Exadata Technical Deep Dive: Architecture and Internals

September 18–22, 2016
San Francisco

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.

The Exadata Database Machine Vision

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

1. State-of-the-art enterprise-grade hardware, refreshed yearly (processors, flash, disks, network)
2. Sized, tuned and optimized exclusively for Oracle Database workloads (DW, Analytics, OLTP, Mixed)
3. High-powered intelligent storage servers capable of offloading database workloads
4. "Smart" database protocols and optimizations from servers to network to storage
5. One vendor responsible for all hardware, software and customer support

Exadata Unique Intellectual Property

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, Sales*force*, SAS, …

4 OF THE TOP 5

BANKS, TELCOS, RETAILERS RUN EXADATA

On-Premises

Exadata Database Machine

Customer Data Center

Purchased

Customer Managed

Cloud at Customer

Preview:

Exadata Cloud Machine

Customer Data Center

Subscription

Oracle Managed

Public Cloud

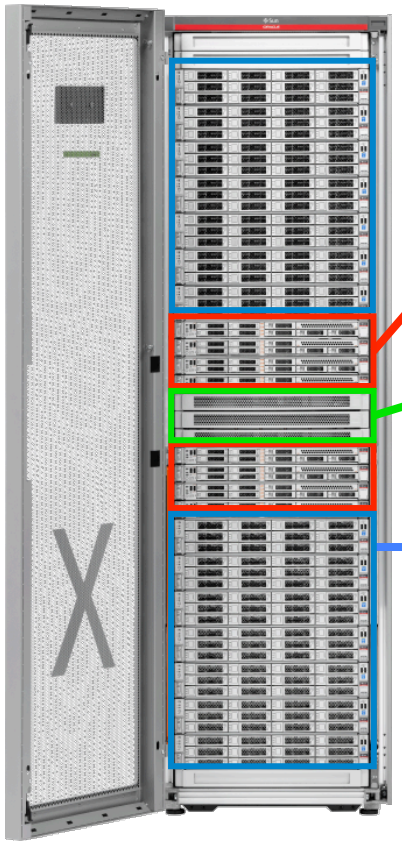
Exadata Cloud Service

Oracle Cloud

Subscription

Oracle Managed

Exadata Database Machine X6-2



- **Scale-Out Database Servers**

  - 2 socket x86 processors
  - **44** CPU cores
  - **256 GB - 1.5 TB** GB 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

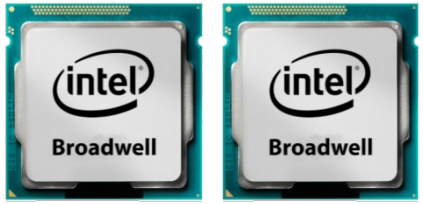
  - **High-Capacity Storage Server**

25.6 TB PCI Flash
20 CPU cores

  - **Extreme Flash Storage Server**

Compute Software

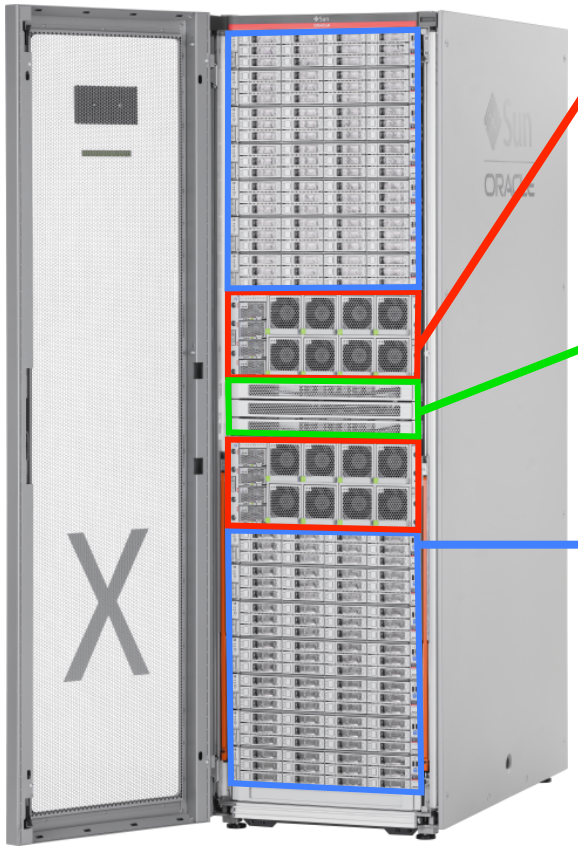
- Oracle Linux 6
- Oracle Database Enterprise Edition
- Oracle VM (optional)
- Oracle Database options (optional)



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

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

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

- **Scale-Out Intelligent Storage**

  - High-Capacity Storage Server

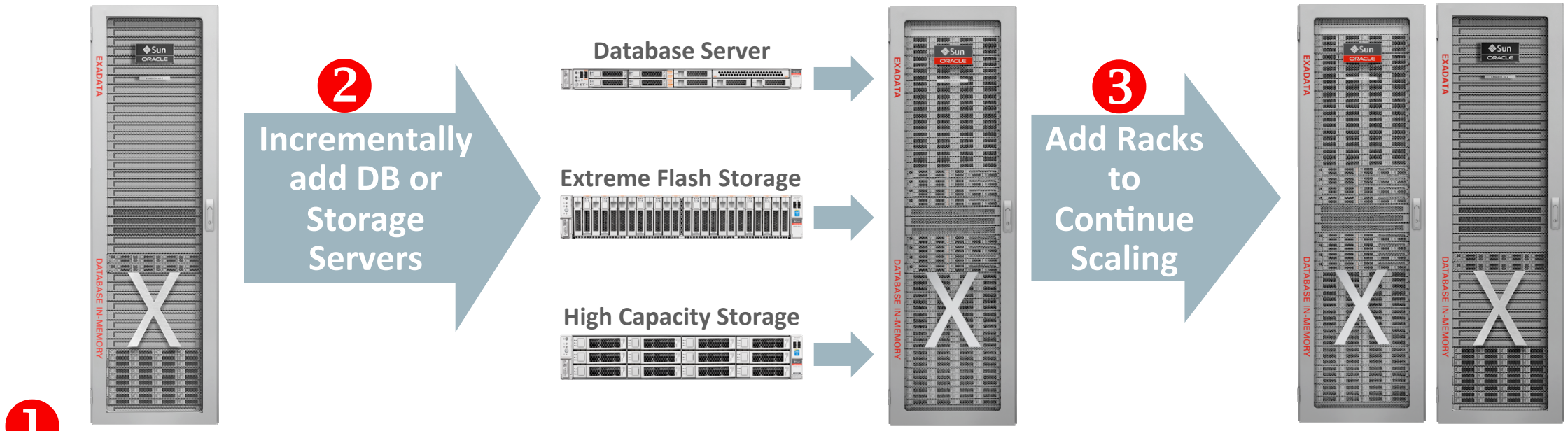
  - Extreme Flash Storage Server

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



- **Enable Database CPU cores as needed with Capacity on Demand**
- **Expand older Exadata machines with new X6-2 servers**

Oracle Database **Exadata Cloud Service**

- **Full Oracle Database with all advanced options**
  - **100% Compatible** with on-premises databases
- **On fastest and most available database cloud platform**
  - Scale-Out Compute, Scale-Out Intelligent Storage, InfiniBand, PCIe Flash
  - **Complete isolation** of tenants with no overprovisioning
- **All Benefits of Public Cloud**
  - Fast, elastic, web driven provisioning
  - Oracle experts deploy and manage infrastructure
  - Monthly or yearly subscription with **online capacity bursting**

Best of On-Premises with Best of Cloud

ORACLE

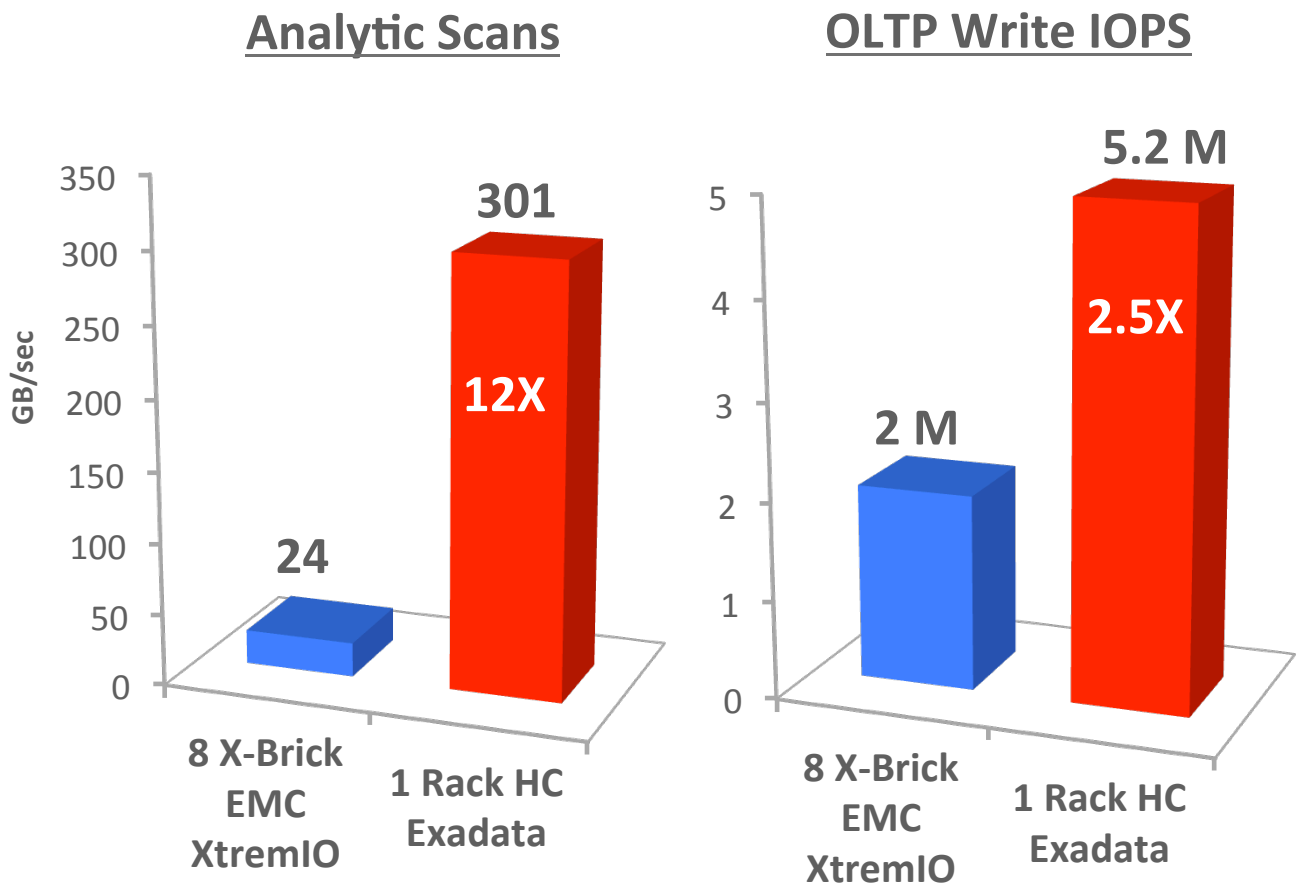
Preview: Oracle Public Cloud Services @ Customer

- Same PaaS and IaaS hardware and software as Oracle Public Cloud
- Managed by Oracle and delivered as a service in your datacenter behind your firewall
- Same cost-effective subscription pricing model as Oracle Cloud
- Helps conform to business and government security requirements
- Connect via fast LAN to existing systems

Exadata X6 is Much Faster and Cheaper than All-Flash EMC

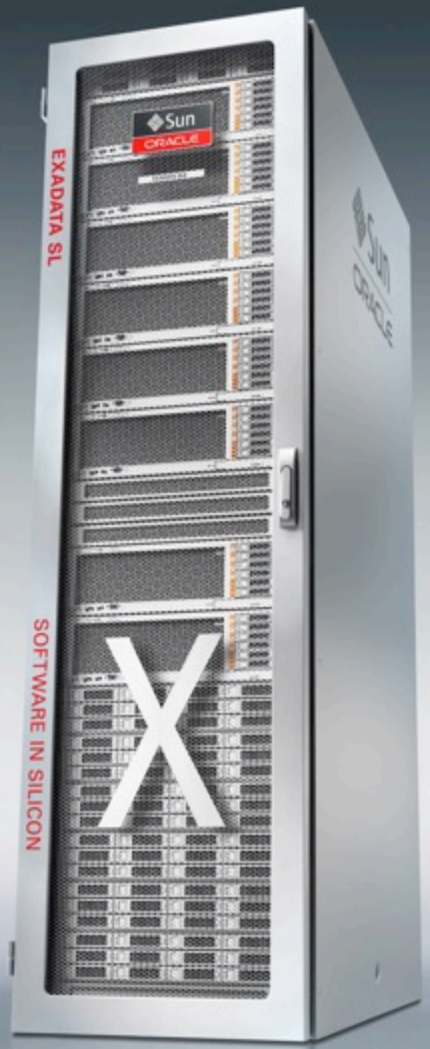
- One **High Capacity** Exadata beats the fastest EMC XtremIO **all-flash** array in every performance metric
  - **12X more throughput**
  - **2.5X more IOPS**
  - **2X faster latency**

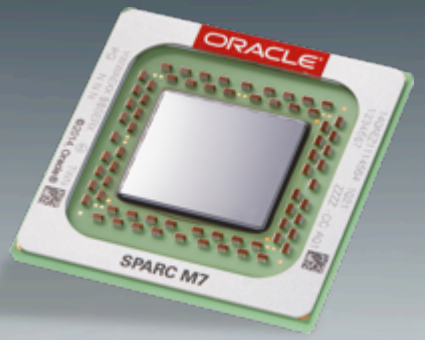
EMC 8 X-Brick XtremIO: $7.8 M
Exadata X6-2 Full Rack: $1.1 M



EMC Performance does not scale higher - Exadata scales by adding racks

Preview: Exadata SL6





Linux on SPARC

Software in Silicon

ORACLE

Database Intelligence Extended into CPU Chip

SPARC M7 Software in Silicon

- Traditional DB algorithms too complex for chips
- Big Change: In-memory algorithms are much simpler
- 5 years ago Oracle initiated a revolutionary project
  - Build fastest ever microprocessor
    - Most processing cores (32)
    - Most concurrent threads (256)
    - Fastest Memory Bandwidth (160 GB/sec)
  - Add **In-Memory DB** operations directly on chip

ORACLE

In-Memory Algorithms Natively Implemented in Silicon

SQL in Silicon

DB Acceleration

Capacity in Silicon

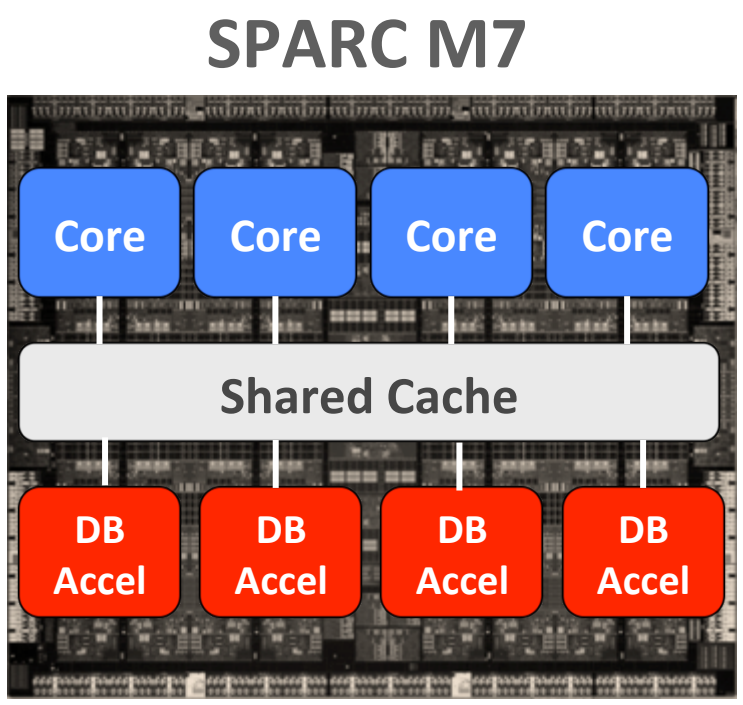
Decompression Engines

Silicon Secured Memory

Fine-Grained Memory Protection

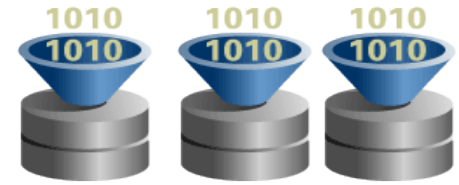
Database Software Already Available

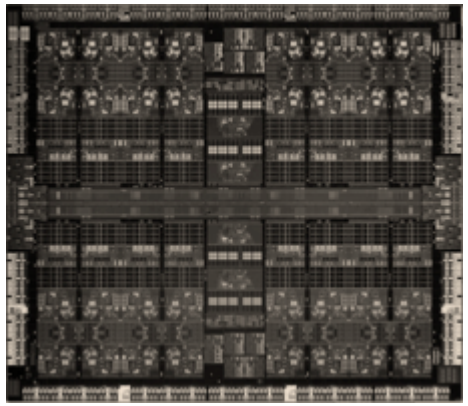
SQL in Silicon: Database In-Memory Acceleration Engines



- SIMD Vectors instructions are fast, but were designed for graphics, not database
- New SPARC M7 chip has 32 optimized database acceleration engines (DAX) built on chip
- Independently process streams of columns
  - E.g. find all values that match 'California'
  - Up to 170 Billion rows per second!
- Like adding 32 additional specialized cores to chip
  - Using less than 1% of chip space

Capacity in Silicon: Decompression Engines





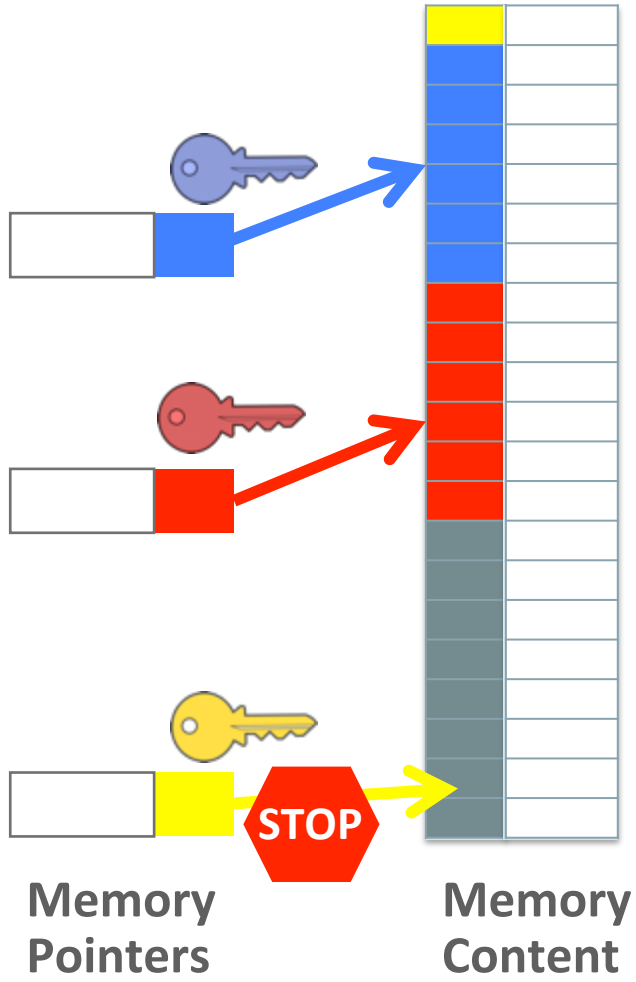
Doubles Memory Capacity

- Compression is key to putting more data in-memory
- Decompression is far more important for databases than compression
  - Data is loaded once, queried many times
- Bit pattern decompression in normal cores is slow
  - 64 CPU cores needed to decompress at full memory speed
- SPARC M7 adds 32 optimized decompress engines
  - Run bit-pattern decompress at memory speed

ORACLE

Silicon Secured Memory: Fine Grained Memory Protection

- Database In-memory places terabytes of data in memory
  - More vulnerable to corruption by bugs/attacks than storage
- SPARC M7 locks memory as it is allocated so only the owner can access it
  - Hidden "color" bits added to pointers (key), and content (lock)
  - Pointer color (key) must match content color or program is aborted
  - Hardware support eliminates performance impact
- Helps prevent access off end of structure, stale pointer access, malicious attacks, etc. plus improves developer productivity

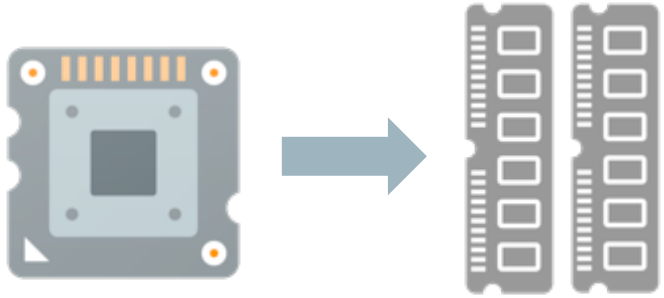


Exadata SL6: Exadata with Ultra-fast SPARC Linux Servers

- Identical to Exadata with x86 Database servers replaced by SPARC T7-2 servers
  - Ultra-fast 32-core SPARC M7 Processors
  - Two-socket T7-2 Servers
- Same elastic configurations as Exadata X6-2
- Storage servers identical as Exadata X6-2
- Runs same Oracle Linux as Exadata X6-2
  - Oracle Linux (UEK2) – single domain configuration
- Runs Oracle Database 12.1.0.2

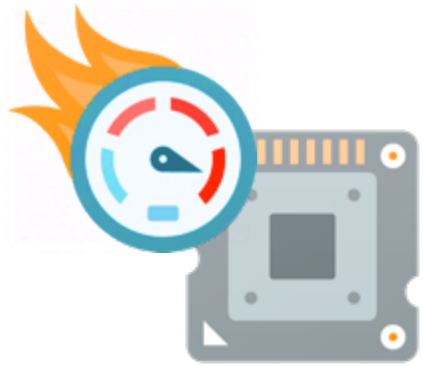
Preview: Exadata SL6

World's Fastest and Most Secure Linux Database Machine



Massive Memory Bandwidth

2.2x Intel x86



Fastest Database Processor

1.9x Intel x86

Silicon Secured Memory

End to End Database Security

Exadata Smart System Software

ORACLE

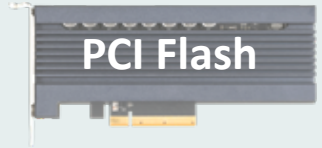
Smart System Software Highlights

Smart Analytics

- Move **queries to storage**, not storage to queries
- Automatically **offload and parallelize** queries across all storage servers
- **100X** faster analytics

Smart Storage

- **Hybrid Columnar Compression** reduces space usage by **10X**
- Database-aware **Flash Caching** gives speed of flash with capacity of disk



Smart OLTP

- **Special InfiniBand protocol** enables highest speed, lowest latency OLTP
- Ultra-fast transactions using DB optimized **flash logging** algorithms
- **Fault-tolerant In-Memory DB** by mirroring memory across servers

Smart Consolidation

- **Workload prioritization** from CPU to network to storage ensures QoS
- **4X** more Databases in same hardware

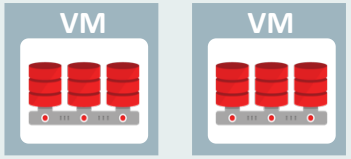
Smart System Software Introduced in 2015

Smart Analytics

- **5X** faster scans by converting data to **Columnar** format in Flash Cache
- **3X** faster **JSON/XML** by offloading to storage servers

Smart Consolidation

- Zero overhead **VMs**
- **Snapshots** for test/dev
- Set flash cache min size per DB to ensure QoS
- InfiniBand partitioning
- IPv6 for Ethernet

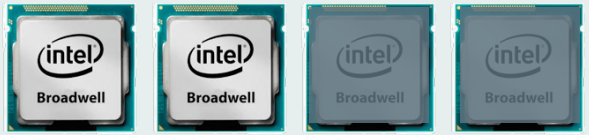


Smart OLTP

- **3X** faster OLTP messaging using **direct DB to InfiniBand** access
- **Instant** detection of node failure
- **Sub-second** capping of I/O latency by rerouting I/Os to faster storage

Smart Licensing

- **Capacity-on-Demand** reduces license cost by disabling unneeded cores
- **Trusted Partitions** limit license scope of specialized options



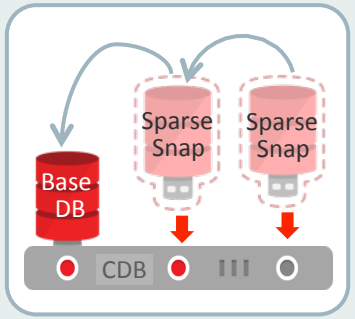
Preview: New Smart System Software

Smart Analytics

- **Database In-Memory** columnar format in storage server
- **Aggregation** in storage
- **Set membership** using new type of storage index

Smart Consolidation

- Hierarchical snapshots
- **2X** application connections*
- Automated **VLAN** creation*
- Add extra 10g Ethernet Card
- 64GB DIMMs for 2X Memory



Smart OLTP

- **Smart Fusion Block Transfer** eliminates log writes when moving blocks between nodes*
- Automated rolling upgrade across full stack
- **2X** faster **disk recovery**

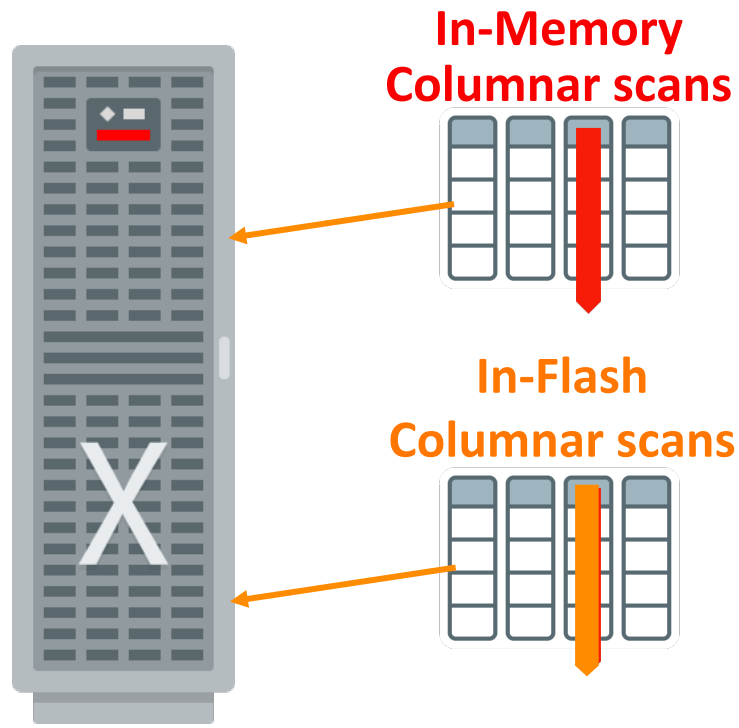
Smart Availability

- **Short Range Stretch (Extended) clusters**
- **4X** faster software updates*
- **High redundancy** Quorum disks on Quarter and Eighth racks*
- Storage Index preserved on rebalance*

***Already Released**

Upcoming: In memory format in Columnar Flash Cache

- 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



Upcoming release of Exadata Software

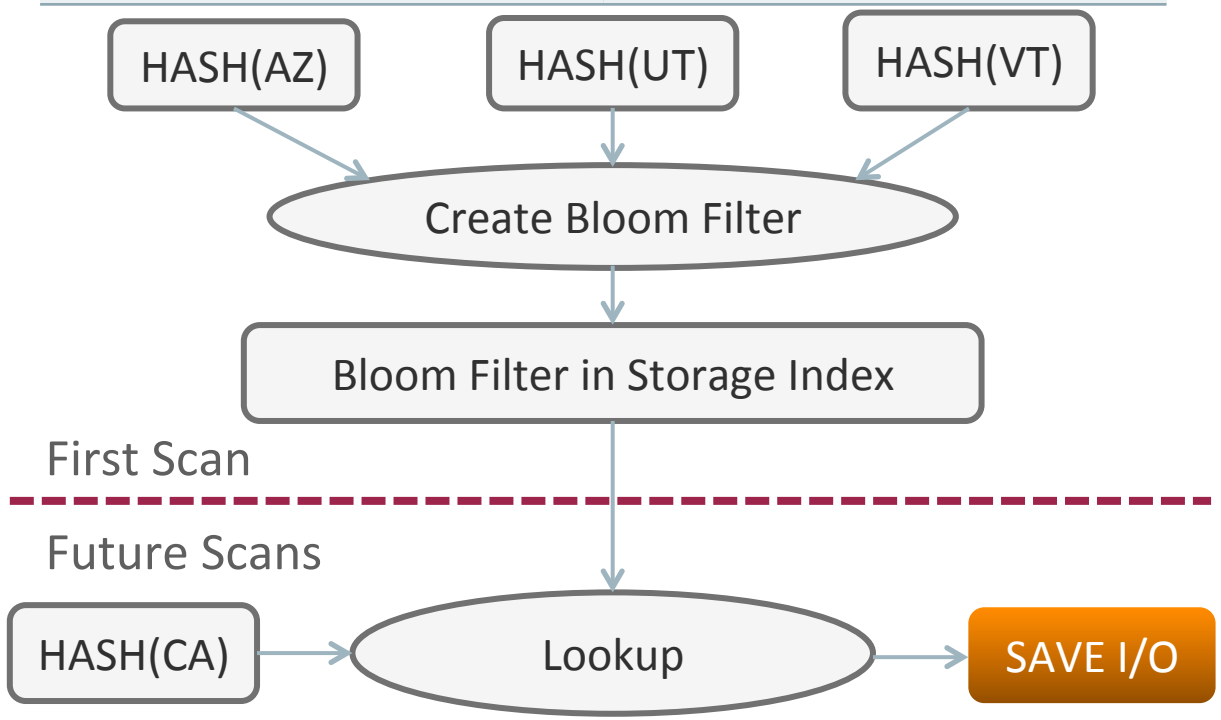
Upcoming: Storage Index Set Membership

- Storage Index
  - Currently contains up to 8 columns of min/max summary
  - Created automatically and kept in memory
  - Used to skip performing I/Os
- What about queries with low cardinality columns?

```
select name, address from travels
where origin='Sierra Leone' and dest='CA'
```

- Traditional min/max not good enough
- Database gathers stats and find that column has less than 256 distinct values
- Database requests storage to compute bloom filter
- Storage will compute distinct values and create a bloom filter
- Smart Scans check value 'CA' against bloom filter and saves performing I/O

ORIGIN	DEST	NAME	ADDRESS
Sierra Leone	AZ	Alice	…
Sierra Leone	UT	Bob	…
Sierra Leone	VT	John	



Upcoming release of Exadata Software

Upcoming: Join and Aggregation Smart Scan

- Extend In-Memory Aggregation technique into storage
- Find Sales per country

```
SELECT /*+ VECTOR_TRANSFORM  */ country_id, sum(amount_sold) amount_sold
 FROM customers, sales
 WHERE customers.cust_id = sales.cust_id
 GROUP BY customers.country_id
 ORDER BY customers.country_id;
```

- Storage cells scanning sales fact table will return tuples

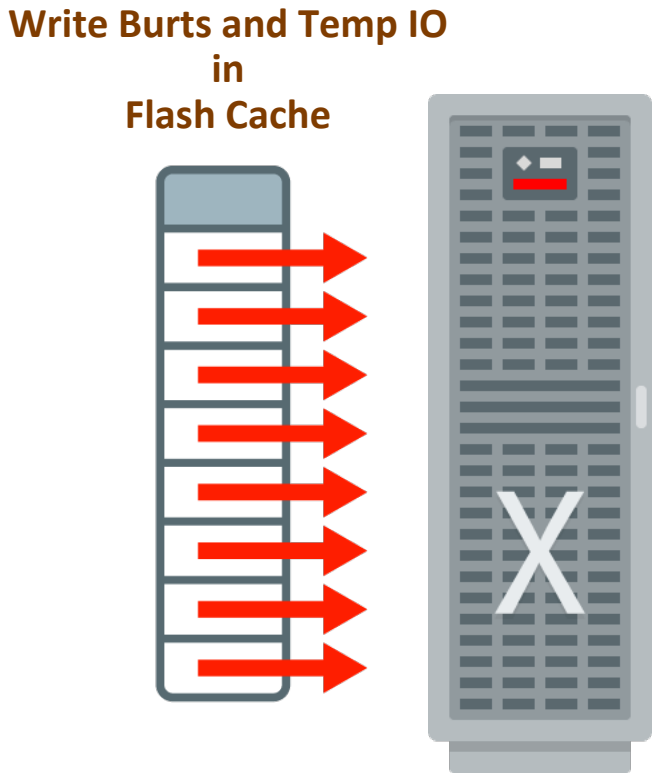
  { `country_id, sum_amount_sold` }

- Join and Aggregation offloaded to the storage server

12.2 Database and 12.2 Exadata Storage Server Software

Upcoming: Smart write bursts and temp IO in flash cache

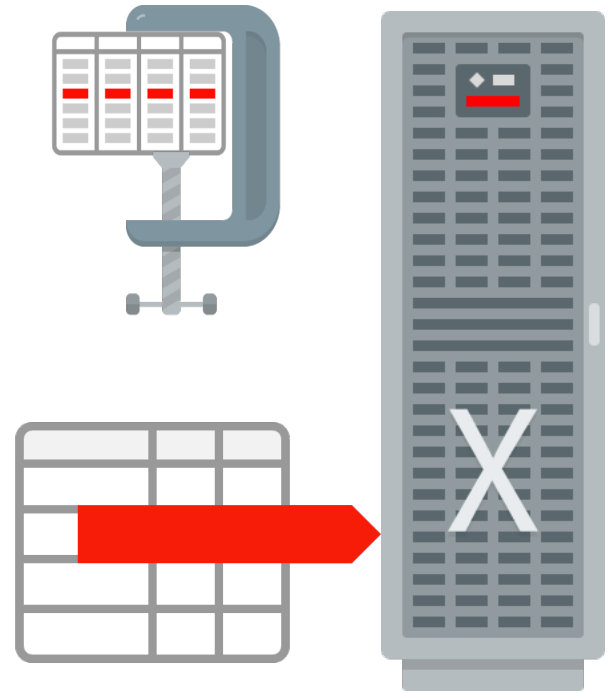
- Write throughput of four flash cards has become greater than the write throughput of 12-disks
- When database write throughput exceeds the throughput of disks, smart flash cache intelligently caches writes
- When queries write a lot of temp IO and it is bottlenecked on disk, smart flash cache intelligently caches temp IO
  - Writes to flash for temp spill reduces elapsed time
  - Reads from flash for temp reduces elapsed time further
- Smart flash cache prioritizes OLTP data and does not remove hot OLTP lines from the cache
- Smart flash wear management for large writes



Upcoming release of Exadata Software

Upcoming: Smart Analytics Software Features

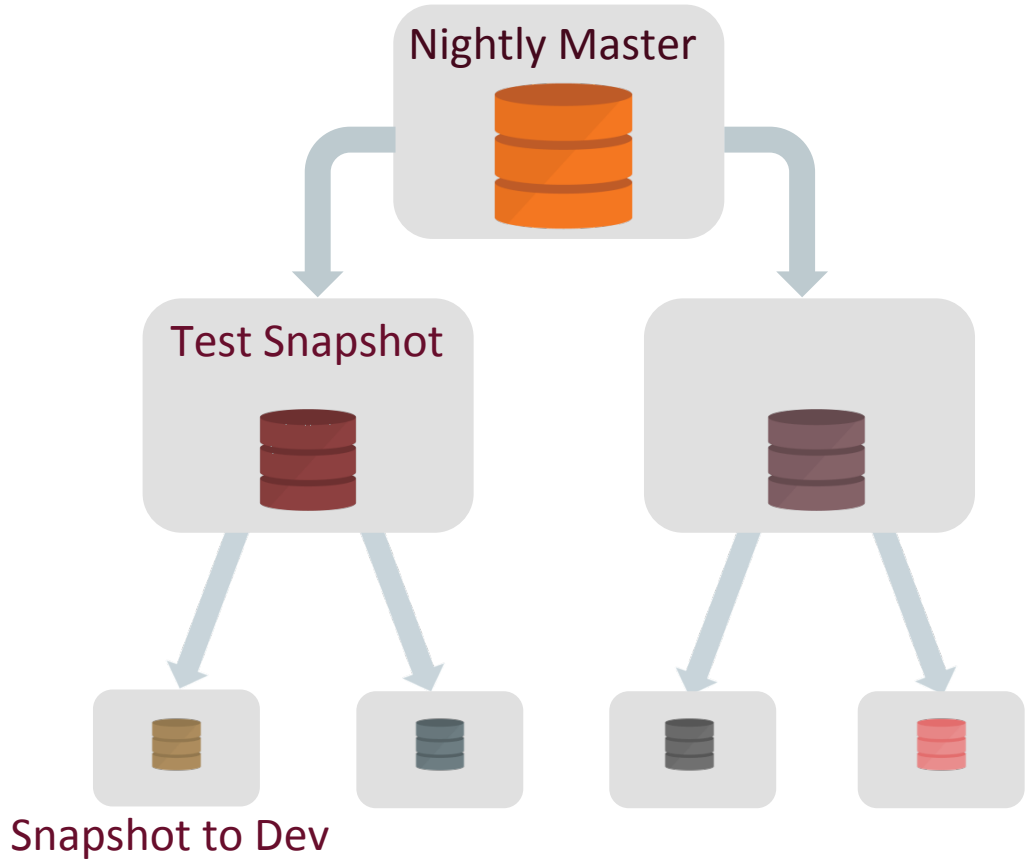
- Compressed Index Fast Full Scan
- Smart Scan VIEWs with LOBs, XML and JSON
  - not just tables
- AWR Enhancements
  - Diff report for Exadata section
  - Flash Cache Metrics
  - More granular histograms
- Up to 25% reduction in Storage Server CPU for SPARC SuperCluster during Smart Scans
  - Reduces endianness conversion overhead



12.2 Database and 12.2 Exadata Storage Server Software

Upcoming: Snapshots

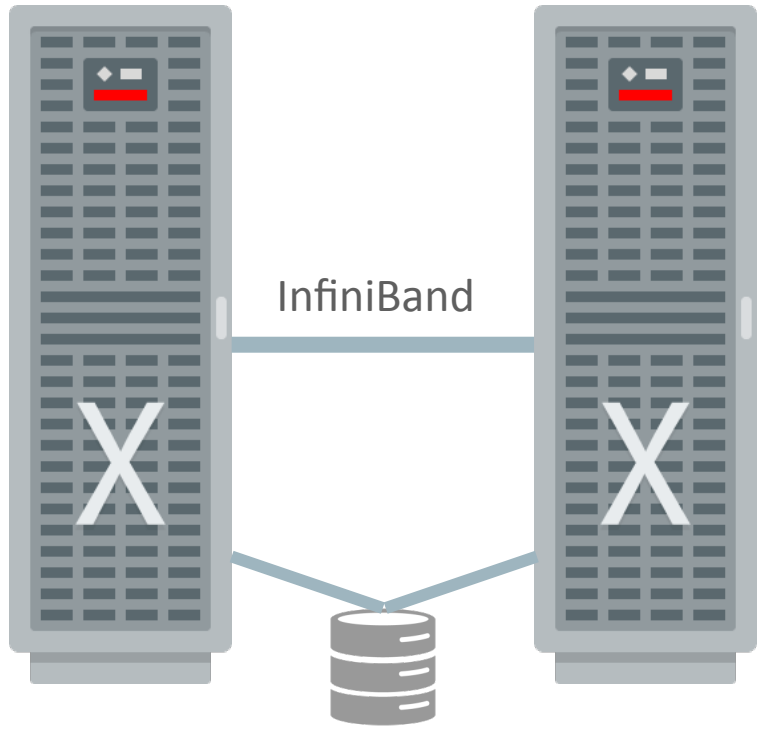
- Hierarchical Snapshots
  - Create snapshots of databases on previously created snapshots
  - Use case example
    - Development releases nightly build of the database
    - Tester creates a snapshot for himself and finds a bug
    - Tester creates a snapshot of his snapshot
    - Tester provides the new copy back to development for analysis
  - Syntax and technology remain unchanged
  - Works with pluggable and non-pluggable databases
- Sparse backup of snapshots
  - RMAN backs up the modified blocks and not the unchanged blocks from parent



12.2 Database and 12.2 Exadata Storage Server Software

Upcoming: Extended Distance Clusters

- Two sites and a quorum site
- InfiniBand connected for high performance
  - 100m optical cables in 2016 (best for fire cells)
- Created using ASM Extended Diskgroups
  - Nested failure groups
- Compute nodes at each site read data local to that site
- Data is written to all sites
- Smart Scans scan across cells on both sites increasing throughput
  - Row filtering, column projection, storage index, and flash cache provide extreme performance
- Data Guard continues to be the recommended DR solution



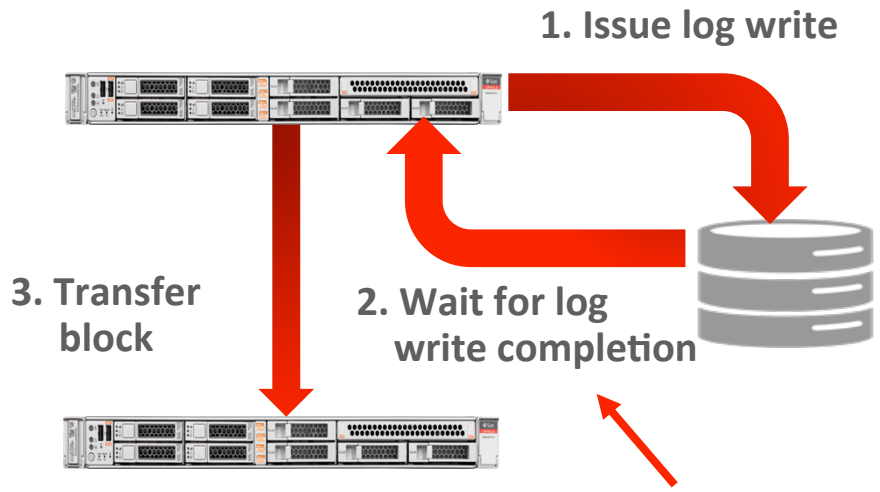
Quorum Failure Group

12.2 Database and 12.2 Exadata Storage Server Software

Smart Fusion Block Transfer

- OLTP workloads can have hot blocks that are frequently updated (e.g. right-growing index)
  - Log file must be written before transferring a hot block between instances so the block can be recovered
  - Adds latency and reduces throughput
- On Exadata, Oracle does not wait for the log write
  - Exadata ensures the log write completes before changes to block on another instance commit, guaranteeing durability
  - Wait for Log I/O during transfer of hot blocks is eliminated
  - Up to 40% throughput and 33% response time improvement in some heavily contended OLTP workloads

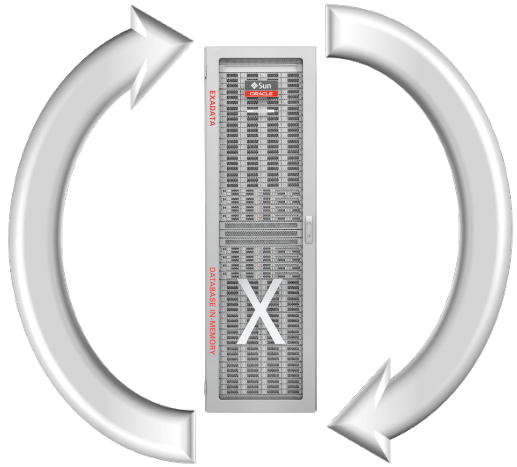
Prior Inter-Instance Block Transfer Protocol



Available with 12.1.0.2 BP12

Upcoming: Super Fast Software Updates

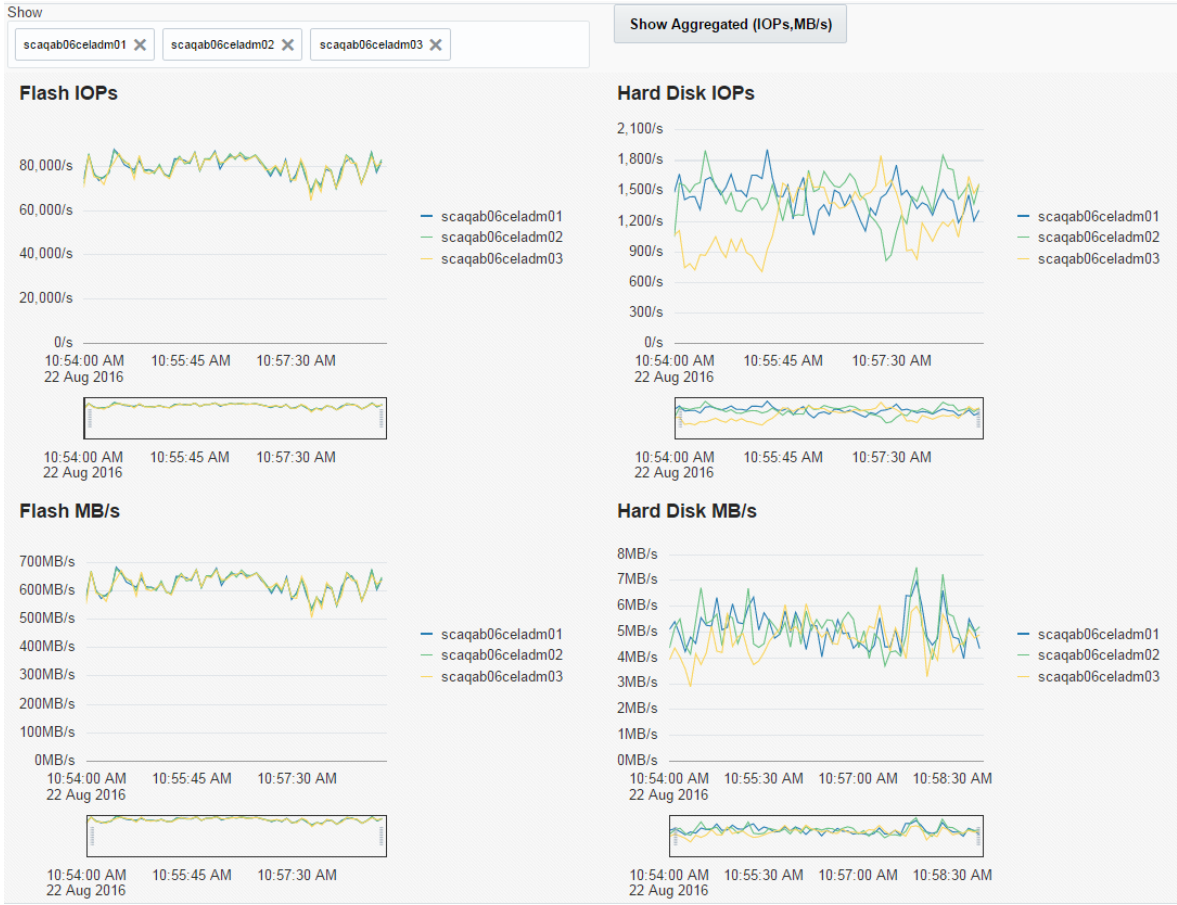
- 4x speed up in Storage Server Software Update
  - Parallel firmware upgrades across components such as hard disks, flash, ILOM/BIOS, InfiniBand card
  - Reduced reboots for Software updates
  - Use kexec where possible
- Manage a Cloud instead of managing a single rack
  - Use single patchmgr utility to upgrade hundreds of racks
- Enable patchmgr to run from a non-Exadata system and run as low privileged user



Upcoming release of Exadata Software

Upcoming: Extreme Manageability

- IPv6 + Virtual machine + VLAN deployments
- Get graphs from Exawatcher
- Make DNS, NTP, and other IP address changes online
- Seamless customer service with Automatic Service Requests sending diagnostic attachments
- Manage Compute nodes using a RESTful service
  - ExaCli enabled for compute nodes in addition to storage cells
- Much faster rebalance with improved flash cache hit ratio during rebalance
- Secure Erase during hardware retirement



Upcoming release of Exadata Software

Exadata Advantages Increase Every Year

Transformational OLTP, Analytics, Consolidation

Cloud Without Compromise

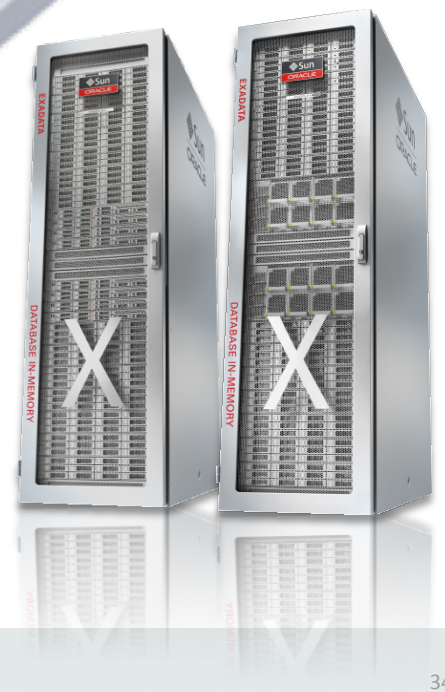
- **Exadata Cloud Service**
- In-Memory Columnar in Flash
- Smart Fusion Block Transfer
- In-Memory Fault Tolerance
- Direct-to-wire Protocol
- JSON and XML offload
- Instant failure detection
- Network Resource Management
- Multitenant Aware Resource Mgmt
- Prioritized File Recovery
- 3D V-NAND Flash
- Software-in-Silicon
- IO Priorities
- Data Mining Offload
- Offload Decrypt on Scans
- Tiered Disk/ Flash

Smart Software

- Database Aware Flash Cache
- Storage Indexes
- Columnar Compression
- PCIe NVMe Flash
- Unified InfiniBand
- Smart Scan
- InfiniBand Scale-Out

Smart Hardware

- DB Processors in Storage
- Scale-Out Storage
- Scale-Out Servers



Integrated Cloud

Applications & Platform Services