z/VM System Limits

Revision 2011-07-22 (BKW)

IBM z/VM Performance Evaluation
Bill Bitner bitnerb@us.ibm.com
Brian Wade bkw@us.ibm.com
Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml: AS/400, DBE, e-business logo, ESCON, eServer, FiCON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/390, System z, VM/ESA, VSE/ESA, WebSphere, xSeries, z/OS, zSeries, z/VM

The following are trademarks or registered trademarks of other companies

Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation
Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
LINUX is a registered trademark of Linus Torvalds
UNIX is a registered trademark of The Open Group in the United States and other countries.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.
SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.
Intel is a registered trademark of Intel Corporation
* All other products may be trademarks or registered trademarks of their respective companies.

NOTES:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

Any proposed use of claims in this presentation outside of the United States must be reviewed by local IBM country counsel prior to such use.

The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.
Acknowledgements

- **Contributors to this material**
  - Bill Bitner
  - Wes Ernsberger (now retired - 😞 - we miss you Wes)
  - Bill Holder
  - Virg Meredith
  - Brian Wade
Agenda

- Describe various limits
  - Architected
  - Consumption
  - Latent
- Show how to keep tabs on consumables
- Discuss limits that may be hit first
Limits

- Processors
- Memory
- I/O
- Others
- Latent limits
Processors

- Processors (architected): 64
  - Includes all engine types (CP, zAAP, zIIP, IFL…)
- Processors (hardware): 80 (z196), 64 (z10 EC), 54 (z9 EC)
- Logical processors in a partition (hardware): 80 (z196), 64 (z10), 54 (z9)
- Logical processors in a z/VM partition (unsupported): 64 (z196, z10), 54 (z9)
- Logical processors in a z/VM partition (support statement): 32
- Master processor (architected): 1
  - 100%-utilized master is the issue
  - z/VM will elect a new master if master fails
- Virtual processors in single virtual machine (architected): 64
  - But $N_{\text{Virtual}} > N_{\text{Logical}}$ is not usually practical
- Number of partitions: 60 (z196, z10, z9)
Processor Scaling

Number of Supported Processors

- 1985
- 1991 - VM/ESA 1.1.1
- 1994 - VM/ESA 1.2.2
- 2004 - z/VM 5.1
- 2007 - z/VM 5.3
Processors: FCX100 CPU

FCX100  Run 2007/09/06 14:00:28       CPU

General CPU Load and User Transactions

From 2007/09/04 09:07:00
To 2007/09/04 10:00:00
For 3180 Secs 00:53:00

z/VM   V.5.3.0 SLU 0701
CPU 2094-700

<table>
<thead>
<tr>
<th>CPU Load</th>
<th>Vector Facility</th>
<th>Status or</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROC TYPE</td>
<td>%CPU</td>
<td>%CP</td>
</tr>
<tr>
<td>P00  IFL</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>P15  IFL</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>P14  IFL</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>P13  IFL</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>P12  IFL</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>P11  IFL</td>
<td>18</td>
<td>2</td>
</tr>
</tbody>
</table>

... truncated ...

1. T/V ~ 18/16 = 1.13  a little CP overhead here
2. Master does not seem unduly burdened
Wait State Analysis by User

From 2007/09/04 09:07:00
To 2007/09/04 10:00:00
For 3180 Secs 00:53:00

<table>
<thead>
<tr>
<th>Userid</th>
<th>%ACT</th>
<th>%RUN</th>
<th>%CPU</th>
<th>%LDG</th>
<th>%GW</th>
<th>%OW</th>
<th>%SI</th>
<th>%MT</th>
<th>%FW</th>
<th>%FI</th>
<th>%EL</th>
<th>%DM</th>
<th>%PGA</th>
<th>%LI</th>
<th>%OTH</th>
<th>Q0</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>E0-3</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>64</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>83</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>29</td>
<td>10</td>
<td>57</td>
<td>0</td>
<td>211</td>
</tr>
<tr>
<td>TCP1P</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RSCSDNS</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SNMPD</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>98</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SZVAS001</td>
<td>100</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>97</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>12</td>
<td>85</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

1. %CPU wait is very low – nobody is starved for engine
2. %TIW is “test idle wait” – we are waiting to see if queue drop happens
Memory (1 of 3)

- **Central storage**
  - CEC limit for a partition: 512 GB-HSA (z9), 1 TB (z10, z196)
  - Supported for a z/VM partition: 256 GB
  - Unsupported for a z/VM partition: 512 GB

- **Expanded storage**
  - CEC limit for a partition: 16 TB
  - Supported for a z/VM partition: 128 GB
  - Unsupported for a z/VM partition: 600 – 700 GB

- **Virtual machine size (hardware):**
  - z196, 16 TB; z10, 8 TB; z9, 1 TB
  - On z990 and z900, 256 GB
Memory (2 of 3)

- **Instantiated guest real limit imposed by PTRM space limits (architected):** 8 TB
  - 16 4-GB PTRM spaces; each PTRM space can map 512 GB of guest real

- **Virtual to real ratio (practical): about 2:1 to 3:1**
  - Assumes guests that tend to use all of their memory
  - Some performance-sensitive production workloads will require 1:1
  - If you really, really do your homework on your paging subsystem you can push this up somewhat
  - VMRM-CMM can help with this too (it encourages Linux to Diag x’10’ its guest real)
  - Many factors come into play here:
    - Relative mix of active and idle guests
    - Workload’s or SLA’s sensitivity to delays
    - Exploitation of shared memory
Memory (3 of 3)

- Paging space (architected) (optimal when <= 50% allocated):
  - 11.2 TB for ECKD
  - 15.9 TB for Emulated FBA on FCP SCSI
- Paging volumes: 255
- Concurrent paging I/Os per paging volume: 1 for ECKD, >1 for EDEV (have observed 1.6)
- System Execution Space (SXS) (architected): 2 GB
  - For practical purposes it is 2 GB, but there are structures in the space placed above 2 GB
- DCSS size (architected):
  - Each segment can be up to 2047 MB
  - Segments can map into > 2 GB, starting in z/VM 5.4
- Minidisk Cache (architected): 8 GB
  - Practically somewhat less, ~ 2 GB
Memory Scaling

Effective Real Memory Use Limits

- 2001 - z/VM 3.1: Increase
- 2005 - z/VM 5.2: Further increase
- 2007 - z/VM 5.3: Significant increase
Page Slots: FCX146 AUXLOG

FCX146  Run 2007/09/06 14:00:28         AUXLOG

Auxiliary Storage Utilization, by Time

From 2007/09/04 09:07:00
to   2007/09/04 10:00:00
for   3180 Secs 00:53:00

<table>
<thead>
<tr>
<th>Interval</th>
<th>Total Used</th>
<th>Total Used</th>
<th>Total Used</th>
<th>&lt;--- Spool Files ---&gt;</th>
<th>&lt;Average MLOAD&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>End Time</td>
<td>Slots</td>
<td>% Slots</td>
<td>Slots</td>
<td>% Slots</td>
<td>Paging Spooling</td>
</tr>
<tr>
<td>09:08:00</td>
<td>87146k</td>
<td>44%</td>
<td>5409096</td>
<td>52%</td>
<td>54.02</td>
</tr>
<tr>
<td>09:09:00</td>
<td>87146k</td>
<td>44%</td>
<td>5409096</td>
<td>52%</td>
<td>1.02</td>
</tr>
<tr>
<td>09:10:00</td>
<td>87146k</td>
<td>44%</td>
<td>5409096</td>
<td>52%</td>
<td>1.02</td>
</tr>
<tr>
<td>09:11:00</td>
<td>87146k</td>
<td>44%</td>
<td>5409096</td>
<td>52%</td>
<td>1.02</td>
</tr>
<tr>
<td>09:12:00</td>
<td>87146k</td>
<td>44%</td>
<td>5409096</td>
<td>52%</td>
<td>1.02</td>
</tr>
</tbody>
</table>

1. This system is using 44% of its page slots.
2. 87146k slots / (256 slots/MB) = 332 GB of paging space.
3. 332 GB * 44% = 146 GB in use.
Load and Performance of CP Owned Disks

From 2007/09/04 09:07:00
To 2007/09/04 10:00:00
For 3180 Secs 00:53:00

---

Page / SPOOL Allocation Summary

<table>
<thead>
<tr>
<th>PAGE slots available</th>
<th>87146k</th>
<th>SPOOL slots available</th>
<th>540906</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAGE slot utilization</td>
<td>44%</td>
<td>SPOOL slot utilization</td>
<td>52%</td>
</tr>
<tr>
<td>T-Disk cylinders avail.</td>
<td>...</td>
<td>DUMP slots available</td>
<td>0</td>
</tr>
<tr>
<td>T-Disk space utilization</td>
<td>...%</td>
<td>DUMP slot utilization</td>
<td>...%</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>&lt; Device Descr. -&gt;</th>
<th>&lt;---------- Rate/s -----------&gt;</th>
<th>User</th>
<th>Serv</th>
<th>MLOAD</th>
<th>Block</th>
<th>%Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>Area</td>
<td>Used</td>
<td>&lt; -Page--&gt;</td>
<td>&lt; -Spool--&gt;</td>
<td>SSCH</td>
<td>Inter</td>
</tr>
<tr>
<td>Addr</td>
<td>Devtyp</td>
<td>Serial</td>
<td>Type</td>
<td>Extent</td>
<td>%</td>
<td>P-Rds</td>
</tr>
<tr>
<td>F088</td>
<td>3390</td>
<td>VS2P49</td>
<td>PAGE</td>
<td>0-3338</td>
<td>45</td>
<td>2.6</td>
</tr>
<tr>
<td>F090</td>
<td>3390</td>
<td>VS2P69</td>
<td>PAGE</td>
<td>0-3338</td>
<td>45</td>
<td>2.7</td>
</tr>
</tbody>
</table>

1. Interesting fields: slot utilization, MLOAD, Queue Lngth.
2. See a wait queue? See also %PGW and %LDG in FCX114 USTAT.
**V:R Ratio and Segment Tables: FCX113 UPAGE**

<table>
<thead>
<tr>
<th>Userid</th>
<th>Owned</th>
<th>Reads</th>
<th>Write</th>
<th>Steals</th>
<th>&gt;2GB</th>
<th>&gt;MS</th>
<th>X&gt;MS</th>
<th>X&gt;OS</th>
<th>VBS</th>
<th>Resrvd</th>
<th>R&lt;2GB</th>
<th>R&gt;2GB</th>
<th>L&lt;2GB</th>
<th>L&gt;2GB</th>
<th>XSTOR</th>
<th>DASD</th>
<th>Size</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;System</td>
<td>.0</td>
<td>1.7</td>
<td>1.1</td>
<td>4.1</td>
<td>0</td>
<td>2.4</td>
<td>3.7</td>
<td>1.4</td>
<td>122050</td>
<td>0</td>
<td>2347</td>
<td>106962</td>
<td>6</td>
<td>24</td>
<td>12240</td>
<td>179131</td>
<td><strong>1310M</strong></td>
<td><strong>212</strong></td>
</tr>
<tr>
<td>ABCDEFGH</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>0</td>
<td>.1</td>
<td>.0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>483</td>
<td>254</td>
<td>32M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DATAMOVA</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.5</td>
<td>.5</td>
<td>.0</td>
<td>147</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>220</td>
<td>368</td>
<td>32M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DATAMOVB</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.6</td>
<td>.6</td>
<td>.0</td>
<td>192</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>220</td>
<td>366</td>
<td>32M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DATAMOC</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.6</td>
<td>.6</td>
<td>.0</td>
<td>191</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>220</td>
<td>369</td>
<td>32M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DATAMOD</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.6</td>
<td>.6</td>
<td>.0</td>
<td>189</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>220</td>
<td>362</td>
<td>32M</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Resident guest pages = (2347 + 106962) * 212 = 88.3 GB
2. \[V:R = \frac{(1310 \text{ MB} \times 212)}{91 \text{ GB}} = 2.98\] (FCX103 shows 91 GB central)
3. Segment table pages: hard to say. Conservatively:
   \[212 \text{ guests} \times (4 \text{ ST/guest} \times 4 \text{ pg/ST}) = 13 \text{ MB}\]
## PTRM Space: FCX134 DSPACESH

<table>
<thead>
<tr>
<th>拥有者</th>
<th>数据空间名称</th>
<th>Pgstl</th>
<th>Pgrds</th>
<th>Pgwrt</th>
<th>X-rds</th>
<th>X-wrt</th>
<th>X-mig</th>
<th>总计</th>
<th>Resid R&lt;2GB</th>
<th>Lock L&lt;2GB</th>
<th>Count</th>
<th>Lockd</th>
<th>XSTOR</th>
<th>DASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;System</td>
<td>FULL$TRACK$CACHE$1</td>
<td>0.026</td>
<td>0.016</td>
<td>0.001</td>
<td>0.015</td>
<td>0.026</td>
<td>0.000</td>
<td>103k</td>
<td>51</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>ISFCDATASPACE</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>524k</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>PTRM0000</td>
<td>4.257</td>
<td>0.492</td>
<td>0.442</td>
<td>3.957</td>
<td>4.036</td>
<td>0.000</td>
<td>1049k</td>
<td>386k</td>
<td>15885</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>REAL</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>24M</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>SYSTEM</td>
<td>0.080</td>
<td>0.001</td>
<td>0.034</td>
<td>0.079</td>
<td>0.080</td>
<td>0.000</td>
<td>524k</td>
<td>45</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>0</td>
</tr>
</tbody>
</table>

1. PTRM space = 386,000 pages = 1.47 GB of PGMBKs.
2. This maps 128 * 1.47 GB = 188.5 GB of guest storage.
3. z/VM 5.3 and later can have >2 GB of PGMBKs.
Available List Management, by Time

<table>
<thead>
<tr>
<th>Interval</th>
<th>&lt;Low&gt;</th>
<th>&lt;High&gt;</th>
<th>&lt;Available&gt;</th>
<th>&lt;Obtains/s&gt;</th>
<th>&lt;Returns/s&gt;</th>
<th>&lt;Empty&gt;</th>
<th>&lt;Scan1-&gt;</th>
<th>&lt;Scan2-&gt;</th>
<th>&lt;EmScan-&gt;</th>
<th>Scan</th>
<th>&lt;Empty&gt;</th>
<th>&lt;Scan&gt;</th>
<th>&lt;EmScan&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:08:00</td>
<td>20 7680 5820 13480</td>
<td>6665 15122 353.3 838.5 353.2 1007</td>
<td>0 0</td>
<td>0 43091</td>
<td>3 26491</td>
<td>0 0</td>
<td>3 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:09:00</td>
<td>20 7680 5820 13480</td>
<td>3986 5496 163.1 640.2 108.9 442.7</td>
<td>0 0</td>
<td>1 14528</td>
<td>0 0</td>
<td>0 0</td>
<td>1 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:10:00</td>
<td>20 7681 5820 13481</td>
<td>6622 9542 222.4 556.1 257.0 598.3</td>
<td>0 0</td>
<td>0 30103</td>
<td>2 8868</td>
<td>0 0</td>
<td>1 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:11:00</td>
<td>20 7681 5820 13481</td>
<td>4982 6710 292.1 615.2 248.8 535.6</td>
<td>0 0</td>
<td>0 21246</td>
<td>0 8547</td>
<td>1 3989</td>
<td>1 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:12:00</td>
<td>20 7681 5820 13481</td>
<td>4769 1560 284.9 946.9 254.4 830.0</td>
<td>0 0</td>
<td>0 18253</td>
<td>0 22438</td>
<td>2 656</td>
<td>1 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Pct ES = 88% generally this system is tight on storage
   (might just be intense use of VDISKs – where would you look?)
2. Scan fail >0 generally this system is tight on storage
3. Times Empty = 0 this indicates it isn’t critical yet
1. How we touch guest pages: (1) 64-bit; (2) AR mode; (3) SXS.
2. There are 524,288 pages in the SXS.
3. This system has 466,000 SXS pages available on average.
### MDC: FCX178 MDCSTOR

<table>
<thead>
<tr>
<th>Interval</th>
<th>Actual</th>
<th>Mn</th>
<th>Max</th>
<th>Page</th>
<th>Steal</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:57:41</td>
<td>5839k</td>
<td>82738</td>
<td>1354k</td>
<td>0</td>
<td>7864k</td>
</tr>
<tr>
<td>09:58:11</td>
<td>5838k</td>
<td>119813</td>
<td>1932k</td>
<td>0</td>
<td>7864k</td>
</tr>
<tr>
<td>09:58:41</td>
<td>5838k</td>
<td>119825</td>
<td>1932k</td>
<td>0</td>
<td>7864k</td>
</tr>
<tr>
<td>09:59:11</td>
<td>5838k</td>
<td>119825</td>
<td>1932k</td>
<td>0</td>
<td>7864k</td>
</tr>
<tr>
<td>10:00:11</td>
<td>5838k</td>
<td>119837</td>
<td>1932k</td>
<td>0</td>
<td>7864k</td>
</tr>
</tbody>
</table>

Xstore not used for this configuration so edited out from report.

Add up the pages in main storage for this run and you get about 8 GB in use for MDC.
MDC Spaces: FCX134 DSPACESH

You’ll see the address spaces used for MDC (track cache)
Values here are zero for page counts, ignore.
Reorder Processing - Background

- **Page reorder** is the process of managing user-frame-owned lists as input to demand scan processing.
  - It includes resetting the HW reference bit.
  - Serializes the virtual machine (all virtual processors).
  - In all releases of z/VM

- It is done periodically on a virtual machine basis.
  - Even if the system is not paging.

- The cost of reorder is proportional to the number of resident frames for the virtual machine.
  - Roughly 130 ms/GB resident on z10
  - Delays of ~1 second for guest having 8 GB resident
  - This can vary for different reasons +/- 40%
Reorder Processing - Diagnosing

- **Performance Toolkit**
  - Check resident page fields (“R<2GB” & “R>2GB”) on FCX113 UPAGE report
    - Remember, reorder works against the resident pages, not total virtual machine size.
  - Check Console Function Mode Wait (“%CFW”) on FCX114 USTAT report
    - A virtual machine may be brought through console function mode to serialize reorder. There are other ways to serialize for reorder and there are other reasons for CFW, so this is not conclusive.

- **REORDMON**
  - Available from Bill Bitner or http://www.vm.ibm.com/download/packages/
  - Works against MONWRITE data for all monitored virtual machines
  - Works in real time for a specific virtual machine
  - Provides how often reorder processing occurs in each monitor interval
Reorder Processing - Mitigations

- Try to keep the virtual machine as small as possible.
- Virtual machines with multiple applications may need to be split into multiple virtual machines with fewer applications.
- APAR VM64774 is now available
  - Implements a flexible SET REORDER function
I/O (1 of 3)

- **Number of subchannels in a partition (aka device numbers) (architected):** 65,535
- **CHPIDs per server:**
  - z9 EC or z10 EC: 1024 ESCON, 336 FICON Express 4, 96 OSA Express 3, 16 HiperSockets
  - z196: 240 ESCON, 336 FICON Express 8, 336 FICON Express 4, 96 OSA Express 3, 32 HiperSockets
- **Device numbers per disk volume**
  - Without PAV, 1
  - With PAV or HyperPAV, 256 (base plus 255 aliases, but can use only 7 aliases for ESCON)
- **Virtual devices per virtual machine:** 24576 (24K)
- **Concurrent real I/Os per ECKD disk volume:** 1 usually, but more with PAV or HyperPAV if of guest origin
- **Concurrent real I/Os per chpid (aka “open exchange limit”):**
  - 1 for ESCON
  - 32 for FICON Express
  - 64 for FICON Express 2 and later
- **I/O rates:**
  - Fastest FICON is an 8 Gb/sec link (translates to about 640 MB/sec)
  - About 1-2 msec per I/O are required for a nominal DASD I/O from a z9 to a 2107 (aka rates of 500-1000/sec/device)
- **Ref:** [http://www-03.ibm.com/systems/z/hardware/](http://www-03.ibm.com/systems/z/hardware/)
I/O (2 of 3)

- **ECKD volume sizes**
  - Largest ECKD minidisk that can contain a CMS file system (architected): 32768 cylinders (22.5 GB)
  - Largest ECKD volume, period: 65536 cylinders (43 GB)

- **EFBA volume sizes**
  - Largest EFBA minidisk that can contain a CMS file system (architected): 381 GB (tough to beat for archiving)
    - Practical limit is 22 GB due to CMS in-memory file system structures under 2 GB, unless very few, very large files
  - Largest EDEV CP can use: 1024 GB (but PAGE, SPOL, DRCT must be below 64 GB line on volume)
  - Largest EDEV, period: \(2^{32}\) FB-512 blocks (2048 GB)

- **VDISK size (architected):** 2 GB minus eight 512-byte blocks
- **Total VDISK (architected):** 2 TB
- **Single VSWITCH OSAs:** 8
I/O (3 of 3)

- **Number of files on a user’s accessed CMS disks or directories**
  - 262,144 (16 MB / 64 bytes per FST)
  - You cannot use all of your <16MB storage for FSTs anyway
  - Files residing in SFS DIRCONTROL directories in a data space do not charge to this limit because those FSTs aren’t below 16 MB in CMS storage

- **SFS limits**
  - 32767 storage groups
  - \(2^{32}-1\) blocks (16 TB) per file pool, storage group, file space, file
  - 2 GB of data per DIRCONTROL-directory-in-data-space
  - No architected limit on numbers of files or users
  - See Appendix B in CMS File Pool Planning, Administration, and Operation
DASD I/O: FCX108 DEVICE

<table>
<thead>
<tr>
<th>Addr Type</th>
<th>Label / ID</th>
<th>Links ths</th>
<th>I/O Avoid</th>
<th>Pend Disc Conn Serv Resp CU</th>
<th>Qued Busy READ Cyl s SSCH Set/s Dly/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>F024</td>
<td>3390</td>
<td>VS2426</td>
<td>1 4</td>
<td>12.9 147.0 .2 .7 .4 1 3.1 .3 .0 2 91 193 0 . . . . .</td>
<td></td>
</tr>
<tr>
<td>OC20</td>
<td>CTCA</td>
<td></td>
<td>...</td>
<td>1 12.6 ... .3 .2 .6 1 1.1 .0 .0 1 ... 0 ... ...</td>
<td></td>
</tr>
<tr>
<td>F685</td>
<td>3390</td>
<td>VS2W01</td>
<td>290 4</td>
<td>11.8 .3 .2 .0 .3 .5 .5 .0 .0 1 84 89 0 ... ...</td>
<td></td>
</tr>
<tr>
<td>F411</td>
<td>3390</td>
<td>VS2613</td>
<td>1 4</td>
<td>10.6 .5 .2 .3 .4 .9 .9 .0 .0 1 1 1303 0 ... ...</td>
<td></td>
</tr>
</tbody>
</table>

1. Interesting columns: Avoid, Serv, Req, Qued.
2. Req. Qued > 0 is a good trigger for looking at PAV or HyperPAV.
3. For queuing on page or spool, use FCX109 DEVICE CPOWNED.
How z/VM Exploits PAV For Its Guests

- Guests do minidisk I/O
- Old: we serialized the resultant real I/Os
- New: we use the volume’s PAV aliases to drive the real I/Os concurrently
- One wait queue for the volume
- Many device numbers for the volume

Note: z/VM 5.2 plus VM63855 (UM31771) (May 2006)
FICON Open Exchange Limit

- **Parallel and ESCON:** 1 I/O at a time on a chpid
  - Pending time >0 could mean chpid contention
  - Controller disconnect was a good thing, and so they did
- **FICON:** 64 (was 32) I/Os at a time on a chpid
  - Pending time >0 probably now means slow IR
  - Little motive for controller to disconnect anymore
    - Controller cache miss is still a good reason
- **Calculating “open exchange level” is not easy**
- **Very seldom is this an issue anyway**
Other

- **Number of spool files (architected):**
  - 9999 per user
  - 1.6 million spool files per system
    - 1024 files per warm start block * (180 * 9) warm start blocks

- **Number of logged-on virtual machines (approximate):** about 100,000 (per designers)
## Metrics for Formal Spin Locks

**FCX265**  **CPU 2094**  **SER 19B9E**  **Interval 02:31:51 - 12:34:01**  **GDLVM7**

<table>
<thead>
<tr>
<th>Interval</th>
<th>Locks Average /sec</th>
<th>Pct</th>
<th>Locks Average /sec</th>
<th>Pct</th>
<th>Locks Average /sec</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>End Time</td>
<td>LockName</td>
<td>/sec</td>
<td>usec</td>
<td>Spin</td>
<td>/sec</td>
<td>usec</td>
</tr>
<tr>
<td></td>
<td>SRMATDLK</td>
<td>1.9</td>
<td>.539</td>
<td>.000</td>
<td>1.9</td>
<td>.539</td>
</tr>
<tr>
<td></td>
<td>RSAAVCLK</td>
<td>.0</td>
<td>2.015</td>
<td>.000</td>
<td>.0</td>
<td>2.015</td>
</tr>
<tr>
<td></td>
<td>FSDVMLK</td>
<td>.0</td>
<td>24.97</td>
<td>.000</td>
<td>.0</td>
<td>24.97</td>
</tr>
<tr>
<td></td>
<td>SRMALOCK</td>
<td>.0</td>
<td>.000</td>
<td>.000</td>
<td>.0</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>HCPTRQLK</td>
<td>4.1</td>
<td>.195</td>
<td>.000</td>
<td>4.1</td>
<td>.195</td>
</tr>
<tr>
<td></td>
<td>SRMSLOCK</td>
<td>34.0</td>
<td>1.096</td>
<td>.001</td>
<td>32.7</td>
<td>1.037</td>
</tr>
</tbody>
</table>

This is really for our use. Just look at T/V.
Latent Limits

- Sometimes it’s not an architected limit
- Sometimes it’s just “your workload won’t scale past here, because…”
- In our studies of z/VM 5.3, we found these kinds of latent limits:
  - Searching for a below-2-GB frame in lists dominated by above-2-GB frames (storage balancing functions)
  - Contention for locks, usually the scheduler lock
- These kinds of phenomena were the reasons we published the limits to be 256 GB and 32 engines
  - We wanted to publish supported limits we felt would be safe in a very large variety of workloads and environments
  - Some of our measurement workloads scaled higher than this
Other Notes on z/VM Limits

- **Sheer hardware:**
  - z/VM 5.2: 24 engines, 128 GB real
  - z/VM 5.3: 32 engines, 256 GB real
  - System z: 65,000 I/O devices per partition

- **Mad-scientist stuff we’ve tried:**
  - 54 engines
  - 1 TB / 128 GB with 100 10 GB Linux Apaches
  - 440 GB / 20 GB with 8 1 TB thrasher guests
  - 256 GB of thrashers in 3 GB of central

- **Utilizations we routinely see in customer environments**
  - 85% to 95% CPU utilization without worry
  - Hundreds of thousands of XSTORE pages per second
  - Tens of thousands of DASD pages per second

- **Our limits tend to have two distinct shapes**
  - Slow rolloff: CPUs
  - Fast rolloff: storage

![Graph showing performance vs. utilization with two curves: one precipitous (e.g., storage) and one gradual (e.g., CPUs).]
Keeping Tabs on Consumption Limits

- **Processor**
  - CPU utilization: FCX100 CPU, FCX126 LPAR, FCX144 PROCLOG, FCX114 USTAT
  - Good article on CPU at http://www.vm.ibm.com/perf/tips/parinfo.html

- **Memory & Paging**
  - Page slots in use: FCX146 AUXLOG
  - Paging I/O: FCX109 DEVICE CPOWNED
  - V:R Memory ratio: FCX113 UPAGE
  - PTRM space consumed: FCX134 DSPACESH
  - Storage in use for segment tables: FCX113 UPAGE
  - Consumption of SXS space: FCX261 SXSAVAIL
  - MDC: FCX178 MDCSTOR, FCX134 DSPACESH
  - Consumption of real memory: FCX103 STORAGE, FCX254 AVAILLOG
  - Consumption of expanded storage: FCX103 STORAGE
  - Good article on paging at http://www.vm.ibm.com/perf/tips/prgpage.html

- **I/O**
  - Guest DASD I/O: FCX108 DEVICE
  - Concurrency on FICON chpids: FCX131 DEVCONF, FCX215 INTERIM FCHANNEL, FCX168 DEVLOG, FCX232 IOPROCLG
What Consumption Limits Will We Hit First?

- **Guest-storage-intensive workload:**
  - page slots on DASD... at 5-6 TB things start to get interesting... mitigate by paging to SCSI
  - paging I/O concurrency – only 255 at a time – mitigate by paging to SCSI
  - utilization on paging volumes and chpids -- watch for MLOAD elongation -- mitigate by spreading I/O
  - reorder processing – use more and smaller guests
  - mitigation by application tuning… perhaps smaller guests
  - segment table constraints: probably an issue at 50% (128 TB of logged-on guest real) … not anytime soon

- **Real-storage-intensive workload:**
  - Ability of the system to page will limit you: ensure adequate XSTORE and paging capacity
  - You can define > 256 GB of real storage, but we are aware that some workloads cannot scale that high
  - Mitigation by application tuning or by using CMM

- **CPU-intensive workload:**
  - FCX100 CPU, FCX126 LPAR, and FCX 114 USTAT will reveal CPU limitations
  - You can define > 32 engines, but we are aware that some workloads cannot scale that high
  - Mitigation by application tuning

- **I/O-intensive workload:**
  - Device queueing: consider whether PAV or HyperPAV might offer leverage
  - Chpid utilization: add more chpids per storage controller

- Ultimately partitions can be split, but we would prefer you not have to do this (too complicated)

- Without trend data (repeated samples) for your workloads it is difficult to predict which of these limits you will hit first
Summary

- **Knowing limits:**
  - Real resource consumption
  - Limits to managing the virtualization of real resources

- **Measuring limits:**
  - Knowing where to watch for these limits
  - Including these in capacity planning

- **Managing limits:**
  - Tuning and configuring
  - Planning for growth
End of Presentation