Achieving Memory Level Performance: Secrets Beyond Shared Flash

Kothanda (Kodi) Umamageswaran
Vice President, Exadata Development

Gurmeet Goindi
Exadata Product Management
Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.
Did You Miss the Storage Revolution?
Good Chance Your Storage Vendor Did Too

• Incumbent storage vendors have decades old investment in legacy protocols keeping them from adopting new technologies

• PCIe Flash with NVMe interface is a new interface that realizes full flash potential

• PCIe/NVMe storage architectures are orders of magnitude faster than what you probably use today

• Available now with Oracle Exadata and SuperCluster
Solid State Media is Very Different Than Spinning Disk

• Compared to Spinning Disk, Flash
  – Is many orders of magnitude faster
  – Has many orders of magnitude higher bandwidth
  – Has extremely low latency
  – Has wearing issues as it ages, but technology is catching up
  – Is expensive, but the price gap is shrinking

• Every storage vendor has some flash based solution for your Database

Q: Will my database realize the full benefit of flash technology?
A: It will depend on how fast you can move the data from the flash to the database
SCSI Access Model

- SCSI was designed for tapes and HDDs
- HDDs are sequential whereas Flash devices are massively parallel
- Traditional IO stack is optimized for spinning media
  - 512 Byte block size transfers
  - Flash and databases do 4KB/8KB IOs
- Using legacy interfaces like **SCSI** fundamentally bottlenecks flash drives
PCI Express Vs SAS Connectivity

• PCI Express is orders of magnitude faster than SAS, and is getting faster
• PCI Express has the same characteristics as Flash
  – High Throughput
  – Low Latency
• Using legacy interconnects like SAS fundamentally bottlenecks flash drives

PCIe has 13x throughput of SAS

![Throughput comparison chart]

- SAS 6 Gbps: 0.6 GB/s
- SAS 12 Gbps: 1.2 GB/s
- PCIe 3.0 x4: 4 GB/s
- PCIe 3.0 x8: 8 GB/s
PCI Express Flash with NVMe Interface

- Non Volatile Memory Express is a brand new grounds up interface designed for flash
- NVMe is inherently parallel
- NVMe provides native atomic IO size affinity for databases
- NVMe IO stack massively reduces CPU utilization and latency

PCI Express Flash with NVMe Interface is the right choice for your Database
Exadata is Leading NVMe Adoption

Thousands of Exadata and SuperCluster systems shipped with NVMe Flash since 2014

1st NVMe Drive by Samsung

Exadata X5-2
Industry’s First Enterprise System with NVMe

Facebook launches
Lightning based on NVMe

Exadata X6-2
2nd Generation NVMe with 3DV NAND

Exadata X7
3rd Generation NVMe

1st NVMe Drive by Intel

Exadata Cloud Service uses NVMe in Public Cloud

EMC Announces DSSD D5 with NVMe

EMC Closes DSSD division

Copyright © 2017, Oracle and/or its affiliates. All rights reserved.
Shared Storage Has Many Advantages over Local Storage

- Much better space utilization
- Much better security, management, reliability
- Enables DB consolidation, DB high availability, RAC scale-out
- Shares storage performance
  - Aggregate performance of shared storage can be dynamically used by any server that needs it
NVMe PCI-e Flash Disrupts the Storage Array Model
New improvements are causing **100X bottlenecks** across shared storage stack

Latest PCIe Flash
5.5 GB/sec

SAN Link = 40Gb
5 GB/sec
Less than 1 Flash card

Leading All Flash Array
38 GB/sec
Less than 5 Flash card

All-Flash Storage Array IO Path: many steps, each adds **latency** and creates **bottlenecks**
Exadata Achieves Memory Performance with Shared Flash

- **Exadata X7 delivers 350GB/sec flash bandwidth** to any server
  - Approaches 800GB/sec aggregate DRAM bandwidth of DB servers

- **Must move compute to data to achieve full flash potential**
  - Requires owning full stack, can’t be solved in storage alone

- **Fundamentally, Storage Arrays can share flash capacity but not flash performance**
  - Even with next gen scale-out, PCIe networks, or NVMe over fabric

- **Shared storage with memory level bandwidth** is a paradigm change in the industry
  - Get near DRAM throughput, with the capacity of shared flash
Getting Memory performance with Shared Flash using Smart Software
Oracle’s Infrastructure Innovations in Flash

- Oracle Exadata V2: First to bring flash storage to the database market
- Oracle Exadata X3: Doubled flash capacity
- Oracle Exadata X4: 100GB/s throughput scans in a single rack
- Oracle Exadata X5: Lowest latency NVMe and increases scans to 263GB/s
- Oracle Exadata X5: Hot-pluggable NVMe server for the database
- Oracle Linux: First Linux vendor with production NVMe drivers
- Oracle Exadata X6: Highest throughput over 300GB/s, over 5 Million IOPs
- Oracle Exadata X7: Highest throughput over 350GB/s, nearly 6 Million IOPs
Oracle’s Software Innovations in Flash

- Exadata Smart Flash Cache
- Exadata Smart Flash Log
- Exadata Smart Flash Cache Scan Awareness
- Exadata Smart File Initialization
- Exadata Columnar Flash Cache
- Exadata Flash Cache Space Resource Management
- Smart write burst and temp IO in Flash Cache
Exadata Smart Flash Cache – Completely Automatic

- Understands different types of I/Os from database
  - Skips caching I/Os to backups, data pump I/O, archive logs, tablespace formatting
  - Caches Control File Reads and Writes, file headers, data and index blocks
  - Enables more space for relevant user data

- Immediately adapts to changing workloads

- Write-back flash cache
  - Caches writes from the database not just reads

- Doesn’t need to mirror in flash for read intensive workloads
  - Flash arrays store both mirror copies always in flash increasing your cost

- Smart Scans can run at the throughput of flash drives
  - Flash arrays need lots of servers with lots of processes and still cannot match Smart Scan throughput of single query

- Provides performance of flash at cost of disk
Exadata Smart Flash Log

- Outliers in log IO slow down lots of clients
- Outliers from any one copy of mirror slow down all the foregrounds
  - Database wait time goes up by #foregrounds * Stall time
  - Backlog doesn’t clear immediately like an accident on the freeway and increases “log file sync” waits
- Performance critical algorithms like space management and index splits are sensitive to log write latency
- Legacy storage IO cannot differentiate redo log IO from others
- UPS protected cache in traditional storage seems to work initially until the cache is overwhelmed by other writes
  - Measure log file latency with full backup or a data load running
Exadata Smart Flash Log – Completely Automatic

• Smart Flash Log uses flash as a parallel write cache to disk controller cache
• Whichever write completes first wins (disk or flash)
• Reduces response time and outliers
  – “log file parallel write” histogram improves
  – Greatly improves “log file sync”
• Uses almost no flash capacity (< 0.1%)
• Network resource management provides priority for redo log I/Os across the network
• OLTP workloads transparently accelerated and provide predictable response times
Exadata Columnar Flash Cache – Completely Automatic

- Hybrid Columnar Compression balances need for OLTP and Analytics
- As CPUs get faster want even faster scans
- Smart Flash Cache automatically transforms blocks from hybrid columnar to pure columnar for analytics during flashcache population
- Dual format representation for single row lookups
- Only selected columns read from flash during a query
- Up to 5x query speedup
Flash Cache Space Resource Management

- Flash Cache is a shared resource
- Database as a Service creates need for efficient resource sharing
- Specify minimum (flashCacheMin) and maximum (flashCacheLimit) sizes, or fixed allocations (flashCacheSize), a database can use in the flash cache
  
  ```sql
  ALTER IORMPLAN -
  dbplan=((name=sales, flashCacheSize=100G), -
  (name=finance,flashCacheLimit=100G, flashCacheMin=20G), -
  (name=schain, flashCacheSize=200G))
  ```

- Container database resource specified at the storage
- Pluggable database container resource limits expressed as percentages in the container database
- Database and Pluggable database I/O resource management is unique to Exadata
- Predictable performance for database queries – no more noisy neighbor
Write bursts and temp IO in flash cache – Completely Automatic

• Write throughput of four flash cards has become greater than the write throughput of 12-disks

• When database write throughput exceeds the throughput of disks, smart flash cache intelligently caches writes
  – Schema changes during application upgrades rewrite entire tables in some packaged applications
  – Large database consolidations can have write bursts at the same time

• When queries write a lot of temp IO and it is bottlenecked on disk, smart flash cache intelligently caches temp IO
  – Writes to flash for temp spill reduces elapsed time
  – Reads from flash for temp reduces elapsed time further

• Smart to prioritize OLTP data and does not remove hot OLTP lines from the cache

• Smart flash wear management for large writes

• Much faster scans and disk writes
Exadata Smart Flash Benefits

• Automatic Database Aware Flash Cache
• Smart Flash Logging avoids redo log outliers automatically
• Smart Flash Cache Scan provides subset scanning and is table scan resistant
• Smart File Initialization creates a file by writing meta-data to flash cache
• Smart Columnar Flash Cache extends columnar benefit to storage automatically
• Smart Flash Cache Space Resource Management provides granular control
• Smart write burst and temp IO in Flash Cache
The Next Big Thing:

In-Memory Performance in Storage
Analytics: Exadata Brings In-Memory Analytics to Storage

• With Exadata Flash throughput approaching memory throughput, SQL bottleneck moves from I/O to CPU

• Exadata automatically transforms table data into In-Memory DB columnar formats in Exadata Flash cache
  – Enables fast vector processing for storage server queries

• **Uniquely** optimizes next generation Flash as memory
  – Works for both row format OLTP databases, and Hybrid Columnar Compressed Analytics databases
In-Memory Columnar Formats in Flash Cache (12.2.1.1.0)

3 - 4x Overall Analytics Performance Improvement

25.6 TB Flash x 3 = 76.8 TB (or more)

IMC (In-Memory Columnar) data

HCC/OLTP compressed/Uncompressed Data

Copyright © 2017, Oracle and/or its affiliates. All rights reserved.
OLTP: Exadata Brings In-Memory OLTP to Storage

- Exadata Storage Servers add a memory cache in front of Flash memory
  - Similar to current Flash cache in front of disk

- Cache is additive with cache at Database Server
  - Only possible because of tight integration with Database

- **2.5x Lower latency for OLTP IO** – 100 usec

- Up to **21 TB of DRAM for OLTP acceleration** with Memory Upgrade Kit
  - Compare to 5TB of flash in V2 Exadata
**In-Memory OLTP Acceleration – Journey of a Database Block**

1. **DB reads a block**

   - **In-Memory OLTP Cache**
   - **Flash Cache**
   - **Hard Disk Drive**

   Data initially resides on hard disk

   Exadata Serves the Block from Storage

---

**Oracle Confidential – Internal**

Copyright © 2015, Oracle and/or its affiliates. All rights reserved.
In-Memory OLTP Acceleration – Journey of a Database Block

2. Flash Cache Gets Populated
In-Memory OLTP Acceleration – Journey of a Database Block

1. Database Server
2. Storage Server

3. Database evicts the block

- DB Buffer Cache
- In-Memory OLTP Cache
- Flash Cache
- Hard Disk Drive

Exadata Caches the block in In-Memory OLTP Cache
In-Memory OLTP Acceleration – Journey of a Database Block

1. Database writes a block
2. Block is written to the In-Memory OLTP Cache
3. In-Memory OLTP Cache writes block to Flash Cache
4. Database reads the same block again
   - Exadata serves the block from In-Memory OLTP Cache with 100us latency

Database Server

Storage Server

DB Buffer Cache
In-Memory OLTP Cache
Flash Cache
Hard Disk Drive
In-Memory OLTP Acceleration

Data is never in DB Buffer Cache or In Memory OLTP Cache at the same time
Cache is **Additive** with Cache at Database Server

- Blocks in buffer cache will not be cached in Storage Server In-Memory OLTP Cache
  - Client read hits in Storage Server In-Memory OLTP Cache will evict the blocks from the cache
  - Client read misses in Storage Server In-Memory OLTP Cache will populate flash cache, but not Storage Server In-Memory OLTP Cache

- Blocks evicted from buffer cache globally will be populated into Storage Server In-Memory OLTP Cache
  - Storage Server will read the blocks from flash cache and populate into In-Memory OLTP Cache

- Elimination of Context Switches Further Reduces Latency
  - 100usec read latency
Identifying Workloads for In-Memory OLTP Acceleration

**Buffer Pool Advisory**

- Only rows with estimated physical reads > 0 are displayed
- Ordered by Block Size, Buffers For Estimate

<table>
<thead>
<tr>
<th>P</th>
<th>Size for Est (M)</th>
<th>Size Factor</th>
<th>Buffers (thousands)</th>
<th>Est Phys Read Factor</th>
<th>Estimated Phys Reads (thousands)</th>
<th>Est Phys Read Time</th>
<th>Est %DBTime for Rds</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>1,600</td>
<td>0.10</td>
<td>189</td>
<td>3.13</td>
<td>197,019</td>
<td>1</td>
<td>93139.00</td>
</tr>
<tr>
<td>D</td>
<td>3,200</td>
<td>0.19</td>
<td>377</td>
<td>2.51</td>
<td>158,033</td>
<td>1</td>
<td>74601.00</td>
</tr>
<tr>
<td>D</td>
<td>4,600</td>
<td>0.23</td>
<td>566</td>
<td>2.21</td>
<td>138,174</td>
<td>1</td>
<td>65646.00</td>
</tr>
<tr>
<td>D</td>
<td>6,400</td>
<td>0.33</td>
<td>754</td>
<td>1.96</td>
<td>124,414</td>
<td>1</td>
<td>50637.00</td>
</tr>
<tr>
<td>D</td>
<td>8,000</td>
<td>0.49</td>
<td>943</td>
<td>1.77</td>
<td>111,227</td>
<td>1</td>
<td>52375.00</td>
</tr>
<tr>
<td>D</td>
<td>9,600</td>
<td>0.58</td>
<td>1,131</td>
<td>1.60</td>
<td>100,589</td>
<td>1</td>
<td>47324.00</td>
</tr>
<tr>
<td>D</td>
<td>11,200</td>
<td>0.68</td>
<td>1,329</td>
<td>1.48</td>
<td>92,883</td>
<td>1</td>
<td>43665.00</td>
</tr>
<tr>
<td>D</td>
<td>12,800</td>
<td>0.78</td>
<td>1,598</td>
<td>1.32</td>
<td>83,466</td>
<td>1</td>
<td>30165.00</td>
</tr>
<tr>
<td>D</td>
<td>14,400</td>
<td>0.88</td>
<td>1,897</td>
<td>1.16</td>
<td>74,193</td>
<td>1</td>
<td>34790.00</td>
</tr>
<tr>
<td>D</td>
<td>16,000</td>
<td>0.97</td>
<td>2,186</td>
<td>1.04</td>
<td>65,174</td>
<td>1</td>
<td>30507.00</td>
</tr>
<tr>
<td>D</td>
<td>17,600</td>
<td>1.07</td>
<td>2,493</td>
<td>1.00</td>
<td>62,954</td>
<td>1</td>
<td>29455.00</td>
</tr>
<tr>
<td>D</td>
<td>19,200</td>
<td>1.17</td>
<td>2,838</td>
<td>0.87</td>
<td>54,679</td>
<td>1</td>
<td>25524.00</td>
</tr>
<tr>
<td>D</td>
<td>20,800</td>
<td>1.26</td>
<td>3,251</td>
<td>0.81</td>
<td>51,182</td>
<td>1</td>
<td>23863.00</td>
</tr>
<tr>
<td>D</td>
<td>22,400</td>
<td>1.36</td>
<td>3,649</td>
<td>0.76</td>
<td>48,036</td>
<td>1</td>
<td>22370.00</td>
</tr>
<tr>
<td>D</td>
<td>24,000</td>
<td>1.46</td>
<td>2,928</td>
<td>0.72</td>
<td>45,262</td>
<td>1</td>
<td>21052.00</td>
</tr>
<tr>
<td>D</td>
<td>25,600</td>
<td>1.56</td>
<td>3,017</td>
<td>0.68</td>
<td>42,027</td>
<td>1</td>
<td>19891.00</td>
</tr>
<tr>
<td>D</td>
<td>27,200</td>
<td>1.65</td>
<td>3,205</td>
<td>0.64</td>
<td>40,304</td>
<td>1</td>
<td>10736.00</td>
</tr>
<tr>
<td>D</td>
<td>28,800</td>
<td>1.75</td>
<td>3,394</td>
<td>0.61</td>
<td>38,491</td>
<td>1</td>
<td>17837.00</td>
</tr>
<tr>
<td>D</td>
<td>30,400</td>
<td>1.85</td>
<td>3,582</td>
<td>0.59</td>
<td>36,558</td>
<td>1</td>
<td>17081.00</td>
</tr>
<tr>
<td>D</td>
<td>32,000</td>
<td>1.96</td>
<td>3,771</td>
<td>0.56</td>
<td>35,412</td>
<td>1</td>
<td>16375.00</td>
</tr>
</tbody>
</table>
A Choice of Exadata Deployment Models

On-Premises

X7-2
X7-8

Customer Data Center
Purchased
Customer Managed

Cloud at Customer

Customer Data Center
Subscription
Oracle Managed

Public Cloud Service

Oracle Cloud Subscription
Oracle Managed
Exadata Advantages Increase Every Year

Dramatically Better Performance and Cost

Smart Software
- Database Aware Flash Cache
- InfiniBand Scale-Out
- Smart Scan
- Columnar Compression
- Smart Scan
- InfiniBand Scale-Out
- IO Priorities
- Data Mining Offload
- Offload Decrypt on Scans
- Network Resource Management
- Multitenant Aware Resource Mgmt
- Prioritized File Recovery
- In-Memory Fault Tolerance
- Direct-to-wire Protocol
- Instant failure detection
- Exadata Cloud Service
- In-Memory Columnar in Flash
- Smart Fusion Block Transfer
- 3D V-NAND Flash
- Software-in-Silicon
- Tiered Disk/Flash
- DB Processors in Storage
- Unified InfiniBand
- PCIe NVMe Flash
- In-Memory OLTP Acceleration
- Hot Swappable Flash
- 25 GigE Client Network
- In-Memory OLTP Acceleration
- Hot Swappable Flash

Smart Hardware
- Scale-Out Storage
- Scale-Out Servers

Copyright © 2017, Oracle and/or its affiliates. All rights reserved.
Integrated Cloud
Applications & Platform Services