Exadata with Persistent Memory: An Epic Journey
Jia Shi, Senior Director of Exadata Development, Oracle
What is an Exadata?
Exadata Vision

- **Ideal Database Hardware** – Scale-out, database optimized compute, networking, and storage for fastest performance and lowest cost

- **Smart System Software** – Specialized algorithms vastly improve all aspects of database processing: OLTP, Analytics, Consolidation

- **Automated Management** – Automation and optimization of configuration, updates, performance, and management culminating in Fully Autonomous Infrastructure and Database
Exadata X8M-2

Scale-Out 2-Socket Database Servers

- 100Gb **RDMA** over Converged Ethernet (RoCE) Network Fabric

Scale-Out **Intelligent** 2-Socket Storage Servers

- **2** socket Xeon processors
  - 48 cores per server
  - **384 GB - 1.5 TB** DRAM

- **168** TB disk capacity
  - **25.6 TB PCI NVMe Flash**
  - **1.5 TB Persistent Memory**
  - **32 cores for SQL offload**

- **51.2 TB PCI NVMe Flash**
  - **1.5 TB Persistent Memory**
  - **32 cores for SQL offload**

Extreme Flash

High Capacity

25/10 GigE external network

100 Gb/s RoCE network fabric
Exadata X8M-8

Scale-Out 8-Socket Database Servers
- 8 socket Xeon processors
- 192 cores
- 3-6 TB DRAM

Same Storage, Network and Software as X8M-2

100 Gb RDMA over Converged Ethernet (RoCE) Network Fabric

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

- 100 Gb/s RoCE network fabric
- 25/10 GigE external network

- 1.8 TB disk capacity
- 25.6 TB PCI NVMe Flash
- 1.5 TB Persistent Memory
- 32 cores for SQL offload
- 51.2 TB PCI NVMe Flash
- 1.5 TB Persistent Memory
- 32 cores for SQL offload
## Exadata X8M: Scalability

<table>
<thead>
<tr>
<th>Eighth Rack</th>
<th>Quarter Rack</th>
<th>Elastic</th>
<th>Multi-Rack</th>
<th>Elastic</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Eighth Rack" /></td>
<td><img src="image2" alt="Quarter Rack" /></td>
<td><img src="image3" alt="Elastic" /></td>
<td><img src="image4" alt="Multi-Rack" /></td>
<td><img src="image5" alt="Elastic" /></td>
</tr>
</tbody>
</table>

**X8M-2**

**X8M-8**
What Happened In The Last Decade?
# The Flash Revolution (2009-2019)

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X8M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Xeon E5430 Harperton</td>
<td>Xeon E5540 Nehalem</td>
<td>Xeon X5670 Westmere</td>
<td>Xeon E5-2690 Sandy Bridge</td>
<td>Xeon E5-2697v2 Ivy Bridge</td>
<td>Xeon E5-2699v3 Haswell</td>
<td>Xeon E5-2699v4 Broadwell</td>
<td>Xeon 8160 Skylake</td>
<td>Xeon 8260 Cascade Lake</td>
<td>Xeon 8260 Cascade Lake</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X8M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flash Cache (TB)</td>
<td>0</td>
<td>5.3</td>
<td>5.3</td>
<td>22.4</td>
<td>44.8</td>
<td>89.6</td>
<td>179.2</td>
<td>358</td>
<td>358</td>
<td>358</td>
</tr>
<tr>
<td>Scan Rate (GB/s)</td>
<td>14</td>
<td>50</td>
<td>75</td>
<td>100</td>
<td>100</td>
<td>263</td>
<td>301</td>
<td>350</td>
<td>560</td>
<td>560</td>
</tr>
<tr>
<td>Read IOPS (M)</td>
<td>.05</td>
<td>1</td>
<td>1.5</td>
<td>1.5</td>
<td>2.66</td>
<td>4.14</td>
<td>5.6</td>
<td>5.97</td>
<td>6.57</td>
<td>16</td>
</tr>
</tbody>
</table>
Is the Flash Revolution Plateauing?

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X8M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flash Cache (TB)</td>
<td>0</td>
<td>5.3</td>
<td>5.3</td>
<td>22.4</td>
<td>44.8</td>
<td>89.6</td>
<td>179.2</td>
<td>358</td>
<td>358</td>
<td>358</td>
</tr>
<tr>
<td>Scan Rate (GB/s)</td>
<td>14</td>
<td>50</td>
<td>75</td>
<td>100</td>
<td>100</td>
<td>263</td>
<td>301</td>
<td>350</td>
<td>560</td>
<td>560</td>
</tr>
<tr>
<td>Read IOPS (M)</td>
<td>.05</td>
<td>1</td>
<td>1.5</td>
<td>1.5</td>
<td>2.66</td>
<td>4.14</td>
<td>5.6</td>
<td>5.97</td>
<td>6.57</td>
<td>16</td>
</tr>
</tbody>
</table>

Flash SCSI

Flash NVMe

16M IOPS 19 usec latency
What Is The Secret Sauce In X8M?
Do You Recognize The Dynamic Duo?
Intel® Optane™ DC Persistent Memory

- Capacity, performance, price between DRAM and flash
- Reads at memory speed – much faster than flash
- Writes survive power failure unlike DRAM
- Requires sophisticated algorithms for PMEM data integrity upon failure
  - Call special instructions to flush data from CPU cache to PMEM
  - Complete or backout sequence of writes interrupted by a crash
Remote Memory Direct Access (RDMA)

- **Database Server**
  - Memory Region
  - CPU
  - RDMA Write

- **Storage Server**
  - Memory Region
  - CPU
  - RDMA Read
New RoCE Internal Network Fabric

- Exadata 100Gb/s RoCE enables RDMA on Ethernet fabric
- Exadata runs RDMA on InfiniBand since 2008 for high throughput and low latency
- What is the new trick?
What can they do together?

PMEM

RDMA
PMEM in Conventional Storage

- Persistent Memory usage with conventional storage:
  - Database issues read I/O call to OS
  - OS sends message to storage
  - Storage CPU issues read to Persistent Memory
  - Storage CPU sends reply to Server OS
  - Server OS wakes up Database

- Overwhelmed by high cost network and I/O software, interrupts, and context switches

- Performance benefit from PMEM is erased
Conventional Read

Conventional Storage Performs Server-side Cache Read

DB Sends READ Request to Storage [Disk, Offset]

Database Server

Storage Sends Data to DB

Storage Looks up [Disk, Offset] -> Location on Flash, Issues Read to Flash

Flash Cache

... Flash Cache Line

... Flash Cache Line

Storage Server
Dissect the Exadata Flash IO Latency

Database Server

Storage Server

Flash Read Raw Latency: <100 us

Database 8K Read End-to-end Latency: ~200 usec

Context Switch: 10s of us
What if we drop in PMEM as is?

Database 8K Read
End-to-end Latency: ~100 usec

Over 90% of the Time *Wasted.*

Database Software

Kernel/OS (Database Server)

Storage Server Software

Context Switch: 10s of us

Context Switch: 10s of us

PMEM Read
Raw Latency: ~1 us
A Radical Approach – RDMA to PMEM

Database Server

Storage Server

Database 8K Read
End-to-end Latency: <19 usec

10x Faster than Exadata X8

RDMA

Context Switch: 10s of us

Context Switch: 10s of us

Database Software

Kernel/OS (Database Server)

Kernel/OS (Storage Server)

Storage Server Software

PMEM

10x Faster than Exadata X8
Ultra Fast RDMA Read

New Disruptive Technology Enables PMEM Cache Read via RDMA

RDMA to Fetch Data From PMEM

Database Server

Storage Server

PMEM Cache

PMEM Cache Line

PMEM Cache Line
Exadata X8M With Persistent Memory Data Accelerator

World’s First and Only Shared Persistent Memory Optimized for Database

- Exadata Storage Servers transparently add Persistent Memory Accelerator in front of Flash memory
- Database uses RDMA instead of I/O to read remote PMEM
  - Bypasses network and IO software, interrupts, context switches
- PMEM Automatically tiered and shared across DBs
  - Using as a cache for hottest data increases effective capacity 10x
- Persistent Memory mirrored automatically across storage servers for fault-tolerance
- 16Million IOPS, <19us latency for 8K I/Os from the database

Enabled with Exadata System Software 19.3 and Database Software 19c
Conventional Log Write

- DB Sends Requests to Storage
- Storage Writes to Flash Log and Sends Ack

Storage Server Issues Write to Flash and HDD Simultaneously

DB Sends Log Write request to Storage

Storage Server Sends Ack to DB
What about Redo Log Writes to PMEM?

Database Server

Storage Server

Database Software

Kernel/OS (Database Server)

Kernel/OS (Storage Server)

Storage Server Software

PMEM Write Raw Latency: <10 us

Database Log Write End-to-end Latency: ~100 usec

Context Switch: 10s of us

Context Switch: 10s of us

Over 90% of the Time *Wasted.*
Ultra Fast RDMA Log Write

New Disruptive Technology Enables Durable PMEM Log Write via RDMA

RDMA Write to Persist Redo Log

Database Server

PMM Log Buffer

PMM Log Buffer

PMM Log Buffer

Storage Server Performs Flash Log Write in the Background

Storage Server
Ultra Fast RDMA Log Write

Storage Server Crash Safe!

Storage Server Performs PMEM Log Recovery

Database Server

Storage Server

PMEM Log Buffer

PMEM Log Buffer

PMEM Log Buffer

PMEM Log Buffer
Exadata X8M Persistent Memory Commit Accelerator

- Log Write latency is critical for OLTP performance
  - Faster log writes means faster commit times
  - Any log write slowdown stalls the whole database

- Automatic Commit Accelerator
  - Database issues one-way RDMA writes to PMEM on multiple Storage Servers
  - Bypasses network and I/O software, interrupts, context switches, etc.
  - Up to 8x faster log writes

Enabled with Exadata System Software 19.3 and Database Software 19c
## An Epic Journey

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X8M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V1 – X8M Growth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>V1</th>
<th>V2</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X8M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage (TB)</td>
<td>168</td>
<td>336</td>
<td>504</td>
<td>504</td>
<td>672</td>
<td>1344</td>
<td>1344</td>
<td>1.68</td>
<td>2.35</td>
<td>2.3 PB</td>
</tr>
<tr>
<td>Flash Cache (TB)</td>
<td>0</td>
<td>5.3</td>
<td>5.3</td>
<td>22.4</td>
<td>44.8</td>
<td>89.6</td>
<td>179.2</td>
<td>358</td>
<td>358</td>
<td>358 TB</td>
</tr>
<tr>
<td>CPU (cores)</td>
<td>64</td>
<td>64</td>
<td>96</td>
<td>128</td>
<td>192</td>
<td>288</td>
<td>352</td>
<td>384</td>
<td>384</td>
<td>384</td>
</tr>
<tr>
<td>Max Mem (GB)</td>
<td>256</td>
<td>576</td>
<td>1152</td>
<td>2048</td>
<td>4096</td>
<td>6144</td>
<td>12288</td>
<td>12288</td>
<td>12</td>
<td>12 TB</td>
</tr>
<tr>
<td>RDMA Network (Gb/s)</td>
<td>20</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>200 Gb/s</td>
<td></td>
</tr>
<tr>
<td>Scan Rate (GB/s)</td>
<td>14</td>
<td>50</td>
<td>75</td>
<td>100</td>
<td>100</td>
<td>263</td>
<td>301</td>
<td>350</td>
<td>560</td>
<td>560 GB/s</td>
</tr>
<tr>
<td>Read IOPS (M)</td>
<td>.05</td>
<td>1</td>
<td>1.5</td>
<td>1.5</td>
<td>2.66</td>
<td>4.14</td>
<td>5.6</td>
<td>5.97</td>
<td>6.57</td>
<td>16 M</td>
</tr>
</tbody>
</table>

### Notes
- V1 – X8M Growth
- X8M Growth
Thank you

Please visit www.snia.org/pmsummit for presentations