z/VM Guest Performance

WAVV 2002 - Fort Mitchell, Kentucky

Bill Bitner
VM Performance
IBM Endicott
607-752-6022
bitnerb@us.ibm.com
Last Updated: April 9, 2002

Abstract
Is VM good for Guest Performance? Bill Bitner of IBM's VM Performance Evaluation department answers that question with "It depends!". This presentation will look at what it depends on, what is meant by performance, and factors that include V=R/F/V guests, enhanced minidisk cache, virtual disk in storage, CCW translation, and hardware assists.

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, VSE/ESA, CICS, z/VM

Credits
I'd also like to acknowledge some of the people behind this presentation: Dr. Wolfgang Kraemer, Greg Kudamik, Wes Ernsberger, Frank Brice, Steve Wilkins, Bill Stephens, Bill Guzior, my former manager Doug Morrison, and countless others. In particular, a lot of this material is based on Dr. Kraemer's VSE/ESA Performance Considerations package. To get copies of the VSE/ESA Performance Papers, IBMers can request the VE13PERF and VE21PERF packages from IBMVSE tools disk. These papers are also available through the VSE Home Page.

Overview

- Is VM Good for Guest Performance?
  - It depends.
- What does it depend on?
  - On what you mean by "performance"
  - Using V=R/F
  - Running VM on Native or LPAR
  - Virtual disk in storage
  - Enhanced minidisk cache
  - Hardware
- Presentation originally for VSE Guests

The overhead of running VSE on VM is a topic that has been discussed for years. While it is tempting to give this presentation a marketing spin, I will try to resist that temptation. I believe each customer needs to answer the question of running VSE with VM for themselves. In the process they will have to answer other questions, such as "What is performance?". In this presentation, we will attempt to cover the various trade offs that need to be made in coming to a conclusion to this question. Also, remember there are nonperformance factors in answering this question.

The Short Answer is Yes

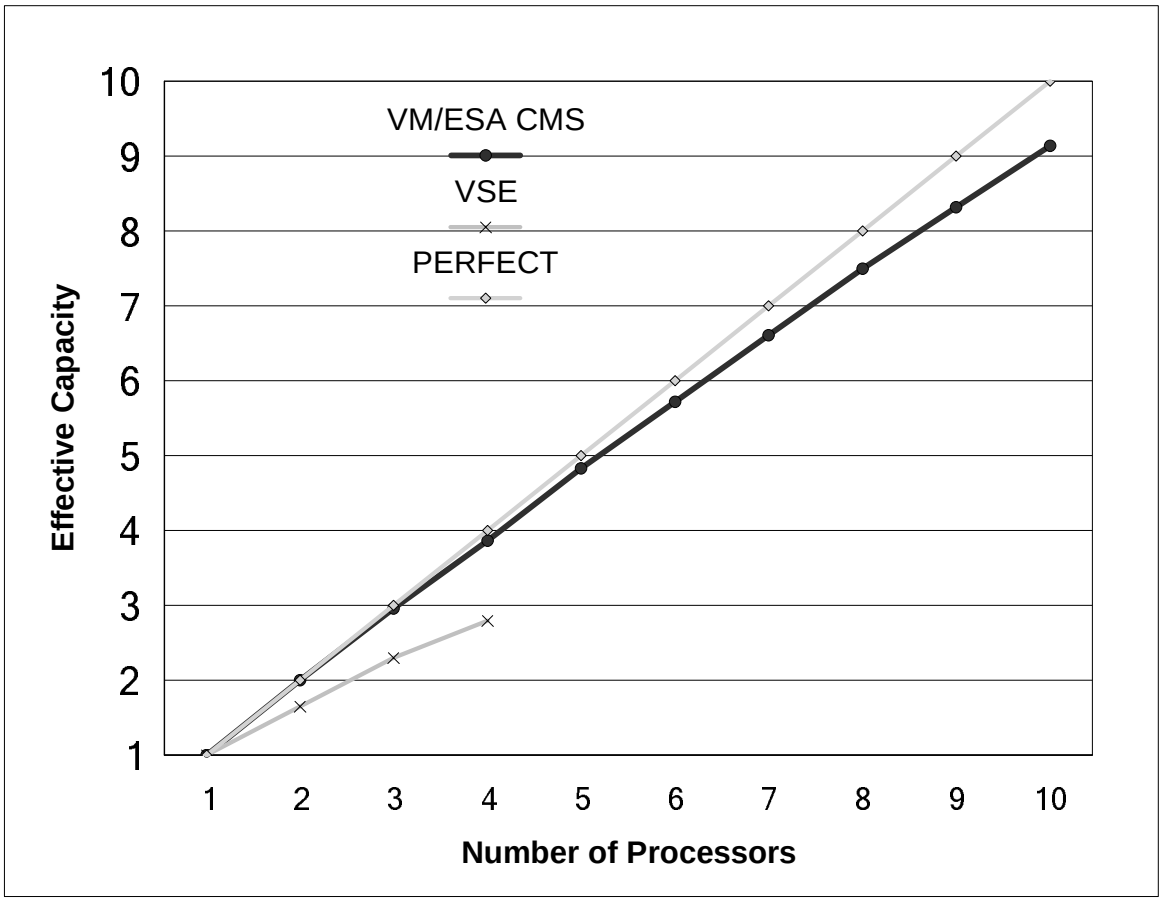
- Performance Value-Add of VM
  - Extend capacity of single Guest, by running multiple Guests
    - N-way considerations
    - Consolidate unused capacity
  - VM/ESA extensions to scheduling such as limit shares
  - Resource sharing
    - Real storage is shared for V=V guests
    - Channels are shared without EMIF
    - DASD devices can be split up into minidisks

The short answer to "Is VM good for VSE performance?" is YES. Listed are some areas where VM/ESA adds value to VSE systems from a performance perspective. VM/ESA has traditionally extended the capacity of a single VSE by providing the ability to run multiple VSE images. While the Turbo Dispatcher provides multiprocessing support, many customers still need VM/ESA to fully exploit the large N-ways in the IBM processor line. VM/ESA's MP support in conjunction with the extensive scheduling features make it very powerful. VM/ESA allows for efficient sharing of storage for V=V guests and the virtualization of many resources.

The Short Answer is Yes...

- VM paging benefits - demand and block paging, use of expanded storage.
- HW exploitation - greater N-way, expanded storage
- VM features
  - Virtual disk in storage for lockfile
  - Enhanced minidisk cache
  - DB2 for VM Guest Sharing for VSE

VM/ESA paging provides the best of both worlds. High-speed demand paging with expanded storage and block paging to the slower DASD. VM/ESA also allows for several performance features. These include virtual disk in storage, enhanced minidisk cache, and VM data spaces. The latter is exploited by DB2 Server for VSE and VM (SQL/DS) for great performance improvements.



This graph shows potential MP factors when the N in N-way increases. In a perfect world, when you added another processor, you would get another full processors worth of work. However, due to MP effects in both hardware and software, that does not happen. The perfect line above shows 10 times the capactiy with a 10-way as a 1-way. One of the VM LSPR workloads tracks fairly close to that line, but still loses about 10% at the 10-way level. VSE cannot fully exploit large n-ways at this time due to larger software MP effects and workload environments.

Nonperformance Reasons

- CMS features
- Isolate production and test
- Problem Determination
- Migration vehicle
- Resource Management
- Accounting
- Other PPs (e.g. OV/VM)

Besides performance, there are other reasons to use VM/ESA with VSE. CMS is a great application development and test platform. The virtual machine model of VM allows for easy migration, test, and isolation. The two systems compliment each other very well with their products and applications. VM offers OV/VM and ADSM. Many customers are also using VM as a network hub which then connects the various guests by virtual channel to channels.

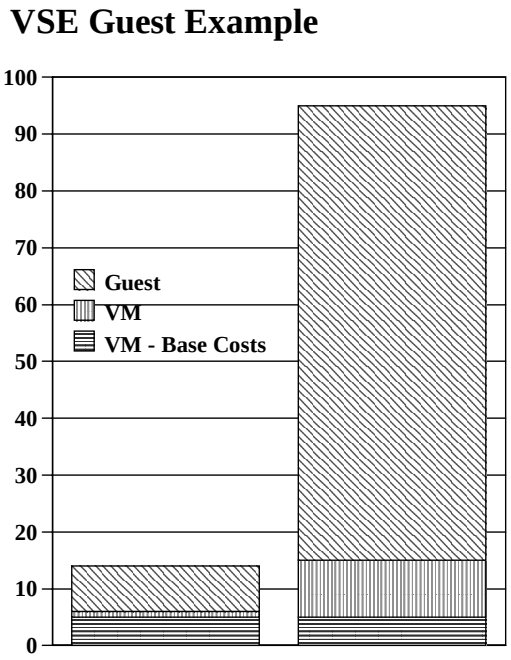
What do you mean by "Performance"?

- Critical to answering the original question.
- Typically one of the following:
  - ITR = Internal Throughput Rate = a measure of work per CPU second.
  - ETR = External Throughput Rate = a measure of work per wallclock second.
  - CPU Utilization = how busy processor is; tied to ITR.
  - Response Time (Elapsed Time) = how long jobs take; tied to ETR
  - Interactive Users vs. Batch Work
  - How many phone calls you get

It is critical to be clear about the meaning of "Performance". When I hear people criticize the performance of VM/VSE, I am amazed that they consider CPU utilization to be the only performance indicator. In general there are two schools of thought, one that looks at CPU utilization and one that looks at response time. The internal throughput rate, or ITR, is a measure of commands per CPU second. Another way of thinking of this is how many commands could be completed if the processor was running at 100%. ITRs can be used to compare processor performance. When done properly, there should be an implied response time limit as well. External transaction rate (ETR) is a measure of commands complete per wallclock time. You will probably also need to determine the priority of batch versus interactive users.

CPU Usage by VM/ESA

- Base costs and background work
  - Scheduling and dispatching
  - Accounting
  - Monitor
- Costs proportional to Guest requests or requirments of VM



One mistake people make is by trying to determine the overhead of running VSE with VM with a trivial test case. There are some base costs to running VM/ESA such as infrastructure, scheduling, dispatching, accounting, and monitor. Many of these are a constant cost. That is, the CPU they require stays the same no matter how busy the system. So if you run a trivial VSE workload as a test, you'll see VM/ESA being a larger percentage of the total CPU usage.

CPU Usage - SIE

- Used by z/VM to run a guest
- Exits from SIE indicate work for VM
- Hardware assists can help avoid SIE exits
- Most common reasons for exiting SIE
  - I/O processing
  - Page fault resolution
  - Instruction simulation
  - Minor time slice expires
  - Loaded wait state

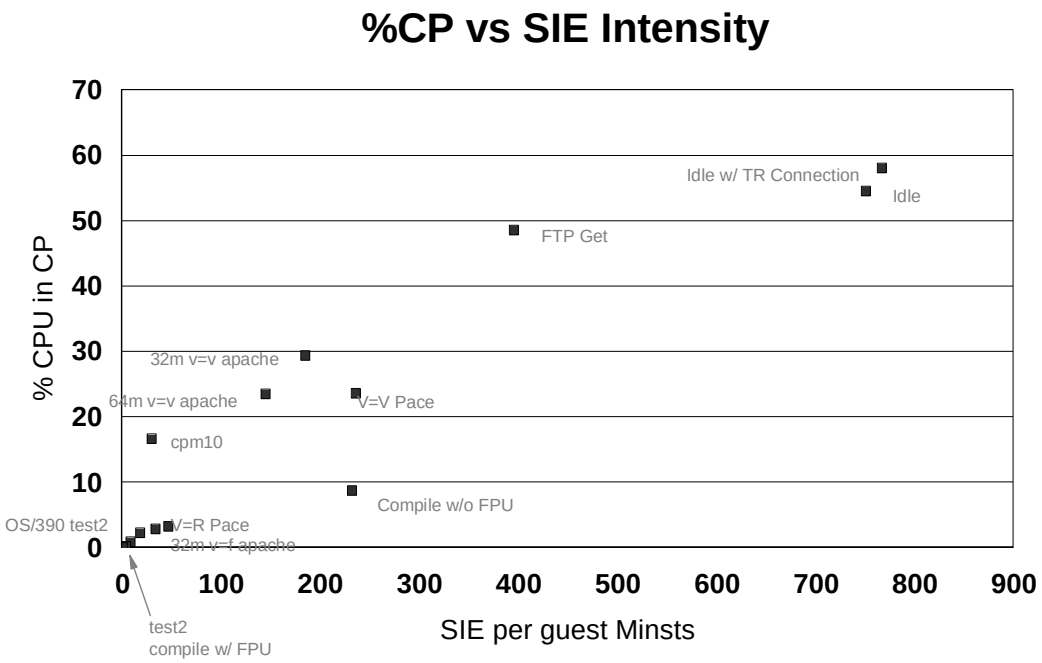
VM/ESA uses the SIE (start interpreted execution) instruction to run virtual processors. The overhead and function processing costs in the VM control program are tied to how often we exit from SIE (via intercept or interrupt). This is a case where the hardware assists play a significant role. Listed are the four most common reasons for exiting SIE. I/O Processing tends to be the most significant. VM/ESA gets involved with all V=V I/O and some V=R/F I/O. VM/ESA also gets involved for page fault processing for V=V guests. SIE is also exited for certain instruction simulation such as unassisted SIGPs and IUCV. VM/ESA will also get control when the minor time slice expires.
The cost of entry and exit from SIE is also processor dependent as some machines have implemented the SIE instruction more efficiently.

Exits from SIE

- Data in memory techniques avoid I/O.
- I/O Assist avoids SIE exit to handle:
  - I/O interrupt processing
  - CCW translation from virtual to real addresses
- CCW translation bypass for V=R guest.
- Minor time slice: SET SRM DSPSLICE
- Avoid Paging
  - V=R/F
  - Reserved pages for V=V
  - Sufficient storage

Minimizing I/O in the guest, by using larger buffers or data-in-memory techniques, can lower VM/ESA overhead. Page fault overhead can be minimized by adding storage or reserving pages as appropriate. The scheduling overhead can be adjusted with the SET SRM DSPSLICE command. However, caution should be used when adjusting the minor time slice. While increasing it may lower the VM/ESA overhead, it also lowers the ability of VM/ESA scheduling to adjust to system changes. We have seen scenarios where the ITR improves, but ETR gets worse when increasing the minor time slice. Note also that dedicated virtual machines get a 500 millisecond dispatch time slice.

VM Overhead Cloud Chart



VM I/O Processing

- I/O Assist
  - V=R/F Guests
  - Dedication Devices
- CCW Translation Bypass
  - Dedicated and full pack minidisks
  - Only some V=R I/O
  - SET NOTRANS ON (SET CCWTRAN OFF)
- Fast CCW translation
  - Applies only to select DASD I/O

I/O Assist (I/O Passthru or SIE Assist) avoids the need for CP involvement for I/O processing. Only dedicated devices of V=R or V=F guests are eligible for I/O passthru. Refer to "VM/ESA Performance, SC24-5642-01" Chapter 3, under "I/O Interpretation Facilities". "CCW translation bypass" only applies to V=R guests for dedicated devices (and limited cases for full-pack minidisks). It can be controlled by the SET CCWTRAN command (which is OFF by default for V=R machine). If you SET CCWTRAN ON for a V=R machine, then the benefit of I/O Assist is lost. In the past, CCW translation bypass was often referred to as I/O fast path. "Fast CCW translation" is related to a feature where CP uses a more efficient path in the translation of virtual to real addresses associated with CCWs. Fast CCW translation applies to a subset of DASD I/O.

I/O Considerations

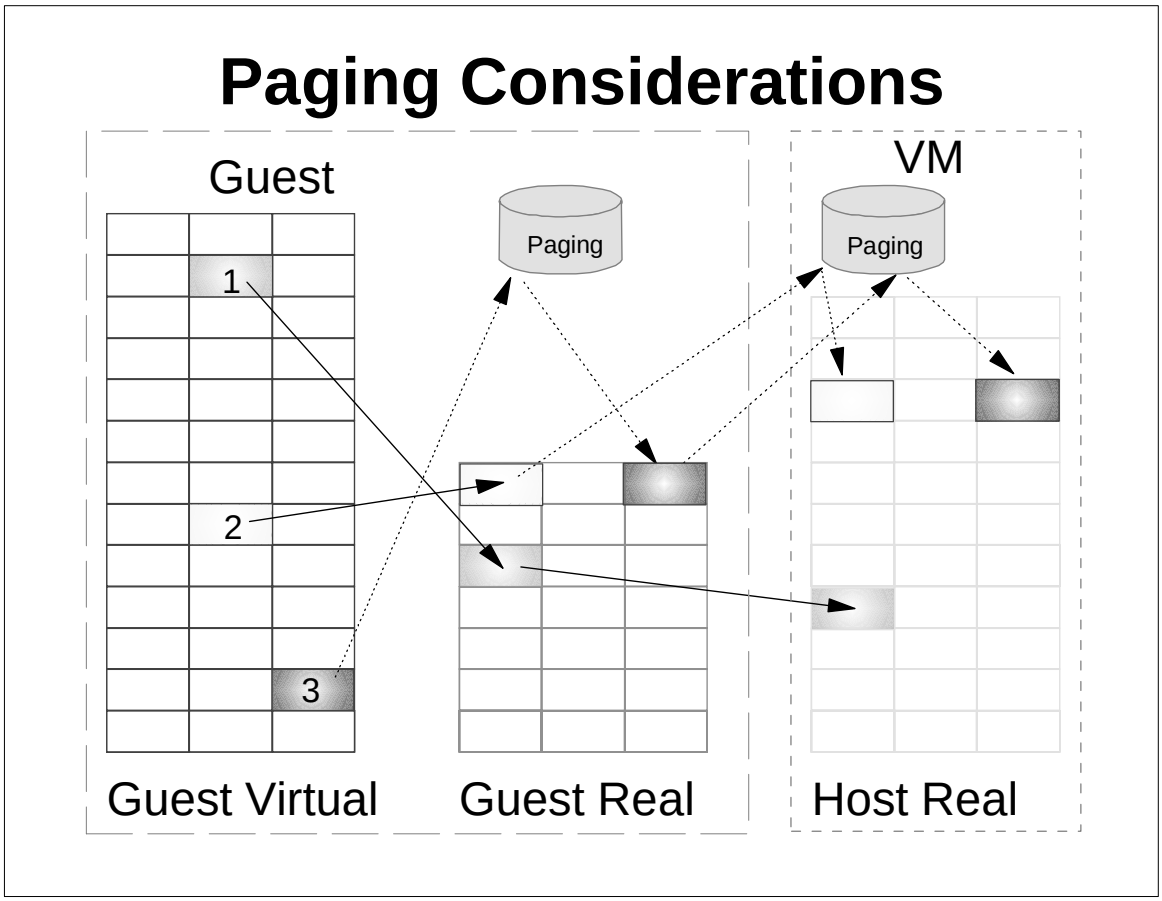
- I/O Assist gives best CPU performance
- Dedicated I/O is not eligible for MDC
- For V=R CCWTRANS OFF makes guest I/O ineligible for fast CCW translation
- For VSE Guests, VSE vdisks are more efficient than VM vdisks.
- Both VM vdisks and MDC require sufficient storage
- MDC read performance is as good as VM vdisk performance

From an ITR view, I/O Assist gives the best performance since it avoids VM processing. However, devices eligible for I/O Assist are not eligible for minidisk cache which may hurt ETR or response time. Remember that the VM overhead results from exiting SIE for processing. Therefore, features in VSE that can be used to avoid I/O that VM sees can be very helpful. VSE virtual disk in storage is a good example. Improving I/O performance often comes at a cost in some other resource. Both virtual disk in storage and minidisk cache require sufficient storage to provide good performance. With VM/ESA you can exploit expanded storage if available. If you are only looking to improve read I/O, then minidisk cache is generally the better feature.

DASD Considerations

- Dedicated Devices
  - Required for full I/O Assist
  - Not eligible for MDC
- Full pack minidisks
  - Can be shared between guests
  - Some I/O assist
  - Define via VOLSER or DEVNO
  - DEVNO not eligible for MDC
- Partial pack minidisks can be shared
- FBA Volumes - should start/end on 64 512-byte block boundaries.

The various types of DASD can influence performance. To get the most from I/O Assist, dedicated devices are best, followed by full pack minidisks. Partial pack minidisks are very flexible, but ineligible for I/O Assist. Remember that dedicated devices are not eligible for minidisk cache. There are also a couple of special guidelines for using minidisk cache with FBA devices dealing with the boundaries.



- It can be confusing to discuss paging of guests that support virtual storage when there is confusion over terms and without pictures. This picture shows storage as it relates to the VM control program point of view. Physical storage or central storage, is labeled host real in the picture above. Guest real is the storage that VSE believes is real even though it is virtual to VM. Guest virtual would be virtual storage from VSE's view point. An VSE application referencing data or instructions might be in one of the three numbered pages in Guest Virtual storage. In Page 1, the page happens to be in guest real, and that guest real page also is resident in Host Real. Therefore, no paging is required at all. For Guest Virtual Page 2, we see it is in Guest Real, but not host real. Therefore, a page fault would occur which VM would need to process. In the case of Guest Virtual Page 3, we see it is not in Guest Real storage. This would require a pagefault at the guest level, which VSE would need to process. If the page selected in Guest Real is not in Host Real, this would result in a page fault at the Host level as well. This is called double paging.

Paging Considerations

- For V=V guests the potential exists for "Double Paging"
- No VM paging for V=R/F
- For VSE guests, the closer the VSE VSIZE is to the defined storage for the virtual machines, the lower the VSE paging.
- PAGEX ON use where appropriate.
- VM can use expanded storage for high speed paging device.

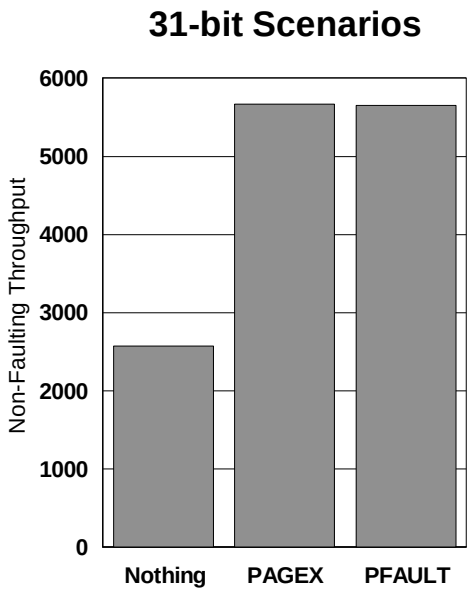
Most people recognize that VM/ESA paging is more advanced than VSE/ESA. One should try to avoid scenarios of double paging. This would happen if a VSE page is not in the VSE space and is paged in to a VSE real address that VM/ESA in turn needs to page in. VM paging is avoided completely for V=R or V=F machines. The closer the VSE VSIZE to the size of the VSE virtual machine, the less VSE system should have to page. If there is no need for paging in VSE, consider the NOPDS option. VSE will use the PAGEX ON option where appropriate. If the configuration includes expanded storage for paging, then let VM/ESA do the paging to here.

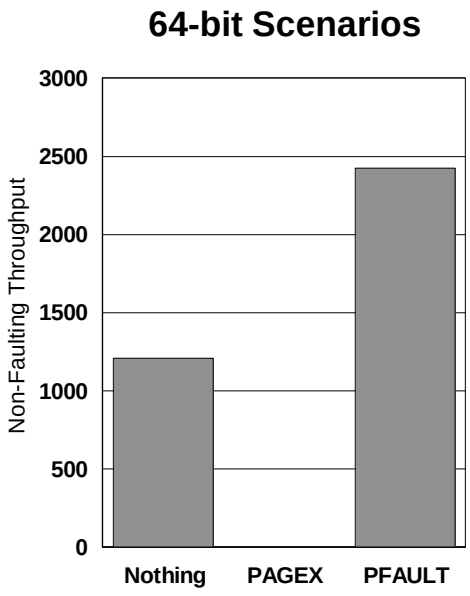
Asynchronous Page Fault Faciltiy

- Ordinarily, page faults serialize the virtual machine. This can be a throughput and response time problem for guest systems
- Enhancements designed for Linux
- PFAULT macro
  - Accepts 64-bit inputs
  - Provides 64-bit PSW masks
- Diagnose x'258'
- Older PAGEX interface limited to 31-bit
- z/VM 4.2.0

- Support was added in z/VM 4.2.0 (and as an APAR to z/VM 4.1.0) for a new or extended asynchronous page fault facility. This has advantages over PAGEX in that it supports 64-bit parameters and also has more flexible interrupt masking.
- The Diagnose x'258' was previously an internal diagnose and has been externalized.
- Changes were also made in Linux for zSeries to exploit this facility.

Page Fault Tests





- Here you see a series of comparisons between the various approaches of handling page faults. The tower marked "nothing" is the case where page faults are handled synchronously. The PAGEX tower is with Linux using the PAGEX handshaking. Since PAGEX is not supported in 64-bit, the tower is missing on the right chart. PFAULT is the new enhanced asynchronous page fault processing.
- These charts show how much additional work can be done by other processes in Linux that are not taking page faults.

V=R/F/V Considerations

- V=R/F potential I/O assist benefit (saves CPU)
- V=F avoids overhead of recovering V=R
- 1 V=R + 5 V=F or 6 V=F
- V=V avoids dedicating storage
- V=R defaults to dedicating processors
- Running z/VM in an LPAR -
  - No V=F, only V=R, but without I/O Assist
  - Often better to use V=V and reserve pages

V=R performance can be lower than V=F performance. Extra processing is required for the recoverability part of V=R support. Preferred guests (V=R/F) on a native VM/ESA avoid VM paging and provide savings with hardware assisted I/O. The total number of preferred guests is still six, even though LPAR can provide more partitions on some processors. If you are running VM/ESA in an LPAR, you need to realize that it changes the characteristics. Both LPAR and VM/ESA use SIE. On older processors (3090E and older), the assists were not available to run SIE on top of SIE. Running this configuration would be very costly. All the current processors have the required interpreted SIE assist for running VM/ESA on LPAR. However, with VM/ESA on an LPAR, only the V=R machine is possible and there is no I/O Assist. In this scenario, you may be better off running the guest as a V=V machine with CP reserved pages.

Virtual MP Support

- Define additional processors dynamically
  - Directory include MACHINE ESA 2
  - CP DEFINE CPU vcpu_addr
- Or put everything in the directory
  - CPU 00 NODEDICATE
  - CPU 01 NODEDICATE
- Detaching vCPU resets virtual machine
- For testing: more virtual than real processors

There are two approaches to creating a virtual MP machine. You can define the virtual processors in the directory so they are available when the virtual machine logs on. Or you can set up the directory so that you can use the DEFINE CPU command to add virtual processors dynamically. Note that detaching a virtual processor resets the virtual machine. Do not define extra virtual processors unless you are going to use them. Defined, but unused, virtual processors can cause performance problems.

Virtual MP Support

- CP commands of interest
  - QUERY VIRTUAL CPUS
  - CPU vcpu_addr cmd_line
  - DEDICATE and UNDEDICATE

This is a list of CP commands that can be useful when using virtual MP machines. The QUERY VIRTUAL CPUS command shows you how may virtual processors you have and their addresses. When setting traces or issuing other commands that affect a virtual processor, you will want to use the CPU command to direct the command at a particular processor or to all virtual processors with the ALL option. Output from CP commands is prefixed with the virtual processor address. The DEDICATE and UNDEDICATE commands can be used to control the dedication of real processors to virtual machines, which can be helpful in virtual UP environments as well.

Virtual MP Tuning

- Share setting is for virtual machine, divided amongst all virtual processors
- Processors can be dedicated
- Mixing dedicated and shared processors is not recommended
- Defined but inactive vCPU (stopped state) makes guest ineligible for I/O assist
- Monitor, INDICATE, RTM for all vCPUs
- Potential for >100% (N*100%)
- Dedicated processor looks 100% busy

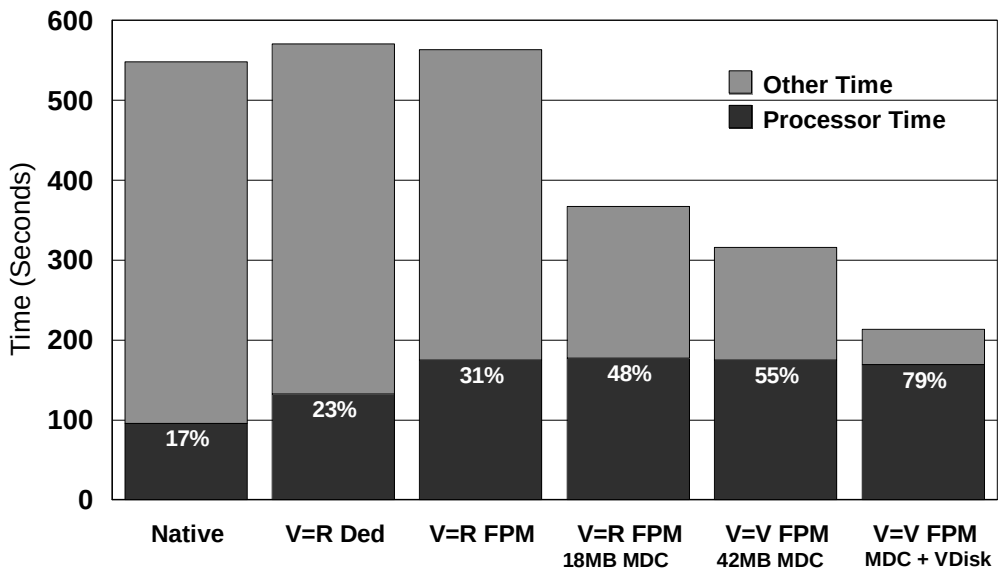
From a tuning perspective, it is important to note that the share value is distributed across the virtual machine. For example, if you have a virtual 4-way and a default share of relative 100, then each virtual processor would be scheduled as if it had a relative 25 share value. A virtual processor in CP stopped state makes it ineligible for I/O Assist. Virtual processors can be dedicated to real processors. I do not recommend mixed environments where a single virtual machine has both dedicated and undedicated virtual processors. This can result in performance anomalies that will be difficult to detect and explain.

VM/ESA Data in Memory Techniques

- VM Data Spaces
  - Exploited by DB2 Server for VSE and VM (SQL/DS)
- VM Virtual disk in storage
  - volatile FBA minidisk
  - private or shareable
  - perfect for lock file
- Minidisk cache
  - Undedicated 3380, 3390, 9345, and RAMAC
  - SSCH, SIO, SIOF and Diagnose I/O
  - Read-once data generally does not benefit
  - New (2.3.0) Record level MDC does not apply to VSE
  - Do not use MDC for VSE lockfile

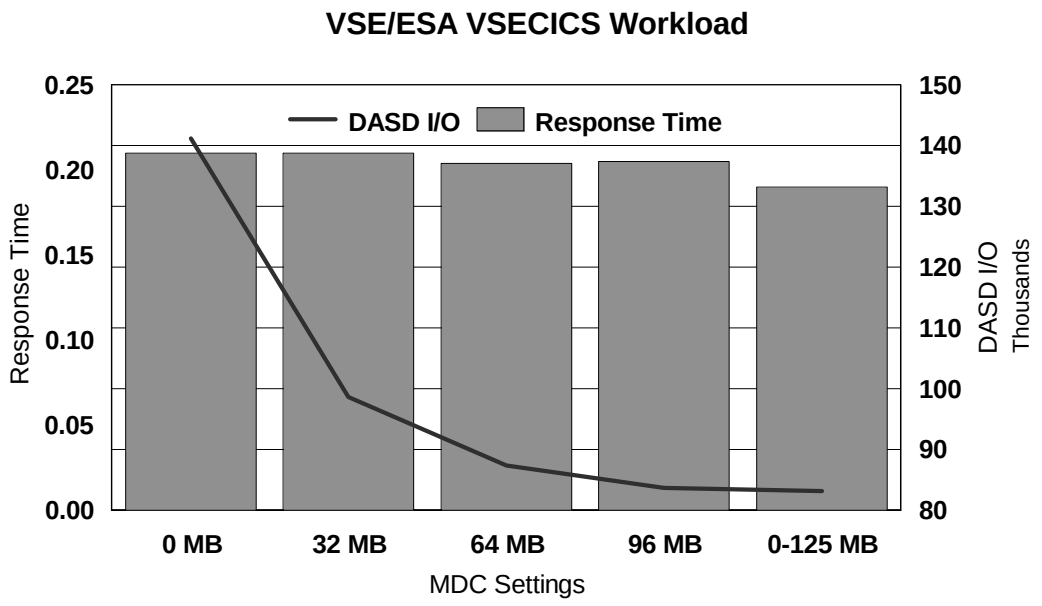
On this foil, we will briefly describe three data-in-memory techniques used by VM/ESA. They are VM data spaces, virtual disks in storage, and minidisk cache. Many products exploit VM data space, of particular interest to VSE customers is DB/2 Server for VM and VSE (SQL/DS). The VM virtual disk in storage feature allows for volatile FBA minidisks that can be defined as shareable or private. This are backed by a VM system utility space. It is ideal for the VSE lock file. Minidisk cache was enhanced in VM/ESA 1.2.2 to be more flexible, allow more types of data, and more types of I/O. The enhancements included a series of CP commands to enable/disable the cache, flush the cache; the ability to use real storage as the cache; the eligibility of almost any type of data; and eligibility of SSCH, SIO, and SIOF I/Os. The minidisk cache is track oriented. One should not suppose that MDC will benefit read-once data, particularly if the reading application has been highly tuned.

Minidisk Cache Benefits Example 1



- The graphs on this foil and the following one show exploitation of minidisk cache by a VSE guest with two different workloads. They bring to light two important facts. First, it is ***possible*** to get better throughput with a guest running under VM/ESA than running native, by trading off processor and storage resources, but it is not ***guaranteed***. Second, it is workload dependent based on I/O characteristics. For the I/O intensive PACEX workload, the new MDC can be very beneficial, especially when combined with virtual disk in storage. Note that when VM/ESA gets more involved, processor utilization can increase. This is a trade-off for better elapsed time.

Minidisk Cache Benefits Example 2



- The VSE CICS workload did not show as significant an improvement as PACEX. The CICS workload is less I/O intensive and has a lower read to write ratio. While there was a significant reduction in the number of DASD I/Os, average response time reduction was not as large. Also, note the diminishing returns as more storage is provided for minidisk cache. From the CICS graph, you can also see the law of diminishing returns. As more and more storage is given to MDC, the rate of improvement lessens.

Linux Guest Guidelines

- Why does my idle Linux consume Processor resources?
  - Timer pops
- How big should my Linux guest be?
  - Not bigger than you need
- Where should Linux swap?
  - Multiple choices: XPRAM, Mdisk, Tdisk, Vdisk
- Should I set QUICKDSP ON for my Linux Guest?
  - Production vs. Test vs Development machines
- See the following URL for other information: http://www.vm.ibm.com/perf/tips/linuxper.html

- See the URL listed at the bottom for more details on these items. There can be a lot of discussion on each of these items and many tend to be related to one another.
- An idle Linux machine tends to never look idle to VM because of the various house keeping tasks that occur: timer pops, network polling, etc.
- Linux tends to use all the storage you give it, so do not give it too much.
- If you are not going to do any significant swapping, then Vdisks are very convenient. However, if you are going to do significant swapping then use minidisks or tdisks.
- QUICKDSP ON for production guests to avoid unwanted stays in the eligible list.

Summary

- Many features to be exploited
- Optimum configuration will depend on
  - What you mean by the term **performance**
  - What resources you have available
- http://www.vm.ibm.com/perf/tips/
  - Common problems/solutions
  - CCW translation matrix
  - VSE Guest Performance
  - Performance related APARs
  - MDC guidelines
  - N-way and CMOS thoughts

I hope this presentation helped generate questions as to how and where VM can be used to help with your guest performance. I'm sure not all your questions were answered here. There is a great deal of information available in manuals, listservers, IBMLINK, and the VM or VSE home pages. Check it out.

I welcome your comments and suggestions on this presentation.