z/VM Paging with SSD and Flash-Type Disk Devices

Session 16452

Bill Bitner
z/VM Development Client Focus and Care
bitnerb@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, ESCO, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/390, 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.
Notice Regarding Specialty Engines (e.g., zIIPs, zAAPs and IFLs):

Any information contained in this document regarding Specialty Engines ("SEs") and SE eligible workloads provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”).

No other workload processing is authorized for execution on an SE.

IBM offers SEs at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.
Agenda

- Choosing the right disk storage
- High Performance Paging Options
- The Need for Faster Paging
- Case Study
Choosing the Right Disk Storage
So Many Choices

ECKD
- Dedicated
- FullPack
- Minidisk
- No MDC
- MDC
- No PAV
- Host PAV
- Guest PAV

SCSI
- Dedicated
- EDEV
- No MDC
- MDC

VDisk

Choices at Linux Level

Other HW Choices
(e.g. Flash, Easy Tier, etc.)
Look at All Attributes – Not Just One
A Few More Things to Think About

- Lack of PAV & HyperPAV support makes the size of ECKD DASD for z/VM paging important. We want to not have I/Os queued up on these volumes but at same time balance this with work to administer the number of DASD volumes involved.

- Even if ECKD volumes could be infinitely large and infinitely fast, we would probably still want as many as there are logical processors for z/VM.

- FCP SCSI for z/VM Paging provides a level of parallelism but at a cost in processor time
  - Greater Bandwidth
  - Higher CPU costs
High Performance Paging Options
Flash vs. SSD and Terminology

- Purists will say “Flash” is not the equivalent of “SSD”
  - Solid State Disk – most often used to describe a device where the access is through an existing disk interface
  - Flash – access to it via direct interface to the memory

- Common Attributes:
  - Storage capacity
  - Write speed
  - Read speed
  - Active power
  - Standby power
  - Write endurance or wear-out
  - Type: NAND, NOR
Flash Express

- Flash Express Feature of zEC12 and zBC12 processors

- PCIe I/O Adapter with NAND Flash SSDs

- Accessed using Extended Asynchronous Data Mover Facility (EADMF)

- RAID 10 mirrored Pairs

- Protected with 128-bit AES encryption

- Maximum of 4 Cards provides 5.6 TB of usable storage

- Not supported by z/VM at this time. 😞
IBM FlashSystem

- IBM FlashSystem V840 - Based off technology from acquired Texas Memory System
  - Uses eMLC Flash

- FCP SCSI Only

- For use as z/VM Paging volumes, must be behind an SAN Volume Controller (SVC)
  - Some models of FlashSystem include SVC

- Various Features:
  - Easy Tier support
  - Compression
  - Data replication
DS8870

- Part of the IBM DS8000® Series
- Can be equipped with SSD drives
- HPFE (High Performance Flash Enclosure) – Newest Option
- ECKD or FCP SCSI
- Lots of features/capabilities
  - RAID 5,6,10
  - Easy Tier
  - GDPS
  - Encryption
  - Etc.
- Maximum configuration 3072 TB
Flash also Available for Other Options

- Storwize V7000
  - Internal flash drives available
  - External IBM FlashSystem Storage

- XIV
  - Flash optimized options
The Need

z/VM

where penguins multiply like rabbits.
I/O Rates and Latency

- Two aspects of paging I/O
  - Overall capacity: IOPS or MB/Second
  - Performance or Latency: delay per page read

- Historically, top-end storage servers could not be saturated by z/VM Paging

- Changing History:
  - Larger amount of memory supported and page rates increasing
    - If z/VM 6.3 allows 4 times number of virtual machines, then 4 times the page rate when you add those virtual machines.
  - Elimination of scaling problems in z/VM that allow greater paging rates
  - Better determination of the actual disk paging bandwidth
    - z/VM 6.3 algorithms changed to better estimate and utilize disk paging bandwidth
As workloads scale up, the paging I/O rates will as well.
There is no PAV or HyperPAV for z/VM paging to help scenario.
Virtual to Real Memory Overcommitment

- One of the factors often forgotten is the performance (capacity and bandwidth) of the paging configuration.

- A 100 GB real memory system with 125 GB of active virtual memory basically means being able to constantly turn over 25 GB
  – Potentially more based on the amount of memory that is changing and resulting in page writes in addition to page reads.

- As virtual machines are delayed for paging, pages that are resident tend to be needed longer, creating more demand and potential spiral-effect.
Paging Best Practices

- All paging volumes should have the same attributes: Size, Performance, etc.

- Do not mix page space with other data types

- Do not mix FCP SCSI and ECKD paging volumes

- Be aware of any shared hardware in the path (channels, control units) and who/what is sharing them

- Follow planning guidelines for amount of space
Case Study
Customer Proof of Concept

- Customer moved paging volumes from multiple z/VM LPARs and CECs to a DS8870

- For this study, everything was placed in one Logical Control Unit (LCU)
  – This is not recommended but for part of the experiment
  – Limiting to one LCU restricted full use of DS8870 cache and processing power

- The area that came under question what happens when you IPL multiple z/VM systems and need to restart 100s and 100s of virtual machines?
Key Observations

- Performance of DS8870 is a significant improvement over spinning disk.

- No single number can really portray that performance.

- Factors that will be examined:
  - Peaks across different LPARs
  - Data per I/O
  - Read / Write Ratio

- Study will explore the data in the different dimensions above.

- Brief comparison to the non SSD storage servers.
**Summary of Systems from April 13th Data**

<table>
<thead>
<tr>
<th>LPAR</th>
<th>z/VM</th>
<th>Page Vols</th>
<th>Memory Virt:Real</th>
<th>IPL Time</th>
<th>Peak Page Rate</th>
<th>Peak Page Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPAR1</td>
<td>6.3.0</td>
<td>43</td>
<td>2.36</td>
<td>15:11:14</td>
<td>73690</td>
<td>16:04:00</td>
</tr>
<tr>
<td>LPAR2</td>
<td>6.3.0</td>
<td>41</td>
<td>2.34</td>
<td>15:51:33</td>
<td>39865</td>
<td>17:10:00</td>
</tr>
<tr>
<td>LPAR3</td>
<td>6.3.0</td>
<td>29</td>
<td>1.76</td>
<td>15:11:30</td>
<td>39800</td>
<td>16:16:00</td>
</tr>
<tr>
<td>LPAR4</td>
<td>6.3.0</td>
<td>49</td>
<td>2.48</td>
<td>15:41:23</td>
<td>93229</td>
<td>16:38:00</td>
</tr>
<tr>
<td>LPAR5</td>
<td>6.3.0</td>
<td>27</td>
<td>2.36</td>
<td>15:11:34</td>
<td>170</td>
<td>21:01:00</td>
</tr>
<tr>
<td>LPAR6</td>
<td>6.3.0</td>
<td>10</td>
<td>0.87</td>
<td>15:11:10</td>
<td>30</td>
<td>15:26:00</td>
</tr>
</tbody>
</table>

- Peak Page Rate: 4KB Pages/Second, includes read and write
- IPL Time: Time at which z/VM system was IPLed, not necessarily when all virtual machines were brought online.
- LPAR5 and LPAR6 are boring, from a performance perspective.
I/O from LPAR5 and LPAR6 are insignificant. LPAR4 Data spans much longer time, though the highest rates are caused by it. First we’ll focus on the different IPL spans.
There is a large spike when enough virtual machines are active that we start to overcommit memory.

For a period of time the system works hard to reach steady state.

The larger the overcommitment, the larger the spike.

Not all peaks occur at same time.

And not all I/O is equal…
Stacking the I/O to get total going to DS8870 gives you a portrait that reflects the peaks of different I/Os.

- The peak of 4983 IOPS is at 16:46:00.
Individual Paging Volumes have consistent usage and characteristics.
To keep sanity, future charts will look at z/VM System wide information.
All the z/VM systems showed similar trends.
- Larger amounts of data are paged out with single host I/Os early in the initial paging after IPL.
- Variation continues to occur, but is most significant 15 to 30 minutes after start of overcommitment.
- Will look at another aspect on next chart.
The read to write ratio changes drastically through the period. In the early stages, data written is orders of magnitude more than data read.
Paging I/O Service time is often under 1 millisecond.

Though there are several spikes where it is higher.

But, remember how the amount of data per I/O and the read/write ratios changed.
Early numbers are anomalies from small denominators, prior to load starting.
Some spikes look scary, but remember scale. Is 1/100th of a millisecond.
Some of the spikes correlate with activity on other LPARs. Remember everything is in one LCU at this time.
Write cache, as expected is almost always 100% (or 0% when no I/O takes place)

Very limited benefit from additional write cache.
No reads prior to overcommitment
- Initially, 100% read hit and then settles down.
- Pct hit shown is of that LPAR’s I/Os, not all I/Os to DS8870
- Potential benefit from additional cache.
Does the new DS8870 Make a Difference?

- Yes!
- How much is always hard to quantify without perfectly controlled environment.
- A few items from LPAR1 (Remember not apples to apples)
  - Shutdown processing is as “interesting” as IPL processing

<table>
<thead>
<tr>
<th>Metric</th>
<th>w/o DS8870</th>
<th>w/ DS8870</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mload (z/VM Measurement of Paging Subsystem Performance)</td>
<td>36.8</td>
<td>3.7</td>
<td>-90%</td>
</tr>
<tr>
<td>Average Queue Length of Paging Volumes</td>
<td>12.9</td>
<td>0.44</td>
<td>-97%</td>
</tr>
<tr>
<td>Average %PGW from State Sampling</td>
<td>6%</td>
<td>0%</td>
<td>-100%</td>
</tr>
<tr>
<td>Average Service Time on Paging Volumes (milliseconds)</td>
<td>17.4</td>
<td>0.21</td>
<td>-99%</td>
</tr>
</tbody>
</table>
Remember Other Spikes on LPAR4?

LPAR4 - Page I/O Rates - April 13th

- Each caused by a different set of guests becoming more active and a flurry of activity.
Selected Page Read size here as Guests (and applications) wait for page reads, not page writes.

- Remember I/O characteristics vary.
**Perfect Storms**

- The z/VM systems are started at different offsets, but spent some time looking at what would happen if they did align.

- It would be significantly more activity, still containable, but I would recommend holding to the staggered start.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Actual Peak</th>
<th>Perfect Storm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host Page I/Os</td>
<td>4983/second</td>
<td>11,963/second</td>
</tr>
<tr>
<td>Paging MB/second</td>
<td>385 MB/Second</td>
<td>963 MB/Second</td>
</tr>
</tbody>
</table>
Summary

- DS8870 with SSD is providing much better I/O performance characteristics compared to spinning disk.
  - Bonus benefit in Processor resource savings

- The process of restarting 100s of Linux guests impacts paging performance significantly, though the characteristics can be different from high paging rates after the system has stabilized.

- Need to continue to track I/O operations and data rates, as well as the normal performance metrics:
  - Page wait
  - Asynchronous Page wait
  - Available List management

- The need for higher IOPS and bandwidth can be important to z/VM
  - Other limits eliminated
  - Higher consolidation workloads
QUESTIONS & DISCUSSION
Please remember to fill out an evaluation. Thanks!