Largest Exadata OLTP Environment

Amit Das, Database Engineering Architect

CON 11256: Mission Critical Oracle Exadata OLTP Deployment at PayPal
Introduction: about our team

• Sehmuz Bayhan – Our visionary director. Executed great changes in lightning speed.
• Saibabu Devabhaktuni – Our fearless leader, around PayPal for at least 8 years.
• Kyle Towle – Our fearless database architect, around PayPal for at least 7 years.
• Dong Wang – Goldengate expert, speaker at multiple conferences, PayPal DBA for going on 6 years.
• John Kanagaraj – Author, Oracle ACE, frequent speaker at Oracle conferences
• Sarah Brydon – One of the very few Oracle Certified Masters.
Who Am I?

- 11 years in Oracle RAC Development team.
- Technical lead for world first Exadata production go-live (Apple), while at Oracle.
- Currently Engineering lead/architect for World largest Exadata OLTP system (PayPal).
- Frequent presenter inside/outside of Oracle.
- Lake fishing.
- World Traveler.
PayPal’s Amazing Growth and Requirements

• **Amazing Growth**
  – Exponential growth in PayPal business year to year

• **Business is growing rapidly**
  – New users, features, transaction
  – New channels: POS, Mobile, etc

• **Massive growth in database demand every year**
  – Not uncommon to see database workloads grow 50-100%
One of the Largest OLTP database on Oracle

- Measured by Executions X Processes (concurrency)

- Fast paced VLDB OLTP environment on Oracle
  - 500+ database instances
  - OLTP databases commonly 10-130 TB
  - 5,000-14,000 concurrent processes
  - 80,000 executions/second, 10GB Redo/Minute

- Continuously growing
  - High growth of PayPal’s business per year ➔ up to 2 X workload increase
  - Tier one databases built to support 300+K execs/sec to support 2012Q4
Architecture

• **Two Data Centers containing**
  – 3 Exadata “Production” Clusters
  – 3 Exadata “Standby/Reporting” Active Data Guard
  – 1 Test/Dev Exadata Cluster

• **Each Exadata Production Cluster contains:**
  – 4 node RAC cluster with 64 Exadata Storage Cells (HP)
  – Two X2-8s, One Consolidated Database with 8 “shards”
  – 2 Full Exadata Storage Expansion Racks
  – 500GB SGA, 120 TB database

• **MAA configuration**
  – RAC, ASM, Flashback Technologies, Active Data Guard, ASM high redundancy, corruption settings, GoldenGate for real time replication to Read Replica and Data Warehouse
PayPal's Critical Application Architecture

**Primary Data Center**

**Mission-critical Databases**

**Production Databases**
- 2 X Exadata X2-8
- 2 X Full Storage Expansion

**GoldenGate Real-time Data Integration** *

**Extreme Performance**
- 300+K executions/sec
- Real Time analysis of 99.99% of critical transactions.
- avg 40 ms response for 99.99%
- 10 X performance compared than pre-Exadata system

**HA and MAA**
- 99.99% Availability
- MAA technologies (RAC, ASM, ADG, Exadata, Flashback, GG)
- All disk groups using high redundancy
- Active Data Guard for auto block corruption repair and DR
- Rolling upgrade using ASM, Exadata, CRS, Data Guard, and GoldenGate

**Test/Dev**

**ETL Targets**

**WAN, 650+ miles (30ms)**

**Data Guard ASYNC Redo Transport**

**DR Data Center**

**Active Data Guard Standby**
- Offload queries and reads
- Corruption Protection
- Symmetric System

**Production and Standby Clusters = 8 Exadata Racks**

**3 identical Architectures = 24 Exdata Racks + Test/Dev Resources supporting our Critical Applications.**
### Disk Group Configuration on Each Cluster

<table>
<thead>
<tr>
<th>Disk Group</th>
<th>Description</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DBFS DG</td>
<td>Will use all 64 Exadata Cells</td>
<td>5.5TB</td>
</tr>
<tr>
<td>2. RECO DG</td>
<td>Will use all 64 Exadata Cells</td>
<td>22TB</td>
</tr>
<tr>
<td>3. DATA_1 DG</td>
<td>Partitioned into 16 cells</td>
<td>32.7TB</td>
</tr>
<tr>
<td>4. DATA_2 DG</td>
<td>Partitioned into 16 cells</td>
<td>32.7TB</td>
</tr>
<tr>
<td>5. DATA_3 DG</td>
<td>Partitioned into 16 cells</td>
<td>32.7TB</td>
</tr>
<tr>
<td>6. DATA_4 DG</td>
<td>Partitioned into 16 cells</td>
<td>32.7TB</td>
</tr>
</tbody>
</table>

**Total Space on each cluster**

Everything is ASM high redundancy 131TB
Migration success story

• Deployed Production Exadata Cluster (4 Exadata Racks) and configured in 4-6 days

• Only 10 Minutes Downtime from non-Exadata to Exadata
  – Migrate the data using extent copy (4 days)
  – Sync up the Data to Exadata using GoldenGate
  – Data Validation (2 days)
  – End to end application switchover to Exadata (10 minutes)
    • Most of the time due to restarting application tier, java clients, and mid-tier services
  – Full performance/throughput requirements met in 10 minutes with partial application availability much sooner
Performance Data

• Performance benefits with Exadata
  – Smart flash log for low latency commits
  – Smart flash cache for low latency reads. KEEP in Flash for critical objects
  – High bandwidth and low latency InfiniBand
  – High scalability and throughput achieved by Exadata overall architecture (RAC, ASM, Exadata and Network)
Reduced Physical Reads on Exadata

11g Physical Read

Kilo-Block/sec

Time of the Day
Reduced Load Average on Exadata

CPU bound and caused high runqueue in 10G
Response Time Reduction due to Exadata

- Average 40ms response
- Running 99% < 110ms
- Running 95% < 90ms
- Independent of site load (consistent performance all day)

Meeting Business Requirements and Response Time SLAs

- Average 160-400ms response
- 5% sender txns between 400ms and 500ms response
- 1% sender txns between 600ms and 700ms response
- Dependent on site load. (performance fluctuates throughout the day)
# Size of Smart Flash Cache Matters

Comparison Test during POC

<table>
<thead>
<tr>
<th>Split</th>
<th>100% Flash Cache Online</th>
<th>50% Flash Cache Online</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.99%</td>
<td>99.99%</td>
</tr>
<tr>
<td></td>
<td>Max (0.01%)</td>
<td>Max</td>
</tr>
<tr>
<td>20</td>
<td>127</td>
<td>311</td>
</tr>
<tr>
<td></td>
<td></td>
<td>253</td>
</tr>
<tr>
<td></td>
<td></td>
<td>442</td>
</tr>
<tr>
<td>21</td>
<td>127</td>
<td>365</td>
</tr>
<tr>
<td></td>
<td></td>
<td>226</td>
</tr>
<tr>
<td></td>
<td></td>
<td>459</td>
</tr>
<tr>
<td>22</td>
<td>126</td>
<td>314</td>
</tr>
<tr>
<td></td>
<td></td>
<td>247</td>
</tr>
<tr>
<td></td>
<td></td>
<td>503</td>
</tr>
</tbody>
</table>
## Smart Flash Cache Benefits (smart log/smart cache)

### AWR with 100% Cache

<table>
<thead>
<tr>
<th>Event</th>
<th>Waits</th>
<th>Time(s)</th>
<th>Avg wait (ms)</th>
<th>% DB time</th>
<th>Wait Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB CPU</td>
<td></td>
<td>2,538</td>
<td></td>
<td>47.35</td>
<td></td>
</tr>
<tr>
<td>Streams miscellaneous event</td>
<td>2,304</td>
<td>1,153</td>
<td>500</td>
<td>21.51</td>
<td>Other</td>
</tr>
<tr>
<td>ARCH wait on ATTACH</td>
<td>1,613</td>
<td>445</td>
<td>276</td>
<td>8.31</td>
<td>Network</td>
</tr>
<tr>
<td>log file sync</td>
<td>621,944</td>
<td>384</td>
<td>1</td>
<td>7.17</td>
<td>Commit</td>
</tr>
<tr>
<td>log file sequential read</td>
<td>894,160</td>
<td>345</td>
<td>0</td>
<td>6.44</td>
<td>System I/O</td>
</tr>
</tbody>
</table>

### AWR with 50% Cache

<table>
<thead>
<tr>
<th>Event</th>
<th>Waits</th>
<th>Time(s)</th>
<th>Avg wait (ms)</th>
<th>% DB time</th>
<th>Wait Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB CPU</td>
<td></td>
<td>2,308</td>
<td></td>
<td>45.54</td>
<td></td>
</tr>
<tr>
<td>log file sync</td>
<td>591,730</td>
<td>1,021</td>
<td>2</td>
<td>20.15</td>
<td>Commit</td>
</tr>
<tr>
<td>Streams miscellaneous event</td>
<td>972</td>
<td>486</td>
<td>500</td>
<td>9.58</td>
<td>Other</td>
</tr>
<tr>
<td>ARCH wait on ATTACH</td>
<td>1,970</td>
<td>386</td>
<td>196</td>
<td>7.62</td>
<td>Network</td>
</tr>
<tr>
<td>log file sequential read</td>
<td>599,578</td>
<td>310</td>
<td>1</td>
<td>6.13</td>
<td>System I/O</td>
</tr>
</tbody>
</table>
Lessons learned

- PayPal specific best practices DB
  - *._gc_policy_time=0 // Disable SPIKEs, due to DRM
  - *._mutex_wait_time=10 // Mutex wait time to 10ms
  - *._sixteenth_spare_parameter='942' // ER 12326358: Will not do hard parse for 2nd time on missing objects
  - *._third_spare_parameter = 0 // Faster RAC reconfiguration, bug 10415371 //
  - *.event='12633340 trace name context forever, level 1' // Bug 12633340:
Lessons learned

• PayPal large memory settings
  – *.db_cache_size=392G
  – *.pga_aggregate_target=128G
  – *.shared_pool_size=60G
  – *.use_large_pages='TRUE'

• Paypal specific cell parameter (maybe remove from slide until integrated into production)
  _cell_buffer_expiration_hours=2400 // ER 14589662 //
  _cell_object_expiration_hours=1200 // ER 14589662 //
Conclusion

• PayPal is very happy with Exadata
  – Exadata is meeting all performance and availability SLAs
• What’s next
  – Interested in X3 Exadata machines and new software capabilities
  – Validating 12c capabilities (Pluggable Database, RAC Flex Clusters)