PayPal™

Largest Exadata OLTP Environment

Amit Das, Database Engineering Architect

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.
  - http://sai-oracle.blogspot.com/
- 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.
- Love fishing

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%

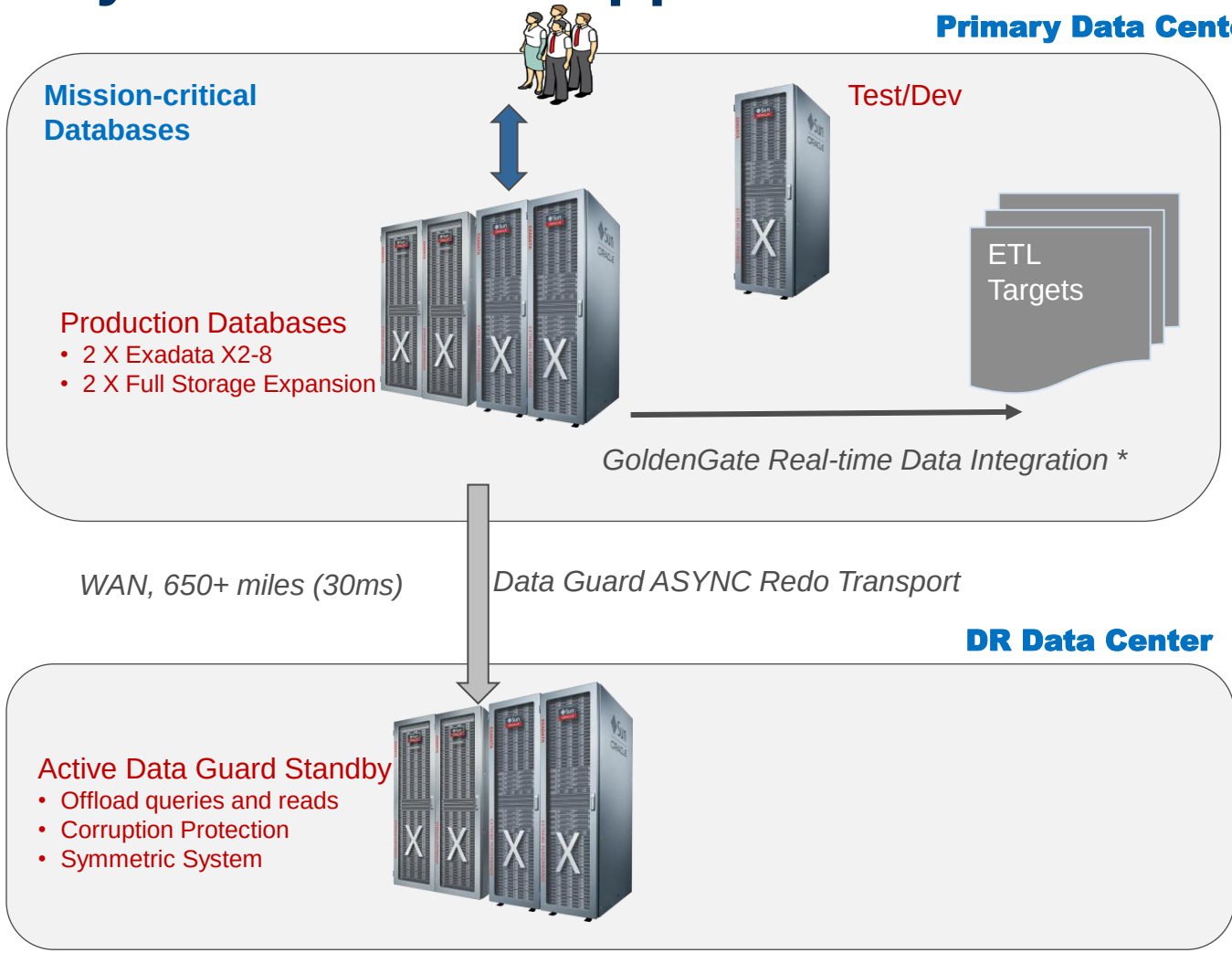
One of the Largest OLTP database on Oracle

- 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 300K+ execs/sec to support the Holiday in 2012

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



Extreme Performance

- 300K+ 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

Production and Standby Clusters = 8 Exadata Racks
3 identical Architectures = 24 Exadata Racks + Test/Dev Resources supporting our Critical Applications.

Disk Group Configuration on Each Cluster

1. DBFS DG will use all 64 Exadata Cells	5.5TB
2. RECO DG will use all 64 Exadata Cells	22TB
3. DATA_1 DG partitioned into 16 cells	32.7TB
4. DATA_2 DG partitioned into 16 cells	32.7TB
5. DATA_3 DG partitioned into 16 cells	32.7TB
6. DATA_4 DG partitioned into 16 cells	32.7TB
Total Space on each cluster Everything is ASM high redundancy	131TB

Migration success story

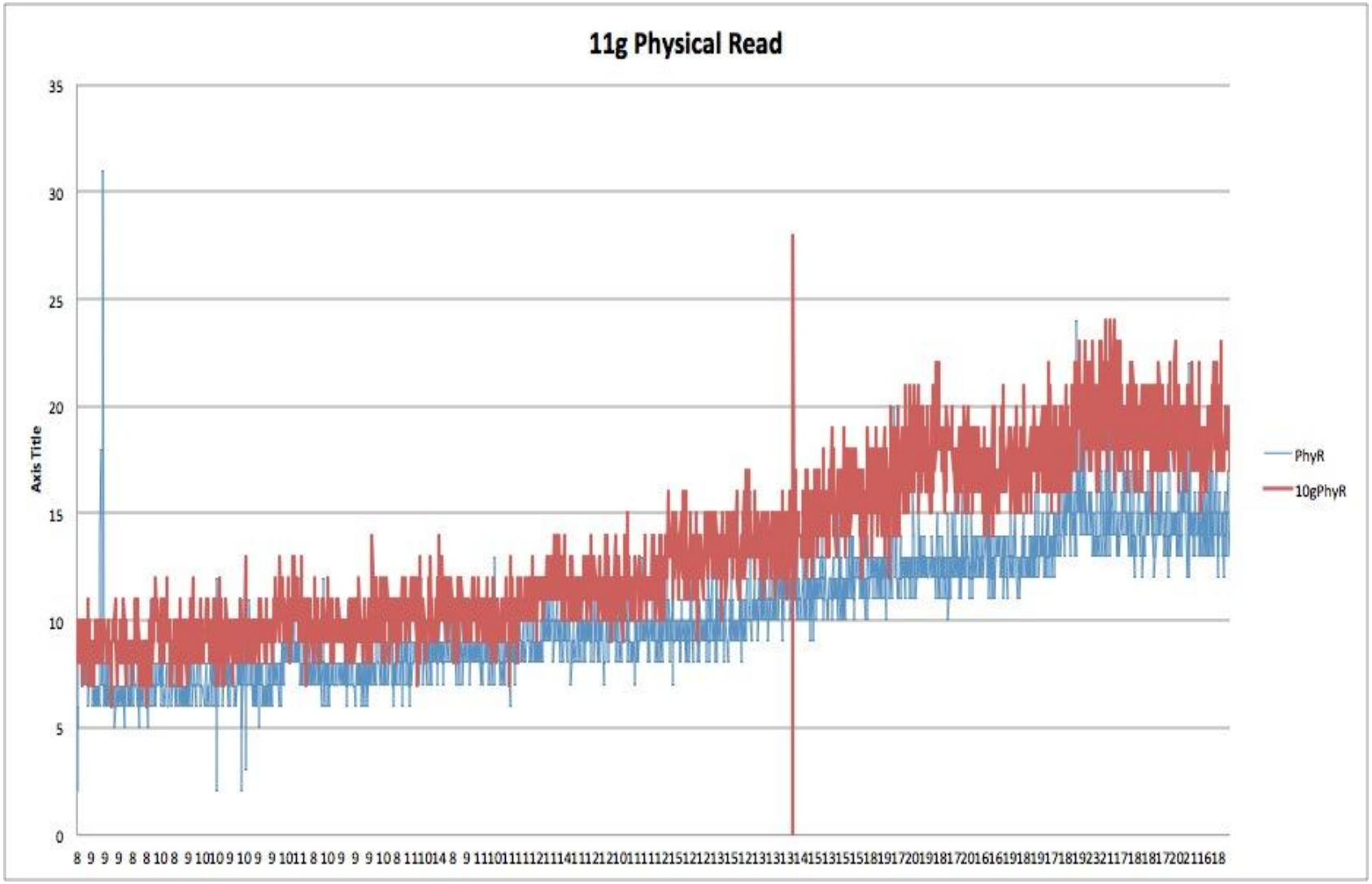
- Deployed Production Exadata Cluster (4 Exadata Racks) and configured in 4-6 days
- Migrate the data using extent copy (4 days)
- Data Validation (2 days)
- Only 10 Minutes Downtime from non-Exadata to Exadata
  - Sync up the Data to Exadata using GoldenGate
  - End to end application switchover to 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)

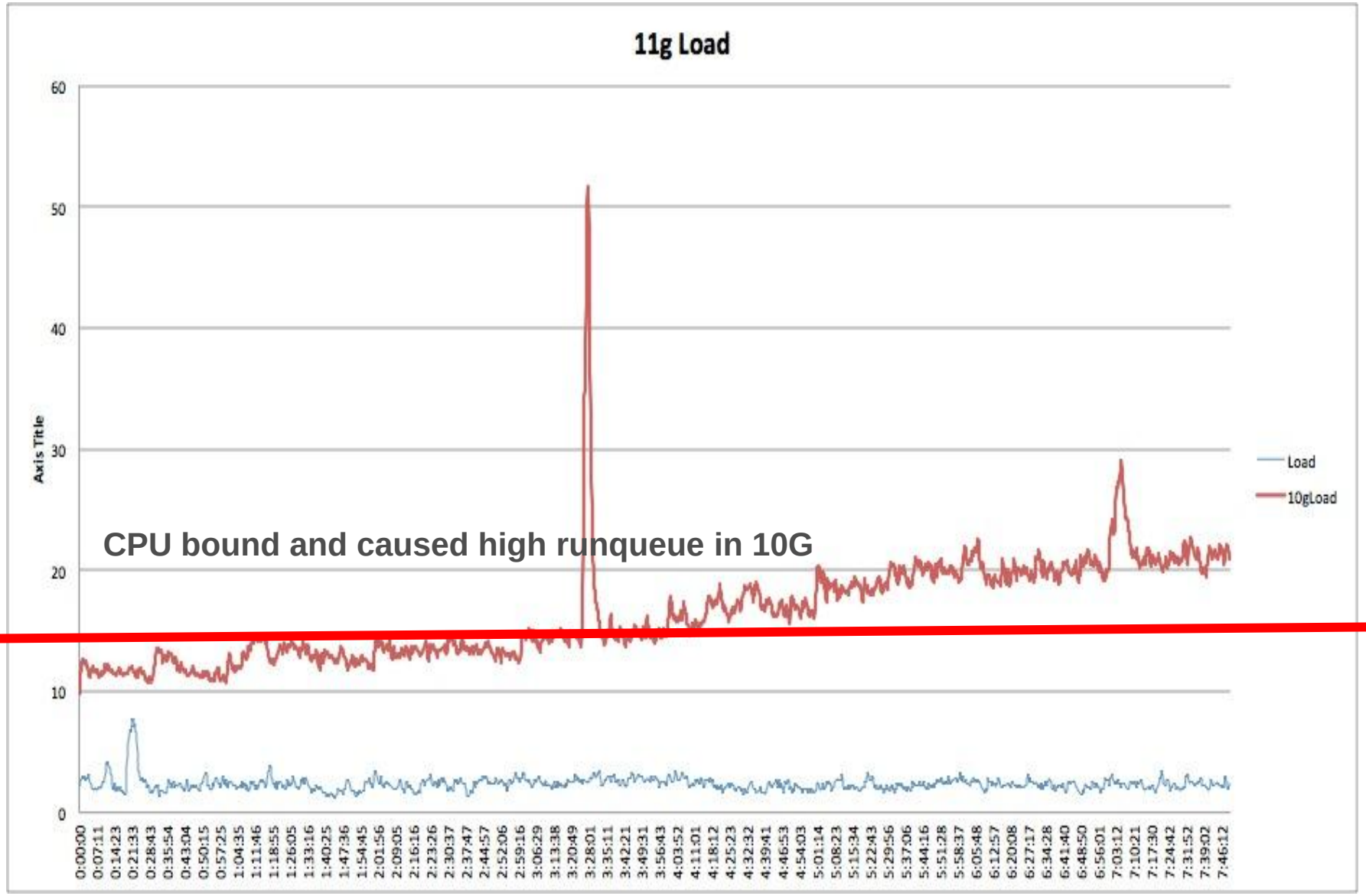
Reduced Physical Reads on Exadata

Kilo-Block/sec

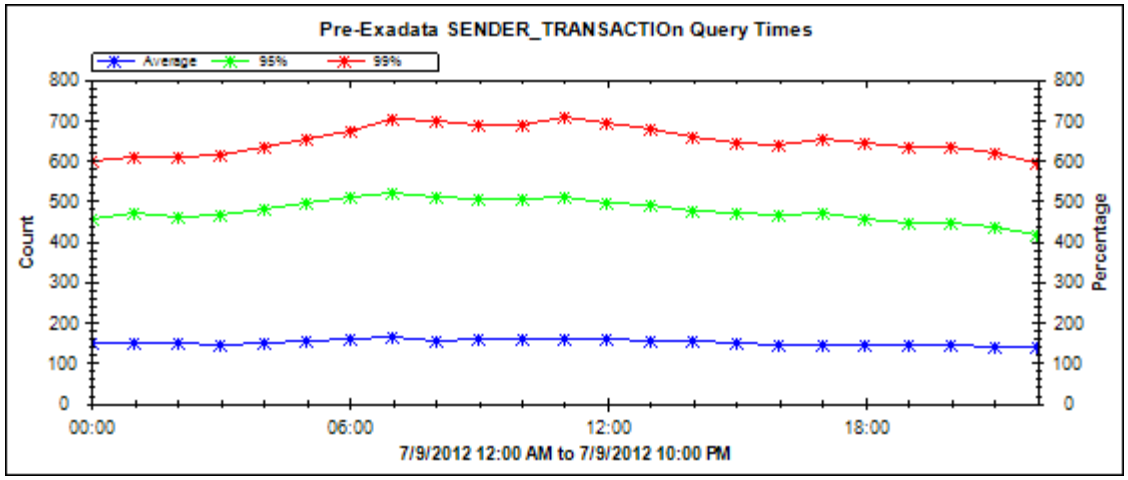


Time of the Day

Reduced Load Average on Exadata



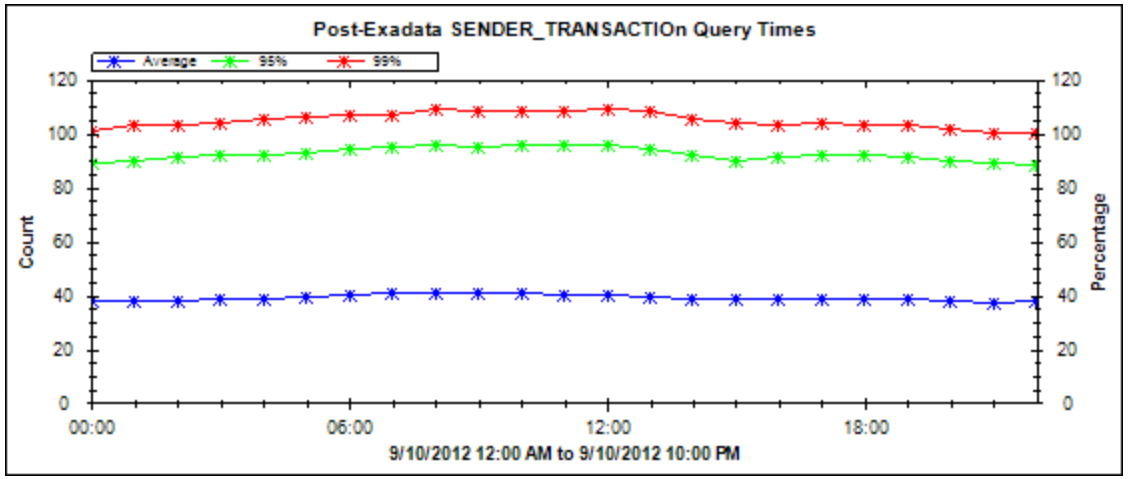
Response Time Reduction due to Exadata



- Average 160-400 ms 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)

- Meeting Business Requirements and Response Time SLAs

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



Size of Smart Flash Cache Matters

Comparison Test during POC

Performance Data in ms				
	100% Flash Cache Online		50% Flash Cache Online	
Split	99.99%	Max (0.01%)	99.99%	Max
20	127	311	253	442
21	127	365	226	459
22	126	314	247	503

Smart Flash Cache Benefits (smart log/smart cache)

AWR with 100% Cache

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

AWR with 50% Cache

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

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 //
  - Double cell failure on HIGH redundancy will cause session SPIKE // 13830962 EXTENDED BROWNOUT WITH DOUBLE CELL FAILURE

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

  _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)

Upcoming session

- See you on IOUG-2013 at Denver (Apr 7-11)
  - **IOUG :#489 : “Internals of Active DataGuard”**
    - **By Saibabu Devabhaktuni, PayPal**
  - **IOUG :#867 :“How to minimize the brownout, due to RAC reconfiguration”**
    - **By Amit Das , PayPal**
  - **IOUG :#861 :“Tech refresh of existing system with ZERO downtime using RAC, ASM Technology”**
    - **By Amit Das , PayPal**