z/VM Performance Introduction
Session 9115

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

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
  - See also session 9166 Thursday 4:30
- 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
  - Absolute Share can be almost as effective
  - A virtual machine should have all dedicated or all shared processors
  - Gets wait state assist and 500ms minor time slice

- 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

- Do not define more virtual processors than are needed.
- Small minor time slice keeps CP reactive.
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 For more background, see http://www.vm.ibm.com/perf/tips/storconf.html
- Exploit shared memory where appropriate
- Size guests appropriately
  - Avoiding over provisioning
  - Do not put them in a high guest paging position
Paging Guidelines

- DASD paging allocations less than or equal to 50%.
  - QUERY ALLOC PAGE
- Watch blocks read per paging request (keep >10)
  - Monitor data (Performance Toolkit FCX103)
- Multiple volumes and multiple paths
- Do not mix Page extents with other extents on same volume.
- Paging to FCP SCSI may offer higher paging bandwidth with higher processor requirements
- 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.
- Set maximum MDC limits
- Better performer than vdisks for read I/Os
Server Machine Guidelines

- Server Virtual Machine (SVM)
  - 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**

<table>
<thead>
<tr>
<th>Queue Name</th>
<th>Type</th>
<th>Priority</th>
<th>Address Range</th>
<th>Age</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDLLIB14</td>
<td>Q3</td>
<td>IO</td>
<td>00002473/00002654</td>
<td>.D.</td>
<td>-0217</td>
</tr>
<tr>
<td>KAZDAKC</td>
<td>Q3</td>
<td>IO</td>
<td>000003964/000003572</td>
<td>....</td>
<td>-0190</td>
</tr>
<tr>
<td>BITNER</td>
<td>Q1</td>
<td>R00</td>
<td>00001073/00001054</td>
<td>.I..</td>
<td>-0163</td>
</tr>
<tr>
<td>LCRAMER</td>
<td>Q3</td>
<td>IO</td>
<td>00003122/00002850</td>
<td>....</td>
<td>.0259</td>
</tr>
<tr>
<td>DSSERV</td>
<td>L0</td>
<td>R</td>
<td>00007290/00007289</td>
<td>....</td>
<td>.3229</td>
</tr>
<tr>
<td>RSCS</td>
<td>Q0</td>
<td>PS</td>
<td>00001638/00001616</td>
<td>.I..</td>
<td>99999</td>
</tr>
<tr>
<td>SICIGANO</td>
<td>Q3</td>
<td>PS</td>
<td>00000662/00000662</td>
<td>.I..</td>
<td>99999</td>
</tr>
<tr>
<td>VMLINUX1</td>
<td>Q3</td>
<td>PS</td>
<td>00018063/00018063</td>
<td>....</td>
<td>99999</td>
</tr>
<tr>
<td>LNXREGR</td>
<td>Q3</td>
<td>PS</td>
<td>00073326/00073210</td>
<td>....</td>
<td>99999</td>
</tr>
<tr>
<td>VMLINUX</td>
<td>Q3</td>
<td>PS</td>
<td>00031672/00031672</td>
<td>....</td>
<td>99999</td>
</tr>
<tr>
<td>TCPIP</td>
<td>Q0</td>
<td>PS</td>
<td>00018863/00018397</td>
<td>.I..</td>
<td>99999</td>
</tr>
<tr>
<td>EDLLNX2</td>
<td>Q3</td>
<td>PS</td>
<td>00032497/00032497</td>
<td>....</td>
<td>99999</td>
</tr>
<tr>
<td>EDLLNX1</td>
<td>Q3</td>
<td>PS</td>
<td>00015939/00015939</td>
<td>....</td>
<td>99999</td>
</tr>
</tbody>
</table>
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

- 3270
- History Files
- Raw Monwrite
- MONWRITE Utility
- MONDCSS Segment
- VMRM
- TCP/IP Network
- DCSS SEGOUT
- *MONITOR System Service
- VM Events
- CP Control Blocks
- Application Data
- 3270
- TEMS
- VM IRA
- LINUX Guest
- OMEGAMON XE
- Browser
- Performance Toolkit
- RAW Monwrite
- History Files
OMEGAMON XE Basic Architecture

- TEP
- TEPS
- TDW
- TEMS

Tivoli Management Services

z/VM command and response flows

Data and synchronization flows

CMD (Guest) → Mon

Extensions

DCSS (Guest) → Linux (Guest)

Linux IRA

Linux IRA

Linux IRA

Linux IRA

TDW

Linux (Guest)
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

- In Performance Toolkit
  - FCX108 DEVICE
  - FCX114 USTAT
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.
  - \(< 1 \text{ ms, unless contention at channels or control units.}\)
- **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
  - 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
  - SC24-6109-03 z/VM 5.3.0
  - links to documents, tools, reference material
  - common problems and solutions
  - guidelines
  - presentations with speaker notes
A Case Study
The Grinch That Stole Performance

From Performance Toolkit USTAT FCX114 Report January 5:

<table>
<thead>
<tr>
<th>%CPU</th>
<th>%LDG</th>
<th>%PGW</th>
<th>%IOW</th>
<th>%SIM</th>
<th>%TIW</th>
<th>%CFW</th>
<th>%TI</th>
<th>%EL</th>
<th>%DM</th>
<th>%IOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>51</td>
<td>8</td>
</tr>
</tbody>
</table>

From Performance Toolkit DEVICE FCX108 Report January 5:

<table>
<thead>
<tr>
<th>Addr</th>
<th>I/O</th>
<th>Avoid</th>
<th>Pend</th>
<th>Disc</th>
<th>Conn</th>
<th>Serv</th>
<th>Resp</th>
<th>Resp</th>
<th>CUWt</th>
<th>Qued</th>
<th>Busy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1742</td>
<td>26.7</td>
<td>0</td>
<td>1.3</td>
<td>18.4</td>
<td>4.7</td>
<td>24.5</td>
<td>69.0</td>
<td>.0</td>
<td>1.2</td>
<td>65.4</td>
<td></td>
</tr>
</tbody>
</table>

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:**

<table>
<thead>
<tr>
<th>Time (msec)</th>
<th>Req.</th>
<th>Rate/s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Addr</th>
<th>I/O Avoid</th>
<th>Pend Disc</th>
<th>Conn Serv</th>
<th>Resp</th>
<th>CUWt</th>
<th>Qued</th>
<th>Busy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec8</td>
<td>41.0</td>
<td>.0</td>
<td>0.3</td>
<td>0.2</td>
<td>2.0</td>
<td>2.6</td>
<td>2.9</td>
</tr>
<tr>
<td>Jan5</td>
<td>26.7</td>
<td>.0</td>
<td>1.3</td>
<td>18.4</td>
<td>4.7</td>
<td>24.5</td>
<td>69.0</td>
</tr>
</tbody>
</table>

**From Performance Toolkit CACHEXT FCX177 Dec. 8th Report:**

<table>
<thead>
<tr>
<th>Hits</th>
<th>Rate/s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cache SCMBK</th>
<th>N-Seq</th>
<th>Seq</th>
<th>FW</th>
<th>Read Tot</th>
<th>RdHt</th>
<th>Wrt</th>
<th>DFW</th>
<th>CFW</th>
</tr>
</thead>
<tbody>
<tr>
<td>53.0</td>
<td>41.0</td>
<td>52.3</td>
<td>0</td>
<td>0.6</td>
<td>99</td>
<td>99</td>
<td>96</td>
<td>96</td>
</tr>
</tbody>
</table>
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
DUPLEX 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 : 0.0ms    Throttle del/s: ...
Status: SHARABLE

Path(s) to device 1742: 0A 2A 4A

<table>
<thead>
<tr>
<th>Device</th>
<th>Overall CU-Cache Performance</th>
<th>Split</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIR</td>
<td>ADDR</td>
<td>VOLSER</td>
</tr>
<tr>
<td>08</td>
<td>1742</td>
<td>USE001</td>
</tr>
</tbody>
</table>
## Performance Toolkit Device Details

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