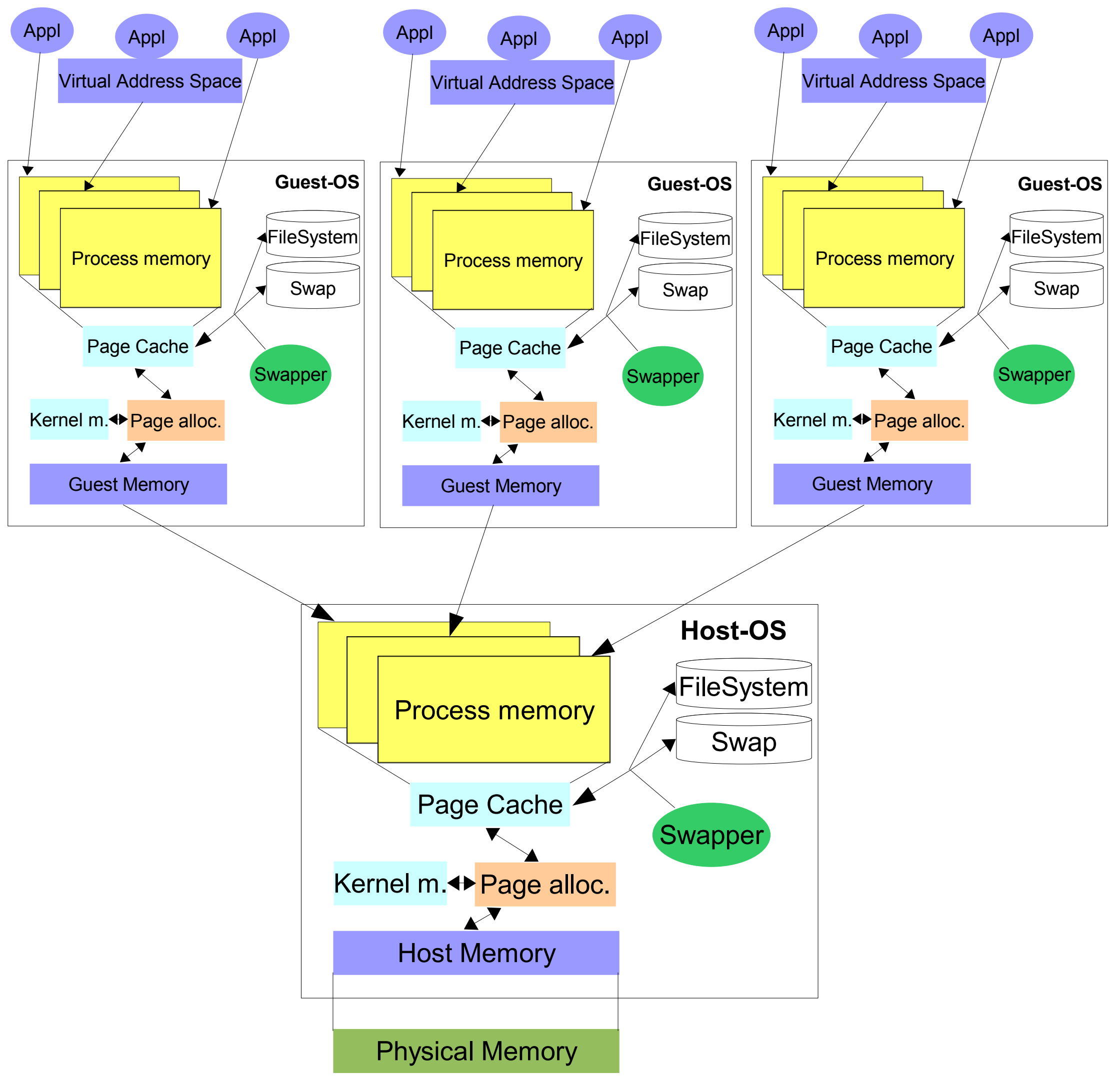
Memory Management in LPAR

- Each LPAR image gets a fixed amount of physical memory
- No memory overcommitment
- Easy .. and boring



Memory management in a virtualized environment

- Guest-OS looks like a process to the host
- Guest real memory = host virtual memory
- To make things interesting
  - Host does memory overcommitment
  - Host does on demand paging just like any OS



Linux as pageable guest: challenges

- Linux is optimized to run on a physical machine
  - Uses all available memory, “free” memory is used for caching
- “Double paging” by both the hypervisor and Linux
  - Each uses the “least recently used” page reclaim algorithm
  - The 2 LRUs will conflict, degrading the performance
- Goal: exchange information between guest and host to optimize the memory management on both level
  - Memory ballooning (CMM1): “quantitative” approach
    - Host instructs each guest to adjust its memory footprint
  - Guest page hinting (CMM2): “qualitative” approach
    - Guest identifies usage characteristics of guest pages
    - Guest obtains host status information, notifications

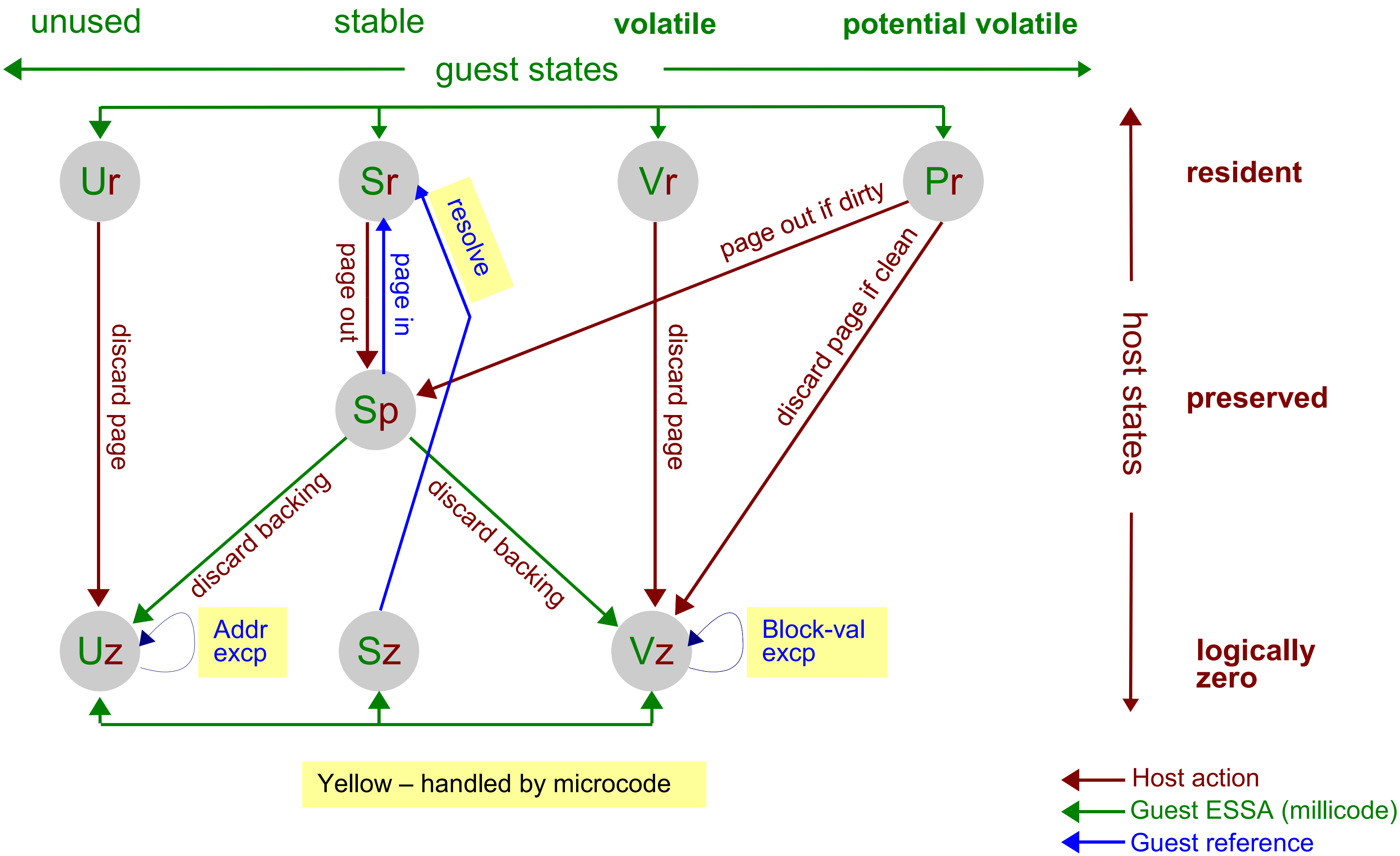
CMM2: guest page hinting

- Basic principle:
  - Pass page usage information from pageable guest to host
  - Allow the host to “steal” pages based on the usage information
  - Deliver “discard faults” if the guest accesses pages removed by the host
- Benefits
  - Host memory management efficiency
    - More intelligent selection of page frames to be reclaimed (unused pages)
    - Reduced reclaim overhead: avoid page writes where possible
  - Guest memory management efficiency
    - Option to avoid double-clearing of pages on reuse
    - Option to favor host-resident pages on allocation requests
  - Reduce guest memory footprint without guest invocation

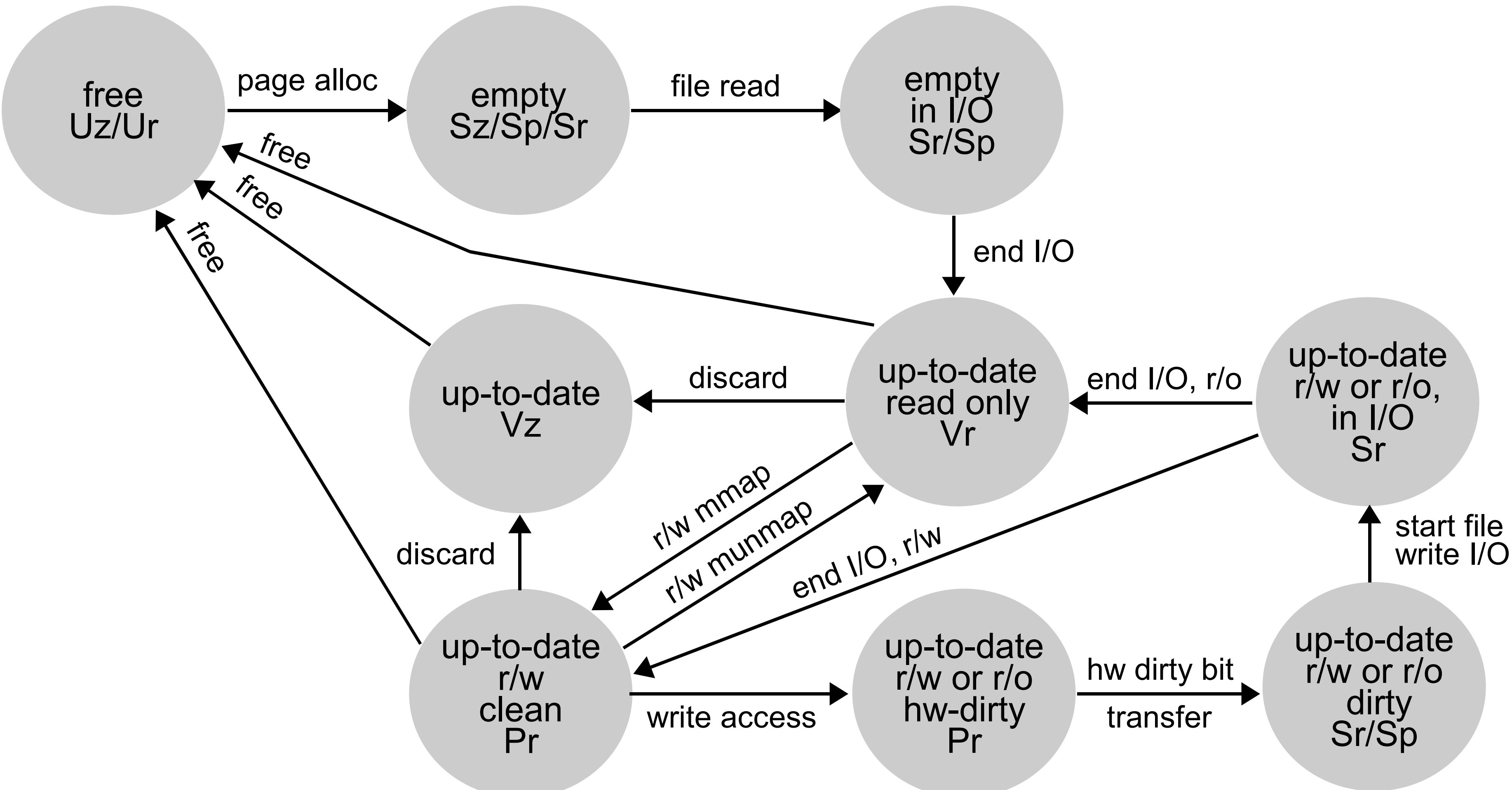
CMM2: guest page hinting - cont

- 4 guest page states
  - Stable (S): page has essential content the guest can't recreate
  - Unused (U): no useful content and any access to the page will cause an addressing exception
  - Volatile (V): page has useful content. The host can discard the page anytime. The guest gets a discard fault on access for discarded pages
  - Potentially Volatile (P): same as (V) but host needs to check the dirty bit
- 3 host page states
  - Resident (r): page is present in host memory
  - Preserved (p): page is not present in host memory but the content is preserved somewhere by the host
  - Zero (z): page is not present in host memory, the content is zero

CMM2: finite state machine



CMM2: life of a file page

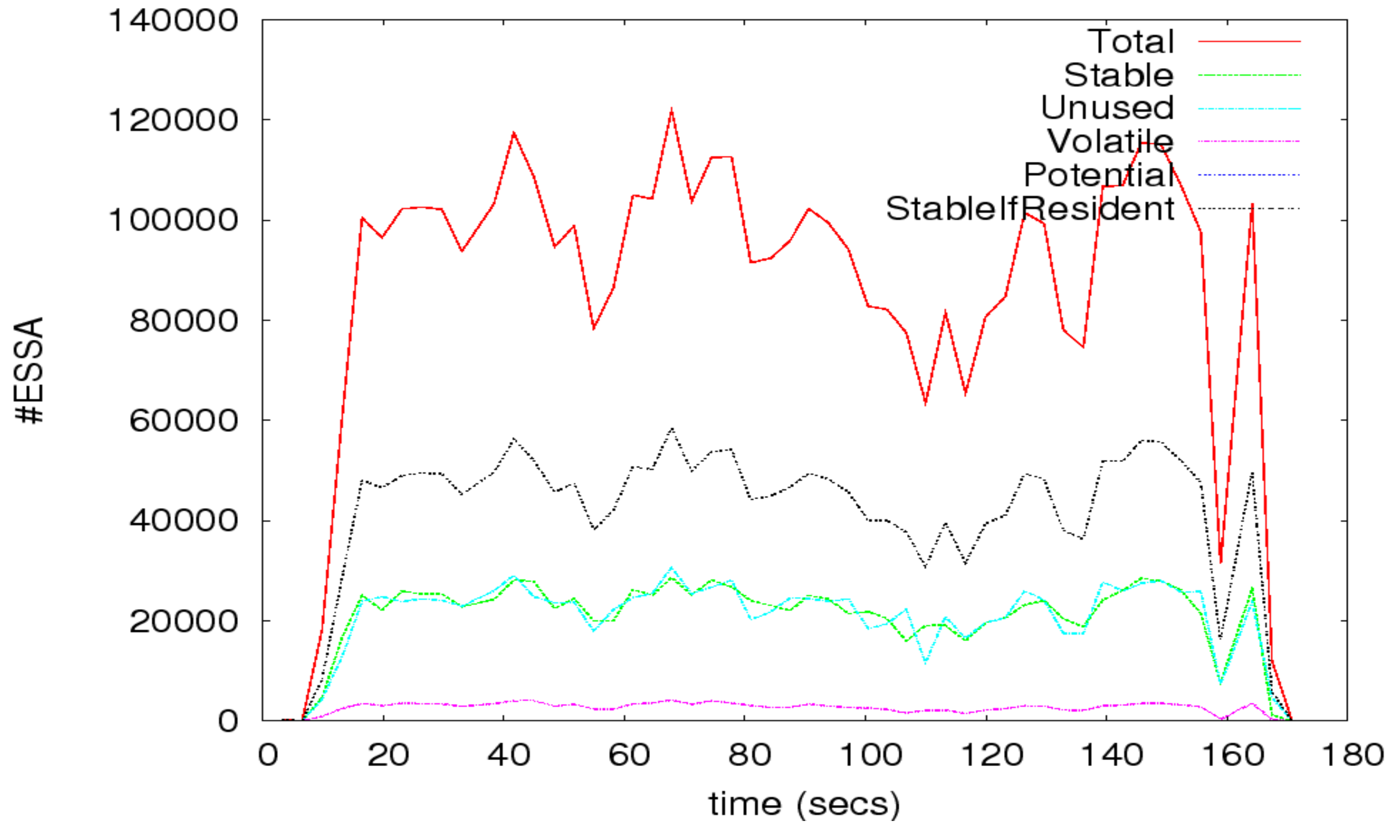


CMM2: guest to host communication

- Guest state changes need to be fast
  - A lot of state changes occur when Linux is working
  - The state change may not cause a SIE break
- “Extract and Set Storage Attribute” instruction
  - ESSA r1,r2,m3
  - r1 (output): receives old pages state
    - 2-bit guest state (Stable, Unused, Volatile, Potentially Volatile)
    - 2-bit host state (resident, preserved, logically zero)
  - r2 (input): contains guest absolute address of target page
  - m3 (immediate operand): specifies operation to be performed “get state”, “set stable”, “set unused”, “set volatile”, “set pvolatile”, “set stable make resident”, “set stable if resident”
  - ESSA is millicoded

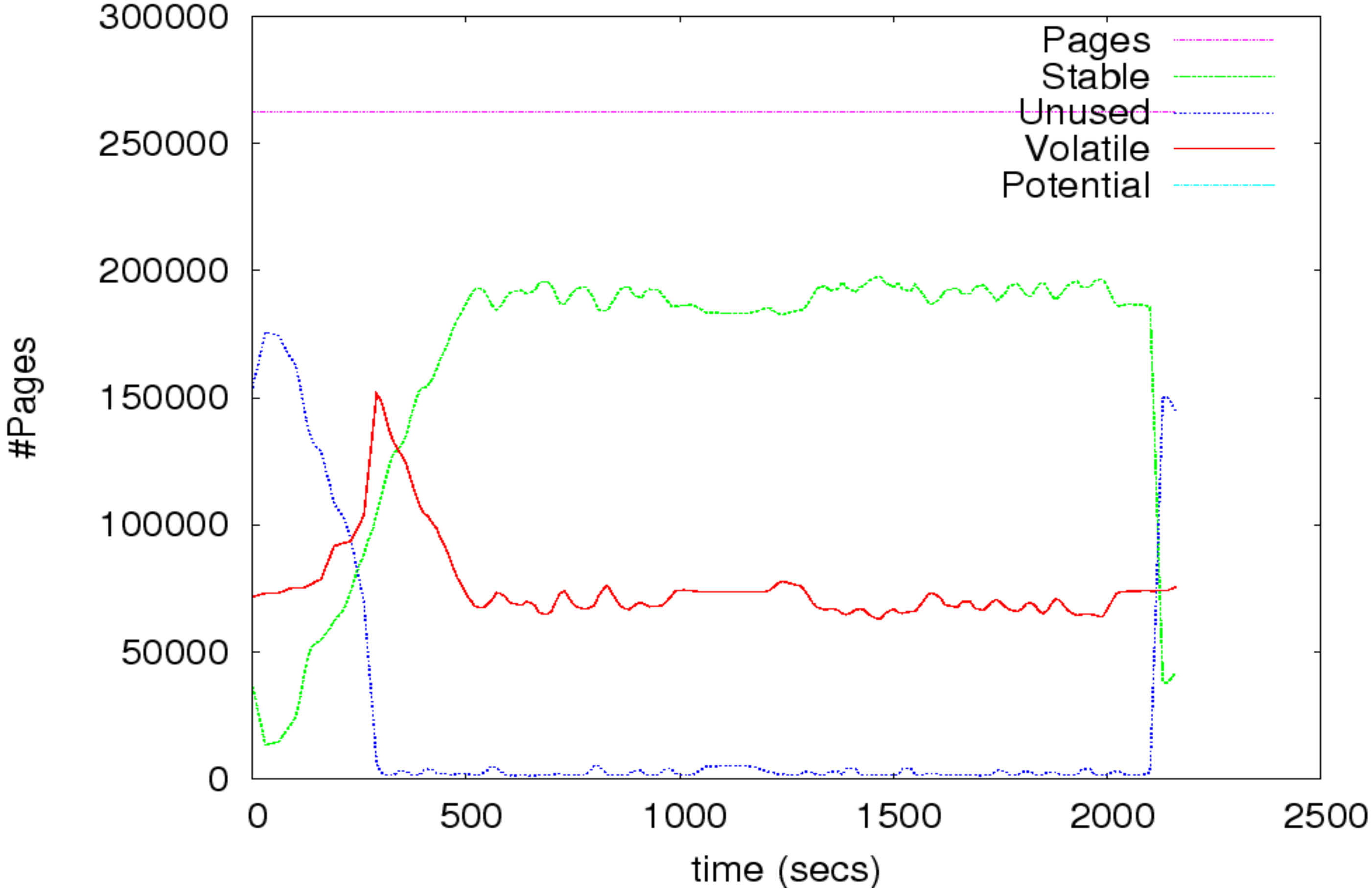
CMM2: state transition overhead

- Benchmark: kernel compile on 4 way without host paging
- ~80-120K ESSA / sec on z9, ~0.25% overhead

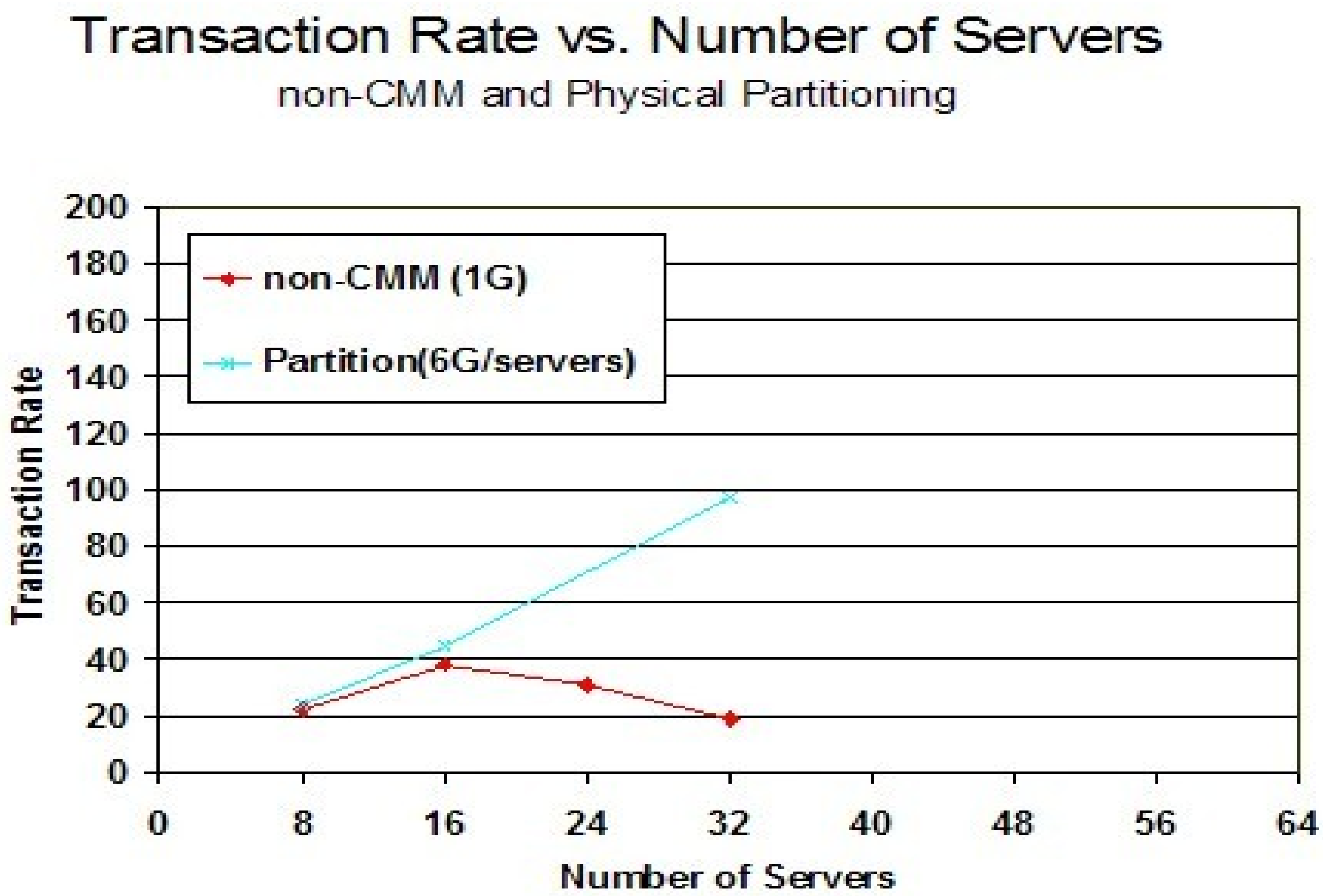


CMM2: State distribution example

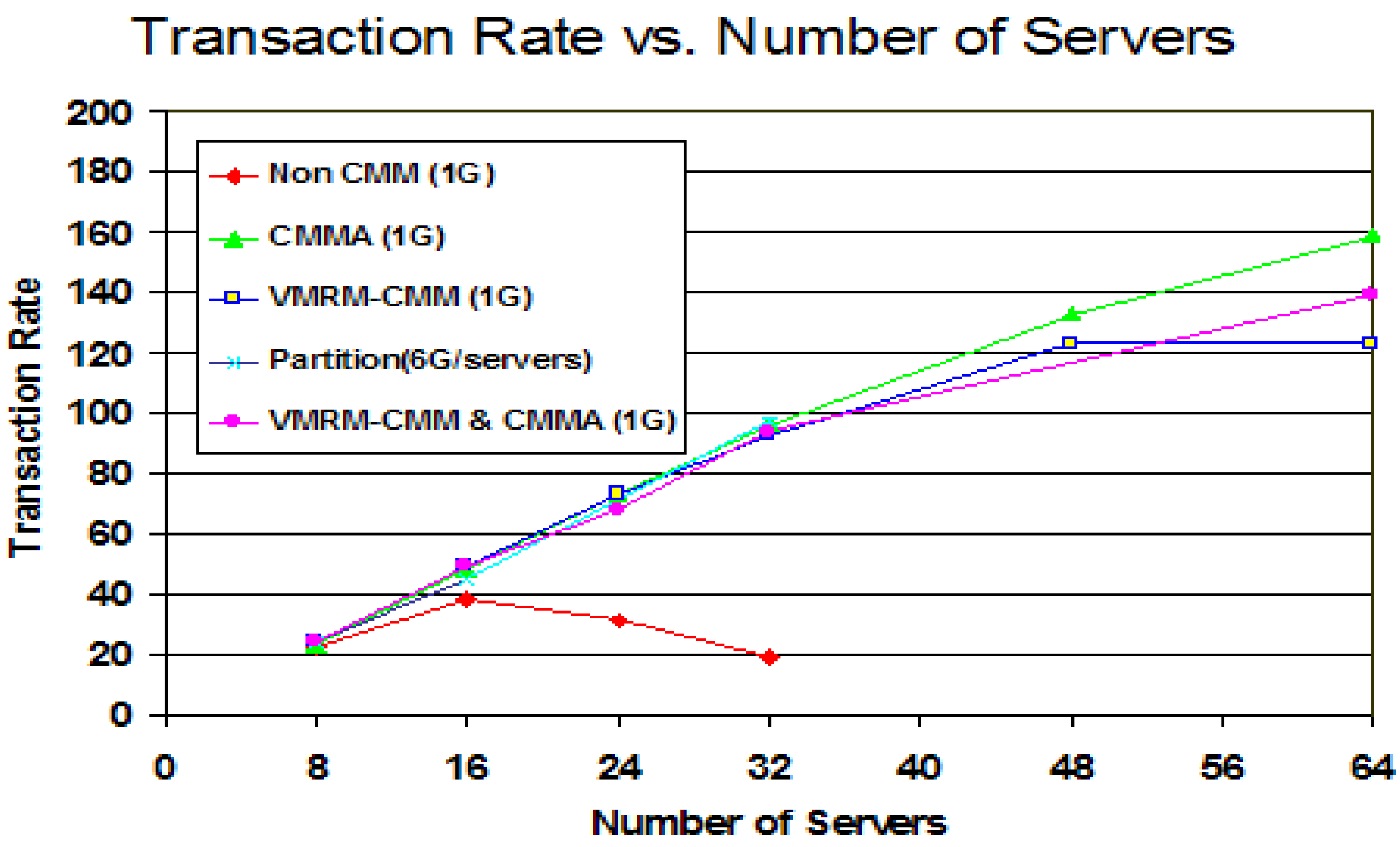
- 1-way 1GB guest running SpecWeb 2005



CMM2: performance results (z9, 6GB real memory)



CMM2: performance results (z9, 6GB real memory)



CMM2: required levels

- Architecture support
  - ESSA millicode instruction has been introduced with System z9
- z/VM support
  - z/VM 5.3 plus APAR VM64265 and VM64297
- Linux support
  - SLES10 SP1 update kernel 2.6.16.53-0.18 or later

- Links:
  http://www.vm.ibm.com/perf/reports/zvm/html/530cmm.html
  http://www.ibm.com/developerworks/linux/linux390/linux-2.6.16-s390-12-october2005.html

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 Linux 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.