VM/ESA 2.3.0
Performance Update

Last Updated September 1, 1998

Bill Bitner
IBM Endicott
1701 North St.
Endicott, NY 13760
607-752-6022
bitner@vnet.ibm.com
USIB1E29 at IBMMAIL

Disclaimer

The information contained in this document has not been submitted to any formal IBM test and is distributed on an "as is" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environment do so at their own risk.

In this document, any references made to an IBM licensed program are not intended to state or imply that only IBM's licensed program may be used; any functionally equivalent program may be used instead.

Any performance data contained in this document was determined in a controlled environment and, therefore, the results which may be obtained in other operating environments may vary significantly.

Users of this document should verify the applicable data for their specific environments.

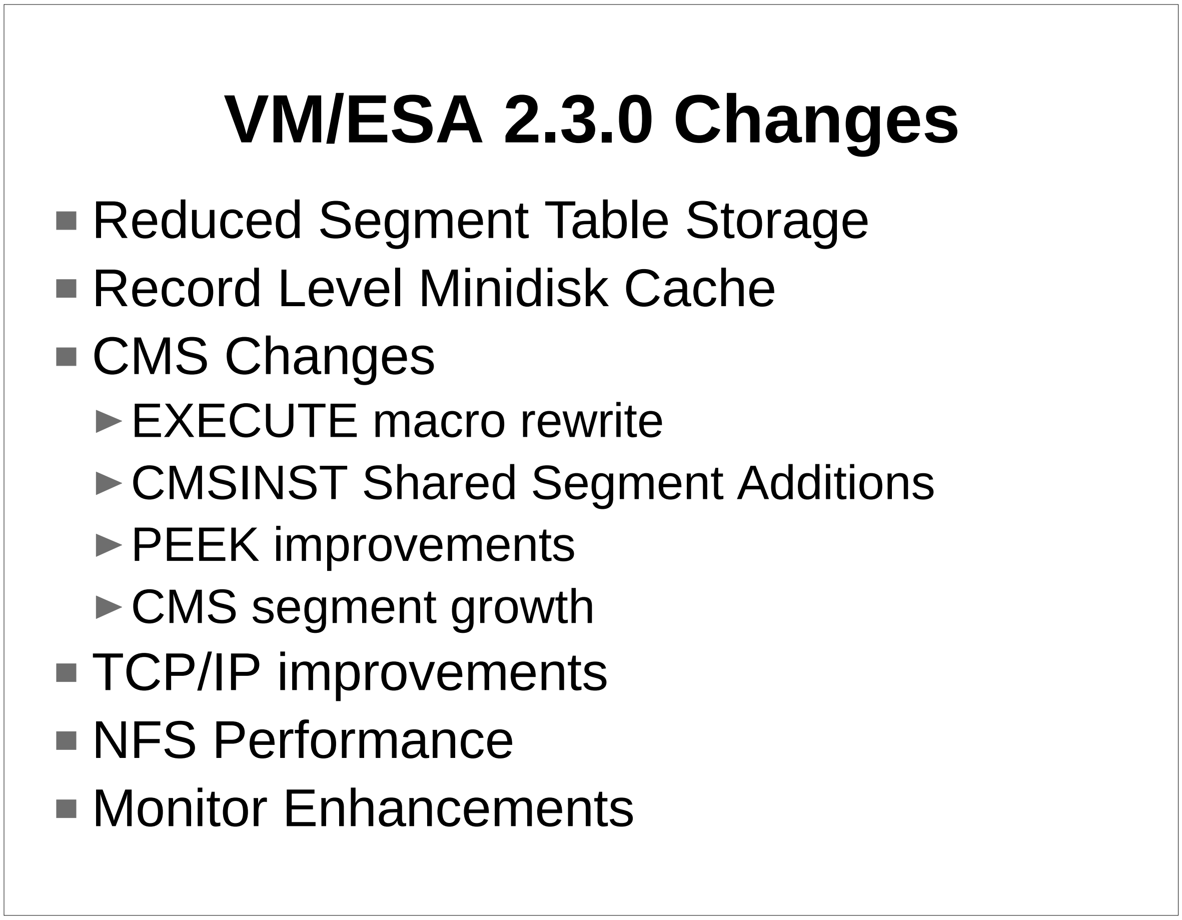
It is possible that this material may contain references to, or information about, IBM products (machines and programs), programming, or services that are not announced in your country or not yet announced by IBM. Such references or information should not be construed to mean that IBM intends to announce such IBM products, programming, or services.

Should the speaker start getting too silly, IBM will deny any knowledge of his association with the corporation.

Trademarks

The following are trademarks of the IBM Corporation:

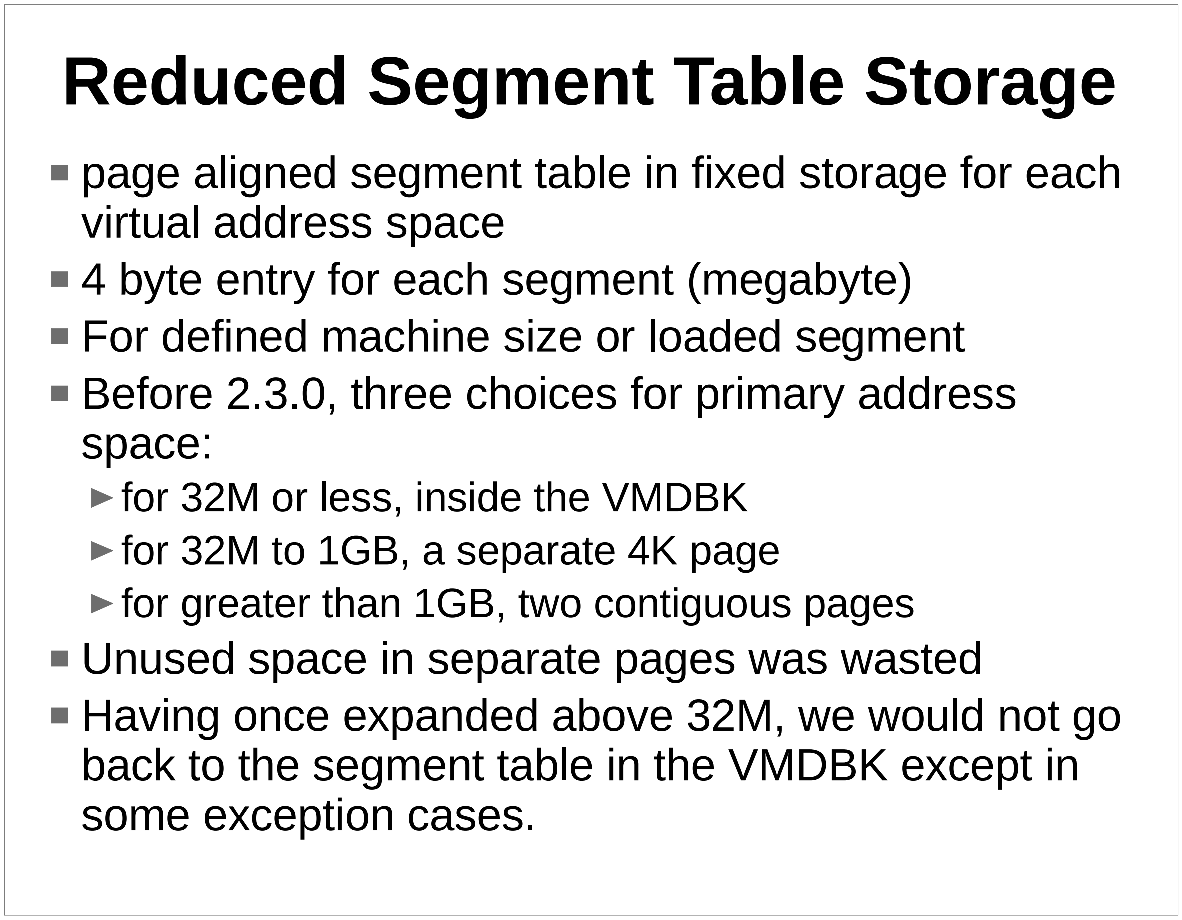
- IBM
- OfficeVision
- VM/ESA



VM/ESA 2.3.0 Changes

- Reduced Segment Table Storage
- Record Level Minidisk Cache
- CMS Changes
  - EXECUTE macro rewrite
  - CMSINST Shared Segment Additions
  - PEEK improvements
  - CMS segment growth
- TCP/IP improvements
- NFS Performance
- Monitor Enhancements

The list above represents most of the performance related changes this release. In addition, there were a number of performance related changes in TCP/IP function level 310 which can be ordered as a feature with VM/ESA 2.3.0.

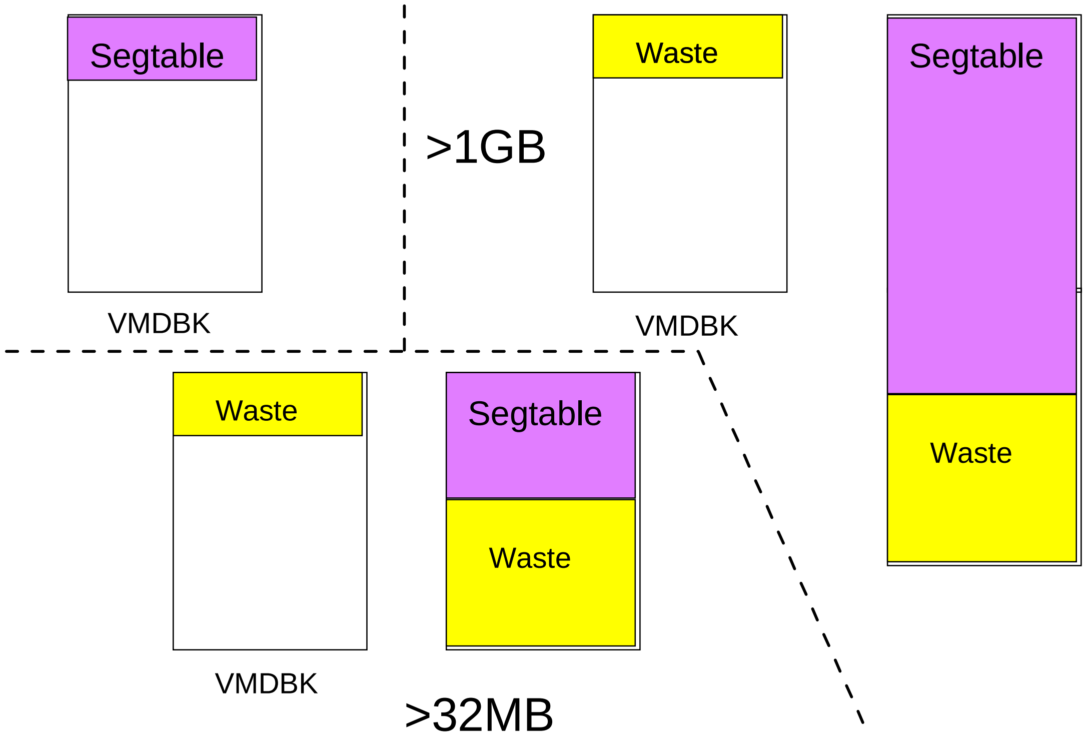


Reduced Segment Table Storage

- page aligned segment table in fixed storage for each virtual address space
- 4 byte entry for each segment (megabyte)
- For defined machine size or loaded segment
- Before 2.3.0, three choices for primary address space:
  - for 32M or less, inside the VMDBK
  - for 32M to 1GB, a separate 4K page
  - for greater than 1GB, two contiguous pages
- Unused space in separate pages was wasted
- Having once expanded above 32M, we would not go back to the segment table in the VMDBK except in some exception cases.

You might recall performance recommendations to avoid defining virtual machines greater than 32 megabytes or to avoid loading segments above the 32 meg line, and if you really had to go above the 32 meg line, try to stay below the 1GB line. This had to do with how the segment table was managed. In previous releases, the segment table was contained in the same frame of real storage as the VMDBK control block, but only if storage was not addressed above the 32M line for that virtual machine. If above 32M, CP would allocate a separate 4K frame for the segment table, and 2 frames if addressing storage above the 1GB line.

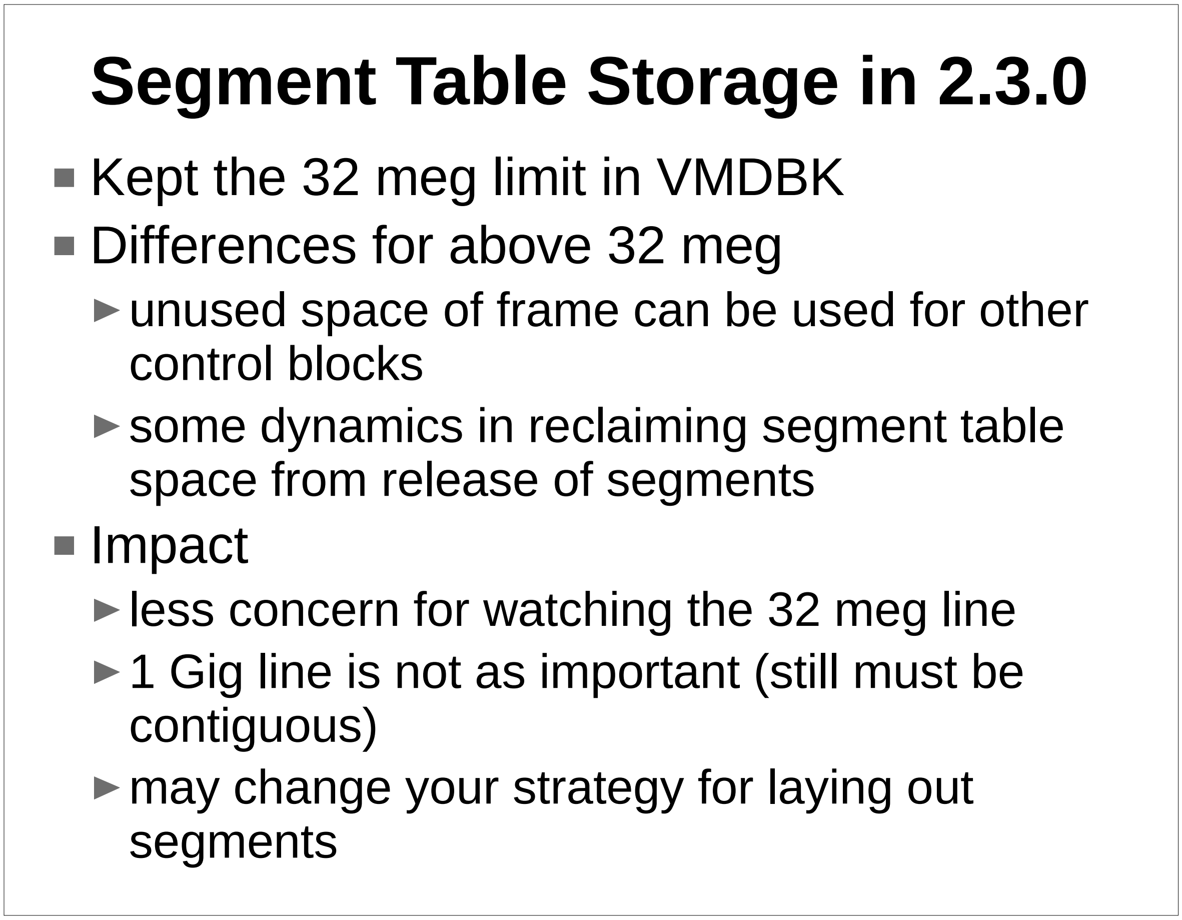
Segment Table prior to 2.3.0



The pictures shown here illustrate the three scenarios of the segment table in releases prior to VM/ESA 2.3.0.. For virtual machines of 32 megabytes or less and with no segments above the 32 megabyte line, the segment table resides in the VMDBK control block.

The bottom picture shows the case of a virtual machine with addressable storage in the range of 32 MB to 1 GB where a new page is obtained from storage and what ever is needed is carved from this page with the rest of the page being wasted.

The right picture shows the case where the virtual machine addresses greater than 1 GB of storage. In this case, two contiguous pages are required for the segment table with the remainder of the second page being wasted storage.

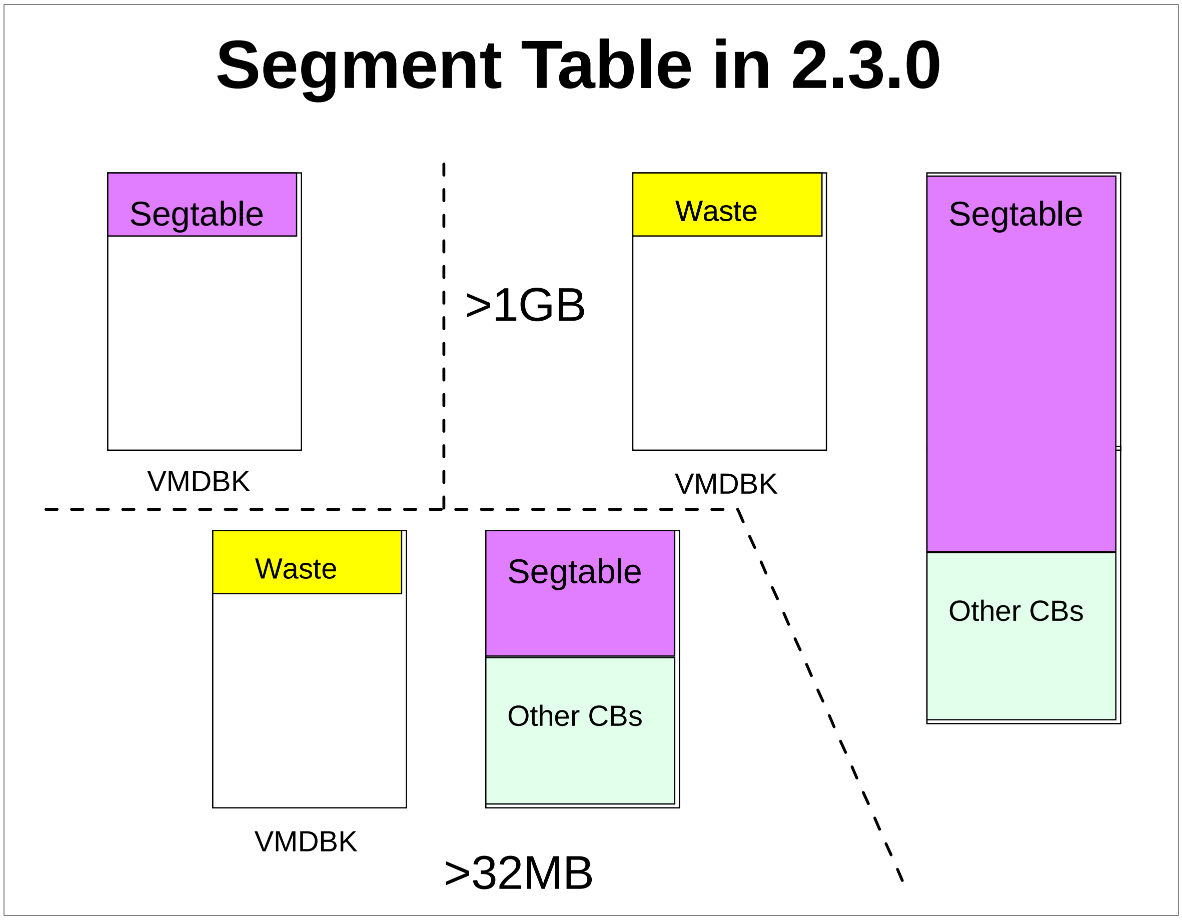


Segment Table Storage in 2.3.0

- Kept the 32 meg limit in VMDBK
- Differences for above 32 meg
  - unused space of frame can be used for other control blocks
  - some dynamics in reclaiming segment table space from release of segments
- Impact
  - less concern for watching the 32 meg line
  - 1 Gig line is not as important (still must be contiguous)
  - may change your strategy for laying out segments

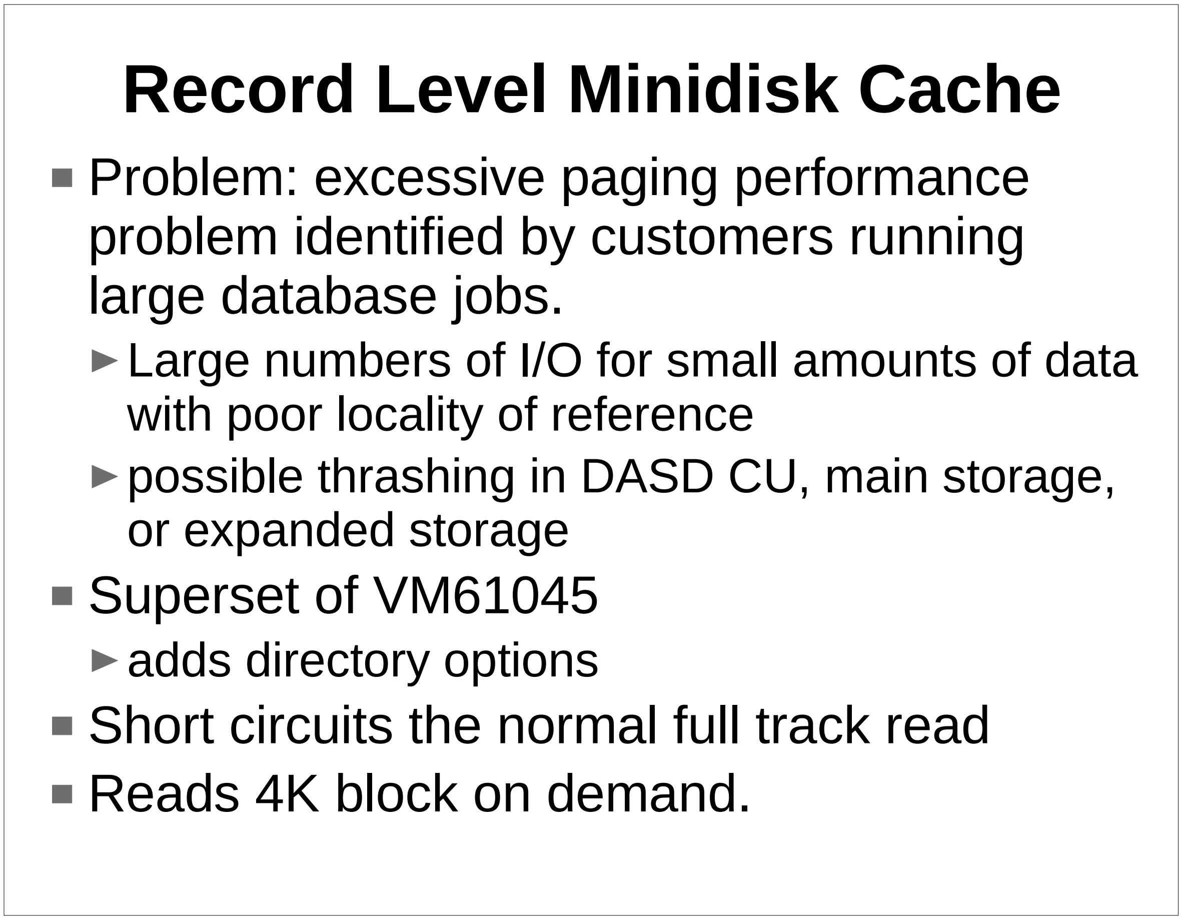
In VM/ESA 2.3.0, CP can still fit a segment table in the VMDBK if no more than 32MB is addressable. However, if it needs to build a segment table outside the VMDBK, it will do so out of CP free storage. The segment table needs to be at start of a page frame. However, CP will use the remaining part of the page to hold other CP control blocks.

This results in less of a concern for defining larger virtual machines and for putting segments above the 32 meg line. However, those that used a strategy of loaded segments starting at 1 gig and working downward will not see a benefit from this. If you use the ESAFREE CP storage analysis tool from the VM download page, you'll want to get the most current version before going to VM/ESA 2.3.0.



Segment Table in 2.3.0

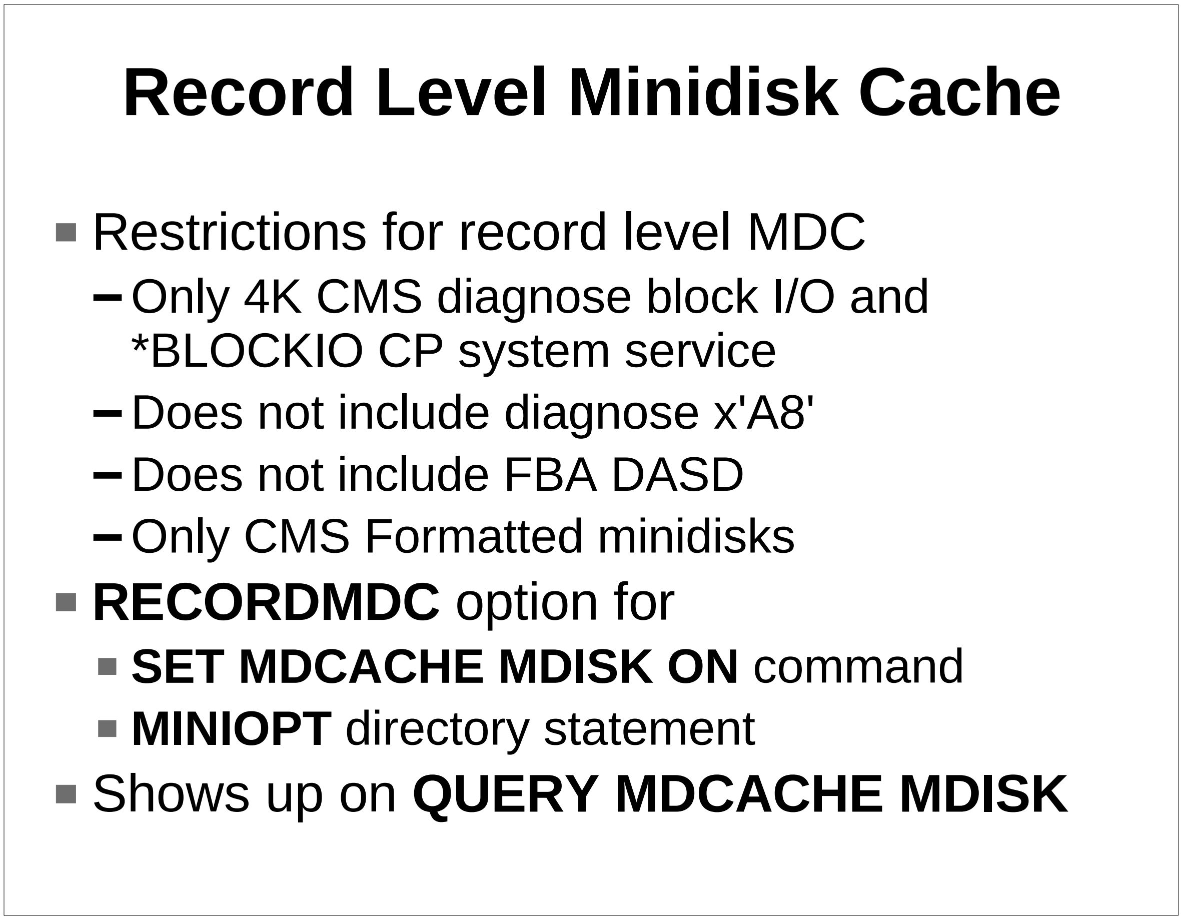
The pictures shown here illustrate the three scenarios of the segment table in VM/ESA 2.3.0.. For virtual machines of 32 megabytes or less and with no segments above the 32 megabyte line, the segment table resides in the VMDBK control block.
The bottom picture shows the case of a virtual machine with addressable storage in the range of 32 MB to 1 GB where a new page is obtained from storage and what ever is needed is carved from this page. However, unlike earlier releases, the remainder of the page could be used for other CP free storage control blocks and therefore is not wasted space.
The right picture shows the case where the virtual machine addresses greater than 1 GB of storage. In this case, two contiguous pages are required for the segment table. Again, any remaining space in the second page is used for other control blocks.



Record Level Minidisk Cache

- Problem: excessive paging performance problem identified by customers running large database jobs.
  - Large numbers of I/O for small amounts of data with poor locality of reference
  - possible thrashing in DASD CU, main storage, or expanded storage
- Superset of VM61045
  - adds directory options
- Short circuits the normal full track read
- Reads 4K block on demand.

Back in VM/ESA 1.2.2, we introduced an enhanced minidisk cache. This MDC would read in an entire track of data into cache on an initial MDC read. For CMS type data, once it was read into the cache, it was treated on a page basis. In most cases, this was good. However, in the case of a database implemented in the CMS file system where there was a great deal of I/O activity for small amounts of data with poor locality of reference, this was very bad. Support for a record level flavor of minidisk cache went out in APAR VM61045. This support and associated enhancements were put into VM/ESA 2.3.0. Basically, the support short-circuits the normal full track read and then reads in a 4K block on demand.



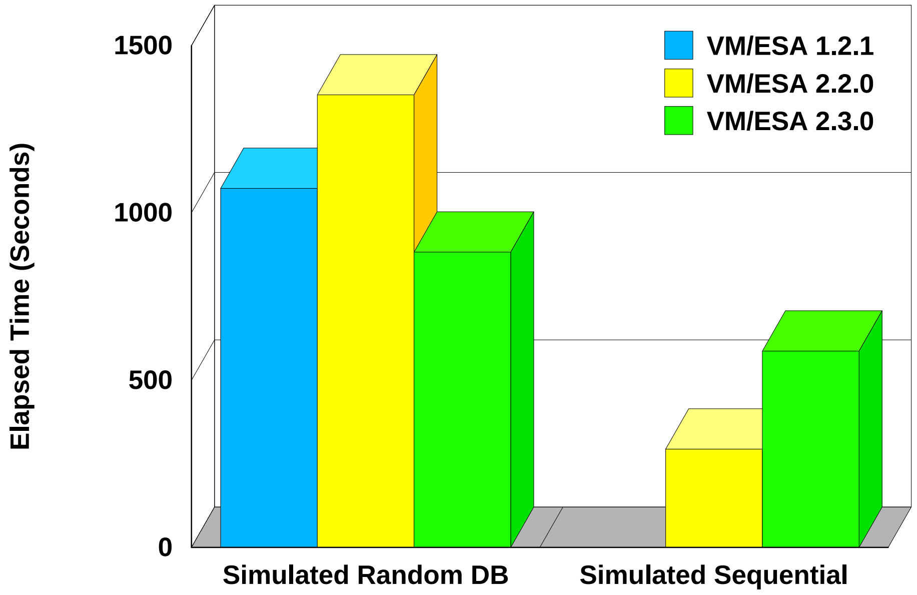
Record Level Minidisk Cache

- Restrictions for record level MDC
  - Only 4K CMS diagnose block I/O and *BLOCKIO CP system service
  - Does not include diagnose x'A8'
  - Does not include FBA DASD
  - Only CMS Formatted minidisks
- **RECORDMDC** option for
  - **SET MDCACHE MDISK ON** command
  - **MINIOPT** directory statement
- Shows up on **QUERY MDCACHE MDISK**

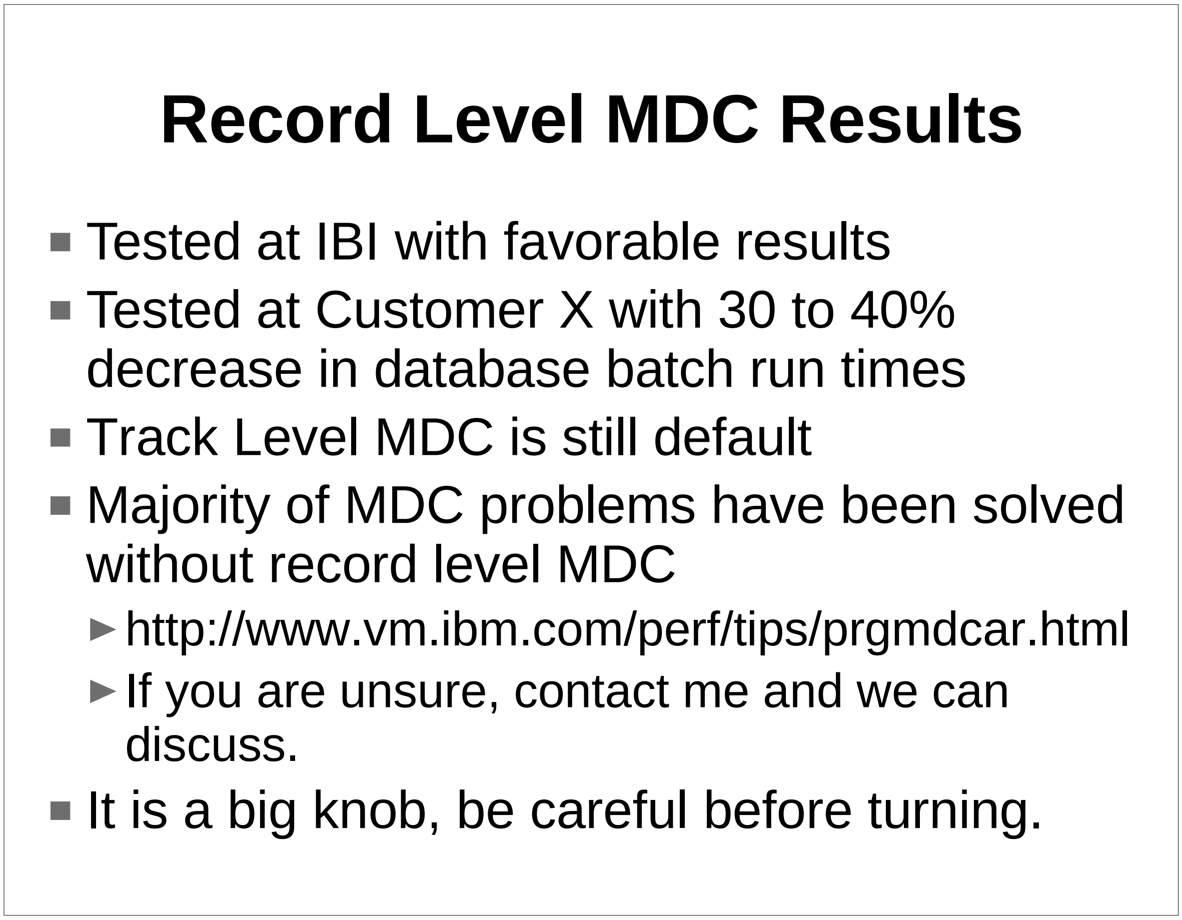
Record level minidisk cache is restricted to a subset of CMS minidisks. FBA minidisks are not supported for record level MDC. Also diagnose x'A8' is not supported.

Record level minidisk cache can be enabled by an option on the SET MDC command or on the MINIOPT directory statement. The directory statement support was not in the APAR version.

Record Level MDC Benchmark



The graphs here show results of early tests of the record level minidisk cache. A workload patterned off a customer database environment was created where the I/O pattern was simulated, but the database processing was not done. This is shown as the Simulated Random DB numbers on the left. In these runs, you see the new record level MDC even out performs the pre-full track cache in 1.2.1. Note however, that more sequential workloads or those with a greater locality of reference will do worse with the record level MDC. Track cache remains the default.

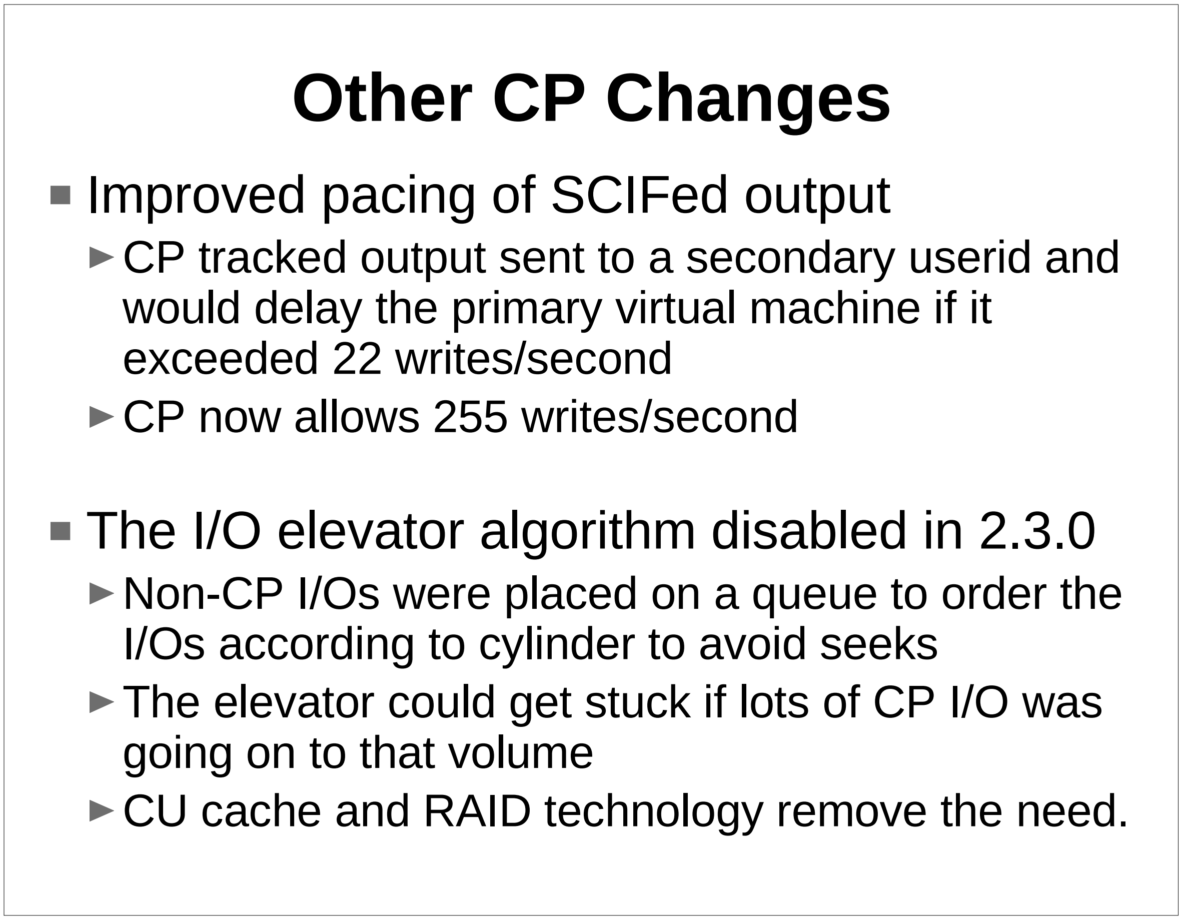


Record Level MDC Results

- Tested at IBI with favorable results
- Tested at Customer X with 30 to 40% decrease in database batch run times
- Track Level MDC is still default
- Majority of MDC problems have been solved without record level MDC
  - http://www.vm.ibm.com/perf/tips/prgmdcar.html
  - If you are unsure, contact me and we can discuss.
- It is a big knob, be careful before turning.

The folks at IBI did some testing for us with favorable results. Also, one of the customers seeing the biggest hit from full track cache, saw an improvement of 30 to 40% when record level cache was used.

Despite the potential for improvement with the record level MDC. If you are seeing poor performance with MDC, you should make sure the system is properly tuned before trying record level MDC.

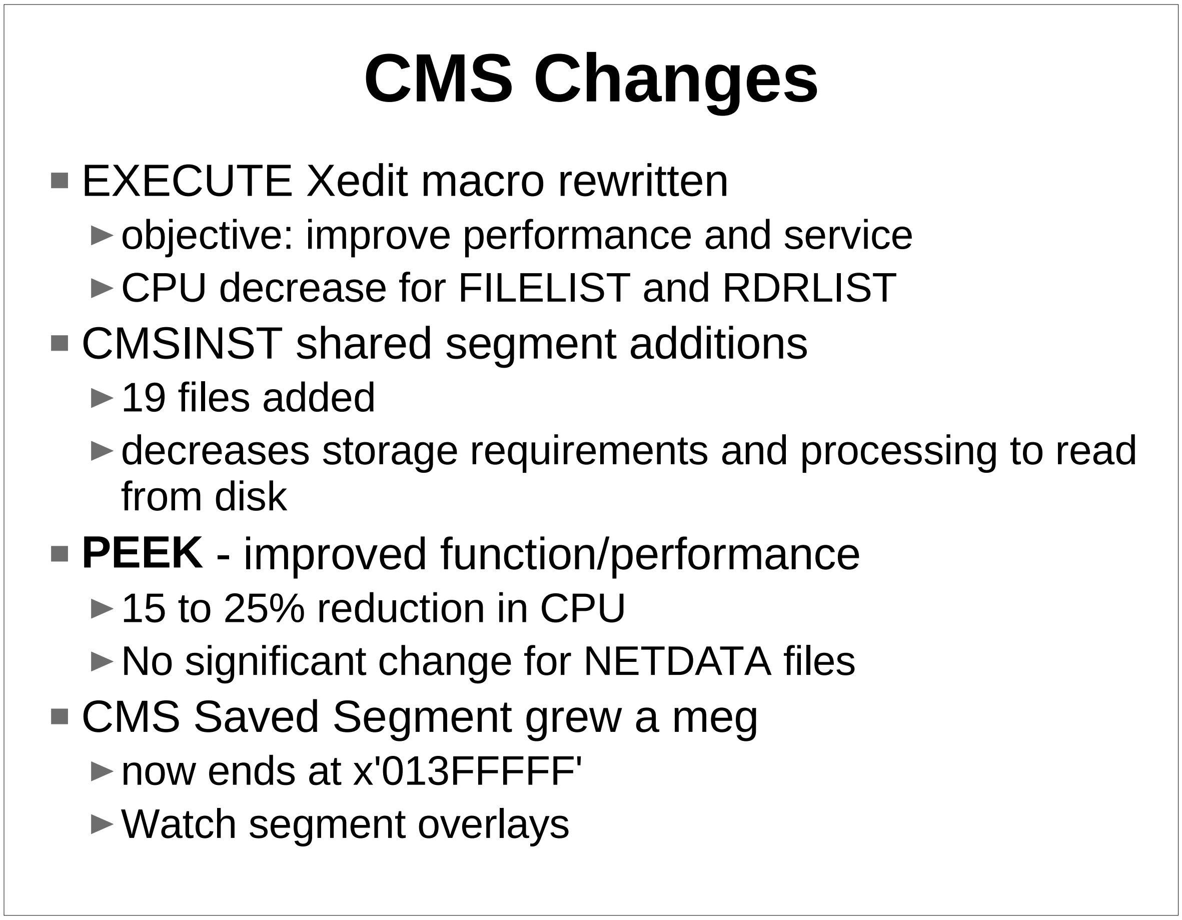


Other CP Changes

- Improved pacing of SCIFed output
  - CP tracked output sent to a secondary userid and would delay the primary virtual machine if it exceeded 22 writes/second
  - CP now allows 255 writes/second

- The I/O elevator algorithm disabled in 2.3.0
  - Non-CP I/Os were placed on a queue to order the I/Os according to cylinder to avoid seeks
  - The elevator could get stuck if lots of CP I/O was going on to that volume
  - CU cache and RAID technology remove the need.

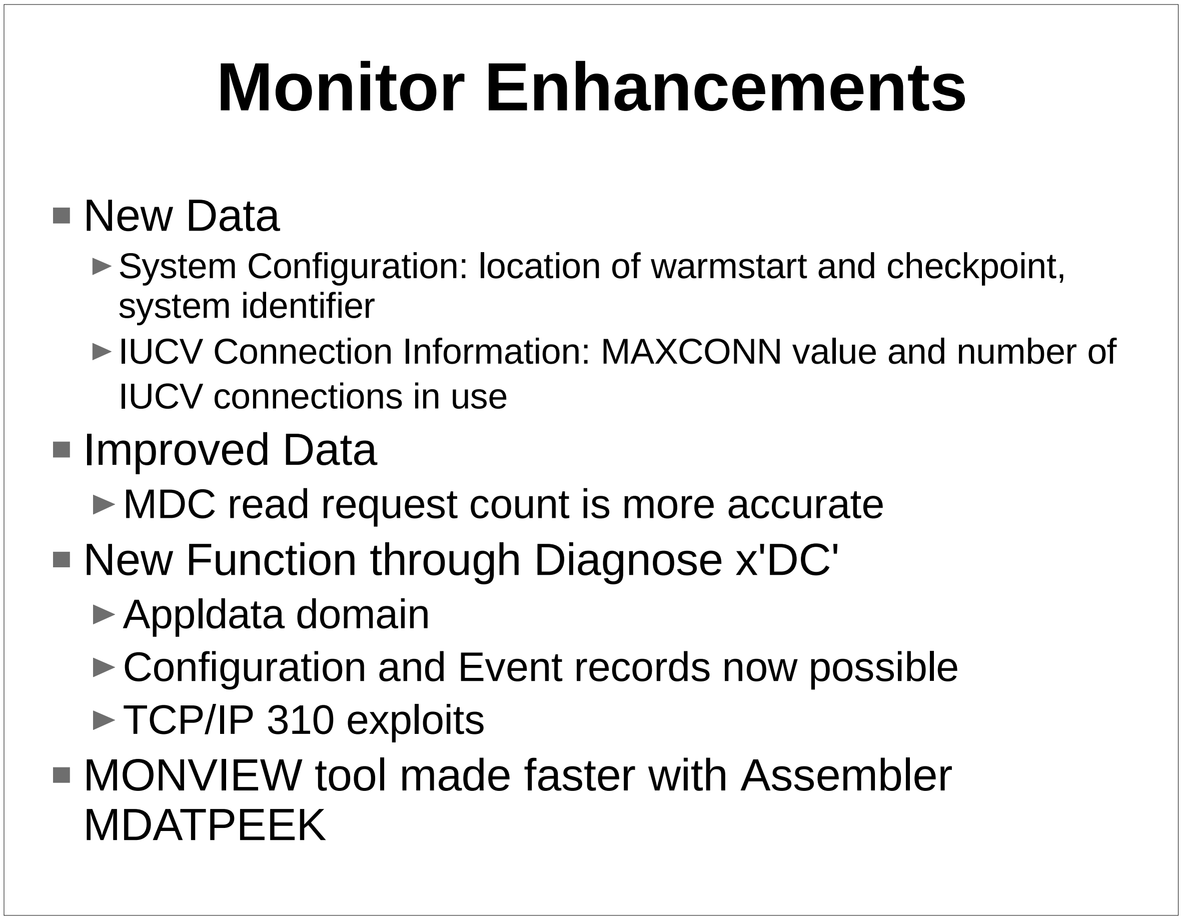
If a server machine is generating a lot of terminal I/Os that go to a secondary console user, HCPRVC tries to avoid having the secondary user overwhelmed. In the past, it track writes per second and if there was more than 22, then it would delay the primary user for an entire second. This is not good if the primary is TCPIP. In 2.3.0, HCPRVC will allow up to 255 writes per second before starting to delay a user.
The elevator algorithm was introduced before CU caching and RAID were used. This algorithm allows I/Os to be sorted in order to minimize DASD seeks. Since CP I/O is not held to the algorithm a user I/O can be left waiting if CP can flood a device. To avoid the hang scenarios, the elevator has been disabled. With current use of cache and RAID, there is little value lost.



CMS Changes

- EXECUTE Xedit macro rewritten
  - objective: improve performance and service
  - CPU decrease for FILELIST and RDRLIST
- CMSINST shared segment additions
  - 19 files added
  - decreases storage requirements and processing to read from disk
- **PEEK** - improved function/performance
  - 15 to 25% reduction in CPU
  - No significant change for NETDATA files
- CMS Saved Segment grew a meg
  - now ends at x'013FFFFF'
  - Watch segment overlays

CMS also saw some performance enhancements. The EXECUTE Xedit macro was rewritten with objectives that included better performance. Our measurements showed an improvement in FILELIST and RDRLIST as a result. Nineteen additional files were added to the CMSINST shared segments. There was also improvements to lower the processor usage for the PEEK command when used for files with the exception of NETDATA type files. You should also watch that you do not have any undesirable segment overlays now that the CMS segment is a meg larger.

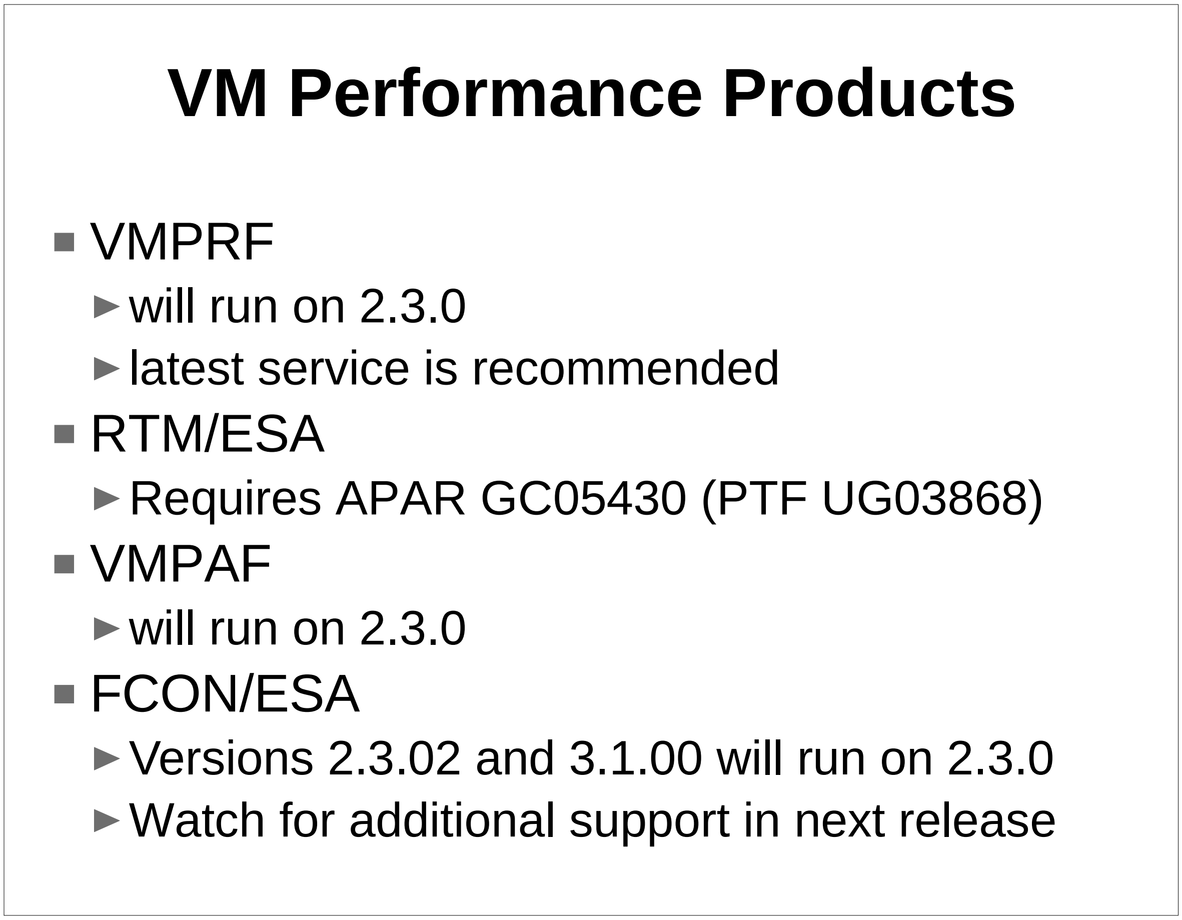


Monitor Enhancements

- New Data
  - System Configuration: location of warmstart and checkpoint, system identifier
  - IUCV Connection Information: MAXCONN value and number of IUCV connections in use
- Improved Data
  - MDC read request count is more accurate
- New Function through Diagnose x'DC'
  - Appldata domain
  - Configuration and Event records now possible
  - TCP/IP 310 exploits
- MONVIEW tool made faster with Assembler MDATPEEK

Monitor was enhanced in for both new data and function. New data includes some configuration fields: the location of checkpoint and warmstart areas, along with the system identifier. The current and maximum allowed number of IUCV/APPC connections was added to the MRUSEACT record. The value reported as the MDC read request count is now more accurate. In the past, there was potential for it to be skewed in either direction.

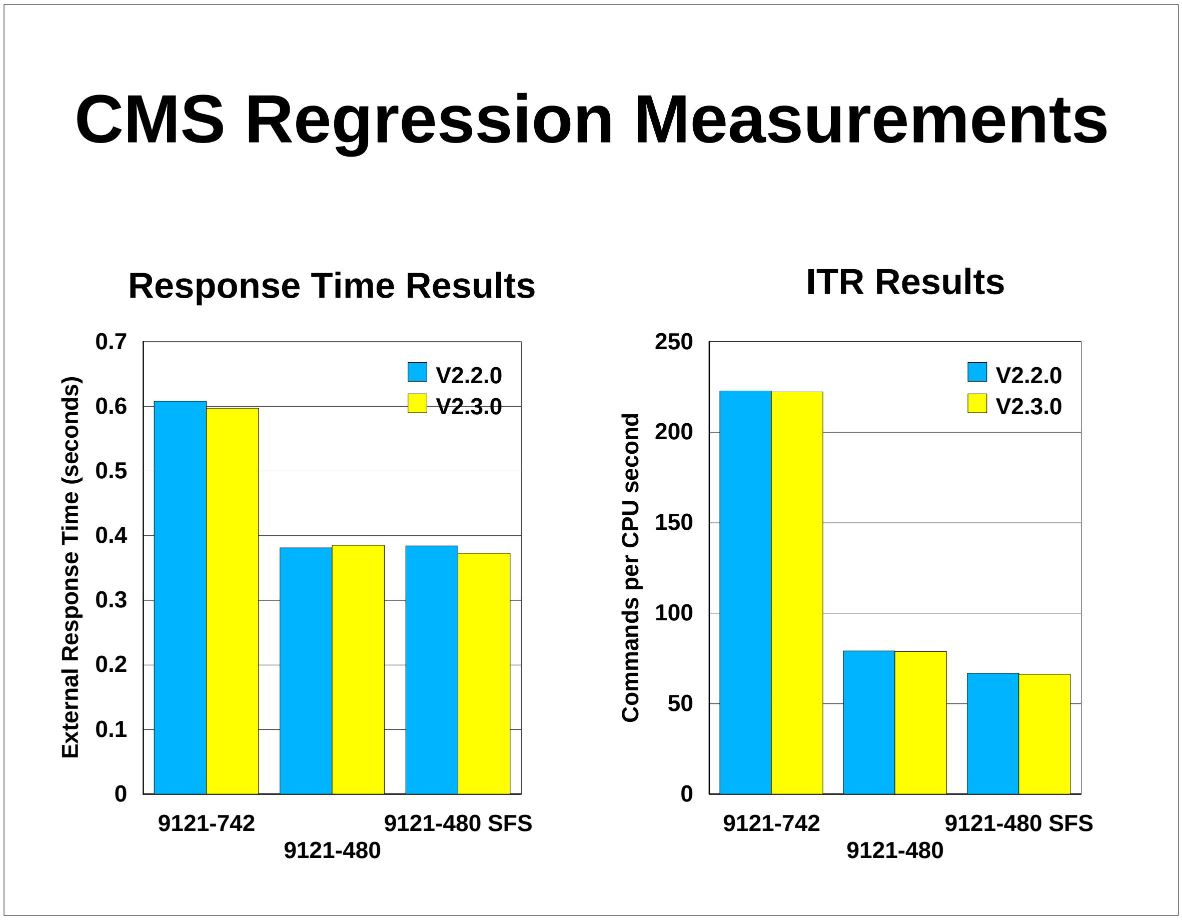
The new function includes the ability for an application to contribute not only sample data, but also event or configuration data to the APPLDATA domain via Diagnose x'DC'. TCP/IP 310 Stack exploits this.The MONVIEW utility is faster for certain scenarios with use of an MDATPEEK stage written in assembler.



VM Performance Products

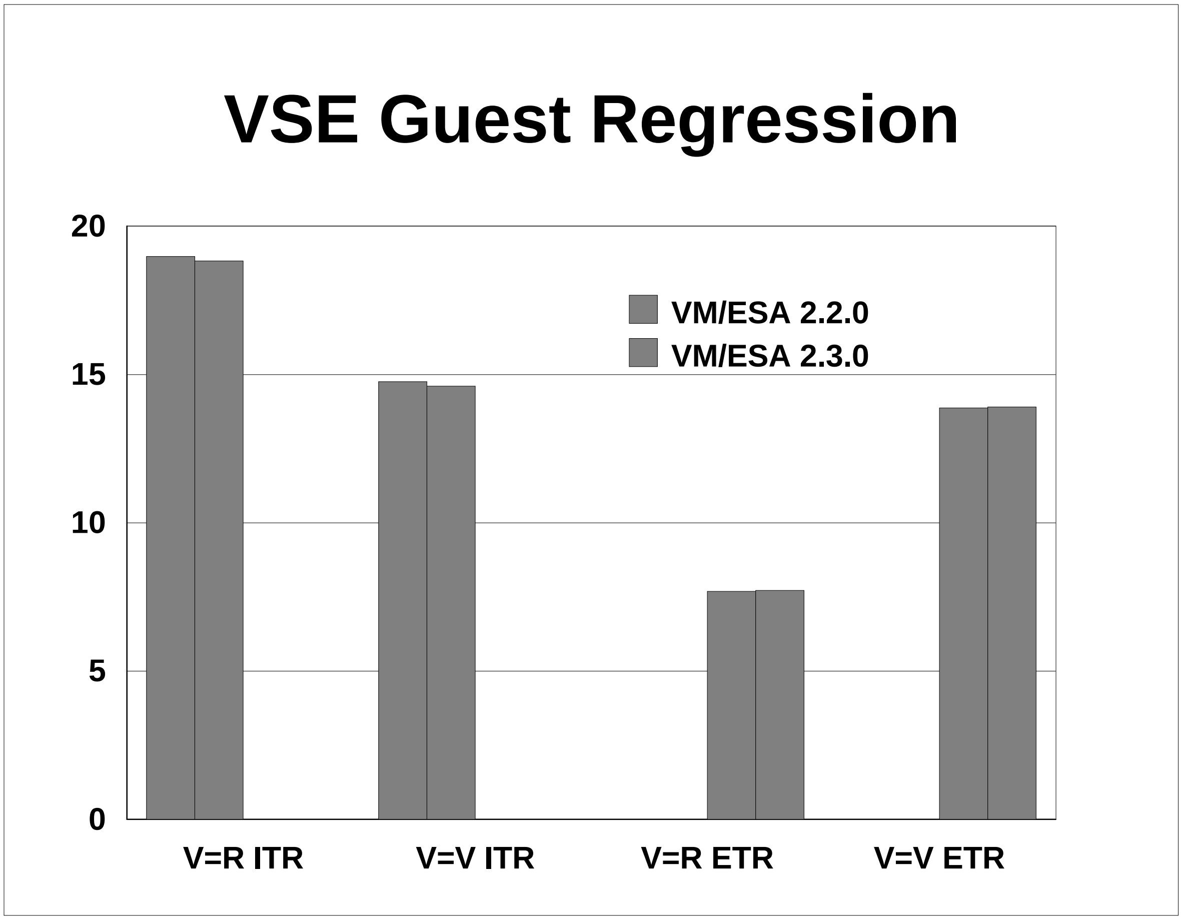
- VMPRF
  - will run on 2.3.0
  - latest service is recommended
- RTM/ESA
  - Requires APAR GC05430 (PTF UG03868)
- VMPAF
  - will run on 2.3.0
- FCON/ESA
  - Versions 2.3.02 and 3.1.00 will run on 2.3.0
  - Watch for additional support in next release

The four IBM VM performance products run on 2.3.0. VMPRF, RTM, and FCON have been enhanced for 2.3.0. For VMPRF and RTM, see the latest service for all the enhancements. There was an error in some early 1998 announcements that talked of the discontinuance of RTM/ESA. This was really suppose to be RTM/370 being withdrawn. RTM/ESA is still available and supported.



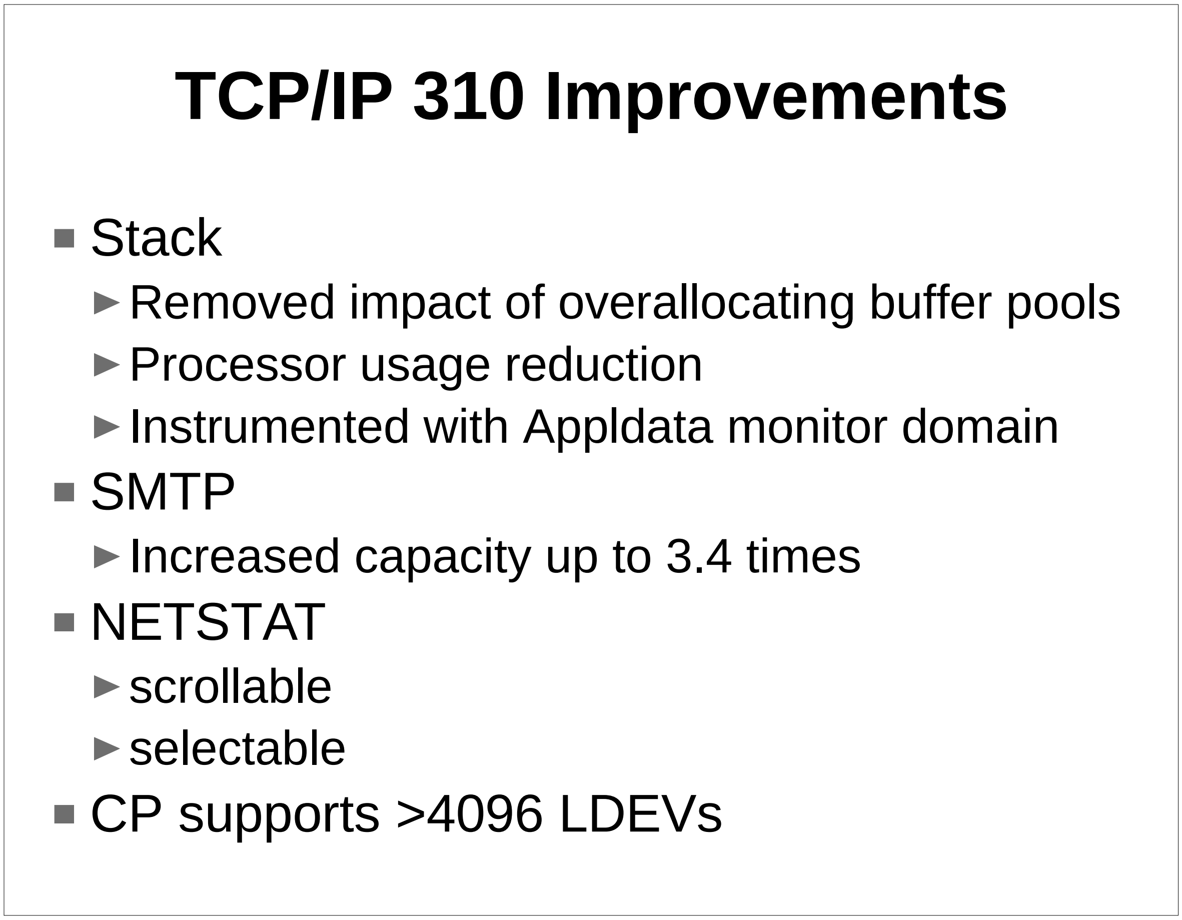
CMS Regression Measurements

Three environments were measured for CMS performance verification. All three showed little change in the performance when comparing 2.2.0 to 2.3.0.



VSE Guest Regression

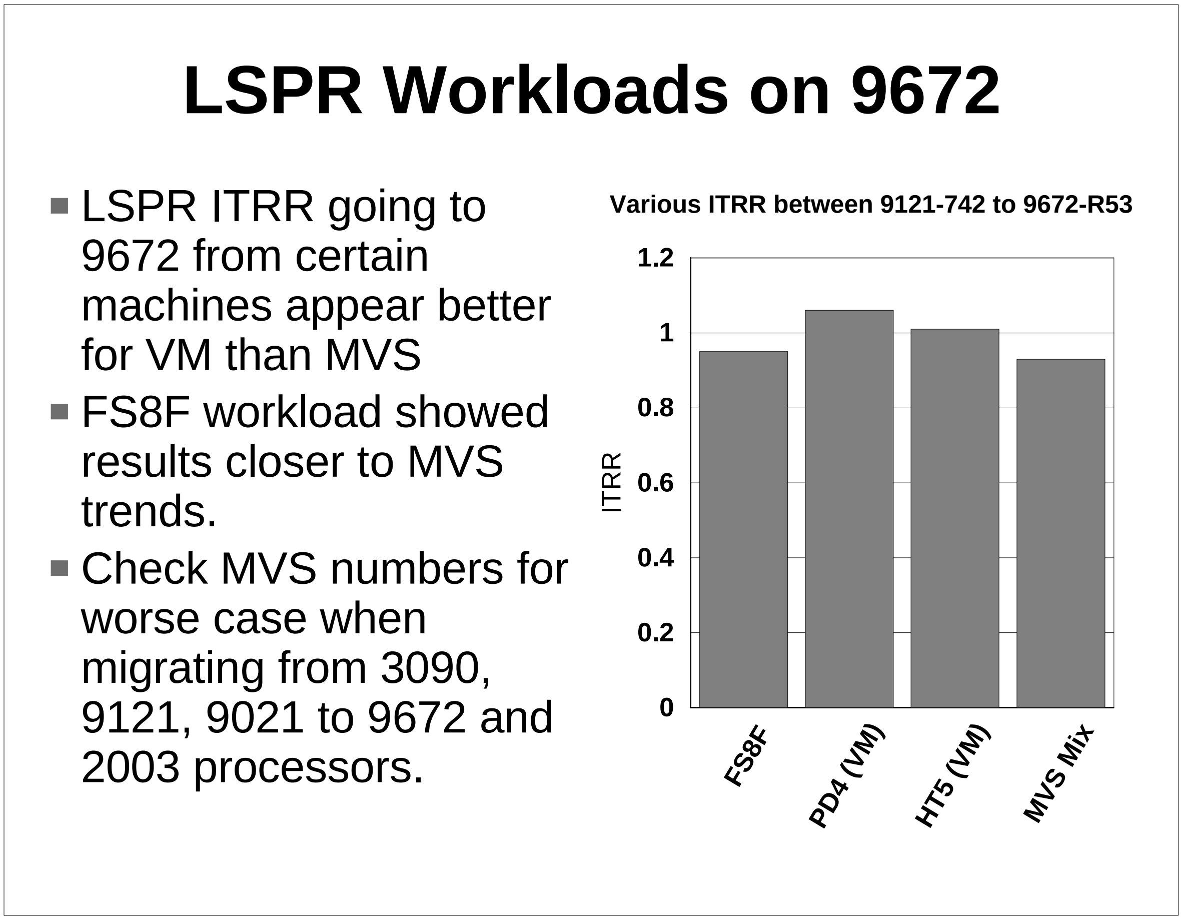
VSE Guest performance also remained roughly equivalent or a small decrease. Both V=R and V=V measurements were made. ITR is Internal Throughput Rate or a measure of commands per CPU second. ETR is External Throughput Rate or a measure of commands per wall clock second.



TCP/IP 310 Improvements

- Stack
  - Removed impact of overallocating buffer pools
  - Processor usage reduction
  - Instrumented with Appldata monitor domain
- SMTP
  - Increased capacity up to 3.4 times
- NETSTAT
  - scrollable
  - selectable
- CP supports >4096 LDEVs

TCP/IP function level 310 had a number of performance improvements. There is a separate presentation on these improvements. The performance enhancements involved decreasing the processor usage by the stack about 2%, making it less sensitive to over committing buffer pool sizes, support for RFC1323 (Long Fat Networks), and instrumenting it. SMTP saw capacity improvements by minimizing synchronous minidisk I/O. In addition, the NETSTAT command is now friendlier by providing selection criteria for certain options and having the interval option present the data in a scrollable and sortable manner.

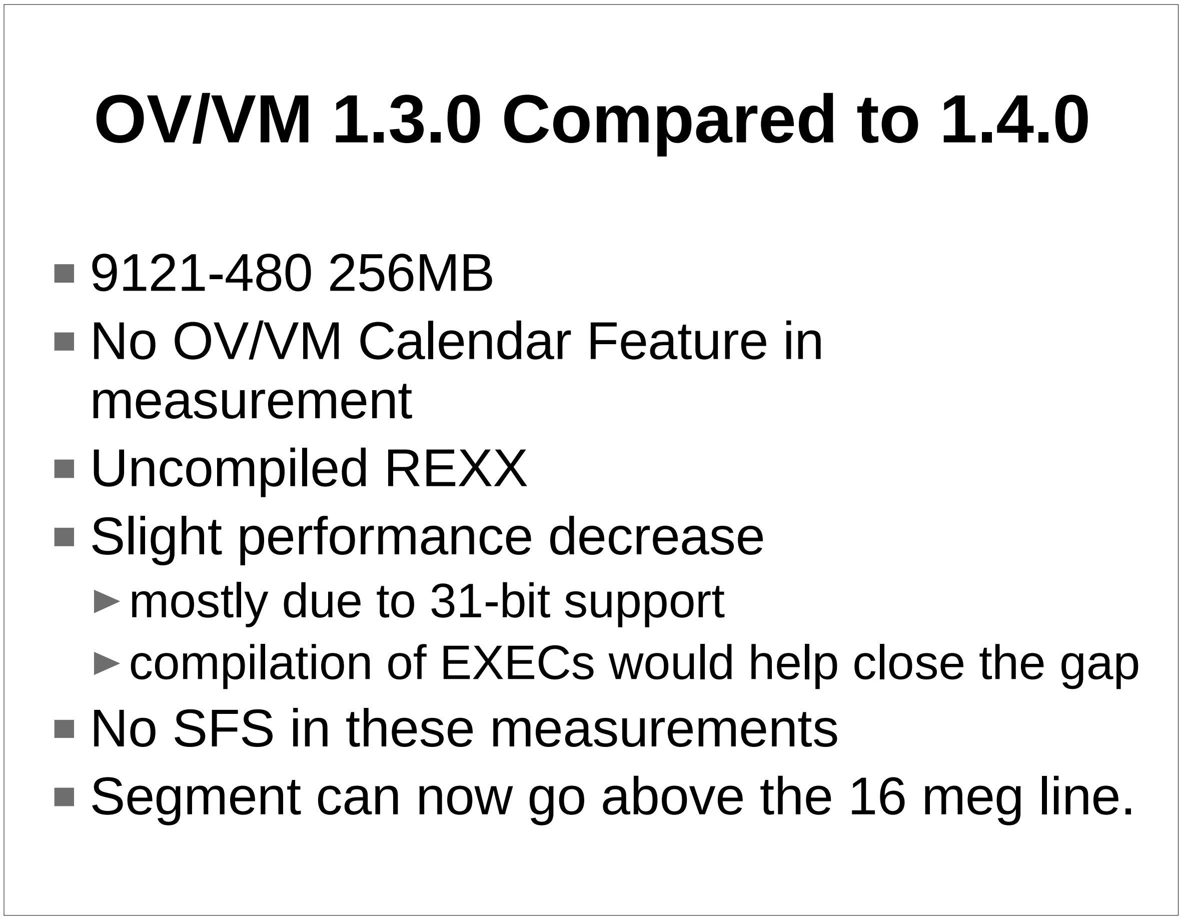


LSPR Workloads on 9672

- LSPR ITRR going to 9672 from certain machines appear better for VM than MVS
- FS8F workload showed results closer to MVS trends.
- Check MVS numbers for worse case when migrating from 3090, 9121, 9021 to 9672 and 2003 processors.

When the Large Systems Performance Reference (LSPR) VM measurements were made on the first 9672s, the results showed that VM did better than MVS on these new processors. However, after some real field experience we found that some customers were seeing ITR ratios more in line with MVS. While testing VM/ESA 2.2.0 we measured the VM development workload (FS8F) on a 9672-R53 and a 9121-742 and compared the results to the LSPR workloads (HT5 and PD4). We have found that FS8F is more in line with the MVS results.

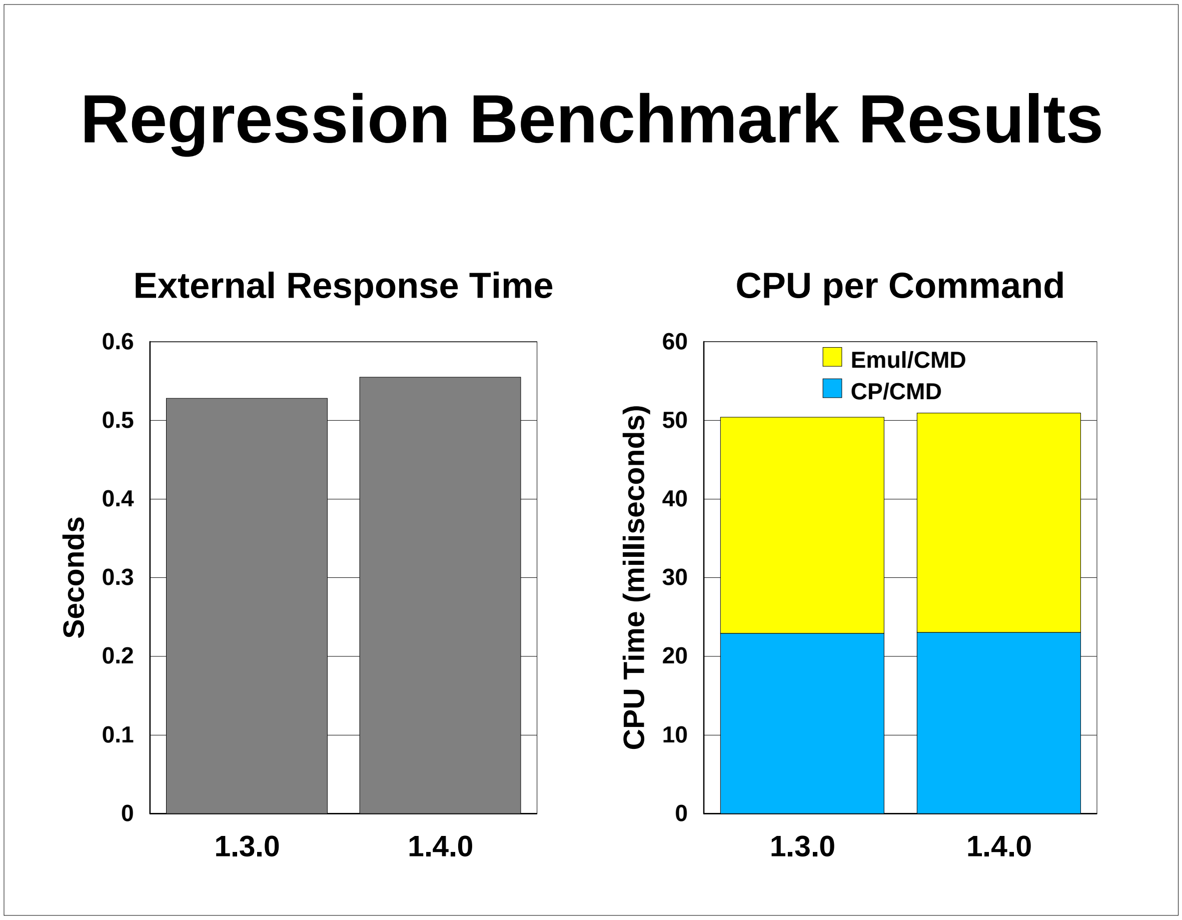
We recommend that when migrating from a 3090, 9121, or 9021 to a 9672 or 2003 processor, that you check the MVS ITR ratios as well as the VM workloads.



OV/VM 1.3.0 Compared to 1.4.0

- 9121-480 256MB
- No OV/VM Calendar Feature in measurement
- Uncompiled REXX
- Slight performance decrease
  - mostly due to 31-bit support
  - compilation of EXECs would help close the gap
- No SFS in these measurements
- Segment can now go above the 16 meg line.

Measurements were made to measure the performance of the new OV/VM release 4 product. This was done on VM/ESA 2.2.0 and the calendar feature was not used. In our environment, we ran without the compiling the OV/VM REXX code. Our measurements showed a slight performance degradation, mostly due to 31-bit support. However, we still saw sub-second response time. The delta in performance would have been smaller had we used compiled REXX. No SFS directories were used in this regression environment.
Also note that in our test environment, we could not exploit the room freed up under the 16MB line.

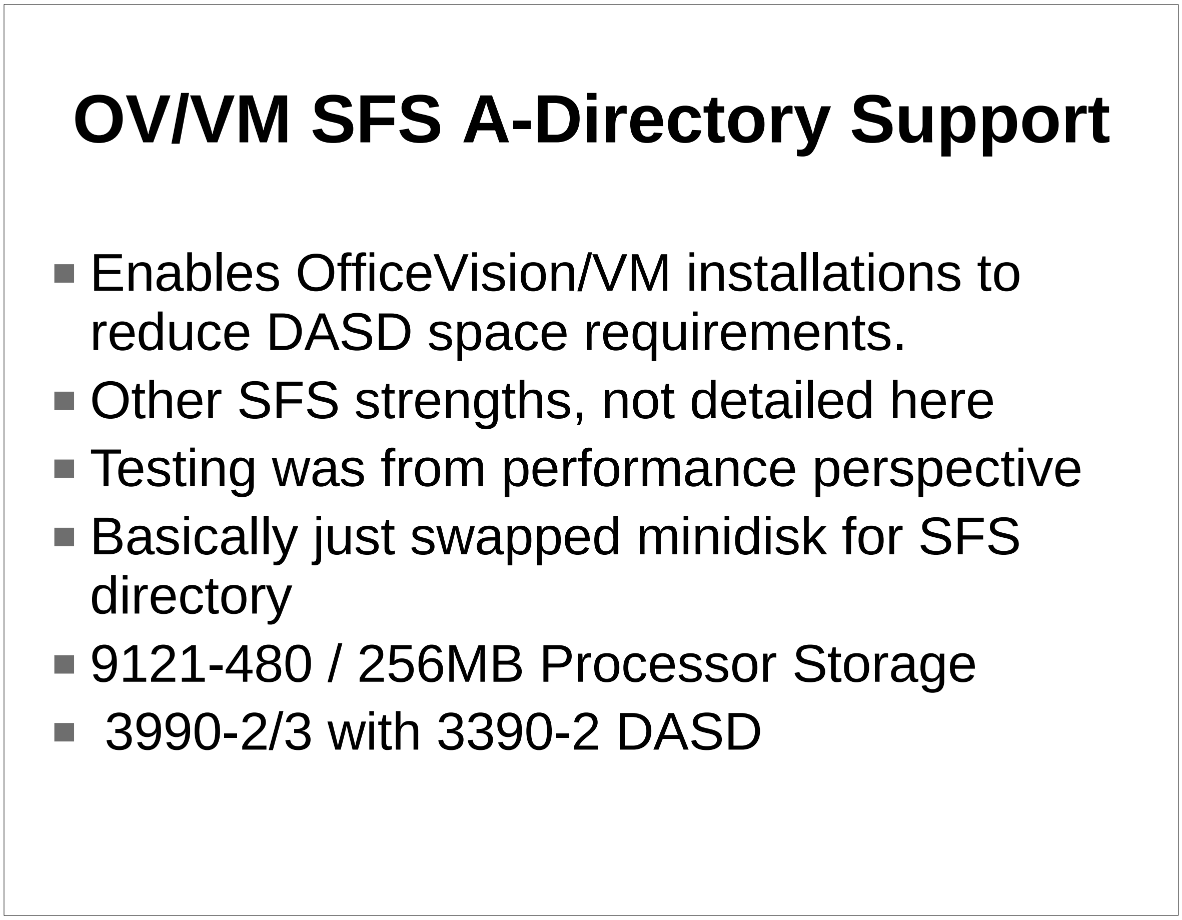


Regression Benchmark Results

External Response Time

CPU per Command

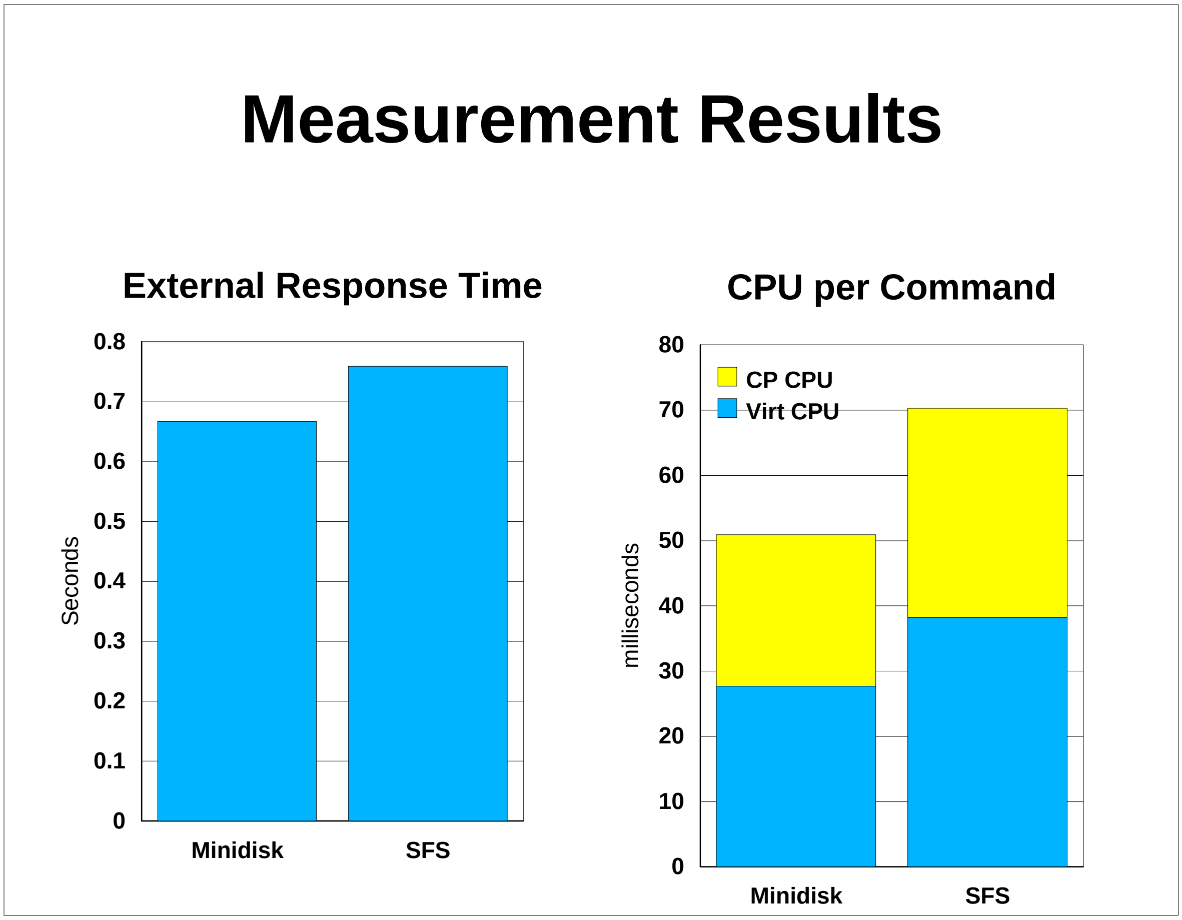
These graphs illustrate the slight increase in response time and processor resources per command when going from OV/VM 1.3.0 to 1.4.0.



OV/VM SFS A-Directory Support

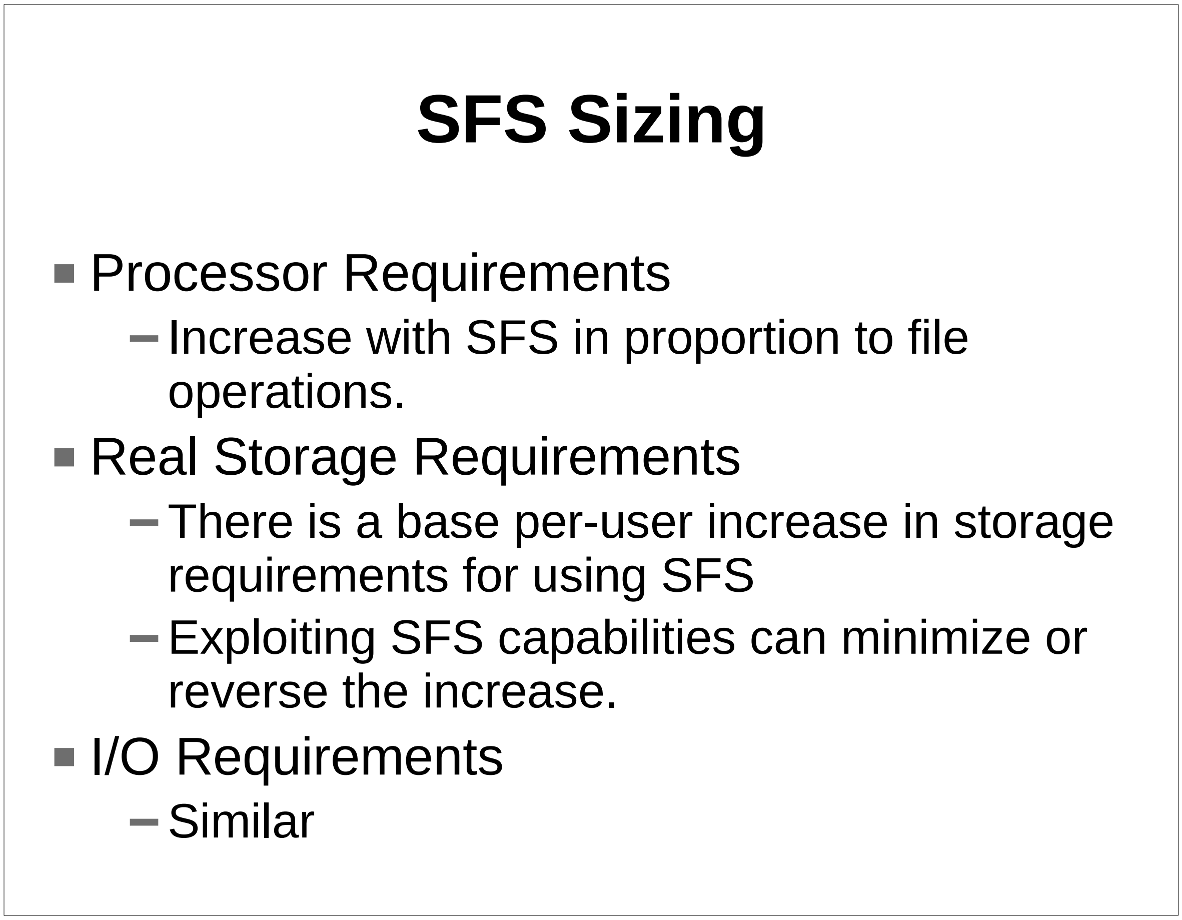
- Enables OfficeVision/VM installations to reduce DASD space requirements.
- Other SFS strengths, not detailed here
- Testing was from performance perspective
- Basically just swapped minidisk for SFS directory
- 9121-480 / 256MB Processor Storage
- 3990-2/3 with 3390-2 DASD

OV/VM now supports the use of an SFS directory as the A-mode. This could help customers reduce DASD space requirements significantly. In addition, there are other SFS strengths that will not be described here. For our measurements we changed our end users from using minidisks to using SFS for their A-mode access. The processor and DASD types are shown.



Measurement Results

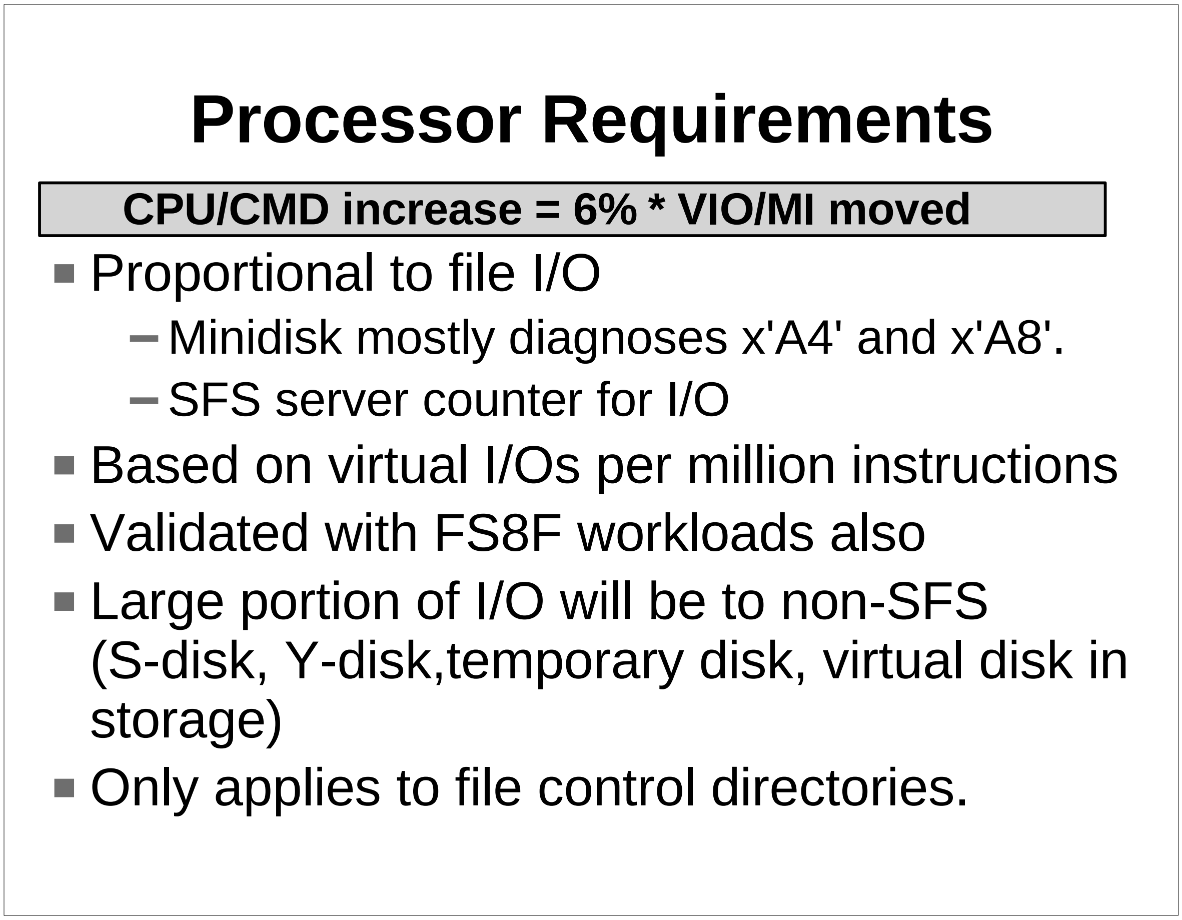
While response time remains sub-second it is slightly higher with SFS. The processor time per command also increased for SFS. About 38% as we projected. Foils that follow we describe how to project the increase in processor requirements as this is workload dependent.



SFS Sizing

- Processor Requirements
  - Increase with SFS in proportion to file operations.
- Real Storage Requirements
  - There is a base per-user increase in storage requirements for using SFS
  - Exploiting SFS capabilities can minimize or reverse the increase.
- I/O Requirements
  - Similar

In general processor requirements for SFS increase in proportion to the amount of file activity that occurs while real storage requirements are basically a fixed cost. The I/O requirements stay the same, except a server machine does the majority of the I/O instead of the end users.

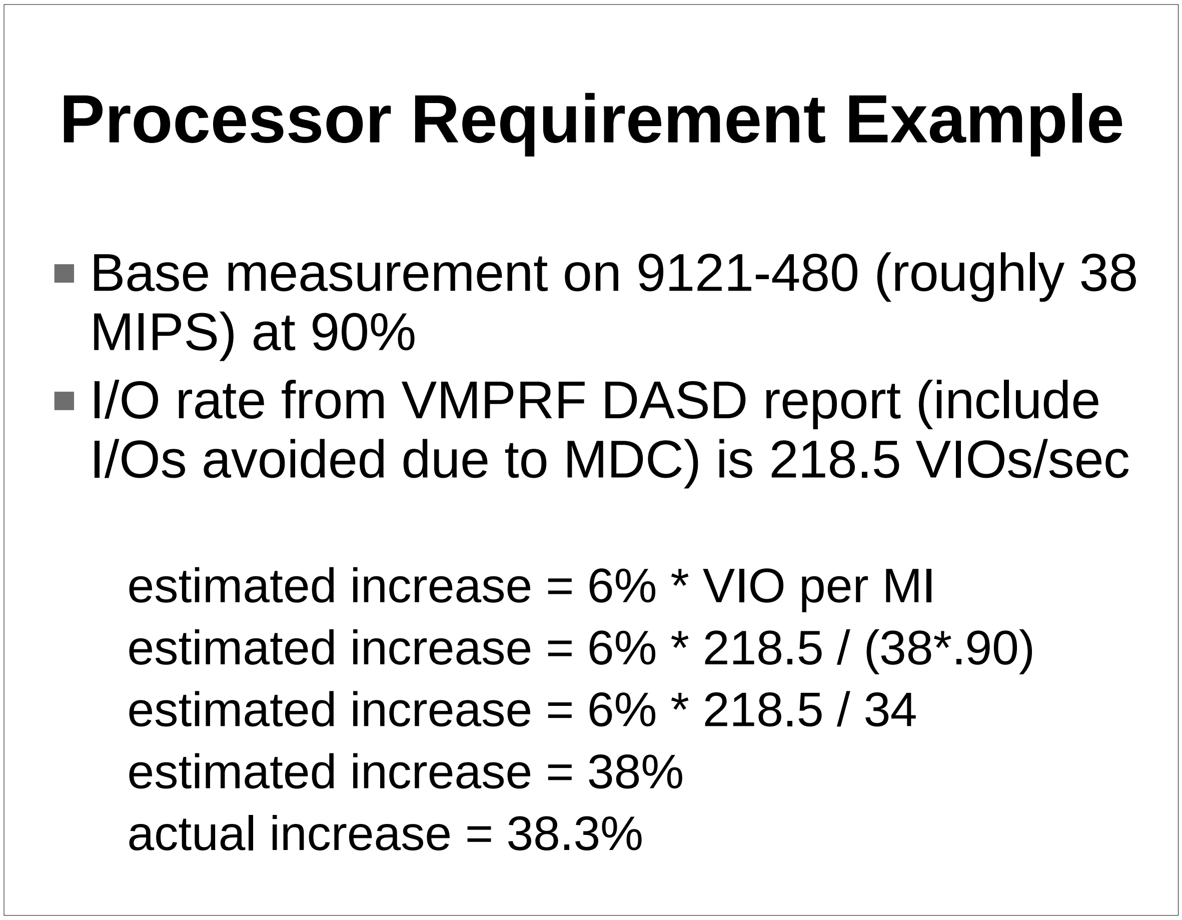


Processor Requirements

CPU/CMD increase = 6% * VIO/MI moved

- Proportional to file I/O
  - Minidisk mostly diagnoses x'A4' and x'A8'.
  - SFS server counter for I/O
- Based on virtual I/Os per million instructions
- Validated with FS8F workloads also
- Large portion of I/O will be to non-SFS (S-disk, Y-disk,temporary disk, virtual disk in storage)
- Only applies to file control directories.

The rule of thumb in the gray box is a formula that can be used to project the increase in CPU usage per command. This is about 6% times the number of virtual I/Os per million instructions executed that are moved from minidisk to SFS. This will become clearer with an example. This does not apply to SFS dircontrol directories backed by VM data spaces.

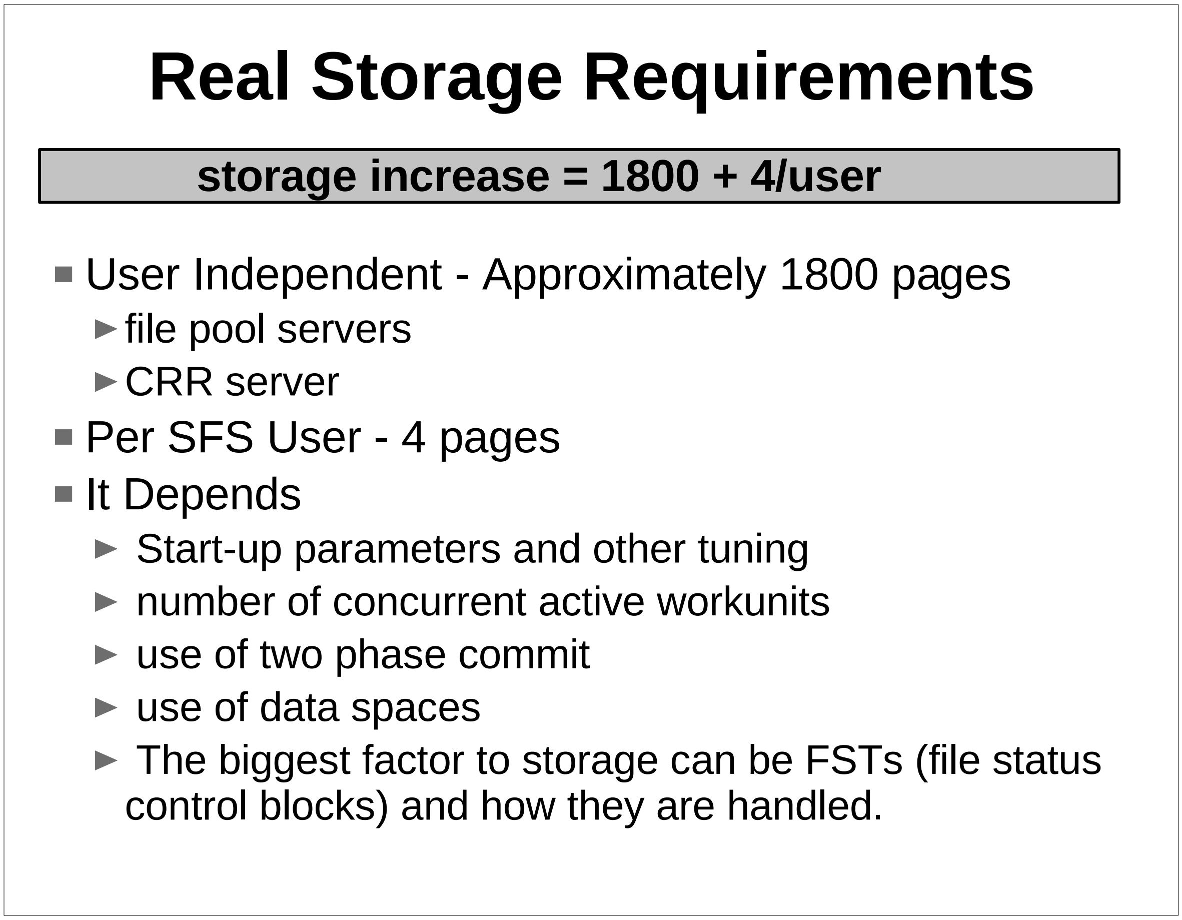


Processor Requirement Example

- Base measurement on 9121-480 (roughly 38 MIPS) at 90%
- I/O rate from VMPRF DASD report (include I/Os avoided due to MDC) is 218.5 VIOs/sec

estimated increase = 6% * VIO per MI
estimated increase = 6% * 218.5 / (38*.90)
estimated increase = 6% * 218.5 / 34
estimated increase = 38%
actual increase = 38.3%

In this example, we have a 9121-480 running at 90% and VMPRF reports the I/O rate to be 217.5 for the volumes we are moving into SFS. So 90% of 38 MIPS is 34. Plugging these numbers in, we get a 38% increase in processor requirements. In reality, we saw 38.3% so the formula was fairly accurate.

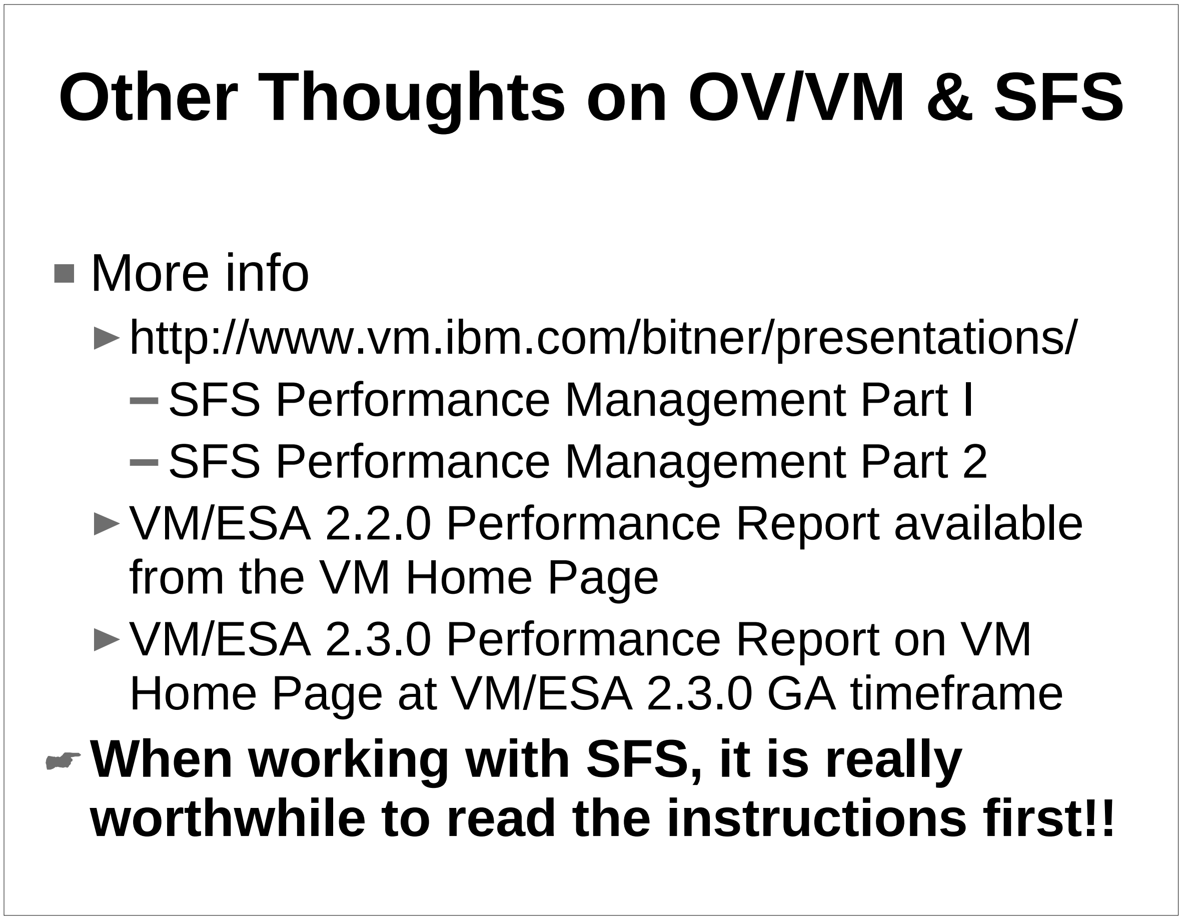


Real Storage Requirements

storage increase = 1800 + 4/user

- User Independent - Approximately 1800 pages
  - file pool servers
  - CRR server
- Per SFS User - 4 pages
- It Depends
  - Start-up parameters and other tuning
  - number of concurrent active workunits
  - use of two phase commit
  - use of data spaces
  - The biggest factor to storage can be FSTs (file status control blocks) and how they are handled.

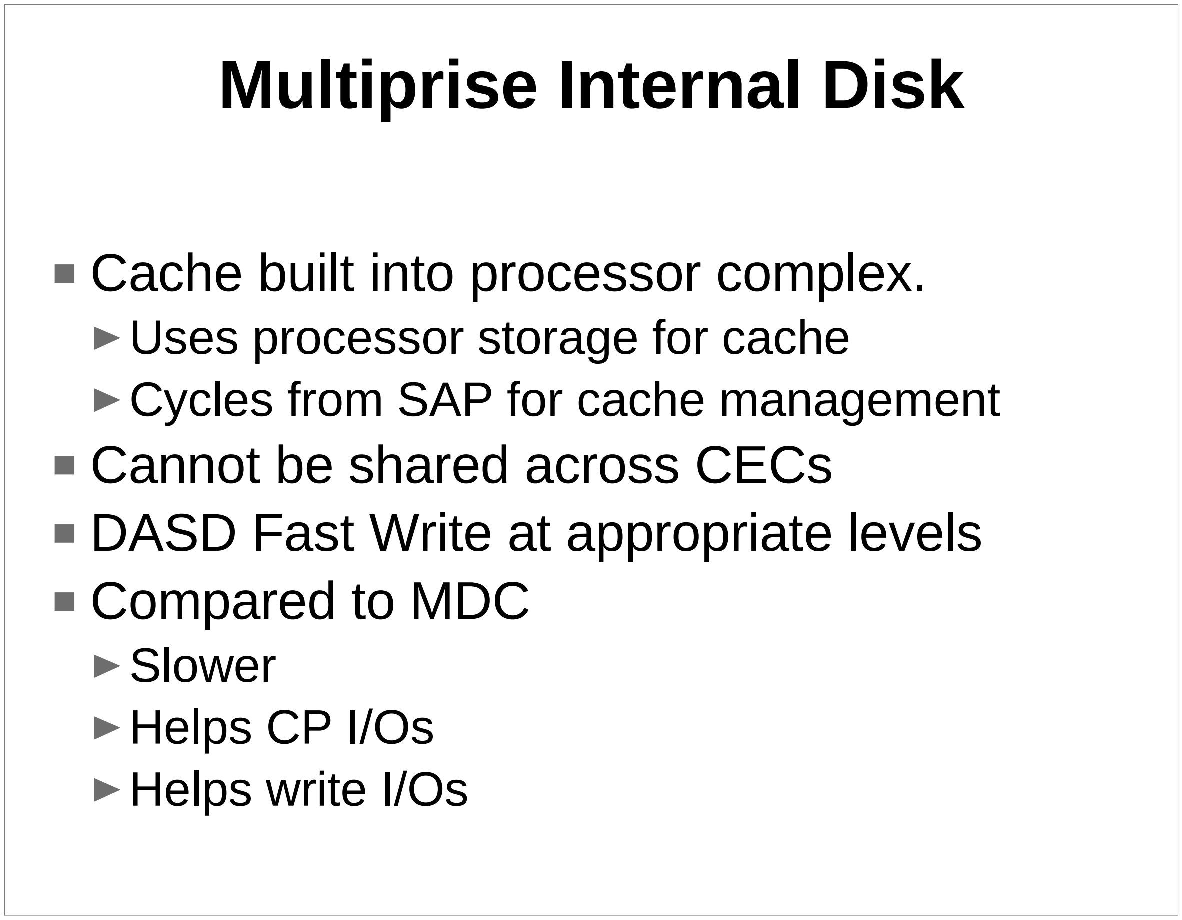
Real storage requirements are easier to size. It is a fixed about of storage based on the number of users that will be connected to the filepool. The gray box gives this rule of thumb. The 4 pages per user include some pages in end users and some in the SFS server machines. There are a number of factors that could change the storage requirements. They are listed here.



Other Thoughts on OV/VM & SFS

- More info
  - http://www.vm.ibm.com/bitner/presentations/
    - SFS Performance Management Part I
    - SFS Performance Management Part 2
  - VM/ESA 2.2.0 Performance Report available from the VM Home Page
  - VM/ESA 2.3.0 Performance Report on VM Home Page at VM/ESA 2.3.0 GA timeframe
- **When working with SFS, it is really worthwhile to read the instructions first!!**

I went through this sizing information fairly quickly. To get additional information on SFS performance management, see the two SFS performance presentations that are on the VM home page. The URLs are given here. I want to really stress the point that doing the homework before you start to use SFS is very important.

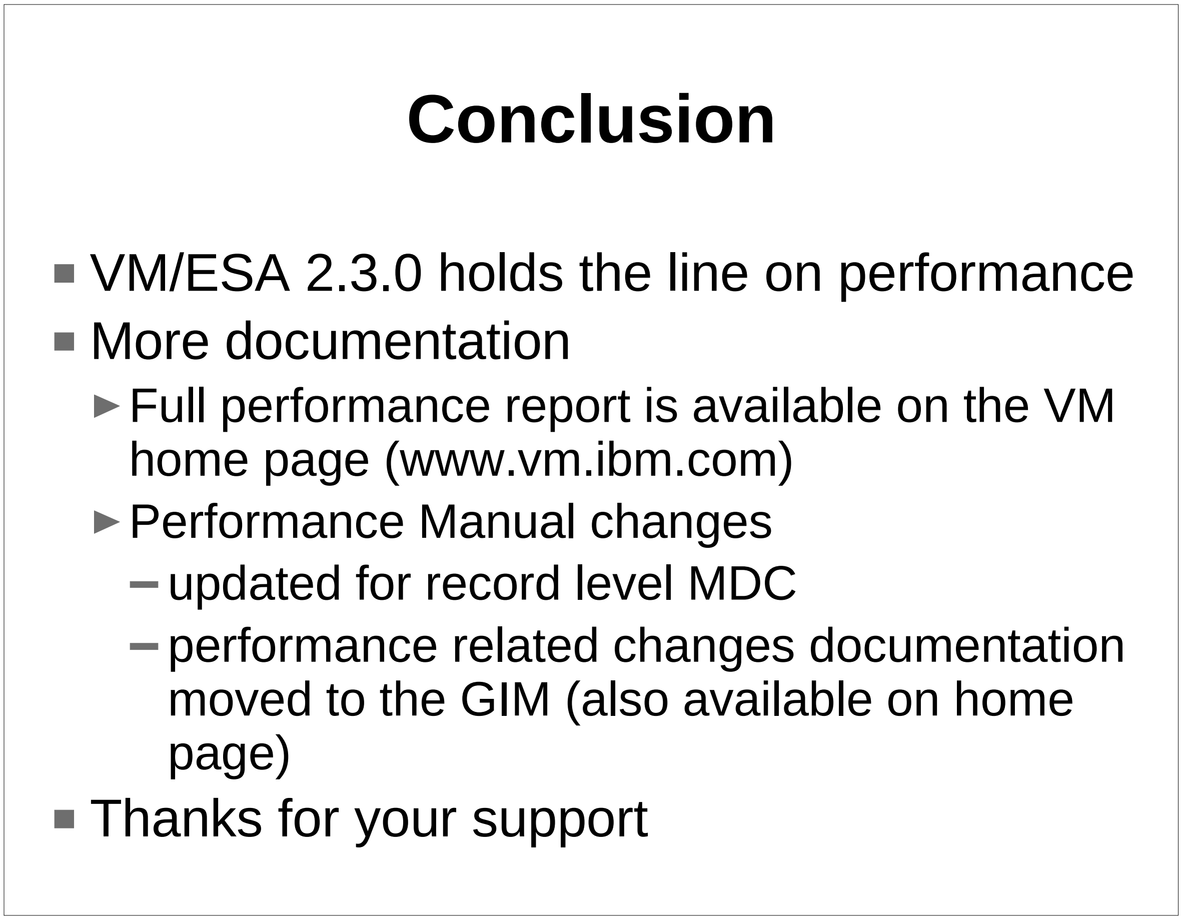


Multiprise Internal Disk

- Cache built into processor complex.
  - Uses processor storage for cache
  - Cycles from SAP for cache management
- Cannot be shared across CECs
- DASD Fast Write at appropriate levels
- Compared to MDC
  - Slower
  - Helps CP I/Os
  - Helps write I/Os

The Multiprise processors can be ordered with Internal Disk hardware. This provides relatively good performing DASD with low costs. However, it is very different than other DASD available for S/390. The cache is taken from processor storage. So now you have three choices on how your processor storage is divided: central storage, expanded storage, and Internal Disk Cache storage. The cache is management by a SAP (system assist processor), instead of by software running on a normal processor unit.

The cache of internal disk is valuable even if you are using minidisk cache (MDC). The internal disk can cache CP I/Os and write I/Os, while MDC cannot. Even though the internal disk uses cycles on a SAP, there is still more overhead than MDC since CP pathlengths to issue the I/O are required.



Conclusion

- VM/ESA 2.3.0 holds the line on performance
- More documentation
  - Full performance report is available on the VM home page (www.vm.ibm.com)
  - Performance Manual changes
    - updated for record level MDC
    - performance related changes documentation moved to the GIM (also available on home page)
- Thanks for your support

For more documentation on VM performance, check our home page. The performance report is also available as a PDF file. There were a few changes made in the Performance manual. Some of the MDC guidelines were enhanced. These are also mirrored on our VM performance home page.

As always, thanks for your support.