



VM Performance Introduction

WAVV 2007

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

VP HA

© 2007 IBM Corporation



Disclaimer

Legal Stuff

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.

Permission is hereby granted to SHARE to publish an exact copy of this paper in the SHARE proceedings. IBM retains the title to the copyright in this paper, as well as the copyright in all underlying works. IBM retains the right to make derivative works and to republish and distribute this paper to whomever it chooses in any way it chooses.

Trademarks

The following are trademarks of the IBM Corporation:

IBM, VM/ESA, z/VM

LINUX is a registered trademark of Linus Torvalds

Overview

- Performance definition
- Guidelines
- Native CP commands
- Other performance tools
- I/O performance concepts
- Case study
- Final thoughts

Definition of Performance

Performance definitions:

- Response time
- Batch elapsed time
- Throughput
- Utilization
- Users supported
- Phone ringing
- Consistency
- All of the above

Performance Guidelines

- Processor
- Storage
- Paging
- Minidisk cache
- Server machines

Processor Guidelines

- Dedicated processors - mostly political
 - ▶ A virtual machine should have all dedicated or all shared processors
- Share settings
 - ▶ Use absolute if you can judge percent of resources required
 - ▶ Use relative if difficult to judge and if lower share as system load increases is acceptable
 - ▶ Do not use LIMITHARD settings unnecessarily
- Small minor time slice keeps CP reactive.
- Do not define more virtual processors than are needed.

Storage Guidelines

- Virtual:Real ratio should be $\leq 3:1$ or make sure you have robust paging system
- Use SET RESERVE instead of LOCK to keep users pages in storage
- Define some processor storage as expanded storage to provide paging hierarchy (even when running a 64-bit CP) (and even when running z/VM 5.2.0)
 - ▶ For more background, see -
<http://www.vm.ibm.com/perf/tips/storconf.html>
- Exploit shared memory where appropriate

Paging Guidelines

- DASD paging allocations less than or equal to 50%.
- Watch blocks read per paging request (keep >10)
- Multiple volumes and multiple paths
- Do not mix with other data types
- In a RAID environment, enable cache to mitigate write penalty.

Minidisk Cache Guidelines

- Configure some real storage for MDC.
- In general, enable MDC for everything.
- Disable MDC for
 - ▶ Minidisks mapped to VM data spaces
 - ▶ write-mostly or read-once disks (logs, accounting)
 - ▶ Backup applications
- In large storage environments, may need to bias against MDC.
- Prior to z/VM 5.2.0, consider disabling Xstore MDC if constrained below 2GB
- Better performer than vdisks for read I/Os

SVM Guidelines

- SVM = Server Virtual Machine
 - ▶ TCP/IP, RACFVM, etc.
- QUICKDSP ON to avoid eligible list
- Higher SHARE setting
- SET RESERVED to avoid paging
- NOMDCFS in directory option
- Ensure performance data includes these virtual machines

CP INDICATE Command

- **LOAD:** shows total system load.
 - ▶ (STORAGE value not very meaningful and was removed in z/VM 5.2.0)
- **USER EXP:** more useful than Indicate User
- **QUEUES EXP:** great for scheduler problems and quick state sampling
- **PAGING:** lists users in page wait.
- **IO:** lists users in I/O wait.
- **ACTIVE:** displays number of active users over given interval

CP INDICATE LOAD Example

INDICATE LOAD

```

AVGPROC-088% 03
XSTORE-000000/SEC MIGRATE-0000/SEC
MDC READS-000035/SEC WRITES-000001/SEC HIT RATIO-099%
STORAGE-017% PAGING-0023/SEC STEAL-000%
Q0-00007(00000)                                DORMANT-00410
Q1-00000(00000)                                E1-00000(00000)
Q2-00001(00000) EXPAN-002 E2-00000(00000)
Q3-00013(00000) EXPAN-002 E3-00000(00000)

PROC 0000-087%                                PROC 0001-088%
PROC 0002-089%
LIMITED-00000
  
```

CP INDICATE QUEUE Example

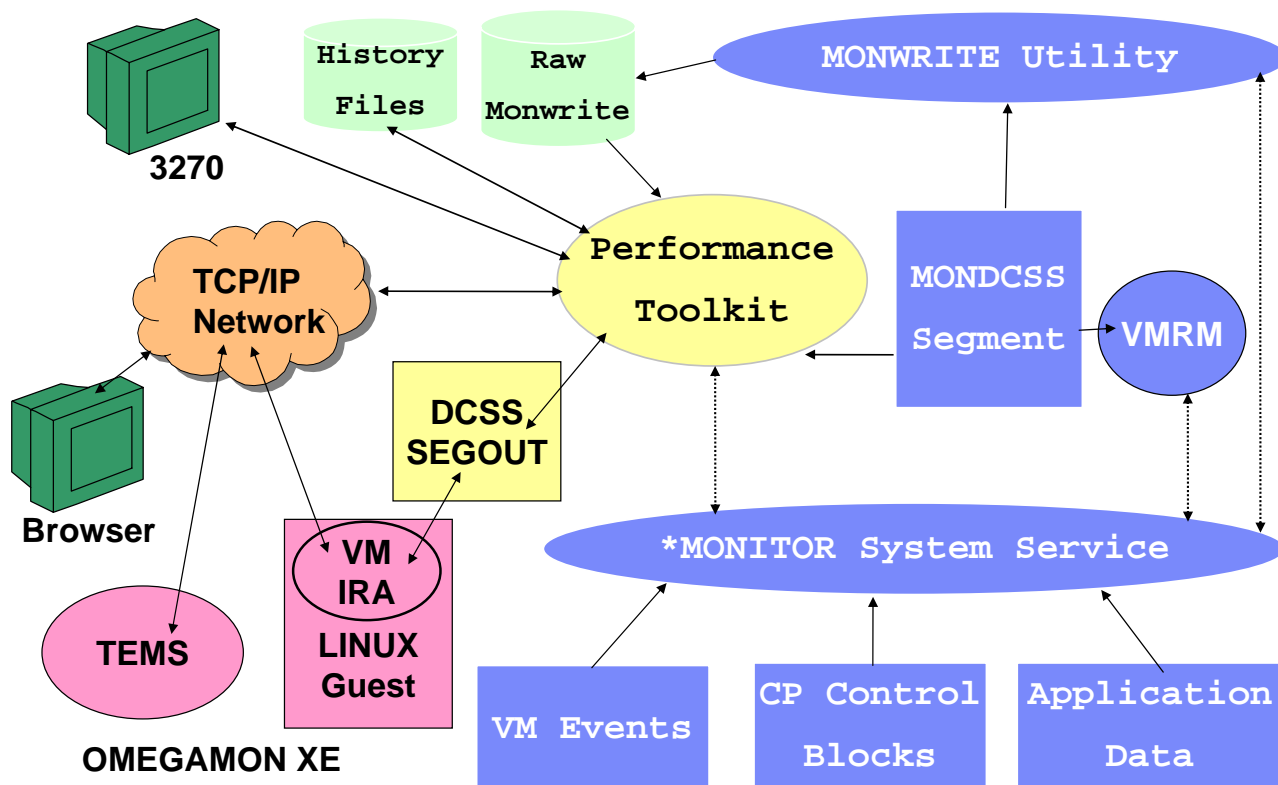
INDICATE QUEUE EXP

EDLLIB14	Q3	IO	00002473/00002654	..D.	-.0217	A00
KAZDAKC	Q3	IO	00003964/00003572	-.0190	A02
BITNER	Q1	R00	00001073/00001054	.I..	-.0163	A01
LCRAMER	Q3	IO	00003122/000028500259	A00
DSSERV	L0	R	00007290/000072893229	A00
RSCS	Q0	PS	00001638/00001616	.I..	99999	A00
SICIGANO	Q3	PS	00000662/00000662	.I..	99999	A00
VMLINUX1	Q3	PS	00018063/00018063	99999	A02
LNXREGR	Q3	PS	00073326/00073210	99999	A02
VMLINUX	Q3	PS	00031672/00031672	99999	A01
TCPIP	Q0	PS	00018863/00018397	.I..	99999	A02
EDLLNX2	Q3	PS	00032497/00032497	99999	A01
EDLLNX1	Q3	PS	00015939/00015939	99999	A02

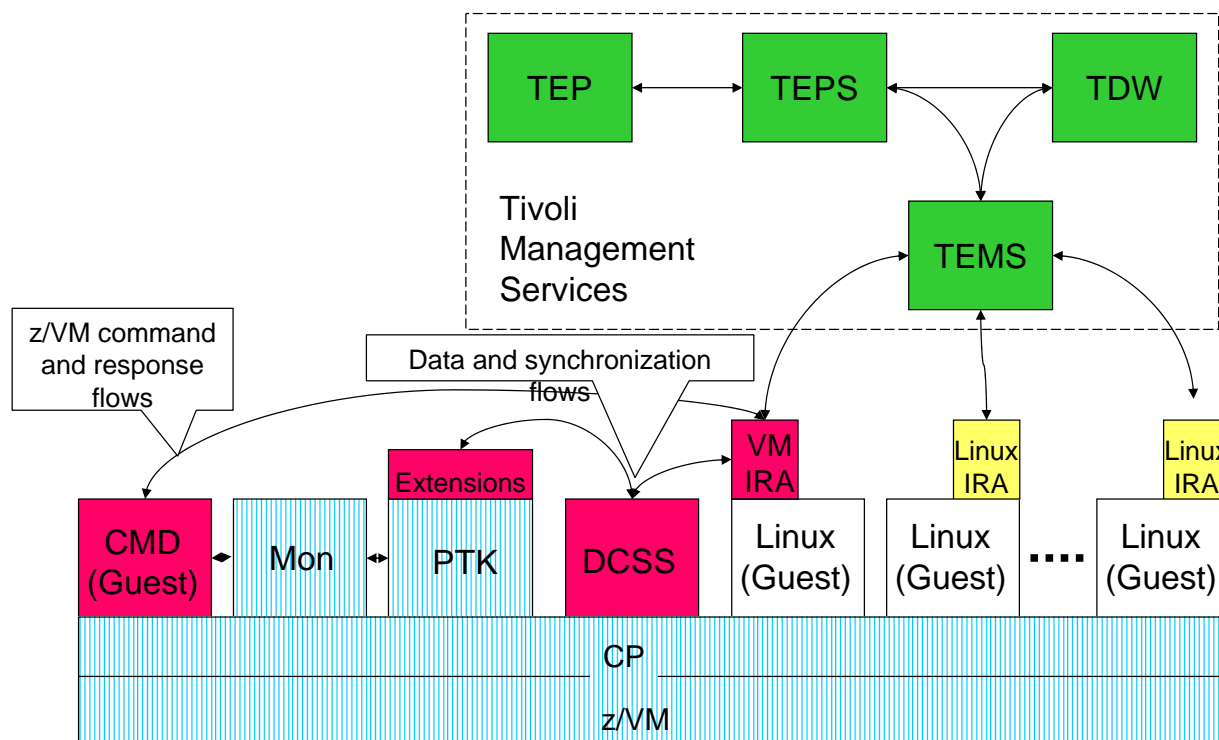
Selected CP QUERY Commands

- Users: number and type of users on system
- SRM: scheduler/dispatcher settings
- SHARE: type and intensity of system share
- FRAMES: real storage allocation
- PATHS: physical paths to device and status
- ALLOC MAP: DASD allocation
- XSTORE: assignment of expanded storage
- MONITOR: current monitor settings
- MDC: MDC usage
- VDISK: virtual disk in storage usage
- SXSPAGES: System Execution Space (5.2.0)

5,000 Foot View



OMEGAMON XE Basic Architecture



State Sampling

- Find the state of given user or device
 - ▶ Consolidation of samples gives useful info
- Low frequency:
 - ▶ INDICATE QUEUES
- High frequency:
 - ▶ Monitor: user, processor, and I/O domains
 - ▶ CP MONITOR SAMPLE RATE

I/O Response Time

$$\text{Resp Time} = \text{Service Time} + \text{Queue Time}$$

$$\text{Service Time} = \text{Pending} + \text{Connect} + \text{Disconnect}$$

- Queue Time: from hi-frequency sampling of queue in RDEV. Reported in monitor.
- Function Pending: time accumulated when a path to device cannot be obtained.
 - ▶ < 1 ms, unless contention at channels or control units.
- Connect: time device logically connected to channel path
 - ▶ proportional to amount of data per I/O

I/O Response Time (*continued*)

- Disconnect: time accumulated when device is logically disconnected from channel while subchannel system is active.
 - ▶ Cache miss
 - ▶ Seek on older devices
 - ▶ CU management
- Device Active: time accumulated between return of channel-end and device-end
 - ▶ Often reported as part of Disconnect Time

Definitions

- WSS = working set size
 - ▶ Comp-Sci Definition: Set of pages a workload needs to run effectively
 - ▶ VM Definition: Estimated working set size based primarily on resident page count
- Transaction
 - ▶ Comp-Sci Definition: End user interaction
 - ▶ VM Definition: transaction ends when scheduler detects end of processing

Other Sources

- Performance Manual - Part of z/VM Library
 - ▶ SC24-5999-03 z/VM 4.4.0
 - ▶ SC24-6109-01 z/VM 5.2.0
- <http://www.vm.ibm.com/perf/>
 - ▶ links to documents, tools, reference material
- <http://www.vm.ibm.com/perf/tips/>
 - ▶ common problems and solutions
 - ▶ guidelines
- <http://www.vm.ibm.com/devpages/bitner/>
 - ▶ presentations with speaker notes

A Case Study

The Grinch That Stole Performance

From Performance Toolkit USTAT FCX114 Report January 5:

```

                                <-SVM and->
%CPU %LDG %PGW %IOW %SIM %TIW %CFW %TI %EL %DM %IOA
   0   0   0   19   2   10   0   3   0  51   8
  
```

From Performance Toolkit DEVICE FCX108 Report January 5:

```

    <-Rate/s-> <----- Time (msec) -----> Req. <Pct>
Addr  I/O Avoid Pend Disc Conn Serv Resp CUWt Qued Busy
1742 26.7  .0  1.3 18.4  4.7 24.5 69.0  .0  1.2 65.4
  
```

Went to check Toolkit CACHEXT FCX177 Report for control unit cache stats, but it didn't exist!

It is a good thing I keep historical data -- let's go back and see what's going on...

When Did We Last See Cache?

From Performance Toolkit DEVICE FCX108 Report:

```

    <-Rate/s-> <----- Time (msec) -----> Req. <Pct>
Addr  I/O Avoid Pend Disc Conn Serv Resp CUWt Qued Busy
Dec8  41.0  .0  0.3 0.2  2.0  2.6  2.9  .0  .0  10.5
Jan5  26.7  .0  1.3 18.4  4.7 24.5 69.0  .0  1.2 65.4
  
```

From Performance Toolkit CACHEXT FCX177 Dec. 8th Report:

```

<----- Rate/s -----> <-----Percent----->
Total Total  Read  Read Write      <----- Hits ----->
Cache SCMBK N-Seq  Seq    FW Read Tot RdHt Wrt DFW CFW
53.0   41.0  52.3   0    0.6  99  99  99  96  96  ..
  
```

Down for the 3-Count

q dasd details 1742

1742 CUTYPE = 3990-EC, DEVTYPE = 3390-06,
VOLSER=USE001

```

CACHE DETAILS:  CACHE NVS CFW DFW PINNED CONCOPY
                  -SUBSYSTEM  F    Y  Y  -    Y    N
                  -DEVICE      Y   -  -  Y    N    N
DEVICE DETAILS: CCA = 02, DDC = 02
DUPLICATE DETAILS: SIMPLEX

```

Pinned data! Yikes! I had never seen that before!

Performance Toolkit Device Details

```

FCX110      CPU 2003      GDLVM7      Interval INITIAL. - 13:08:47      Remote Data

Detailed Analysis for Device 1742 ( SYSTEM )
Device type : 3390-2      Function pend.:      .8ms      Device busy      : 27%
VOLSER      : USE001      Disconnected  : 20.3ms      I/O contention:  0%
Nr. of LINKs: 404      Connected      : 5.4ms      Reserved        : 0%
Last SEEK   : 1726      Service time   : 26.5ms      SENSE SSCH     : ...
SSCH rate/s : 10.5      Response time  : 26.5ms      Recovery SSCH  : ...
Avoided/s   : ....      CU queue time  : .0ms      Throttle del/s: ...
Status: SHARABLE

Path(s) to device 1742:  0A      2A      4A
Channel path status  :  ON      ON      ON

Device      Overall CU-Cache Performance      Split
DIR ADDR VOLSER  IO/S %READ %RDHIT %WRHIT ICL/S BYP/S  IO/S %READ %RDHIT
08 1742 USE001   .0   0     0     0     .0   .0  'NORMAL' I/O only

```

Performance Toolkit Device Details

MDISK	Extent	Userid	Addr	Status	LINK	MDIO/s
101	- 200	EDLSFS	0310	WR	1	.0
201	- 500	EDLSFS	0300	WR	1	.0
501	- 600	EDLSFS	0420	WR	1	.0
601	- 1200	EDLSFS	0486	WR	1	.0
1206	- 1210	RAID	0199	owner		
		BRIANKT	0199	RR	5	.0
1226	- 1525	DATABASE	0465	owner		
		K007641	03A0	RR	3	.0
1526	- 1625	DATABASE	0269	owner		
		BASILEMM	0124	RR	25	.0
1626	- 1725	DATABASE	0475	owner		
		SUSANF7	0475	RR	1	.0
1726	- 2225	DATABASE	0233	owner	366	10.5

Solution

- Use **Q PINNED** CP command to check for what data is pinned.
- Discussion with Storage Management team.
- Moved data off string until corrected.

Pinned data is very rare, but when it happens it is serious.

Some Final Thoughts

- Collect data for a base line of good performance.
- Implement change management process.
- Make as few changes as possible at a time.
- Performance is often only as good as the weakest component.
- Relieving one bottleneck will reveal another. As attributes of one resource change, expect at least one other to change as well.
- Latent demand is real.