Oracle on SSD for Performance

FusionIO and EMC SSD performance for Oracle databases

Presenter: Steve Fluge
Bank of America
Steve Fluge

- Oracle Engineering team at Bank of America (formerly Merrill Lynch)
  - Oracle builds, POCs for new technologies (SSD, etc), establish best practices for Oracle database architectures
  - 15 years experience with Oracle databases as DBA, consultant, engineer
What is the compelling argument for SSD?

PERFORMANCE!

The goal:

How SSD can improve performance of Oracle databases, especially for high demand, I/O bound environments.
Introduction

• Applications which are high demand Tier-0
  – Trading applications
  – Online data services
  – Data marts
  – Etc..

• SSDs can address Tier-0 requirements for performance
SSD storage tested

• Solid State Disk (SSD) technology
  – Tier 0 storage
  – Uses NAND based flash memory
  – Available in single level cell (SLC) and multilevel cell (MLC)
    • SLC is less dense and faster than MLC
    • MLC supports more capacity
  – Capacity
    • FusionIO cards 80G, 160G, 320G, 640G
    • EMC SSD STEC 200G (tested)
SSD storage tested

- FusionIO iodrive card
  - Connects to PCIe bus on the server
    - Limited number of cards depend on # PCI slots available
  - PCI configuration enables very high speed access
    - 100k and higher IOPS on reads and similar numbers for writes
    - Latency is in the microsecond range (vs milliseconds in HDD)
  - Configurable reserve area
    - Results in better performance for writes
    - Reduces amount of useable storage
- NOTE: Available on Windows 64bit, RHEL,SLES only
SSD storage tested

- FusionIO iodrive specs

<table>
<thead>
<tr>
<th>ioDrive capacity</th>
<th>80GB</th>
<th>160GB</th>
<th>320GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAND Type</td>
<td>Single Level Cell (SLC)</td>
<td>Single Level Cell (SLC)</td>
<td>Multi Level Cell (MLC)</td>
</tr>
<tr>
<td>Write Bandwidth</td>
<td>500 MB/s (32K packet size)</td>
<td>670 MB/s (32K packet size)</td>
<td>490 MB/s (64K packet size)</td>
</tr>
<tr>
<td>Read Bandwidth</td>
<td>750 MB/s (32K packet size)</td>
<td>750 MB/s (32K packet size)</td>
<td>700 MB/s (64K packet size)</td>
</tr>
<tr>
<td>IOPS*</td>
<td>119,790 (4K read packet size)</td>
<td>116,046 (4k read packet size)</td>
<td>71,256 (4K read packet size)</td>
</tr>
<tr>
<td></td>
<td>89,549 (75/25 r/w mix 4k packet size)</td>
<td>93,199 (75/25 r/w mix 4k packet size)</td>
<td>67,659 (75/25 r/w mix 4k packet size)</td>
</tr>
<tr>
<td>Access Latency</td>
<td>50μs Read</td>
<td>50μs Read</td>
<td>80μs Read</td>
</tr>
<tr>
<td>Operating Systems</td>
<td>Microsoft &amp; Solaris 10/11, RHEL 4 &amp; 5; SLES 10 &amp; 11</td>
<td>Microsoft &amp; Solaris 10/11, RHEL 4 &amp; 5; SLES 10 &amp; 11</td>
<td>Microsoft &amp; Solaris 10/11, RHEL 4 &amp; 5; SLES 10 &amp; 11</td>
</tr>
</tbody>
</table>

*160GB tested
SSD storage tested

- FusionIO drives are configurable
- Command line utilities include:
  - Fio-status: status of fio devices
  - Fio-format: format usable storage area of device
  - Fio-attach: attach device
  - Fio-detach: detach device
  - Fio-update-iodrive: update drive software
(more)
SSD storage tested

- EMC SSD STEC ST0200
  - SAN Array based storage device
    - Symmetrix DMX4 (tested configuration)
    - Can have lots of drives up to 32 per quadrant
    - Can be used in SRDF configuration
    - Device capacity tested is 200GB
SSD storage tested

- EMC STEC specs

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zeus SSD capacity</td>
<td>Up to 512GB</td>
</tr>
<tr>
<td>NAND Type</td>
<td>SLC/MLC</td>
</tr>
<tr>
<td>Write Bandwidth</td>
<td>115MB/sec</td>
</tr>
<tr>
<td>Read Bandwidth</td>
<td>220MB/sec</td>
</tr>
</tbody>
</table>
| IOPS*                 | Random Reads 45,000/sec
                       | Random Writes 16,000/sec |
| Transfer Rate         | FC 4G/sec (dual port) |
                       | SAS 3G/sec (single port) |
                       | SATA 3G/sec (single port) |
| Interfaces            | FC/SATA/SAS          |

*200GB FC tested
Test configuration

• Platform
  – Dell 24 core Nehalem 32GB memory, 2G HBA cards
  – Linux RHEL5

• Oracle11gR1 with ASM

• Tools
  – Oracle IO Numbers : Orion
    • Simulates Oracle RBMS disk I/O usage and records performance data
  – Benchmark Factory
    • TPC-C and Hardware Scalability tests
Test configuration (cont)

• Orion
  – Oracle database IO simulation tool
  – I/O workload options configured
    • Small IO size = 8kb
    • Large IO size = 1024kb
    • Storage Array simulated type = Raid 0
    • Cache size = 90 GB (SAN Array cache size accounted for)
    • Stripe depth = 1024kb
    • IO types tested = Small Random IOs, Large Random IOs
    • Write = 0 for read intensive tests, write=100 for write intensive test
Test configuration (cont)

- Benchmark Factory
  Workload replay and scalability test tool from Quest includes standard industry benchmarks
    - TPC-C standard benchmark test (OLTP)
    - Scalable Hardware benchmark test
    - Test load up to 800 concurrent users max load
      - 100GB database placed on SSD
      - ASM used with 2 disk groups
Test configuration (cont)

• Sample operational timed tests
  – RMAN backup/restore
  – Exp/Imp
  – Index create
Test Results - Orion

- Graphs will show significant performance increase for SSD technology compared to SAN HDD storage for data transfer and IOPS rates
- Most notable results are Read Intensive test results for large random reads.
- FusionIO cards optimized for random write performance by increasing reserve area
FusionIO shows highest transfer rate 700MBPS
SANSSD performance is limited due to the HBA card speed (2Gbps)
Read intensive test Results
Small sequential reads IOPS

FusionIO shows highest transfer rate 39,000 IOPS
SANSSD performance is limited due to the HBA card speed (2Gbps)
Read intensive test Results
Large Random reads - Latency
Write intensive test Results
Large random writes MBPS

FusionIO shows highest transfer rate 470MBPS
SANSSD performance is limited due to the HBA card speed (2Gbps)
Write intensive test Results
Small random writes IOPS

FusionIO shows highest transfer rate 43,000 IOPS
SANSSD performance is limited due to the HBA card speed (2Gbps)
Test Results for Benchmark Factory

- Graphs will show major difference between SSD results and SAN HDD results
- Includes test data from prior performance tests for other platforms for comparison
Test results – Benchmark Factory
TPC-C comparative results

Oracle TPC-C 400 users

Platform Configuration

<table>
<thead>
<tr>
<th>Platform Configuration</th>
<th>TPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP BL680C</td>
<td>500</td>
</tr>
<tr>
<td>HP BL680c &amp; NAS</td>
<td>700</td>
</tr>
<tr>
<td>IBM 3850</td>
<td>2400</td>
</tr>
<tr>
<td>IBM 3950 NUMA</td>
<td>3400</td>
</tr>
<tr>
<td>Dell &amp; SAN HDD</td>
<td>3800</td>
</tr>
<tr>
<td>Dell &amp; EMC SSD</td>
<td>3900</td>
</tr>
<tr>
<td>Dell &amp; FusionIO SSD</td>
<td>3900</td>
</tr>
</tbody>
</table>
Test results – Benchmark Factory
Scalable Hardware comparative results

Scalable Hardware Read Intensive - 800 users

Platform Configuration

TPS

IBM 3950 NUMA
HP BL680C
IBM 3850
Dell & SAN HDD
Dell & FusionIO SSD
Dell & SAN SSD
Test results – Benchmark Factory
Scalable Hardware comparative results

Scalable Hardware Insert Intensive - 800 users

<table>
<thead>
<tr>
<th>Platform Configuration</th>
<th>TPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP BL680C</td>
<td>12000</td>
</tr>
<tr>
<td>IBM 3850</td>
<td>15000</td>
</tr>
<tr>
<td>IBM 3950 NUMA</td>
<td>17000</td>
</tr>
<tr>
<td>Dell &amp; SAN HDD</td>
<td>20000</td>
</tr>
<tr>
<td>Dell &amp; SAN SSD</td>
<td>38000</td>
</tr>
<tr>
<td>Dell &amp; FusionIO SSD</td>
<td>39000</td>
</tr>
</tbody>
</table>
SSD reserve area and write operations

Question:
Why are the write operations so much slower than read operations?

Answer:
Write intensive operations need to do this when pages are updated:
1. Copy data to reserve area
2. Erase the page(s)
3. Copy the original data plus new data back to the original page(s)
Configuring FusionIO card Reserve area

- SSD devices can experience degraded write performance over time
- Most SSD devices have reserve area, typically 10%
- FusionIO SSD can be configured to increase performance of high random write activity
  - The tradeoff is less available storage

Use a command line utility to format the drive.

Steps:

1. `fio-detach /dev/fct0`
2. `fio-format –s 100G /dev/fct0`
3. `fio-attach /dev/fct0`
Configuring FusionIO card
Reserve area

Example format session:
[root]# fio-detach /dev/fct0
Detaching: [====================] (100%) /
[root]# fio-format -s 100G /dev/fct0
WARNING: formatting will destroy any existing data on the device!
Do you wish to continue [y/n]? y
data channel: geometry: 4096x512x189056 (25 pads, 2 planes, 4 banks)
Creating a device of size 100.00GiB (107.37GB)
Formatting: [====================] (100%) -
Format successful.
[root@]# fio-attach /dev/fct0
Attaching: [====================] (100%) -
fioa
The following chart shows the relationship between write intensive data transfer rates and the amount of storage in the available (non-reserved) area of FusionIO card (160GB card)
Sample database operations tested and timed. For Export, Import, RMAN backup and restore, disk storage was used to hold the dump files and RMAN backups.

<table>
<thead>
<tr>
<th>Operation</th>
<th>FusionIO</th>
<th>SANSSD</th>
<th>SAN RAID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Export</td>
<td>65 minutes</td>
<td>70 minutes</td>
<td>80 minutes</td>
</tr>
<tr>
<td>Import</td>
<td>40 minutes</td>
<td>60 minutes</td>
<td>120 minutes</td>
</tr>
<tr>
<td>Index creation</td>
<td>20 minutes</td>
<td>30 minutes</td>
<td>45 minutes</td>
</tr>
<tr>
<td>RMAN backup</td>
<td>30 minutes</td>
<td>40 minutes</td>
<td>60 minutes</td>
</tr>
<tr>
<td>RMAN restore</td>
<td>65 minutes</td>
<td>90 minutes</td>
<td>130 minutes</td>
</tr>
</tbody>
</table>
Oracle databases can either be wholly on SSD or have some I/O intensive components placed there.

Use AWR report information to identify objects with the most physical reads/writes.

In addition, candidates for locating to SSD:

- Temporary tablespace
- Undo tablespace
- Redo logs
- Flash recovery area
Database components for SSD

- FusionIO tests
  - Temporary and Undo tablespaces placed on SSD showed 20% improvement in performance for index creation
  - Application Data load. Placing application schema on SSD showed 60% improvement over the original load time
Conclusions

• SSD performance makes it a prime candidate for Tier-0 storage requirements
• FusionIO is shown to perform best based on lab tests for all categories
• For write intensive databases tune FusionIO for better sustained performance (increase reserve area)
• When configuring SAN SSD consider all SAN components to avoid bottlenecks (HBA cards for example)
• Optimal configuration for SSD and Oracle should include identifying objects related to I/O bottlenecks in the database and relocating those to SSD
Conclusions

- Another key to storage in addition to performance is HA capability
  - EMC SSD has advantages in that it is external storage which is critical in an HA architecture
  - FusionIO storage is local, however it can use Oracle DataGuard to address HA requirements
Questions?
Thank you!

Steve Fluge
steve_fluge@ml.com