# VM/ESA Performance Update

Last Updated: 12 November 1999

Bill Bitner
VM Performance
607-752-6022
bitner@vnet.ibm.com

This presentation gives an overview of VM/ESA performance changes for 1999. While I have the honor of giving this presentation, there are many others behind the scenes that make all this happen. This includes Bill Guzior and Wes Ernsberger who make up the rest of the VM Performance team, and many others in development and test. It is a privilege to be part of such a great team.

# Legal Stuff

## 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, VM/ESA
The following are trademarks of Sun Microsystems:
Java, JDK

In addition to the above disclaimers, please also note that some measurements for VM/ESA 2.4.0 presented here may not be the final measurements for the release, but may be from pre-code freeze levels of the software. This is potentially true for other components or products shown in this presentation.

# Introduction

- VM/ESA 2.3.0 has been running well.
- VM/ESA 2.4.0 coming soon
  - Hardware Support
  - Scheduler Enhancements
  - TCP/IP
  - Incremental monitor improvements
- Other work in progress
  - Java
  - ADSM Version 3

VM/ESA has been in the field for over a year now and has shown good performance. The next release, VM/ESA 2.4.0, will have a few more performance changes. We will look at some of these in this presentation. A significant portion of the release involves hardware support, some of which is not interesting from a performance perspective. We will also look at incremental improvements to the CP scheduler, TCP/IP, and monitor. Other performance work from this year, but not necessarily tied to VM/ESA 2.4.0, includes Java and ADSM changes.

# VM/ESA Regression

- CMS Regression V2.3.0 to V2.4.0
  - ITR decreased 0.4 to 0.6%.
  - Response time was equivalent.
- VSE
  - Equivalent performance
- TCP/IP
  - FTP equivalent to FL 310
    - VM FTP client "get" throughput improved 2%
  - Telnet equivalent to FL 310
  - NFS big improvements (see later charts)

Regression performance was not the main objective this release. However, performance for these environments continue to be good. There was a slight decrease in ITR (Internal Throughput Rate) for CMS, while response time was roughly the same. A slight increase in response time for SFS, but close to run variation.
VSE guest performance was equivalent.
TCP/IP performance overall was equivalent. Some improvements and enhancements will be discussed on a later foil.

## Support for FICON and Friends

- Capacity gains:
  - Bandwidth of 100 MB/Sec
- Additional Exploitation:
  - Synchronize Control extends current prefetching
    - Paging, Spooling, and Guest
  - Avoidance of nullification window
    - Requires Enterprise Storage Server (ESS) as well as FICON for support

FICON, or Fibre Connection, is a new channel hardware that impacts performance in multiple ways. There is the pure hardware capacity gains in bandwidth. In addition, there are several new features associated with this support that VM/ESA will exploit. We look at two of them here.
Prefetching has limits to avoid data integrity problems which could occur if the same data was being written and read back to back. New Synchronize Control interface adds bits to indicate that you will not perform any problematic I/O patterns. We originially thought we would put this support in the stand-alone utilities as well, but that support is not provided at this time.
The nullification window involves waiting for acknowledgment handshaking between the control unit and the channel subsystem. An ESS (Enterprise Storage Server) is required to be able to avoid the nullification window.

# 2.4.0 CP Monitor Changes

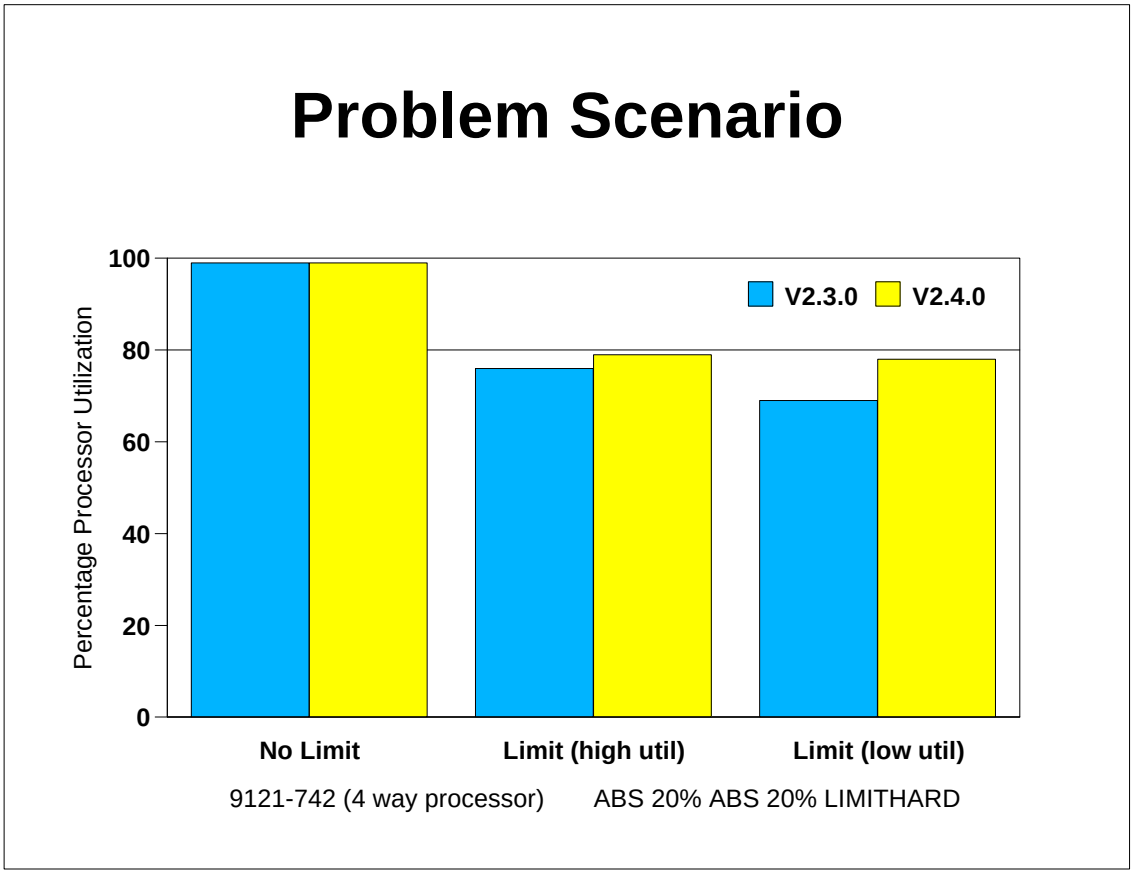
- Enhanced Channel Path Measurement
  - New System Domain Record (per channel)
- Indication of source of device active time (HW or SW)
- Synchronization of SCM block statistics
- CP return free storage requests now accurate
- Support for Parallel Access Volumes on ESS DASD.
- New I/O record for state change events such as PAV in ESS.

With the addition of FICON channels, channel measurement requirements change. The monitor has been enhanced to report on the enhanced channel path measurement data provided by the hardware through new records. There are two changes associated with the subchannel measurement block data. First is an indication whether the device active time component is provided by hardware or calculated by the software. In addition, changes were made to synchronize all the SCM statistics. The CP return free storage request counters had been artificially low since VM/ESA 1.2.0. This was corrected.
A new I/O record exists to record certain state changes of a device, such as a dynamic change in an alias for parallel volume on an ESS DASD device.

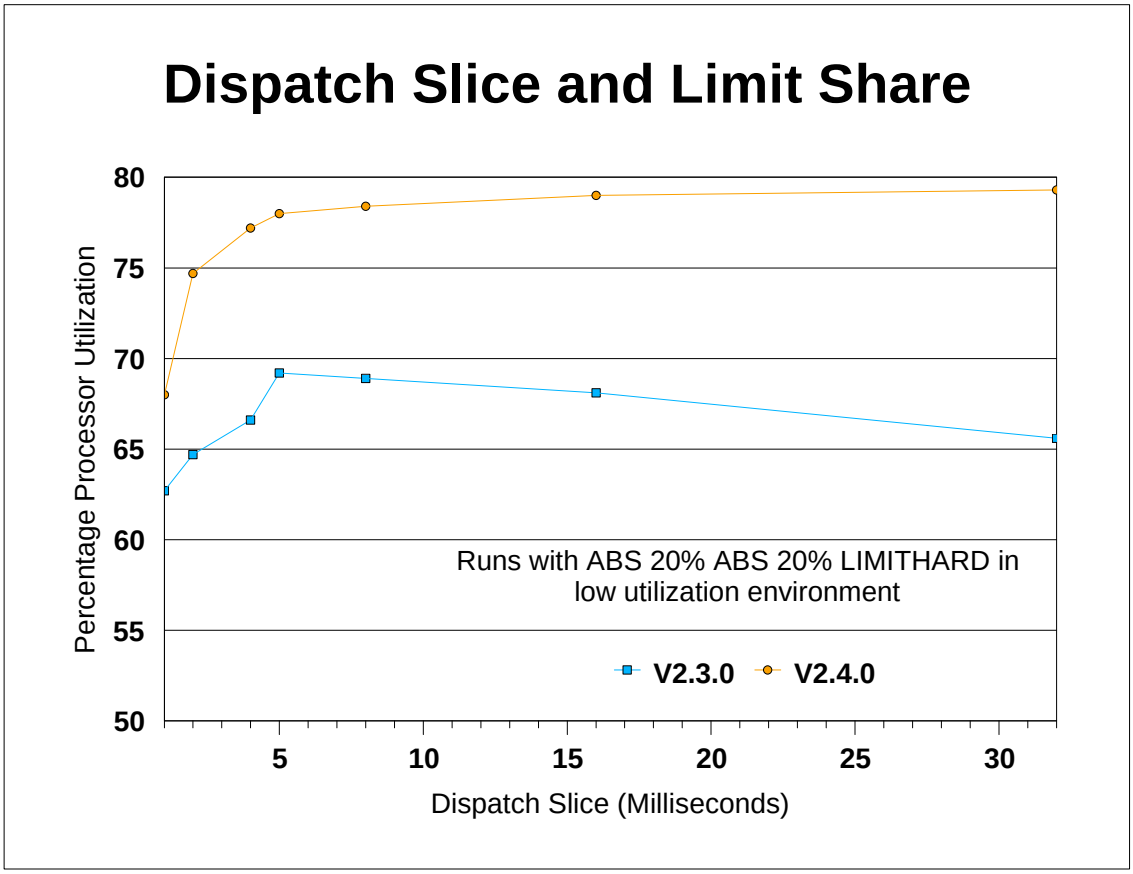
# Improved Limit Shares

- VM/ESA 1.2.2 Introduced Limit Shares
- Two flavors:
  - LIMITHARD - limit regardless of capacity
  - LIMITSOFT - limit unless extra capacity exists
- Worked great... except in
  - Virtual MP environments
  - Low system utilization
- Some minor improvements through service stream
- FIN APAR VM61527 now in VM/ESA 2.4.0

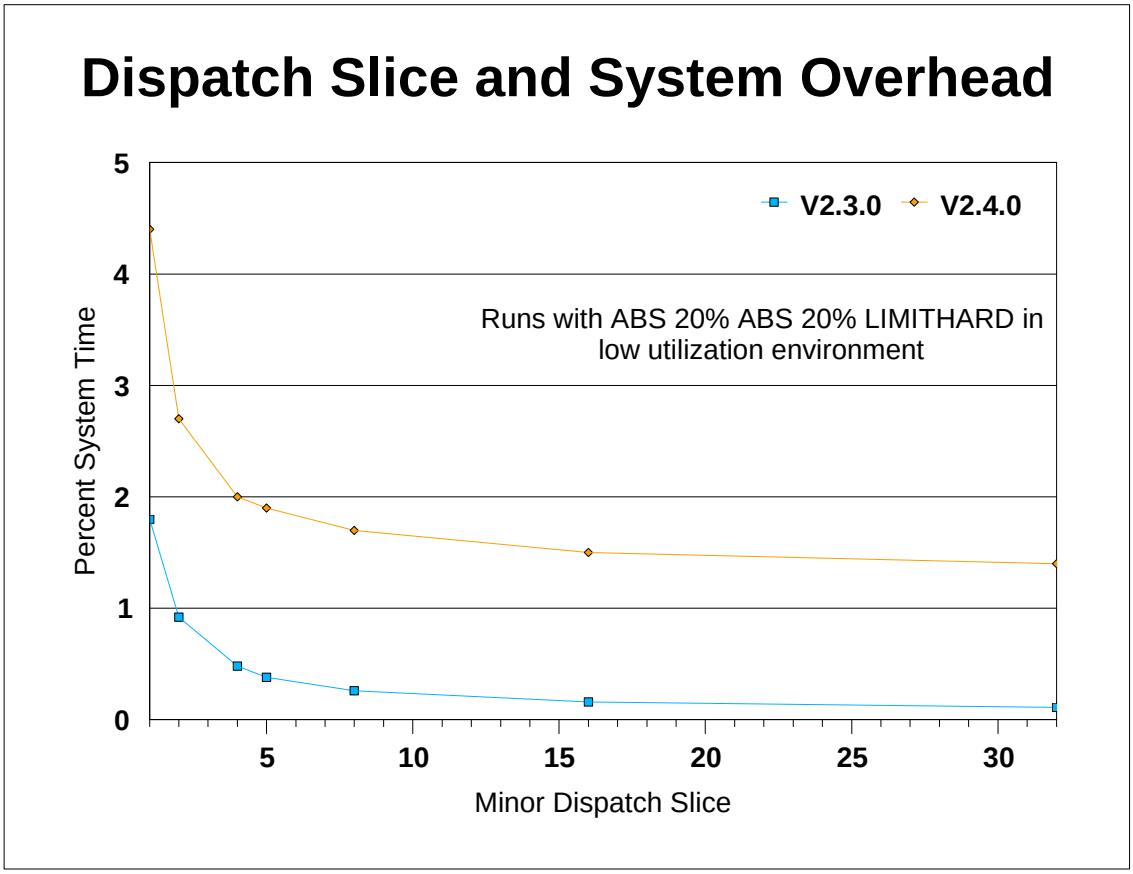
Limit shares were first introduced in VM/ESA 1.2.2. The added function was welcomed by many. However, there were a few anomalies. Some minor improvements were made in the service stream. However, one problem remained: LIMITHARD share users were being held back more than necessary in low utilization environments. APAR VM61527 was opened for this and closed FIN. Code was given to one customer for fixtest and rolled into VM/ESA 2.4.0.



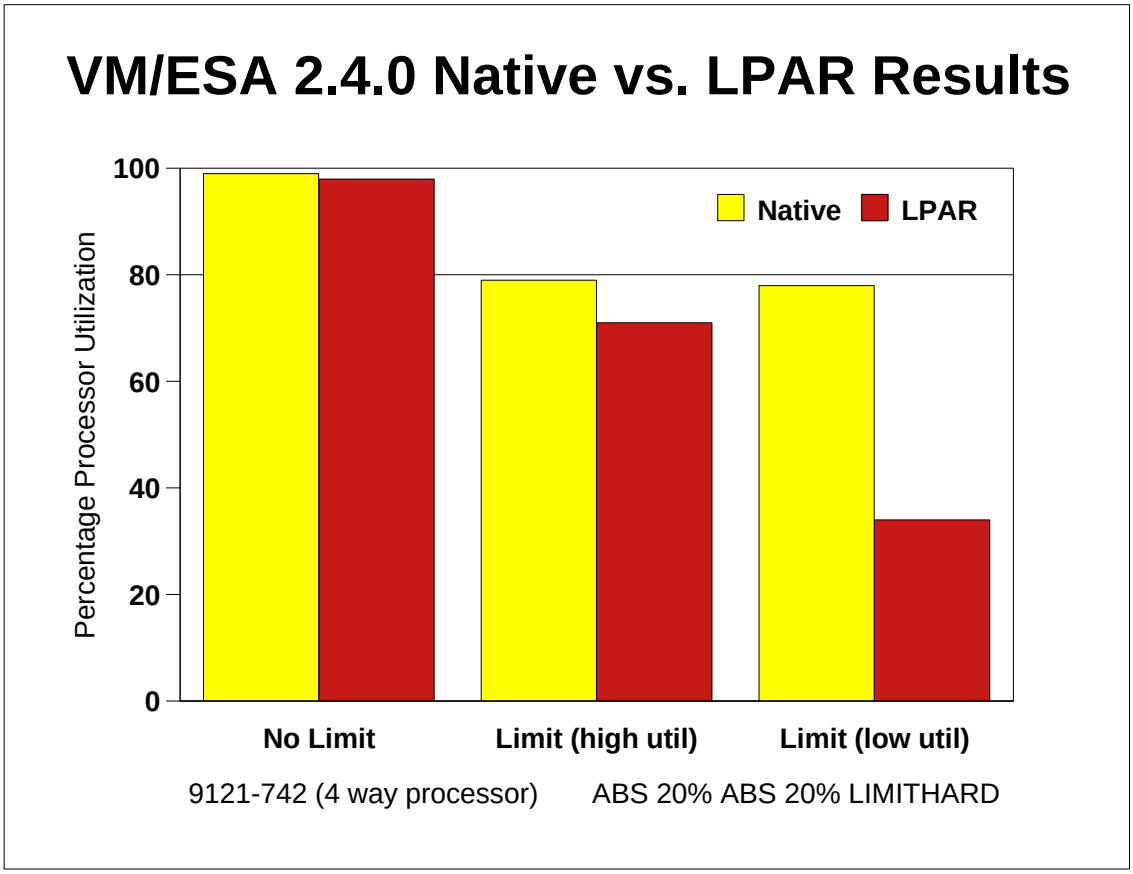

This graph shows the problem scenario and the improvement in VM/ESA 2.4.0. A CMS application that was processor bound ran on a 4-way. Three scenarios are shown here. In the first, no limit share was used and the virtual machine used almost 100% of a processor. A LIMITHARD setting of absolute 20% was set for the next two scenarios with the difference being in how busy the overall system was. In the middle set of bars, the system was running close to capacity while in the 3rd scenario  the system was basically idle except for the test user. In both of these limited cases you see VM/ESA 2.4.0 tracking closer to the limit of 80% of a single processor (80% of 1 processor on a 4-way is 20% of the system).




Experiments were also made with the LIMITHARD setting with various minor dispatch slice times. In VM/ESA 2.3.0, very short or very long minor dispatch slices made the limit tracking less accurate. This has been improved somewhat in VM/ESA 2.4.0.




Another interesting point from these tests with varying minor dispatch slice times was the system overhead. In general, the improved accuracy of LIMITHARD settings costs more in system time. Thought in non-limited cases it stays the same.
You can also see that very small dispatch slices greatly increase the relative overhead. However, the law of diminishing returns applies as very little overhead is saved for values over 8 milliseconds.




Note however, that in LPAR environments, the scheduler limit settings can not be as accurate. Shown here is the same VM/ESA 2.4.0 system running native and in an LPAR. For the LPAR tests, two partitions were defined both as logical 4 ways. The system under test had a weight of 30 with the other system having a weight of 70. As you can see the limited user received less than what they did in native systems.

## Share Capping Summary

- Less restrictive while holding a LIMITHARD in native environments.
- IBM tests show LPAR environment tends to hold user below the LIMITHARD setting.
- One ESP customer LPAR environment shows user getting more than the LIMITHARD.
- Use LIMITHARD with care in an LPAR environment.

The use of LIMITHARD to manage guests at certain MIPS levels has improved greatly, but it is still not an exact science. Care needs to be taken in LPAR environments as the IBM tests proved. In addition, one VM/ESA 2.4.0 ESP customer had an LPAR environment where a limited user was held back, but still got more than the absolute limit share value.

# SFS Performance Improvements

- Recent performance APARs rolled into VM/ESA 2.4.0
  - VM61547 - mitigate "lock out" scenario when deleting very large files (>512KB)
  - VM62008 - follow-on to VM61547
  - VM62086 - mitigate "lock out" scenario for long open-write-close nocommit sequences

There were three SFS performance APARs rolled into the base for VM/ESA 2.4.0. All of them dealt with scenarios where other users of SFS appeared to be "locked out" while a particular task for another user was being processed. VM61547 and VM62008 dealt with the task of deleting large files. The impact is proportional to the file size and is not really noticeable for files under 512 KB. The customer who found the problem was deleting 1 GB files. VM62086 dealt with a different scenario where a large number of file changes were made without a commit being issued.

# TCP/IP Improvements

- Feedback on TCP FL 310 with RFC 1323
  - With fast Ethernet seeing 5 GB/Hour with peaks of 7.9 GB/Hour
  - OSA 2 Fast Ethernet saw a factor of 3 improvement
  - Unlike VM/ESA, some stacks default RFC1323 off.
- APAR PQ18391 - extends TCP Maximum Segment Size (MSS)
- FL 320: TCP Header prediction
  - Lower pathlengths for inbound processing

While we were unable to do many measurements with the TCP/IP RFC 1323 support that went into FL 310, we have gotten positive feedback on this support from customers and other internal locations. Improvements of a factor of 3 in throughput were reported by some.
APAR PQ18391 to FL 310 is in the base for FL 320. It will correct a problem to allow for the full benefit of window scaling. Previously a hard coded limit of 20 segments existed. So if 20 x Maximum segment size is less than the window size, you were constrained in the past.
An enhancement in FL320 is TCP header prediction which will lower pathlengths associated with inbound processing.

# TCP/IP Monitor Improvements

- APAR PQ16942 rolled into FL 320
  - Allow for recording of larger amounts of data on TCB and UCB close records
- FL 320 Changes:
  - count of packets discarded for LAND attack
  - count segment headers predicted correctly
  - TCP close record now includes window scaling factors and local IP address
  - UDP open/close records now created for sessions initiated through sockets interface

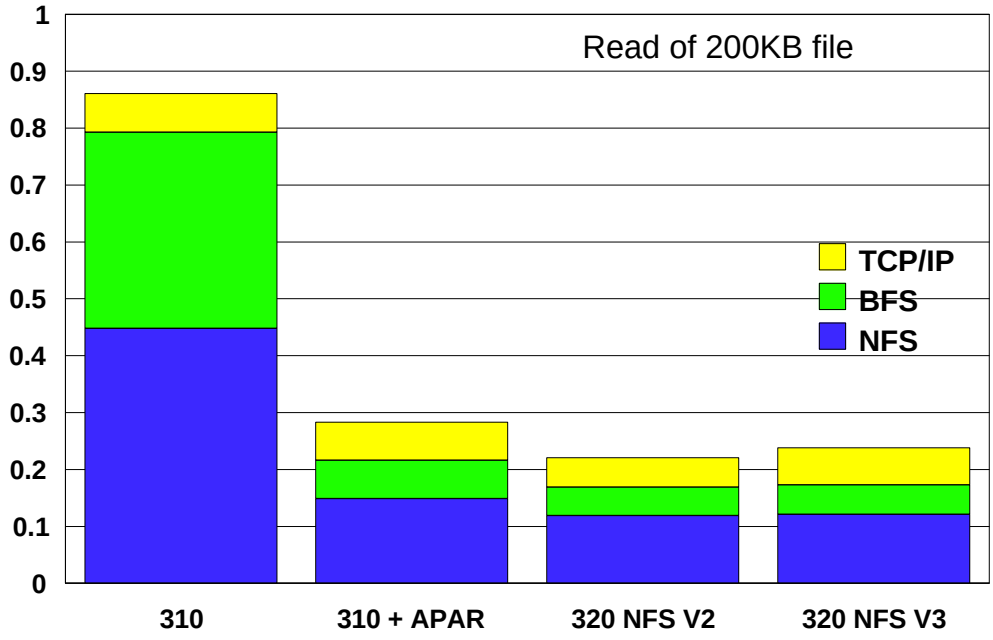
There are some additions to the Stack monitor data. The first is due to a FL310 APAR (rolled into FL 320) that allows for the recording of larger amounts of data in the TCB (TCP) and UCB (UDP) close records. Additionally, FL 320 changes include recording information about packets discarded for denial of service attacks, the number of times segment headers have been predicted correctly, and additional data on TCP close record such as IP address and local window sizes. In addition, problems were corrected that prevented UCB open/close records from being created when the session was initiated through the sockets interface.

# NFS Improvements

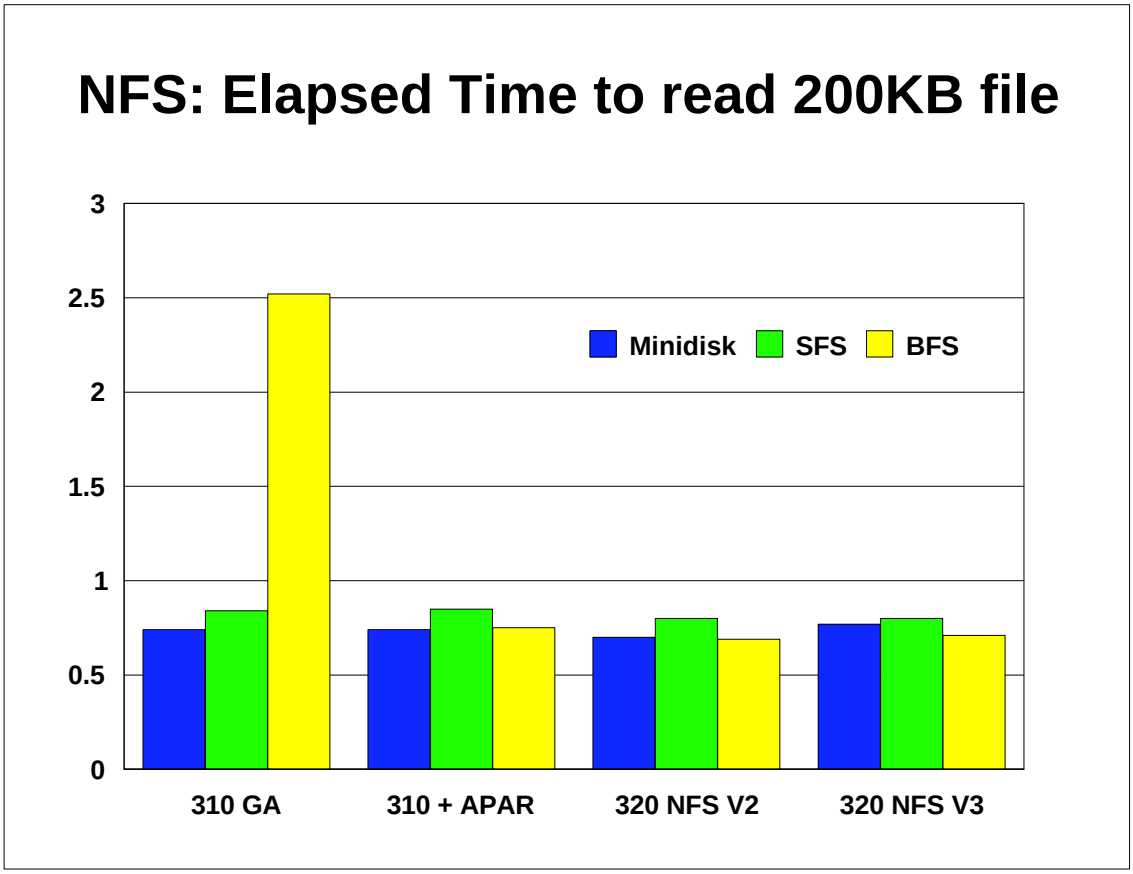
- TCP/IP FL 310:
  - APAR PQ16183 (helps BFS only)
    - improves reading large files
- TCP/IP FL 320:
  - NFS Version 3 Protocol
    - larger block sizes helps large file processing
    - READDIRPLUS helps directory displays
  - allow TCP connections
  - Improvements to BFS interface

NFS Performance continues to be an area where we want to improve. APAR PQ16183 to FL 310 improved read performance for BFS files significantly. This was motivated by network station manager work.
In the new release of NFS, we add NFS version 3 protocol which should improve some scenarios. Also additions in the VM interface for BFS additionally improves things.

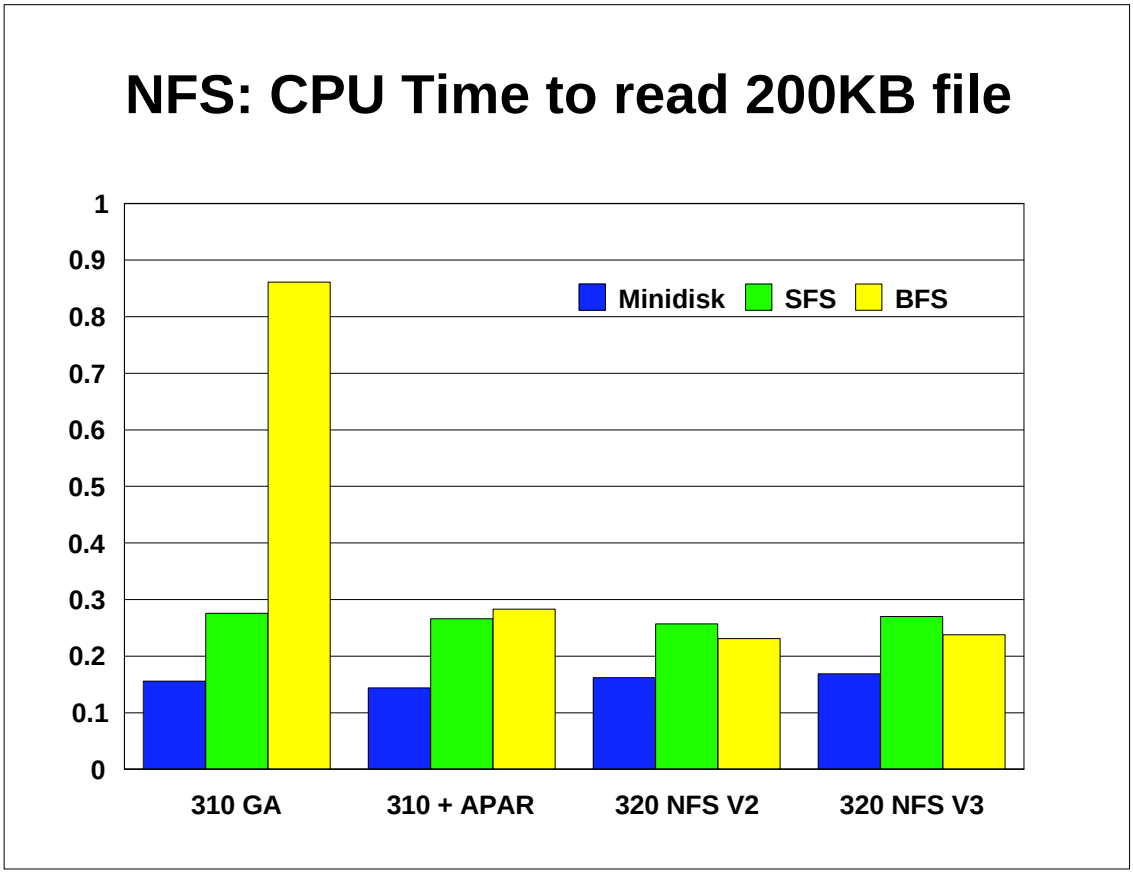
# NFS- BFS: Processor Time Breakdown



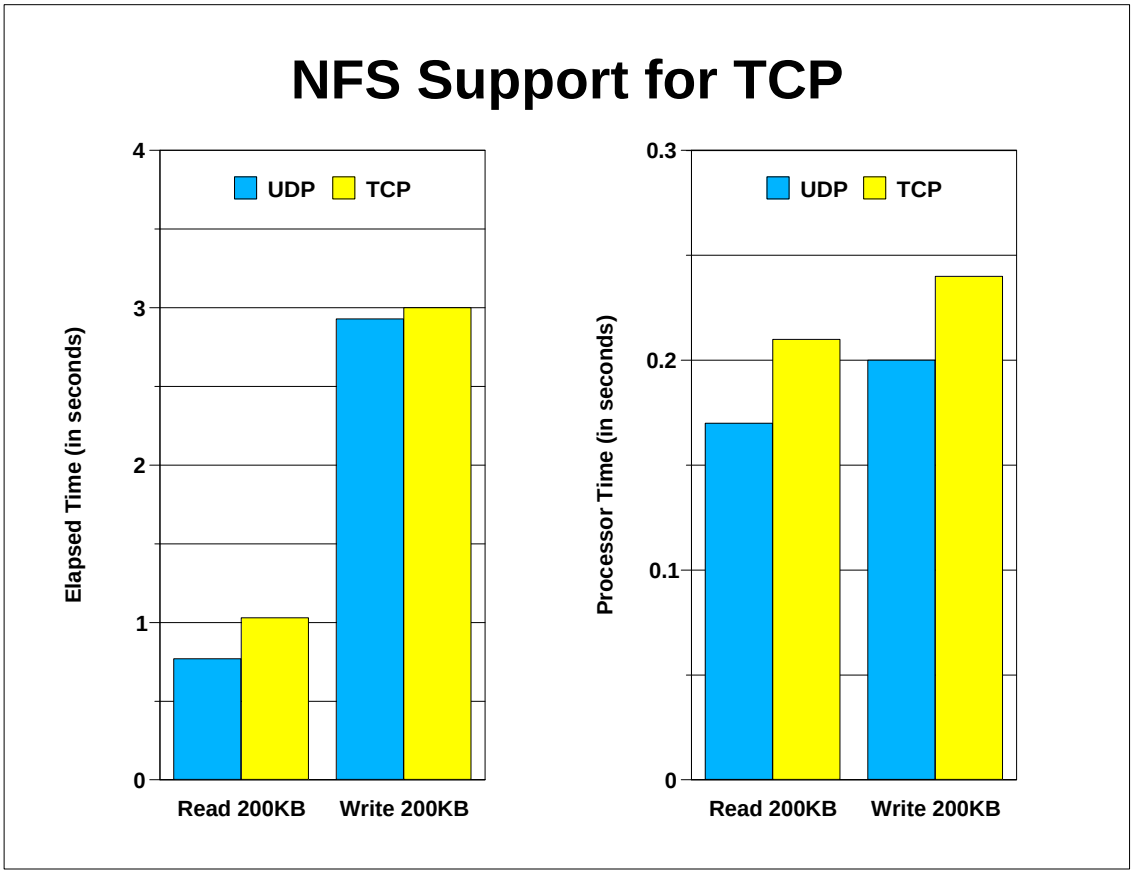

This chart shows the breakdown in processor time to read a 200 KB file from the BFS on VM through NFS. You can see that the TCP/IP portion stays about the same. The bulk of the improvements were in the interface between the NFS machine and the BFS server. The use of NFS version 3 protocol does not improve things in this particular scenario.




This chart shows the elapsed time to read a 200 KB file for various data types (minidisk, SFS, and BFS). Our goal was to bring BFS performance to the level of minidisk. As you can see, we accomplished this.




This chart is similar to the first NFS chart shown, except it only shows the total processor time (not breakdown by component) and includes the three VM sources of data (minidisk, SFS, and BFS). You saw in the previous chart that improvements had been made to make elapsed time for the three environments fairly close. Here you see that there is still additional processor requirements for the file systems that involve a file server virtual machine. However, these processor requirements have been lowered significantly.




As mentioned earlier NFS added support for TCP connections. The two charts on this page show the difference with the NFS for VM/ESA 2.4.0 between UDP and TCP. The chart on the left shows elapsed time. For the read scenario, TCP does add to the elapsed time by about 34%. However, since write processing elapsed time is dictated greatly by the NFS protocol, the use of TCP does not increase the elapsed time for writing files as greatly.

The chart on the right shows the processor time for the same configurations. The processor time increased about 24% for the read case and about 20% for the write case. Unlike elapsed time which is dominated by network communication, you see similar read and write processing increases.

## JAVA

- VM/ESA JDK 1.1.4 level performance challenges:
  - slow execution (no compiler)
  - large cost for initialization
  - while multithreading, it is not multiprocessing
- JIT (Just-In-Time) compiler work in progress
  - Greatly speeds up execution
    - up to 2.5 times improvement compared to no JIT in portable BOB workload (1 thread only)
    - Kernel benchmarks:1 to >50 X faster
  - Does add a hit to initialization

The VM/ESA JDK 1.1.4 level has 3 main challenges. Execution is slow due to a lack of a compiler. There is also a significant cost in I/O and processor time for initialization. Thirdly, while it provides multithreading, it is not multiprocessing and therefore limited to one processor per application. The JIT compiler should greatly speed up execution. In the application like portable BOB workload an 2.5 times improvement was measured. The Microbench kernel benchmark showed individual functions were 1 to 57 times the speed of uncompiled, while the UCSD kernel benchmarks showed 1 to 59 times. So you can see it will be workload dependent.

## JAVA

- Java Initialization
  - Currently 5 seconds on a 19 MIPS/engine box
  - Improved in JDK 1.1.6 with new CMS 15.
- RAWT (Remote Abstract Windowing Toolkit)
  - Will be available with JDK 1.1.6
  - RAWT not recommended for performance sensitive applications.

Some improvements in CMS multitasking for VM/ESA 2.4.0 help improve Java initialization. The RAWT (Remote Abstract Windowing Toolkit) will be available with the JDK 1.1.6 as a requirement to help meet Java compliance. The RAWT is not recommended for performance sensitive applications. A few push buttons may be okay, but do not try to write a new video game in Java for VM with the RAWT.

# ADSM Version 3

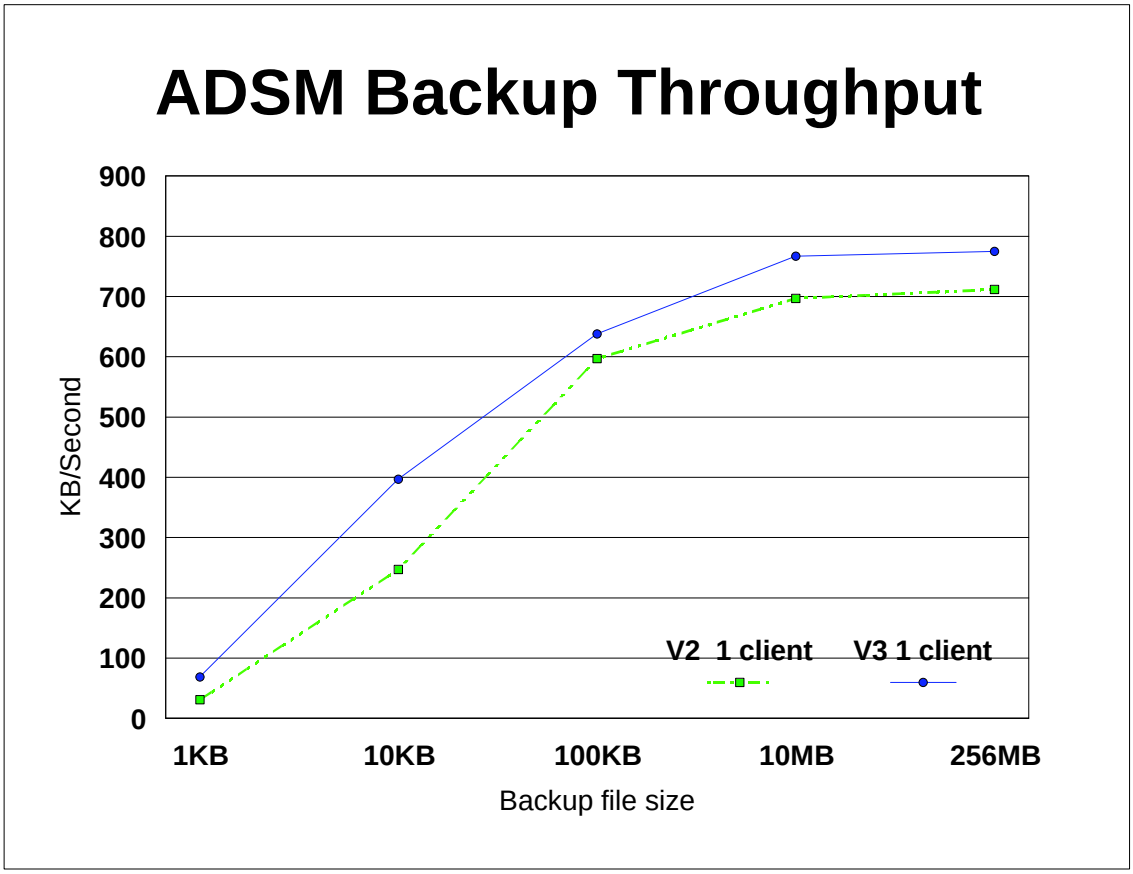
- Much better performance than version 2
- Backup throughput improvements
  - lower processor and DASD I/O requirements
  - smaller files saw larger improvement due server file aggregation item
  - measured throughput improvements of 9% to 126%
- Restore throughput showed little change
  - lower processor and DASD I/O requirements
  - restore throughput much lower than backup

ADSM Version 3 is not part of VM/ESA 2.4.0 but is worth discussing because of its improvements for performance. The performance of backup processing showed the most improvement, with measured throughput increases of 9 to 126%. Much of this was due to lower processor and DASD I/O requirements. The biggest improvements were for small files due to the server file aggregation item.
Restore throughput did not change as dramatically, but did see lower processor and I/O requirements. Restore throughput continues to be lower than backup throughput.

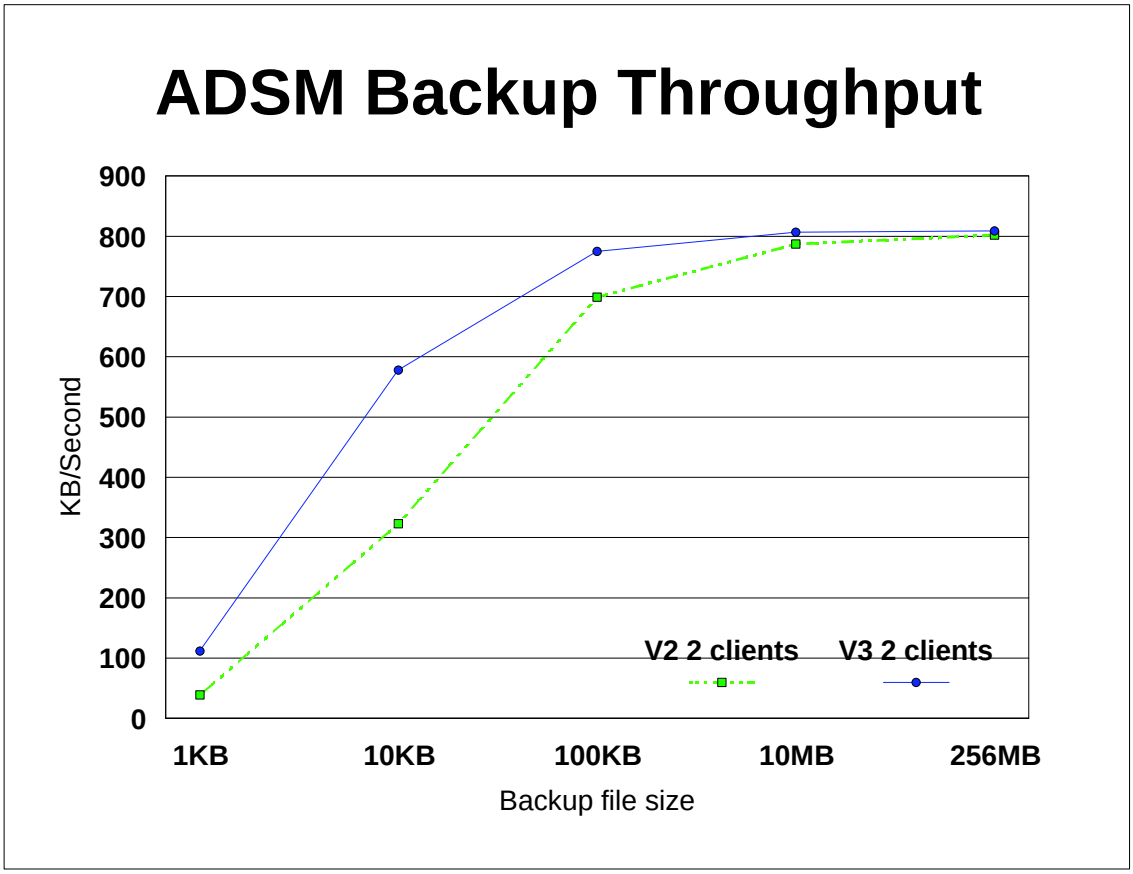
## ADSM Measurement Config

- Server:
  - VM/ESA 2.3.0 on 9121-480
- Clients:
  - AIX 4.1.4 on RS/6000 model 250
  - ADSM clients: version matched server
- Connection:
  - 16 Mbit IBM Token Ring
  - VM connected via 3172-3
  - TCP/IP FL310

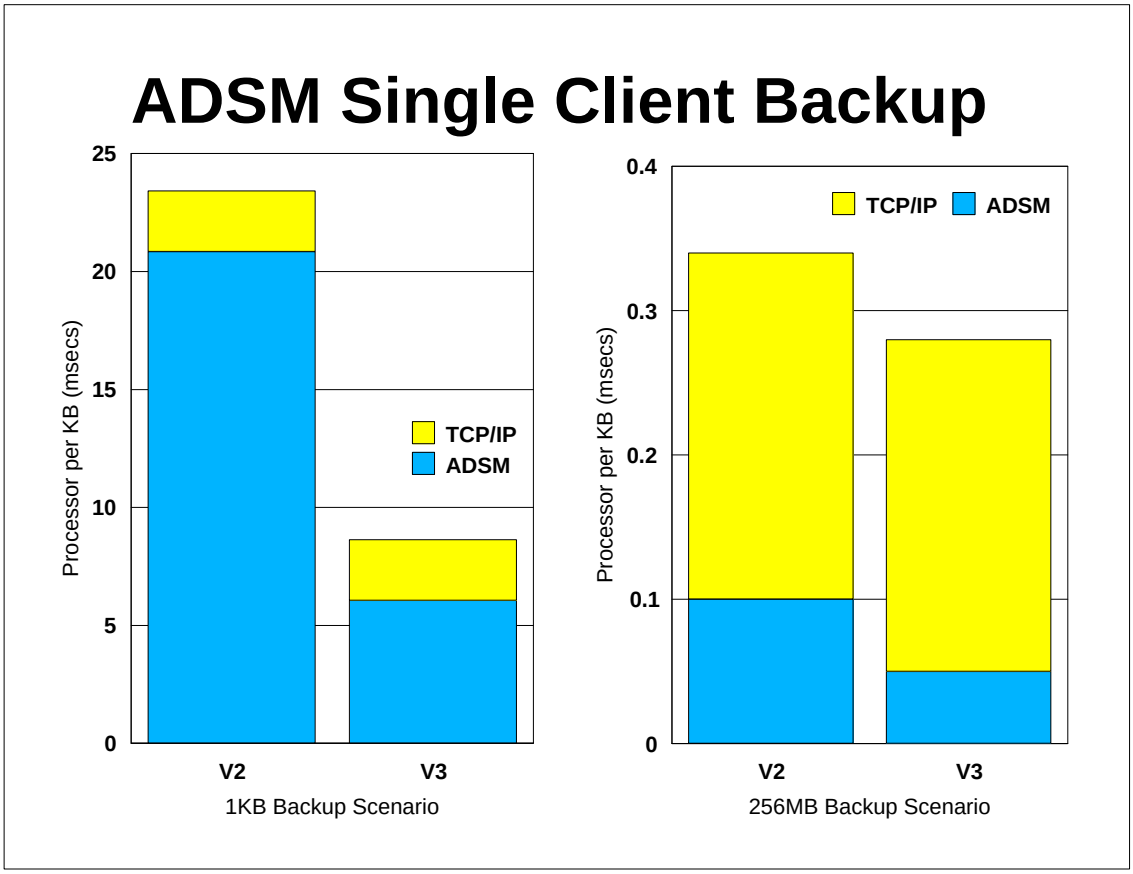
Our measurement configuration included VM/ESA 2.3.0 on a 9121-480 (2-way processor) connected to the network with TCP/IP FL310 to a 16 Mbit IBM Token Ring through the host attached 3172-3.
ADSM clients were run on RS/6000 model 250s with AIX 4.1.4. The ADSM clients were the same version as the server version.




This graph illustrates that the largest improvements in backup throughput were for smaller files. However, the greatest aggregate throughput is for large files. The measurements were made with only original backups, and not subsequent backups. In all the measurements shown here, only a single client was involved.

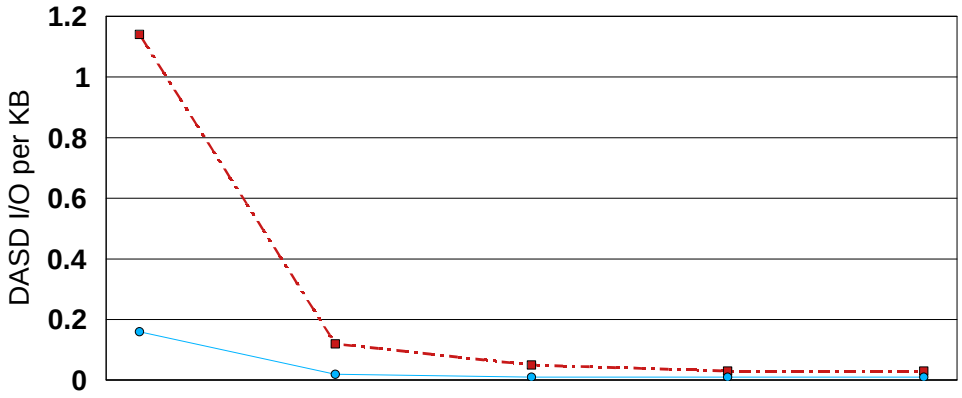



This chart shows backup throughput also, but with 2 clients being used. Remember that in all cases version 3 clients were used. Again you see that ADSM version 3 is the better performer. Limitations on our network and processor speed help lead to the leveling off of performance.



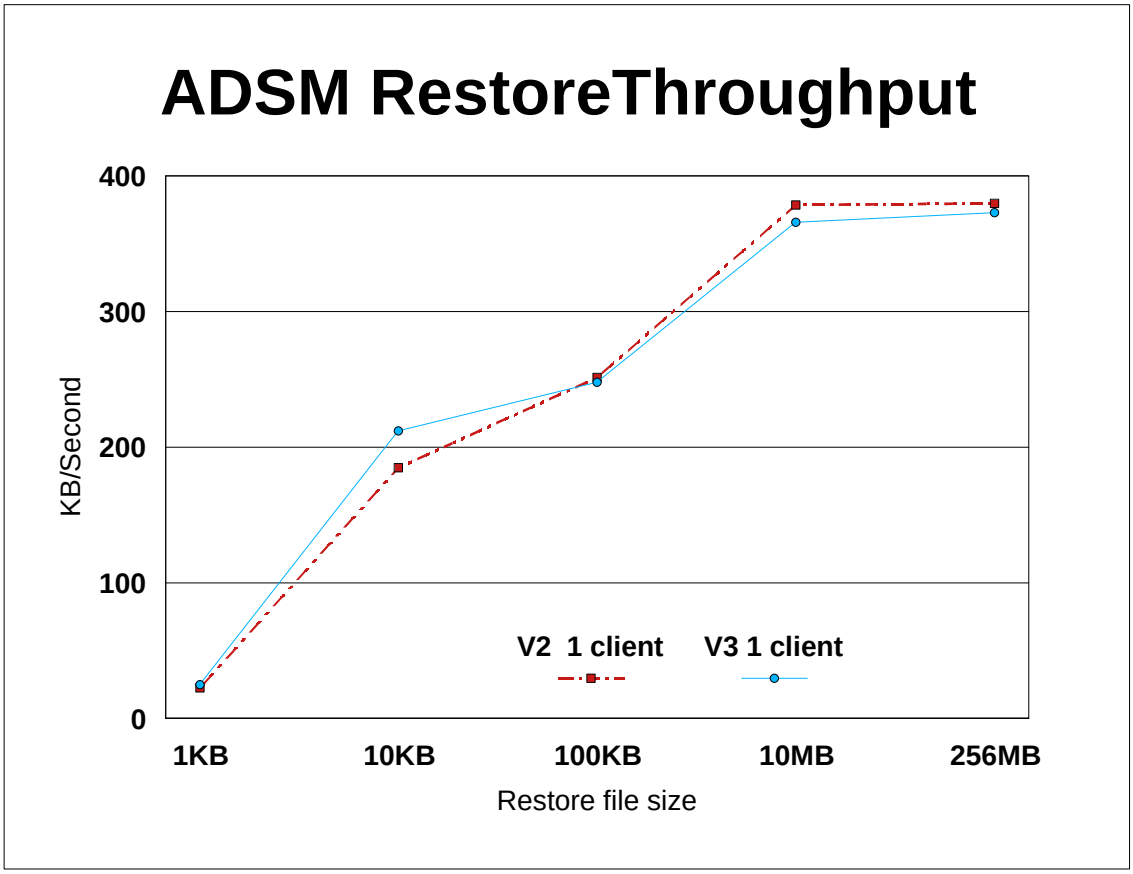

These two charts show the breakdown of processor time per KB moved between the ADSM and the TCP/IP stack machines. The chart on the left is with smaller files and shows that ADSM processor time still is the bulk of the processing even though it was decreased significantly. There was no change in the TCP/IP time. The chart on the right is with 256 MB file backup. You see the cost per KB is much lower than the 1 KB file case. While the ADSM time is lower in version 3, it is not as dramatic as the 1 KB case. Also, we see TCP/IP uses the majority of the processing time. If RFC 1323 were exploited and the new TCP header prediction in the FL 320 stack, these numbers might improve.

# Single Client Backups DASD I/O




| | 1KB | 10KB | 100KB | 10MB | 256MB |
|---|---|---|---|---|---|
| V2 | 1.14 | 0.12 | 0.05 | 0.03 | 0.03 |
| V3 | 0.16 | 0.02 | 0.01 | 0.01 | 0.01 |

The chart and table on this page show how the huge decrease in DASD I/O helps improve performance. Again, you will see that smaller file scenarios drop the I/O rate more, but even the 256 MB file case showed a drop of 67%.




ADSM Restore throughput did not change significantly in version 3 for any file size. These measurements were made with a single client. If you compare back to the throughput graphs for backup, you will note that restore throughput is much lower.

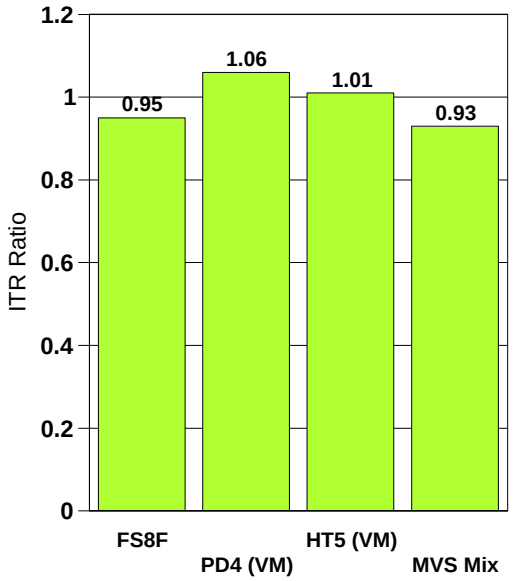
# Dirmaint APAR VM62262

- Performance Improvement APAR
- Avoids need to reopen/reclose files
- Range of change for key indicators:
  - Elapsed time 0 to -46%
  - Virtual I/O 0 to -97%
  - Virtual processor time 1 to -24%

APAR VM6262 was opened based on a customer recommendation. There were some scenarios where dirmaint used PIPE FILERAND in such a way that files were constantly opened, closed, reopened and reclosed. New logic is used to keep a file open for multiple operations. This reduces virtual disk read operations significantly, which results in lower elapsed times especially where MDC is not involved. There can also be some savings in processor usage, but those changes are less significant.

# Previous 9672 Sizing Advice

- 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.



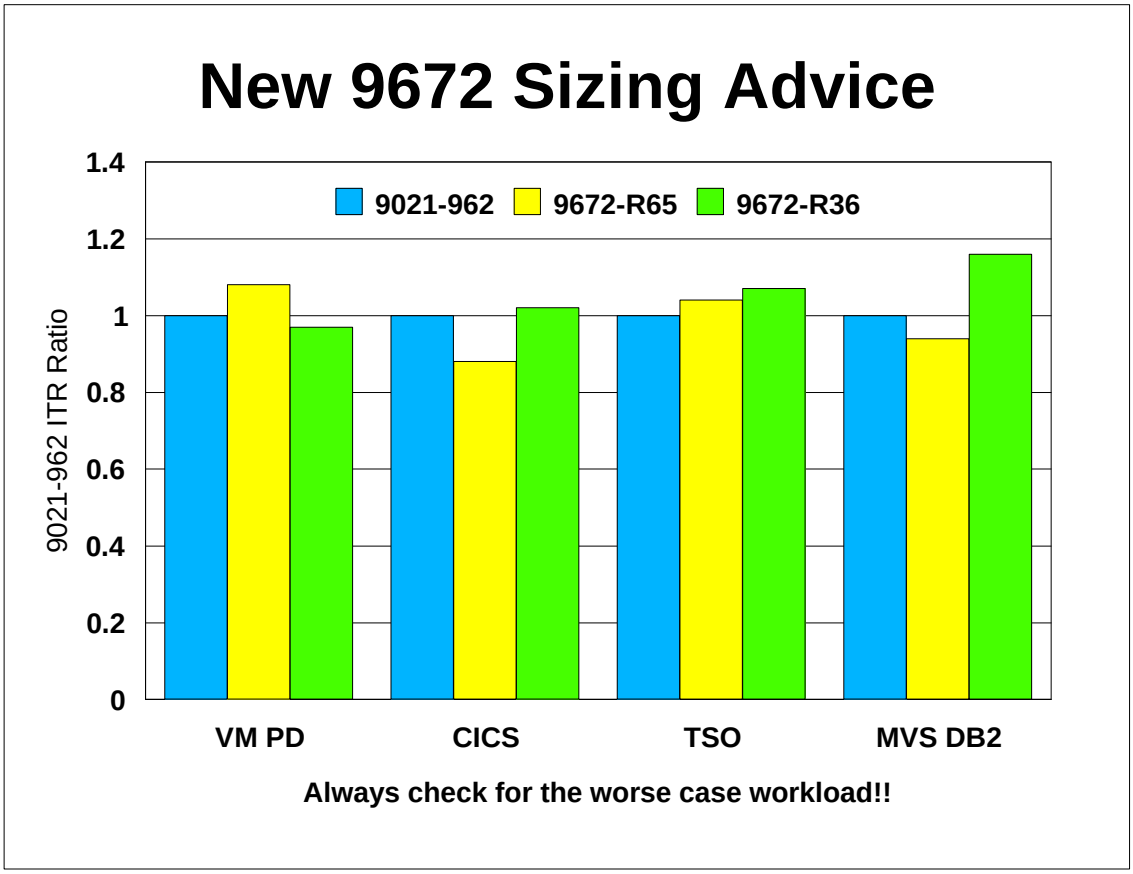

ITR Ratio for 9121-742 to a 9672-R53

If you saw one of my 1998 VM/ESA performance updates, this chart may look familiar. We had done some measurements on a 9672-R53 and 9121-742 with our own FS8F CMS interactive workload and compared it to the LSPR VM and MVS Mix workloads. As you see, the FS8F workload ran more like the MVS mix workload than the PD4 or HT5 workloads of LSPR. I then cautioned people to consider MVS LSPR ITR ratios when doing migration sizing from bipolar to CMOS machines in order to be aware of the worse-case scenario. As we will see on the next couple of charts, that advice needs to be changed slightly.

## G5 Performance Improvements

- First 1 BIPS machine
  - BIPS = billion instructions per second
  - BIP = Baffling Indicator of Performance
- Improvements
  - decrease memory access costs
  - improved processor caching
- Much better performance for workloads with poor locality of reference and/or very short transactions

The G5 family of the IBM 9672 processors broke a new barrier. We now have the first 1 BIPS machine. This was a great accomplishment. The performance improvement came not only from faster chips, but from enhancements to the memory access process and processor caching. This results in much better performance for workloads with poor locality of reference and/or very short transactions. However, some VM workloads have great locality of reference and medium or long transactions.




This chart shows the relative ITR (internal throughput rates) for 3 different processors with 4 different workloads. The processors are a bipolar 9021, a 9672-R65 6-way from the G4 family, and a 9672-R36 3-way from the G5 family. The workloads are all from the LSPR measurement suites. In comparison to the 9021, the VM PD workload shows the G4 doing better and the G5 worse. This is the opposite of what the CICS and MVS DB2 workloads show. The TSO is different from other three workloads. This shows the problem of depending on a single MIPS (or BIPS) number to do processor sizing. Knowing the range of possible performance is important for worse case risk management.

## Summary

- VM/ESA Development team continues to keep an eye on performance
- Full VM/ESA 2.4.0 Performance Report
  http://www.ibm.com/s390/vm/perf/docs/
- Wider scope than traditional "regression" CMS performance
  - scheduler changes
  - hardware support
  - network performance
- For news, keep checking:
  http://www.ibm.com/s390/vm/perf/

As you can see, we have been busy in the world of VM performance, and we plan to stay busy. Some of things we have worked on recently have not been the typical "regression" performance we have done in the past.
To stay in touch with VM performance check out the URL listed above. There are several pages off of performance home page which cover various tips and FAQ, documentation, and performance products.
As always it has been a joy to work with the many customers, coworkers, and vendors in the VM community. Thanks for your support.