How Recent Storage Innovations Can Help Improve Performance and Reliability For Your DB2 Subsystem

Jeffrey Berger
IBM

Session Code: 1413
Date and Time of Presentation  |  Platform: DB2 for z/OS
Agenda

- IBM DS8870
- Solid State Disks
  - Ultra SSD
- High Performance FICON
  - DB2 utilities
  - DB2 disorganized index scans
  - DB2 RID list scans
- How to configure DS8000 storage
- HyperPAV and Extended Addressability Volumes
The 5th Generation DS8000 Disk System – DS8870

- Designed for Enterprise environments with over 5-9’s availability natively
- Designed for Enterprise environments with over 6-9’s availability when DS8000 with Metro Mirror is combined with GDPS/PPRC HyperSwap
## Business Class and Enterprise Class Configuration Options

<table>
<thead>
<tr>
<th>Mode</th>
<th>Processors per CEC</th>
<th>Physical Capacity (max.)</th>
<th>Disk Drives (max.)</th>
<th>System Memory (GB)</th>
<th>Host Adapters (max.)</th>
<th>9xE Attach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>961</td>
<td>2-core</td>
<td>216 TB</td>
<td>144</td>
<td>16/32</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Enterprise Class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>961</td>
<td>4-core</td>
<td>360 TB</td>
<td>240</td>
<td>64</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>961</td>
<td>8-core</td>
<td>2,304 TB</td>
<td>1536</td>
<td>128/256</td>
<td>16</td>
<td>0-3</td>
</tr>
<tr>
<td>961</td>
<td>16-core</td>
<td>2,304 TB</td>
<td>1536</td>
<td>512/1024</td>
<td>16</td>
<td>0-3</td>
</tr>
<tr>
<td>First Expansion Frame</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96E</td>
<td>N/A</td>
<td>504 TB</td>
<td>336</td>
<td>N/A</td>
<td>8</td>
<td>N/A</td>
</tr>
<tr>
<td>Second/Third Expansion Frame</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96E</td>
<td>N/A</td>
<td>720 TB</td>
<td>480</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
SSD History

- 3.5” fiber channel Solid State Disks available for the DS8300 and DS8700
- 2.5” SSA-2 Solid State Disks available for the DS8800
- Easy Tier
- DS8870 improves SSD performance
- Easy Tier – 5th generation
- 40% SSD list price reduction
- Ultra SSD
Easy Tier Server boosts transaction dramatically

Up to $5x$ performance increase for DB2 banking brokerage workload

Base configuration is all-HDD with Easy Tier Server not activated
IBM Power 770 server running AIX with 1 Ultra SSD I/O Drawer
DS8870 146GB 15K drives (RAID 5) with 2 1.3TB database volumes
IBM high-end flash storage

DS8870
Enterprise Storage for Critical Applications

Superior Reliability

Exceptional Performance

Unique Server Integration

Efficiency & Optimization
DS8000 flash evolution
Combining flash optimization plus high-end capabilities
Performance-Cost benefits of flash will replace enterprise usage of spinning drives over time.
DS8800 DB2 Random I/O
short seeks, no cache hits

Response Time (milliseconds)

SSD, 1 DA port
15K, 3 ranks
15K, 12 ranks
15K, 21 ranks
10K, 3 ranks
10K, 12 ranks
10K, 21 ranks
All-flash benefits for transactional (OLTP) workload

Same usable capacity but with...

- 70% reduction in response time
- 80% reduction in drive count
- 41% reduction in raw capacity
- 62% reduction in energy usage
- 33% reduction in floor space

Comparing all-flash DS8870 with all-HDD system boost performance and reduces costs with equivalent $/GB

Source: internal IBM lab measurements
Additional flash news

• 40% list price reduction on all DS8870 SSDs
DS8870 A-frame

- Up to 240 SAS-2 drives in the A-frame

Statement of Direction:
- Up to 4 Ultra SSD drawers connect directly into available PCIe slots
- Each drawer contains 30 SSD drives
- Each drive has 400 GB capacity
- PCIe provides substantial performance improvement
zHPF History

2009

- DS8100/DS8300 with R4.1 or above z10 processor
- Format writes, multi-domain I/O
- QSAM/BSAM exploitation
- Incorrect Length Facility
- z/OS R4.1 or above, EXCPVPR
- Multi-track, but <= 64K
- Media manager exploitation
- z/OS R8 and above
- Multi-track any size

2010

- z196 processor >64K transfers
- IMS, Sort, DSS, etc…
- ISV Exploitation
- z/OS R11 and above, EXCPVPR
- Format writes, multi-domain I/O
- QSAM/BSAM exploitation
- Incorrect Length Facility
- z/OS R11 and above, EXCPVPR
- Multi-track any size

2011

- DS8700/DS8800 with R6.2
- z196 FICON Express 8S

2012

- 100% of DB2 I/O is now converted to zHPF
- IMS, Sort, DSS, etc…
- ISV Exploitation
- EXCP/EXCPVPR Support
zHPF format writes

- Only IBM storage supports zHPF list prefetch

This chart assumes that VPSIZE x VPSEQT is at least 400MB

- Format writes are critical to the performance of LOAD, REORG, REBUILD, RECOVER, DSN1COPY utilities
  - zHPF is especially important for a small page size
zHPF list prefetch

- Only IBM storage supports zHPF list prefetch
Only IBM storage with zHPF can optimize the performance of index scans

zHPF is especially important with SSD
DB2 RID List Scans

- Only IBM storage with zHPF can optimize the performance of RID list scans
- zHPF is especially important with SSD
- Volumes at the middle circumference perform better than inner and outer volumes due to lower seek distances.
- Clustering the data achieves 40% lower response time and 60% higher throughput.
How do I know which are the inner volumes and which are the outer volumes?

If each DS8000 extent pool has one LCU (Logical Control Unit), the volumes in each LCU are allocated from the outer cylinders to the inner cylinders.
HyperPAV

- Principles of PAV
  - Two I/Os to the same volume might not contend with each other
  - Two I/Os to different volumes on the same HDD might contend with each other
- If you define enough PAV aliases and you have sufficient channels, and if you use HyperPAV, the only physical queuing is on the HDD
- The size and number of volumes is irrelevant
How many PAVs do you need?

• With HyperPAV, you *need* roughly 3 or 4 PAVs for each physical disk
  • Example: if you have 240 disks, you need 720 to 960 PAVs
  • Double this to be on the safe side

• Since an LCU is limited to 256 addresses, carving the disks into too many volumes limits the number of PAVs that you can define

• With sufficient PAVs, you will not have IOSQ time unless your physical hardware is saturated
RMF: I/O Queue Activity

<table>
<thead>
<tr>
<th>LCU</th>
<th>CU</th>
<th>DCM GROUP</th>
<th>CHAN</th>
<th>CHPID</th>
<th>% DP</th>
<th>% CU</th>
<th>CUB</th>
<th>CMR</th>
<th>CONTENTION</th>
<th>Q</th>
<th>CSS</th>
<th>HPAP</th>
<th>OPEN</th>
<th>XFER</th>
<th>AVG DELAY</th>
<th>AVG</th>
<th>AVG DATA AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0024</td>
<td>7645</td>
<td>5C</td>
<td>256.38</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.0</td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.54</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>5D</td>
<td>256.38</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5E</td>
<td>256.37</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5F</td>
<td>256.38</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0</td>
<td>256.39</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>256.37</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>256.38</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>256.38</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>2051.0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- HPAV WAIT/MAX shows if your PAV aliases are under-configured or over-configured
DS8000 Extent Pools

Extent pool 1
- LOGCOPY1 and BSDS1
- Database 1 and Image Copy Pool 2

Extent pool 2
- LOGCOPY2 and BSDS2
- Database 2 and Image Copy Pool 1
## DS8000 Extent Pools and Logical Control Units

<table>
<thead>
<tr>
<th>LCU</th>
<th>DS8000 Extent Pool ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>00A0</td>
<td>0</td>
</tr>
<tr>
<td>00A2</td>
<td>2</td>
</tr>
<tr>
<td>00A4</td>
<td>4</td>
</tr>
<tr>
<td>00A6</td>
<td>6</td>
</tr>
<tr>
<td>00A8</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LCU</th>
<th>DS8000 Extent Pool ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>00A1</td>
<td>1</td>
</tr>
<tr>
<td>00A3</td>
<td>3</td>
</tr>
<tr>
<td>00A5</td>
<td>5</td>
</tr>
<tr>
<td>00A7</td>
<td>7</td>
</tr>
<tr>
<td>00A9</td>
<td>9</td>
</tr>
</tbody>
</table>

- Map LCUs and DS8000 extent pools one-to-one
- Even numbered LCUs use one DS8000 cluster and odd numbered LCUs use the other cluster
Configuring DS8000 Storage Extent Pools

• A DS8000 extent pool is the smallest “single point of failure”
  • A DS8000 requires a minimum of two extent pools

• Using Rotate Extents (default), each volume is striped across all of the disks in the extent pool

• If there are only two extent pools, then you have very little flexibility for managing performance and reliability

• Once your extent pools are configured, it is nearly impossible to reconfigure them
How to configure a DS8800 – A Case Study
The single frame DS8800 and DS8870 that SVL uses to evaluate OLTP performance of DB2 for z/OS

- 240 x 15K RPM disks, 300GB each, RAID 5
  - ~53TB capacity
- 8 Channels, 8 Host Adapters
- 10 Extent Pools using Rotate Extent
- 10 Logical Control Units (1 per Extent Pool)
- 32 3390-A volumes (EAV) per LCU
  - 193,662 (174x1113) cylinders per volume
  - 320 volumes total
- 128 PAV aliases per LCU (twice as many as needed)
  - 1280 PAVs total
- HyperPAV enabled
- zHPF enabled (High Performance FICON)
- FlashCopy enabled
### E S S L I N K  S T A T I S T I C S

- **SERIAL NUMBER**: 00000AHM51
- **TYPE-MODEL**: 002107-951
- **CDATE**: 04/23/2013
- **CTIME**: 15.21.33
- **CINT**: 02.13

#### READ OPERATIONS

<table>
<thead>
<tr>
<th>ID</th>
<th>TYPE</th>
<th>RRID</th>
<th>OPS</th>
<th>BYTES</th>
<th>BYTES</th>
<th>RTIME</th>
<th>OPS</th>
<th>BYTES</th>
<th>BYTES</th>
<th>RTIME</th>
<th>---ARRAY---</th>
<th>MIN</th>
<th>RANK</th>
<th>RAID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0002</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0004</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1</td>
<td>7</td>
<td>15</td>
<td>RAID 5</td>
</tr>
<tr>
<td>POOL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>19</td>
<td>15</td>
<td>5700G RAID 5</td>
</tr>
<tr>
<td>0001</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0003</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0005</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1</td>
<td>7</td>
<td>15</td>
<td>2100G RAID 5</td>
</tr>
<tr>
<td>POOL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>19</td>
<td>15</td>
<td>5700G RAID 5</td>
</tr>
<tr>
<td>0004</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0006</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0008</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0009</td>
<td>CKD</td>
<td>1Gb</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

- 30 ranks of 15K RPM 300GB disks divided into 10 extent pools divided among 4 DA loops.
- 3 of the loops are connected to 6 ranks each and the other loop is connected to 12 ranks.
- Each loop has 4 spares (for a total of 16 spares out of 30 ranks). Pools 6 and 7 have no spares.
- Rotate Extents was used to stripe volumes across the ranks in an extent pool.
Extended Address Volumes

- EAV helps relieve constraints to address large capacity needs

- Extended Address Volumes (EAV) enables volumes of more than 65,280 cylinders
  - 223 GB volumes initially supported on z/OS V1.10* and IBM System Storage DS8000
  - IBM storage now supports up to 1TB volumes with z/OS V1.13 and IBM System Storage DS8000 R6.2

- EAV can help simplify storage management.
  - Manage fewer, large volumes as opposed to many small volumes.
  - Avoid multi-volume data sets

- EAV can improve performance

- DS8000 Dynamic Volume Expansion can allow
  - Non-disruptive migration to larger volume sizes
EAV news

- Cheryl Watson’s Tuning Letter, 2013 No. 2, [www.watsonwalker.com](http://www.watsonwalker.com)
- Implementation considerations
  - You need the correct driver level on the DS8000 (driver 7.6.2)
  - Review the required APARs and fixes for both IBM and ISVs
  - Review SHARE presentation 3204 before implementing
  - Review IBM publication: z/OS 1.13 (SC26-7473) – DFSMS Using the New Functions
  - Review the dependencies, coexistence/migration considerations
  - EAV migration tracking facility output (this identifies applications that might fail when data sets are on EAVs)
  - Configure EAV on your DS8000
  - Add EAV to your storage group/pools
  - Enable the use of EAV on your system (IGDSMSxx PARMLIB member, change the default USEEAV(NO) to YES)
  - Migrate data
Reason why I like large volumes

- Too many volumes makes the amount of volume level statistics too voluminous
- I don’t want to waste UCB addresses
- I don’t like multi-volume data sets
  - Spreading the data on many volumes increases the likelihood of increased seek distances
How many volumes do you need?

- A few hundred to a few thousand
- Enough to be able to make volume level performance statistics useful
  - Not so much as to make the volume of statistics overwhelming
- Enough to provide flexible re-assignment of space to different SMS storage groups (which are usually associated with different applications)
- Do not assign so many small volumes to an SMS storage group that it will impact the CPU cost of SMS allocation
  - Nor should you define too many SMS storage groups that necessitate complex ACS routines
Volume sizes

- Migrating to EAV has its challenges, but on an IBM control unit there is no reason not to use the maximum non-EAV size that is a multiple of the DS8000 extent size (1113 cylinders)
  - $58 \times 1113 = 64554$ cylinders = 60GB
- On a DS8000, if the volume size is not a multiple of extent size, part of the extent will be wasted
Volume Sizes.....

• With HyperPAV, you only need 5 I/O addresses for each physical disk to enable the disks to become 80% busy
  • Example: With 240 disks, you need 1200 I/O addresses
• Example: Suppose an LCU consists of 24 disks
  • Example 1a) The LCU has 64 volumes with N cylinders each
    • Since you can only have 196 PAVs, even if the I/Os are evenly distributed across all 64 volumes, you will incur IOSQ time before causing the disks to become 80% busy
  • Example 1b) The LCU has only 16 volumes of 4N cylinders each
    • With only 120 aliases, even if all of the I/Os are to a single volume, you will not have any IOSQ time unless the disks are 80% busy.

• Conclusion: Bigger volumes means better performance
FlashCopy

• System Level Backups
  • Available since DB2 V8 and subsequently enhanced
  • Uses volume level FlashCopy
  • The fastest way to backup or recover the whole DB2 system

• VSAM image copies
  • Available since DB2 10
  • Uses data set level FlashCopy
  • Enables transaction consistent image copies
  • Saves the CPU cost and channel cost of reading the data into DB2 and writing it back out to DASD
  • Faster backup and recovery than traditional image copies
References

Jeffrey Berger
IBM
bergerja@us.ibm.com

Session 1413
How Recent Storage Innovations Can Help Improve Performance and Reliability For Your DB2 Subsystem