Exadata: Delivering Memory Performance with Shared Flash

Kothanda Umamageswaran
Vice President, Exadata Development

Gurmeet Goindi
Technical Product Strategist, Exadata
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

<table>
<thead>
<tr>
<th>SAS 6 Gbps</th>
<th>SAS 12 Gbps</th>
<th>PCIe 3.0 x4</th>
<th>PCIe 3.0 x8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>1.2</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

Throughput comparison chart.
Exadata is Leading NVMe Adoption

**Thousands of Exadata systems shipped with NVMe Flash since 2014**

- **2014**
  - 1st NVMe Drive by Samsung
  - 1st NVMe Drive by Intel

- **2015**
  - Exadata X5-2 Industry’s First Enterprise System with NVMe
  - Facebook launches Lightning based on NVMe

- **2016**
  - Exadata Cloud Service uses NVMe in Public Cloud
  - Exadata X6-2 Second Generation with NVMe
  - EMC Announces DSSD D5 with NVMe

---

EMC DSSD D5 with NVMe

---

Facebook launches Lightning based on NVMe

---

Exadata X6-2 Second Generation with NVMe

---

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

---

1st NVMe Drive by Intel

---

1st NVMe Drive by Samsung
New X6 Super-Capacity and Performance Flash

- 3D V-NAND 3.2TB/card (2X previous card capacity)
  - 48 layer NAND
  - No tradeoffs - faster writes, lower power, higher endurance
- Latest, most modern interface – NVMe (introduced in X5)
- Fastest flash card on market by wide margin
  - Only flash card on market with PCI 8-lane scale bandwidth ~ 5.4GB/sec
  - Highest IOs per second
  - Lowest outliers
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 PCIe Flash Disrupts the Storage Array Model

New improvements are causing **100X bottlenecks** across shared storage stack

- **Latest PCIe Flash**
  - 5.4 GB/sec

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

- **Leading All Flash Array**
  - 24 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 X6 delivers 300GB/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
What is Exadata?
# The Exadata Database Machine Vision

**Best Platform for the Oracle Database – On Premises and in the Cloud**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>State-of-the-art enterprise-grade hardware, refreshed yearly (processors, flash, disks, network)</td>
</tr>
<tr>
<td>2.</td>
<td>Sized, tuned and optimized exclusively for Oracle Database workloads (DW, Analytics, OLTP, Mixed)</td>
</tr>
<tr>
<td>3.</td>
<td>High-powered intelligent storage servers capable of offloading database workloads</td>
</tr>
<tr>
<td>4.</td>
<td>“Smart” database protocols and optimizations from servers to network to storage</td>
</tr>
<tr>
<td>5.</td>
<td>One vendor responsible for all hardware, software and customer support</td>
</tr>
</tbody>
</table>
Proven at Thousands of Critical Deployments since 2008

Half OLTP - Half Analytics - Many Mixed

- Petabyte Warehouses
- Online Financial Trading
- Business Applications
  - SAP, Oracle, Siebel, PSFT, …
- Massive DB Consolidation
- Public SaaS Clouds
  - Oracle Fusion Apps, Salesforce, SAS, …

4 of the Top 5 Banks, Telcos, Retailers run Exadata
Exadata Database Machine Family

- **Exadata X6**
  - X6-2
  - X6-8
  - On-Premises

- **Exadata Cloud Service**
  - Exadata Cloud Service @ Oracle

- **Exadata Cloud Machine**
  - Exadata Cloud Service @ Customer
Exadata Database Machine X6-2

- **Scale-Out Database Servers**
  - 2 socket x86 processors
  - 44 CPU cores
  - 256GB-1.5TB DRAM

- **Fastest Internal Fabric**
  - 40 Gb/s InfiniBand
  - Ethernet external connectivity

- **Scale-Out Intelligent Storage**
  - 12.8 TB PCI Flash
  - 96 TB disk
  - 20 CPU cores
  - 25.6 TB PCI Flash
  - 20 CPU cores

**Compute Software**
- Oracle Linux 6
- Oracle Database Enterprise Edition
- Oracle VM (optional)
- Oracle Database options (optional)

**Storage Server Software**
- Smart Scan (SQL Offload)
- Smart Flash Cache
- Hybrid Columnar Compression
- I/O Resource Management
Exadata Database Machine X6-8

• **Scale-Out Database Servers**
  - 8-socket x86 processors
  - 144 cores
  - 2-6 TB DRAM

• **Fastest Internal Fabric**
  - 40 Gb/s InfiniBand
  - Ethernet external connectivity

• **Scale-Out Intelligent Storage**
  - High-Capacity Storage Server
  - Extreme Flash Storage Server

Large SMP Processor Model
  - Large warehouses
  - Massive database consolidation
  - Big In-Memory databases

**Storage Server Software**
- Smart Scan (SQL Offload)
- Smart Flash Cache
- Hybrid Columnar Compression
- I/O Resource Management

**Same Networking, Storage and Software as X6-2**
Elastic Configurations Incrementally Scale Servers

Achieve any Level of Performance with Minimum Hardware

1. Start Small
   - 2 Database Servers
   - 3 Storage Servers

2. Incrementally add DB or Storage Servers
   - Database Server
   - Extreme Flash Storage
   - High Capacity Storage

3. Add Racks to Continue Scaling
   - Full Rack
     - Enable Database CPU cores as needed with Capacity on Demand
     - Expand older Exadata machines with new X6-2 servers
   - Multi-Rack
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 350GB/s and lowest latency
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 Smart Columnar Flash Cache
- Exadata Smart Flash Cache Space Resource Management
- **Upcoming**: Exadata Smart In Memory Formats in Flash
Exadata Smart Flash Cache

- 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
  - More space for 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
- Smart Scans can run at the throughput of flash drives
  - Compare to: flash arrays that require flash cache in the server doubling cost
  - Compare to flash arrays: 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 affect response time
- 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
- Legacy Storage UPS protected cache seems to work initially until the cache is overwhelmed by other writes
Exadata Smart Flash Log

- 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%)
- OLTP workloads transparently accelerated
Exadata Smart Flash Cache Scan Awareness

- On a traditional cache, if you scan dataset larger than cache size
  - Blocks 0,1,2,3 brought into cache, cache is full
  - Block 20,21,22,23 say replaces 0,1,2,3
- Repeat the same scan
  - Block 0,1, 2, 3 will replace blocks 20,21,22,23
  - Block 20,21,22,23 will again replace block 0,1,2,3
- Traditional caches churn with no actual benefit
- Some implementations call the insertion of new block in the middle scan resistant
Exadata Smart Flash Cache Scan Awareness

- Exadata Smart Flash Cache is scan resistant
  - Ability to bring subset of the data into cache and not churn
  - OLTP and DW scan blocks can co-exist
- Nested scans bring in repeated accesses
  - Repeat, For each item in large table, scan small table
  - Smart enough to pull the small table into flash since it is accessed repeatedly even though the size of large table alone is larger than flash cache
- No need to set “KEEP” attribute in data warehouses
- Scans automatically use flash for extreme performance
Exadata Smart File Initialization

- Combine the benefits of Smart Initialization and Writeback Flash Cache
  - Write file creation meta-data to writeback flash cache
  - Tiny amount of flash space used to cache large portions of initialized data on disk
  - Initialization I/Os to disk deferred or not performed if data loaded

- Create tablespace, file extensions, autoextend show benefit

- Redo log initialization included in Exadata 12.1.1.1.0
  - File creation sped up by over 10x
Exadata Hybrid Columnar Compression

- Hybrid Columnar Compressed Tables
  - New approach to compressed table storage
  - Compressed tables can still be modified using conventional DML operations, such as INSERT and UPDATE
- Useful for data that is bulk loaded and queried
- How it Works
  - Tables are organized into Compression Units (CUs)
  - CUs are larger than database blocks
  - Within Compression Unit, data is organized by column instead of by row
  - Column organization brings similar values close together, enhancing compression
    - Run Length encoding, adding dictionaries and a lot more
- Compression algorithms in traditional storage don’t exploit nature of data
Exadata Smart Columnar Flash Cache

- 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
Smart 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

```
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
Upcoming: In memory format in Columnar Flash Cache

- Exadata PCIe Flash is very fast
  - Smart Scans sometimes limited by CPU not flash
- In-Memory formats used in Smart Columnar Flash Cache
- Enables vector processing on storage server during smart scans
  - Multiple column values evaluated in single instruction
- Faster decompression speed than Hybrid Columnar Compression
- Enables dictionary lookup and avoids processing unnecessary rows
- Smart Scan results sent back to database in In Memory Columnar format
  - Reduces Database node CPU utilization
- In-memory performance seamlessly extended from DB node DRAM memory to 10x capacity flash in storage
  - Even bigger differentiation against all-flash arrays and other in-memory databases
Exadata Smart Flash Benefits

- Smart Flash Cache is database aware
- Smart Flash Logging avoids redo log outliers
- 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
- Smart Flash Cache Space Resource Management provides granular control
- **Upcoming**: Smart Flash cache with in memory formats enables massive capacity for vector processing