Exadata Technical Deep Dive: Architecture and Internals

ORACLE OPEN WORLD

October 1–5, 2017
SAN FRANCISCO, CA

Kothanda (Kodi) Umamageswaran
Vice President, Exadata Development

Gurmeet Goindi
Exadata Product Management

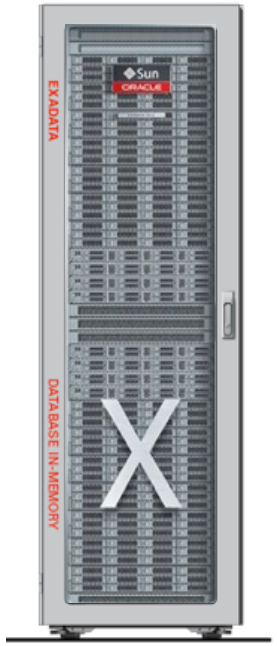
ORACLE

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.

Exadata Vision

Dramatically Better Platform for All Database Workloads



- Ideal Database Hardware - Scale-out, database optimized compute, networking, and storage for fastest performance and lowest costs
- Smart System Software – specialized algorithms vastly improve all aspects of database processing: **OLTP, Analytics, Consolidation**
- Full-Stack Automation – Automation and optimization of: configuration, updates, performance, resource management

Identical On-Premises and in Cloud

Proven at Thousands of Ultra-Critical Deployments since 2008

- **Best for all Workloads**
- **Petabyte Warehouses**
- **Online Financial Trading**
- **Business Applications**
  - SAP, Oracle, Siebel, PSFT, …
- **Massive DB Consolidation**
- **Public SaaS Clouds**

4 OF THE TOP 5

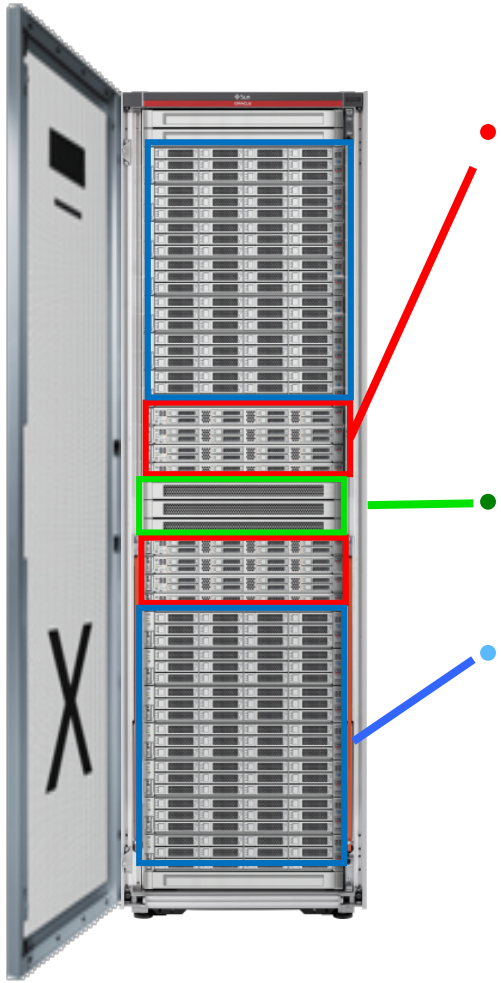
BANKS, TELECOMS, RETAILERS RUN EXADATA

Introducing Exadata X7

Continue Tradition of State-of-the-Art Hardware

ORACLE

Exadata X7 Hardware (changes in red)



- **Scale-Out 2-Socket Database Servers**
  - 20% to 40% faster CPUs – latest 24 core Intel Skylake
  - 150% faster Ethernet – 25 GigE client connectivity
  - 50% more DRAM capacity and throughput

Database Server

- **Ultra-Fast Unified InfiniBand Internal Fabric**

High-Capacity (HC) Storage

- **Scale-Out Intelligent 2-Socket Storage Servers**
  - Intel 10 core Skylake CPUs offload database processing
  - 25% more disk capacity - 10TB Helium Disk Drives
  - 100% more flash capacity - 6.4 TB Hot swappable NVMe Flash

Extreme Flash (EF) Storage

Exadata X7 Database Servers - Details

- X7-2 Database Servers
  - 48 Cores, 2 Twenty-four-Core Intel® Xeon® 8160 Processors
  - 384 GB (12 x 32GB) – Expandable to 1.5 TB
  - 25 Gigabit Ethernet for client connectivity

- X7-8 Database Servers
  - 48 Cores, 8 Twenty-four-Core Intel® Xeon® 8160 Processors
  - 3 TB (48 x 64GB) – Expandable to 6 TB
  - 25 Gigabit Ethernet for client connectivity
  - 2 x 6.4 TB 2.5-inch Flash Accelerator F640 PCIe Drives (Hot-Pluggable)

Exadata X7 Storage Servers - Details

- X7-2 Extreme Flash (EF) Storage Server
  - 2 Ten-Core Intel® Xeon® 4114 Processors, 192 GB DRAM
  - Hot Plug 8 x 6.4TB PCIe rear-mounted flash cards
  - Capacity increases to 51.2TB
  - 2x 150 GB M.2 Drives performing Boot and Rescue Functions

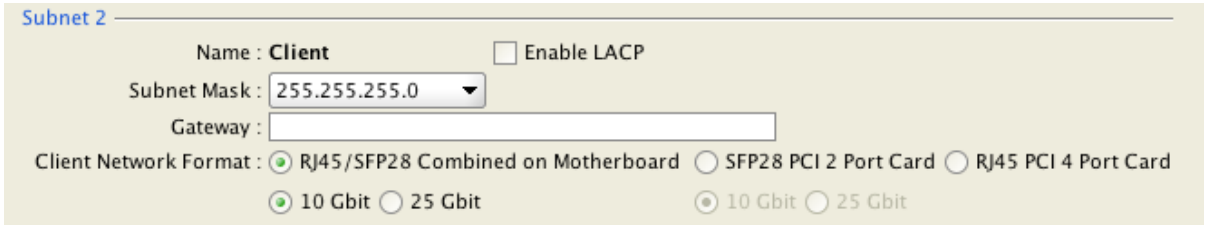
- X7-2 High Capacity (HC) Storage Server
  - 2 Ten-Core Intel® Xeon® 4114 Processors, 192 GB DRAM
  - Hot Plug 4 x 6.4TB PCIe rear-mounted flash cards
  - 12 x 10 TB 7.2K RPM High Capacity SAS (hot-swap) – 3.5” disk size
  - 2x 150 GB M.2 Drives performing Boot and Rescue Functions

Faster Client Connectivity with More Choices

- State-of-the-art **25 Gigabit Ethernet** connectivity from database servers to clients
- Optional dual port **25 Gigabit Ethernet** fiber card

  Or

- Optional quad port 10 Gigabit Ethernet copper card
- All ports can be configured using OEDA
- New management switch



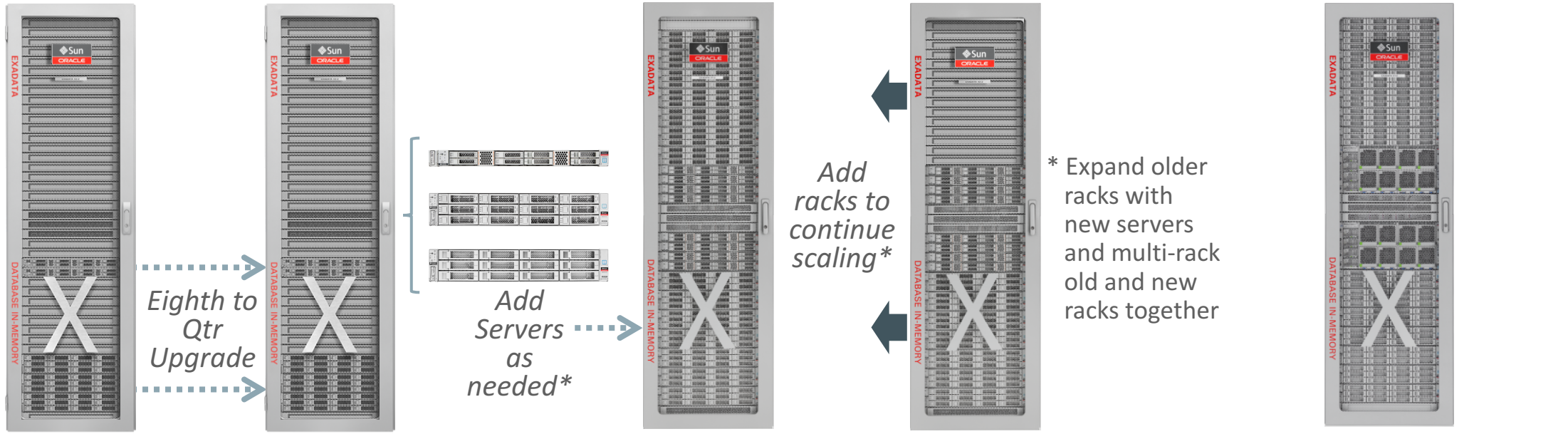
Exadata X7: Biggest Change in Exadata in Many Years

- Intel® Xeon® Processor Scalable Family Skylake
- 50% more memory channels and faster memory
- 50% more default memory, memory expansion in factory
- 150% faster 25Gb/sec Ethernet, more connectivity options
- 100% more capacity, faster, hot swappable NVME PCIe Flash
- New M.2 hot swappable boot drives
- New Ethernet management switch
- 25% larger Disk Drives & new disk controller HBA

ORACLE

Scalable Configurations Right-Size Your Investment

Elastic Hardware Configurations



Capacity-on-Demand Software Licensing

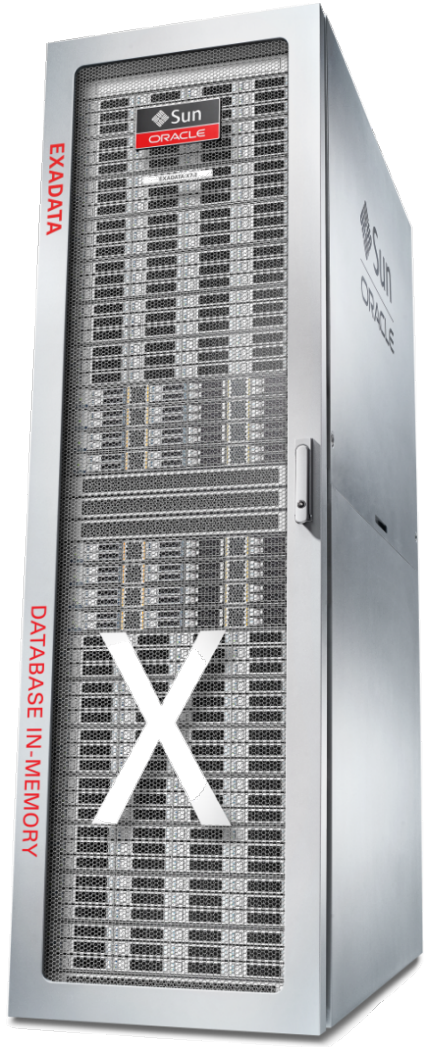
- **Enable compute cores as needed, subject to minimums**
- **License Oracle software for enabled cores only**

* **14 cores minimum** per DB server (max 48 cores)
* **8 cores minimum** per Eighth Rack DB server (max 24 cores)

Breakthrough Database Performance

Exadata X7 Full Rack*

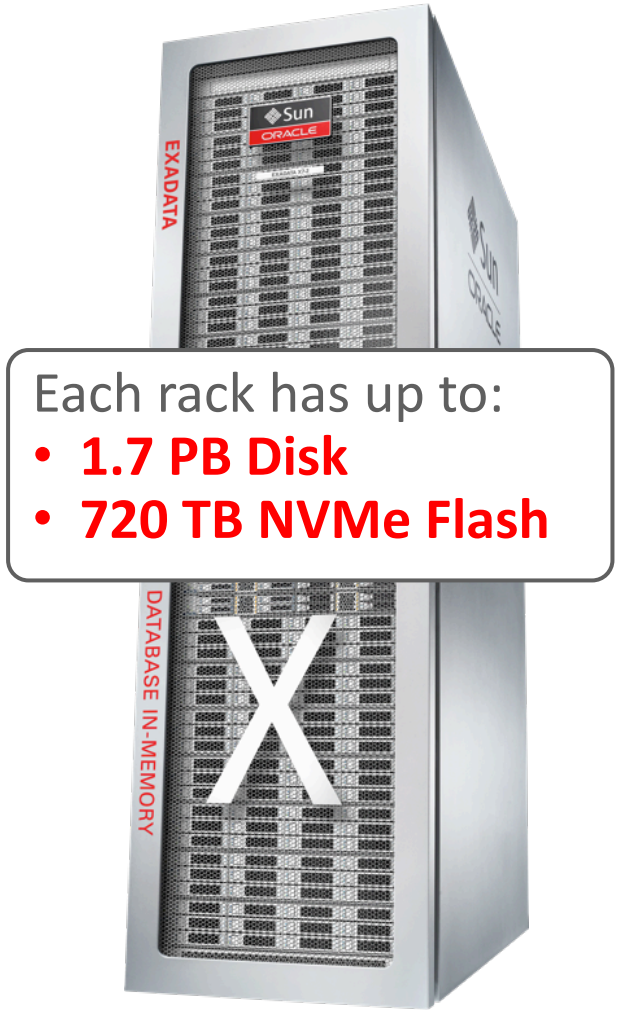
Read	**5.97 million** IO/sec	**350** GB/sec
Write	**5.4 million** IO/sec	n/a
IO Latency	**250** microseconds	n/a



*Exadata Rack with 10 DB servers and 12 Extreme Flash storage servers

Exadata X7-2 and X7-8 Performance Improvements vs X6

- **350 GB/sec** IO Throughput
  - **17%** more (vs Exadata X6)
- **5.97 Million** OLTP Read IOPs
  - **50%** more IOPs (vs Exadata X6) under 250usec = 3.5M
- **40%** CPU improvement for Analytics
- **20%** CPU improvement for OLTP
  - 40% on X7-8
- Dramatically faster than leading all-flash arrays

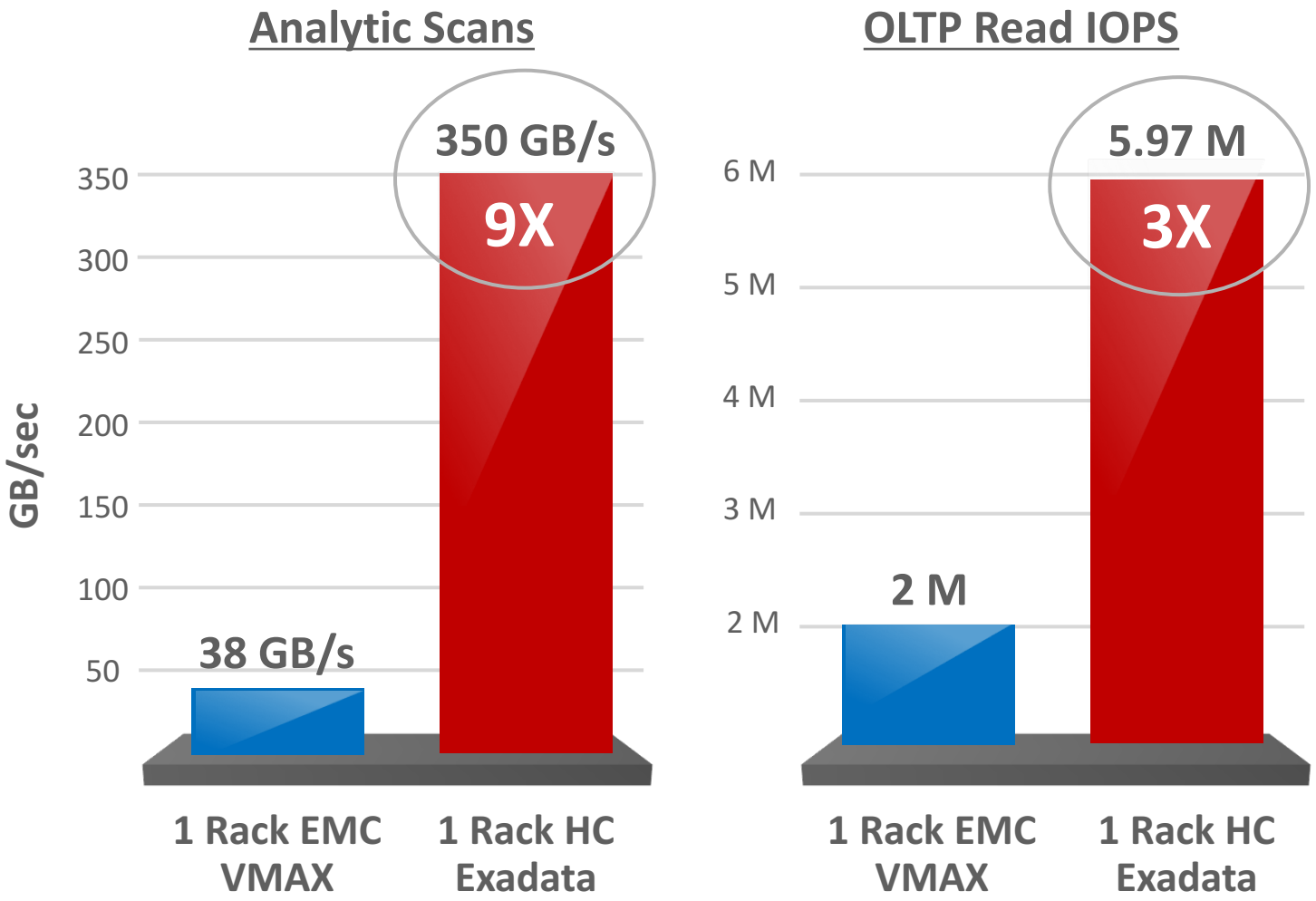


Each rack has up to:
- **1.7 PB Disk**
- **720 TB NVMe Flash**

Exadata X7 I/O is Dramatically Faster than All-Flash EMC

One **High Capacity** Exadata beats the fastest EMC VMAX **all-flash** array in every performance metric

- **9X more throughput**
- **3X more IOPS**
- **2X faster latency**



Exadata Smart Software

Continue Tradition of Adding Major Differentiators

Exadata 12.2.1.1.0 and 18.1.0.0.0 Highlights

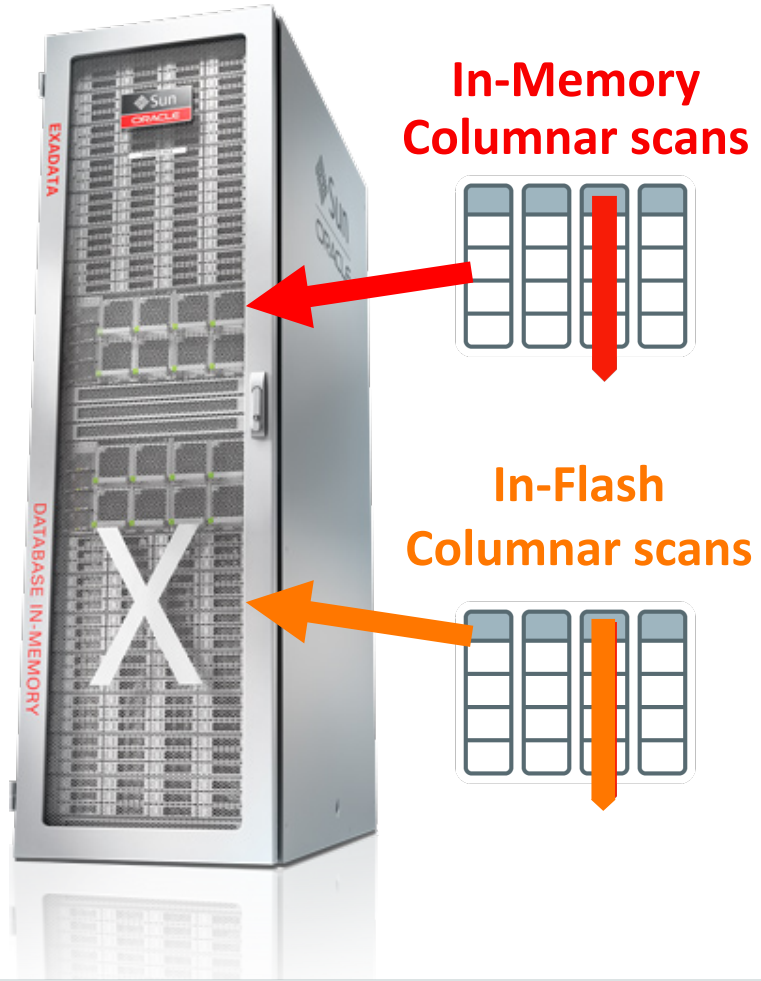
- Over 40 unique software features and enhancements in a year
  - Better analytics, better transaction processing, better consolidation, more secure, faster and more robust upgrades, and easier to manage
- Complete investment protection
  - All new software features work on all supported Exadata hardware generations
- Full storage offload functionality for Database 12.2
  - Database 11.2, 12.1, and 10.2 can coexist along side 12.2 on the same system
- Updated Oracle Linux kernel and Oracle VM improve robustness and scalability
  - Oracle Linux 6.9 with UEK4, Oracle Virtual Machine 3.4.4

ORACLE
EXADATA

Smart Analytics

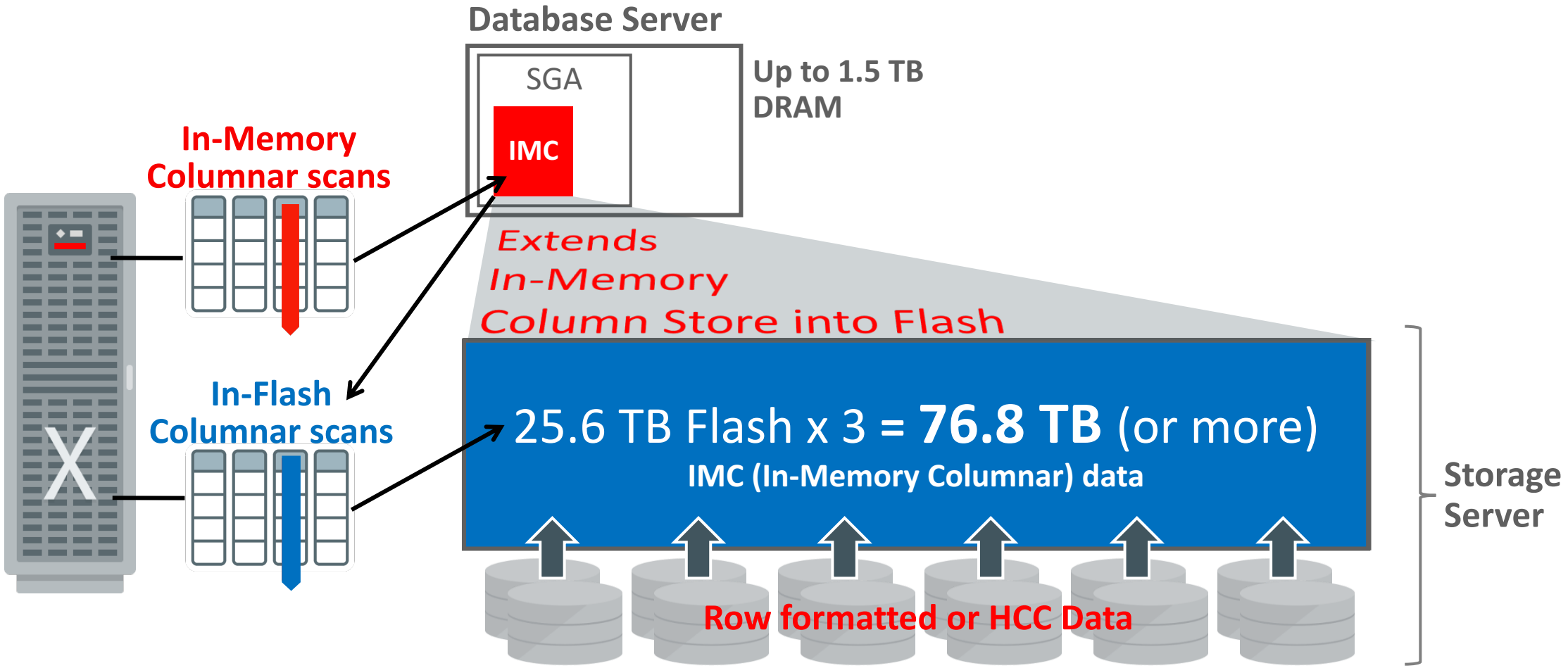
Analytics: Exadata Brings In-Memory Analytics to Storage

- Exadata automatically transforms table data into In-Memory DB columnar formats in Exadata Flash cache
  - Enables fast vector processing for storage server queries
- Faster decompression speed than Hybrid Columnar Compression
- Additional compression for OLTP compressed or uncompressed tables in flash – new in 18.1
- 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
- Uniquely optimizes next generation Flash as memory



Analytics: In-Memory Columnar Formats in Flash Cache

3 - 4x Overall Analytics Performance Improvement

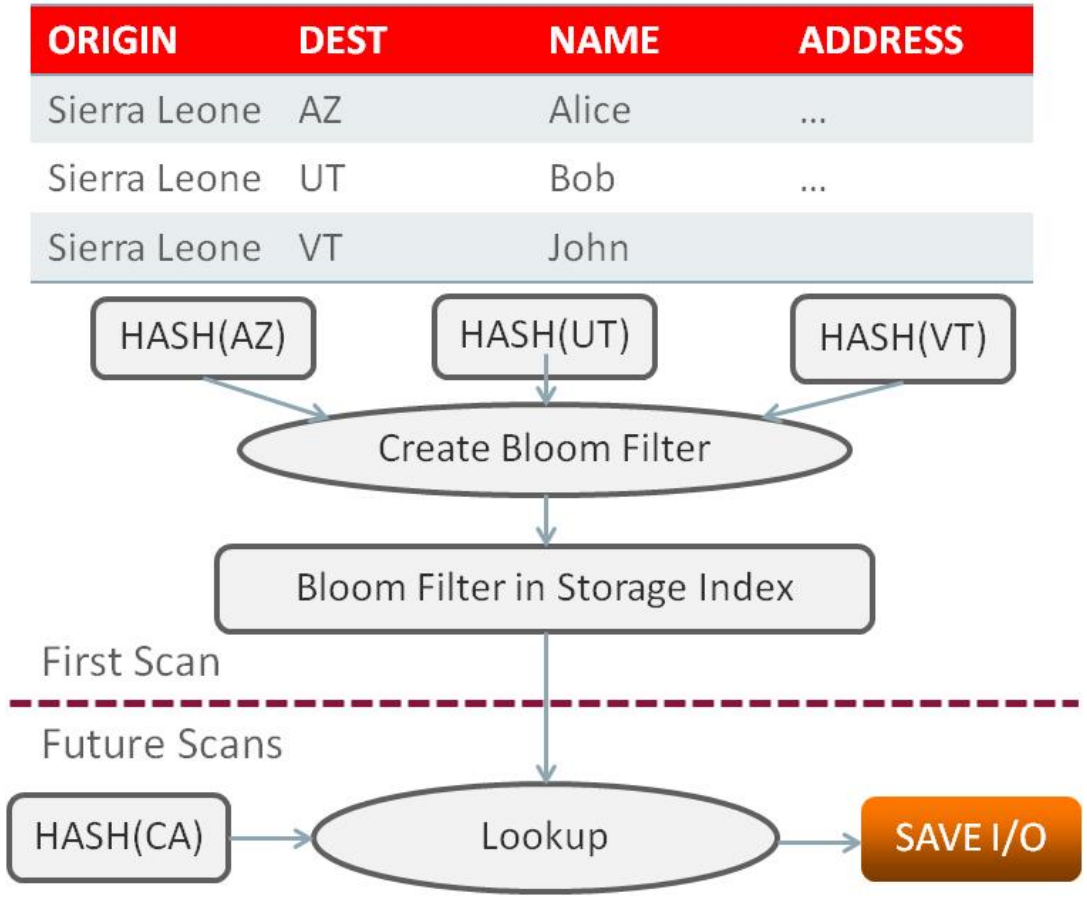


Analytics: Automatic Storage Index Set Membership

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

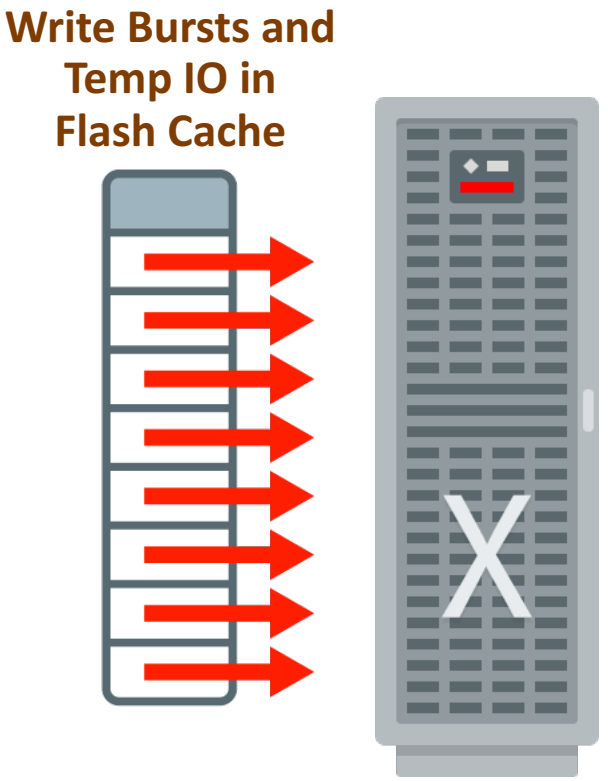
```
select name, address from travels
where origin='Sierra Leone' and dest='CA'
```

- Traditional min/max not good enough
- Database gathers stats and find that column has less than 400 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
- Supports Database 12.1.0.2 and Database 12.2.0.1



Analytics: Automatic Write Bursts and Temp IO in Flash Cache

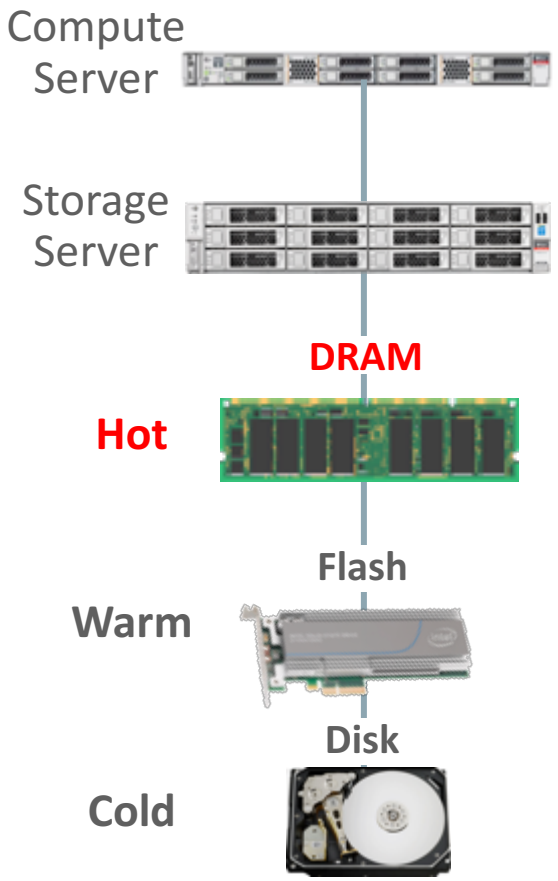
- Write throughput of four flash cards has become greater than the write throughput of 12-disks
- When database write throughput exceeds throughput of disks, Smart Flash Cache intelligently caches writes
- When queries write a lot of temp IO, Smart Flash Cache intelligently caches temp IO
  - Writes 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
- Supports Database 11.2.0.4, 12.1.0.2 and 12.2.0.1



Accelerates Large Joins and Sorts and Large Data Loads

Smart OLTP and Consolidation

OLTP: Exadata Brings In-Memory OLTP to Storage



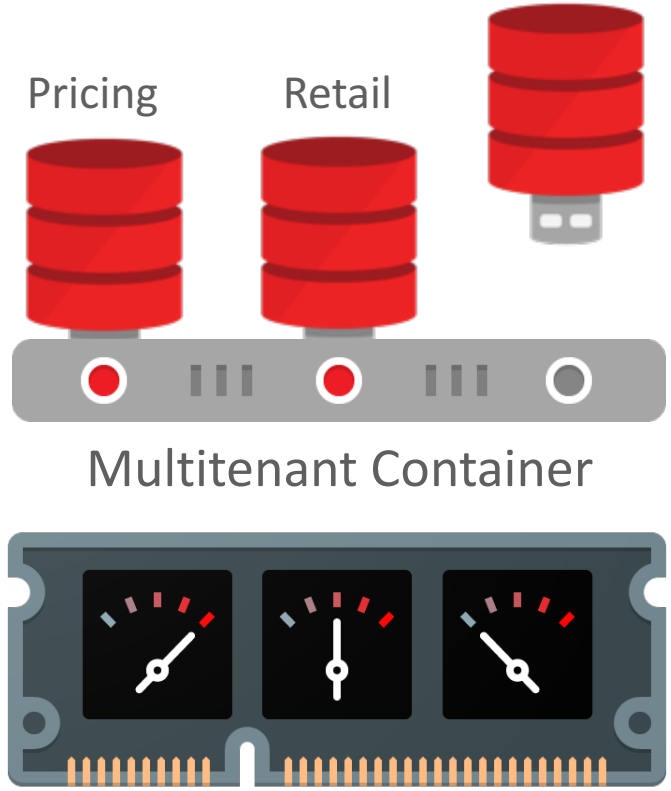
- Exadata Storage Servers add a memory cache in front of Flash memory
  - Similar to current Flash cache in front of disk
- Cache is **additive** with cache at Database Server
  - Only possible because of tight integration with Database
- **2.5x Lower latency for OLTP IO** – 100 usec
- Up to **21 TB of DRAM for OLTP acceleration** with Memory Upgrade Kit
  - Compare to 5TB of flash in V2 Exadata

OLTP: Exadata Commit Cache & RDMA for UNDO

- Exadata Commit Cache
  - In-Memory commit cache per DB Instance that logs commit time
  - DB instance checks against this cache for committed transactions
  - Saves blocks being passed around, no wait for log flush
  - RDMA the commit cache from a remote instance at one shot improving batching
  - Can eliminate up to 60% cache fusion block traffic for batch
- RDMA read for UNDO blocks
  - When a query has to roll a change out of a block to get a consistent version, and the change was made on another instance, an undo block must be fetched from that instance
  - Use RDMA to read UNDO blocks from other instance
  - Reduces latency from 50us to 10us
- Completely automatic and transparent on Exadata

Consolidation: Up To 4,000 Pluggable Databases

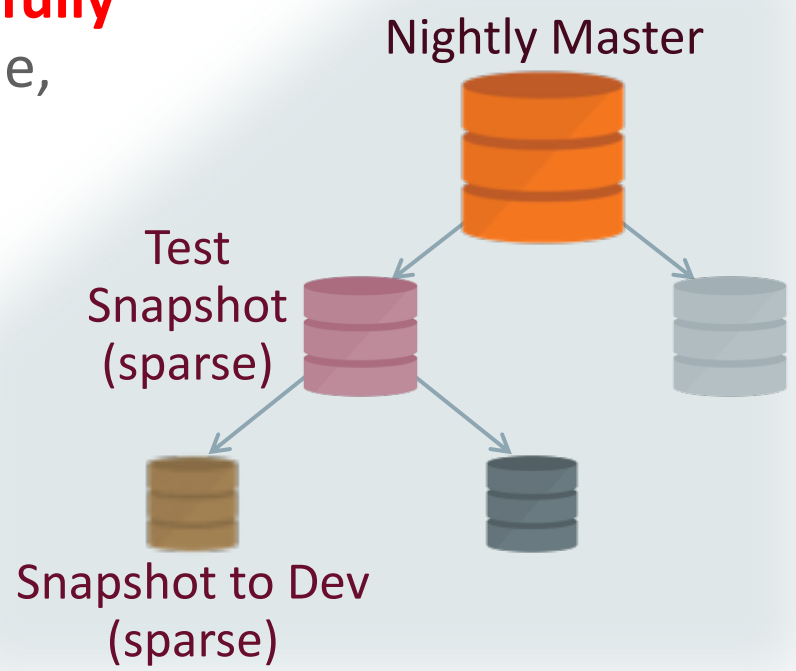
- Exadata offers unique end-to-end resource management and consolidation capabilities
- Multitenant Option now allows greater than 252 Pluggable Databases within a single Oracle Multitenant Container Database
  - Up to thousands
- Exclusively available on Exadata, SuperCluster and DB PaaS



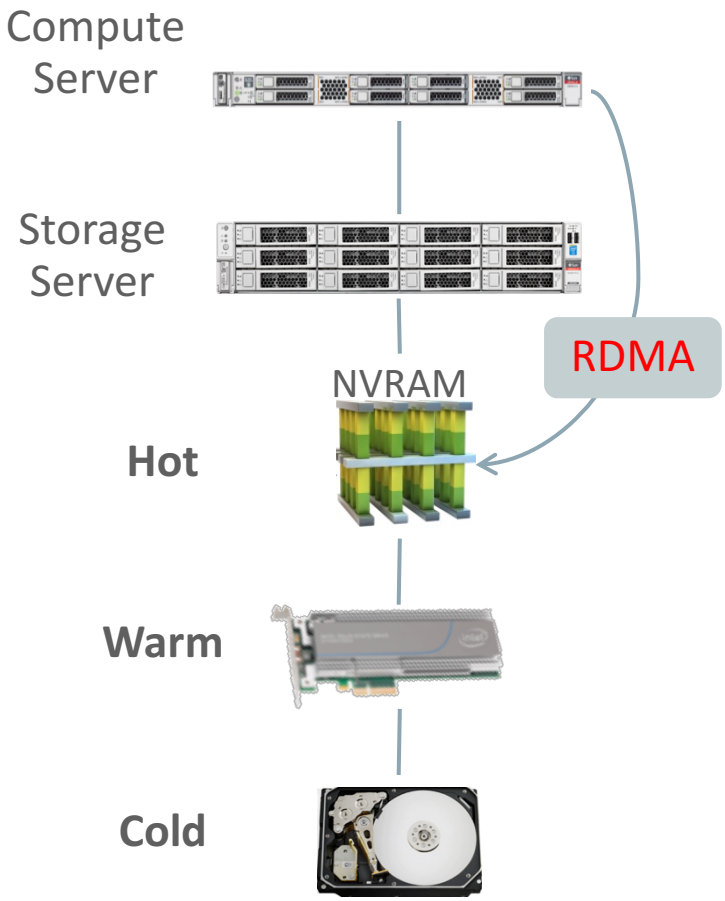
Consolidation: Exadata Hierarchical Snapshots

Develop, Test and Deploy on Exadata for Exadata – Full Lifecycle Value

- Why Exadata Snapshots
  - Ideal for space-efficient read-only or read-write snapshots of an Oracle database that you can use for development, testing, etc.
  - Used by developers and testers who need to validate a **fully functional Exadata environment** (e.g. Smart Flash Cache, Smart Scan, HCC)
- Hierarchical Snapshots (example)
  - Development releases nightly database build
  - Tester creates a snapshot and finds a bug
  - Tester creates a snapshot of her snapshot
  - Provides the new copy to development for analysis



Preview: Non-Volatile Memory Cache in Exadata Storage



- Exadata Storage Servers will add Non-Volatile memory cache in front of Flash memory – Intel 3D X-Point
- **RDMA** enables order of magnitude faster remote access of stored data
  - Direct access to NVRAM gives 20x lower latency than Flash
- NVRAM used as a **cache** effectively increases its capacity by 10x vs using NVRAM directly as expensive storage
  - Cost-effective to run multi-TB databases completely in memory
- Expensive NVRAM shared across servers for lower cost
- NVRAM mirrored across storage servers for fault-tolerance

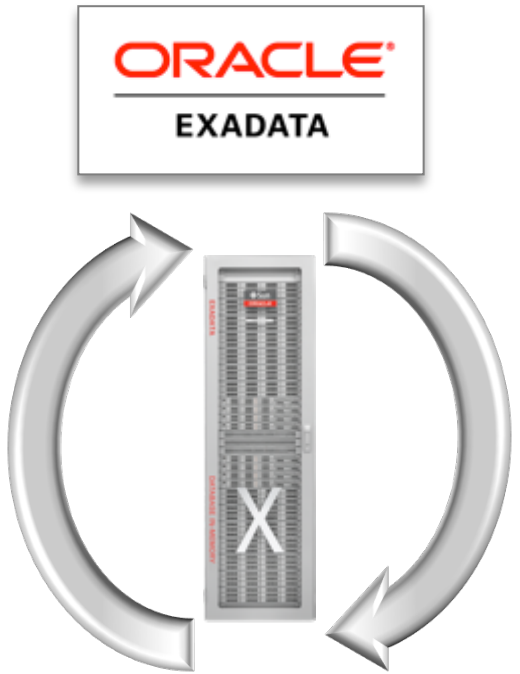
Availability and Manageability

Availability: Efficient Rebalance and Restore

- Intelligent and flexible rebalance power setting
  - Dynamically change ASM_POWER_LIMIT
- ASM rebalance restores redundancy first
  - Drastically reduces secondary failure exposure window
  - Exposed via new REBUILD phase in v$asm_operation
- Exadata leverages flash cache for rebalance reads
  - Improves performance of redundancy restoration by up to 30%

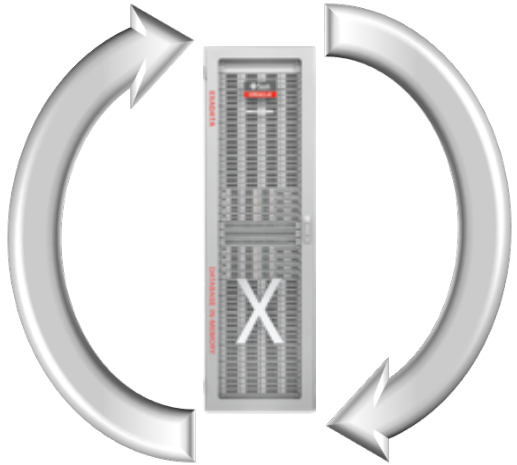
Manageability: Automated Cloud Scale Software Upgrade

- New Upgrade Process in 18.1
  - Point the Storage Server to the Software Store
  - Storage Server downloads new software in the background
  - User schedules time of software upgrade
  - Storage Servers automatically upgrade in rolling fashion online
- Benefits
  - Simpler and faster Cloud and On-Premises upgrades
  - Just need to schedule time of upgrade
  - One software repository for hundreds of machines
- Existing upgrade process continues to work



Manageability: Super Fast and Robust Software Updates

- Oracle Public Cloud is the largest Exadata deployment with hundreds of Exadata Database Machine deployed
  - Each software release goes through thousands of upgrade cycles
  - All upgrade utilities get exercised thousands of times as well
  - Contributes to the robustness of the software release and utilities
- 5x 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
- Database node update is 40% faster in 18.1
- 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



Manageability: Infrastructure improvements

ORACLE EXADATA

- Secure Boot is enabled for X7 Bare Metal
- New Do-Not-Service LED in X7 to prevent multiple failures on a scale out system
- Much faster Ethernet performance with Oracle VM
  - Coupled with X7 hardware running at 25Gb/s
- Oracle Exadata Deployment Assistant (OEDA) provides CLI interface to scale the deployments and script them for the Cloud
- Look for “Exadata Whats New” section in the doc

ORACLE

A Choice of Exadata Deployment Models

On-Premises

X7-2 X7-8

Customer Data Center
Purchased
Customer Managed

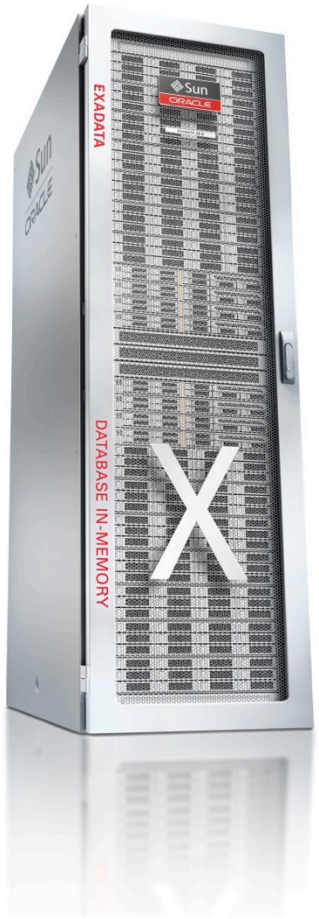
Cloud at Customer

Customer Data Center
Subscription
Oracle Managed

Public Cloud Service

Oracle Cloud
Subscription
Oracle Managed

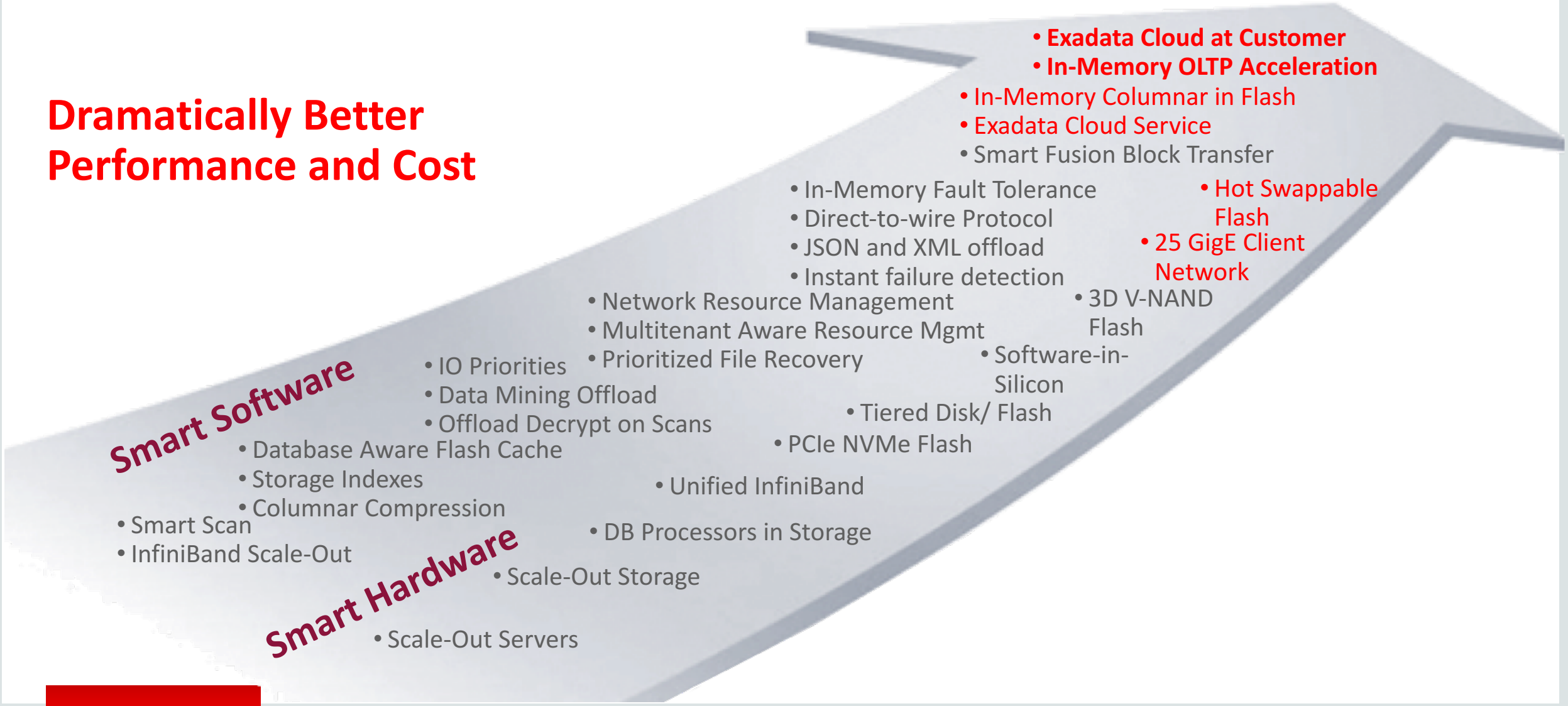
Exadata Innovations Deliver Value to Customers



- Industry's first smart scale-out storage
- Industry's first RDMA and InfiniBand for converged networking
- Industry's first platform to deliver NVMe Flash
- Only Enterprise Storage to make the leap to Public Cloud
- Only Database Machine to make the leap to Public Cloud
- And now:
  - ***Industry first In-Memory Performance in Storage***
  - ***Industry first Mission Critical Cloud at Customer Platform***

Exadata Advantages Increase Every Year

Dramatically Better Performance and Cost



Stay Informed During and After OpenWorld

Twitter: @OracleExadata, @ExadataPM @OracleBigData, @[illegible]ructure Follow #CloudReady

LinkedIn: Oracle IT Infrastructure– Oracle Showcase Page
Oracle Big Data – Oracle Showcase Page

Learn More
oracle.com/exadata

Integrated Cloud

Applications & Platform Services


ORACLE®

Appendix

Smart Integration for Full-Stack Simplicity

Unique Full-Stack Integration

- All layers pre-configured, pre-tuned, pre-debugged
  - DB, OS, drivers, firmware, network, servers, storage

Unique Full-Stack Community

- All users run identical full stack
- You leverage work of others
  - Bank tested full-stack HA
  - Telco tested full-stack scaling
  - Government tested full security

Unique Full-Stack Support

- One Support team expert in and accountable for full stack
- Oracle performs free full-stack updates and 24/7 monitoring

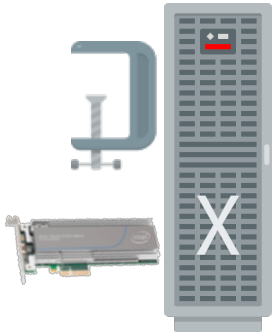
Unique Full-Stack Management

- Full-Stack management tool
- Drill down from DB to storage, and up from storage to DB

Unique Smart Software for Database

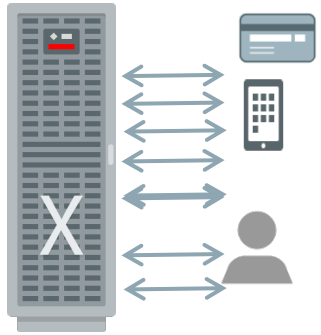
Smart Analytics

- Move queries to storage, not storage to queries
- Automatically offload and parallelize queries across all storage servers
- 100x Faster Analytics



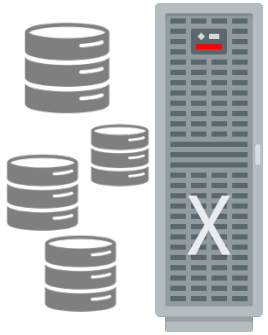
Smart Storage

- Hybrid Columnar Compression reduces space usage by 10x
- Database aware Flash Caching gives speed of flash with capacity of disk
- Storage Index IO reduction



Smart OLTP

- Special InfiniBand protocol enables 3x faster OLTP messaging
- Ultra-fast DB optimized flash logging
- Instant detection of Node Failure and IO issues



Smart Consolidation

- Critical DB messages always jump to head of queue for ultra-fast latency
- CPU, I/O, network prioritized to achieve end-to-end quality of service
- 4x more Databases in same hardware

Dozens of Additional Unique Capabilities

ORACLE

Non-Volatile Memory Opens up New Opportunities

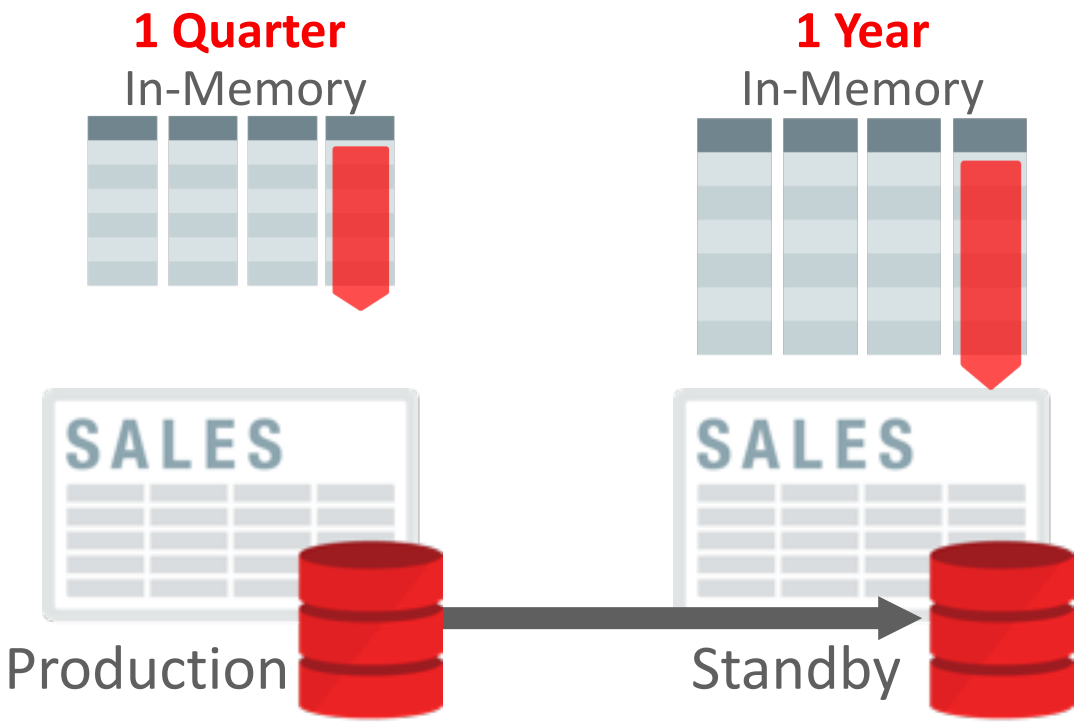
- **High capacity** makes it cost-effective to run multi-TB databases completely in memory
  - The majority of OLTP databases will fit
  - Can combine relatively smaller amount of DRAM for performance with huge Non-Volatile memory
- **Speed** of Non-Volatile memory changes dynamics of storage industry
  - Persistence moves to servers instead of storage arrays
  - Putting Non-Volatile memory behind slow storage network loses much of performance gains
  - More attractive to build **Super fast server storage**, or **super fast server cache** in front of shared disk
- **RDMA operations** enable order of magnitude faster remote access of stored data
  - No software interaction required on remote node, network card directly reads/writes memory

Exadata Full-Stack Automation

Continue Tradition of Database Optimizations

In-Memory Analytics on Active Data Guard Standby

Make Full Use of Memory on Standby – Offload Production Exadata



- **Real-time analytics with no impact on production database**
- **Can populate different data from production database**
  - Use new DISTRIBUTE BY SERVICE to determine where to populate a table
- **Exclusively available on standby Exadata and DB PaaS**

Analytics: Join and Aggregation Smart Scan

NEW IN DB 12.2

- Extend In-Memory Aggregation technique into storage (vector joins and vector aggregation)
- 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 return tuples

  `{country_id, sum_amount_sold }`

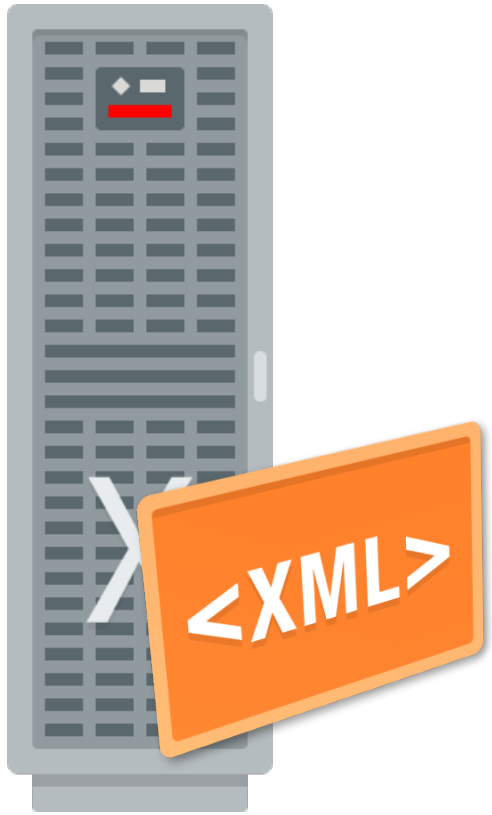
- Join and Aggregation offloaded to the storage server

ORACLE

Analytics: More Smart Scan Enhancements

NEW IN DB 12.2

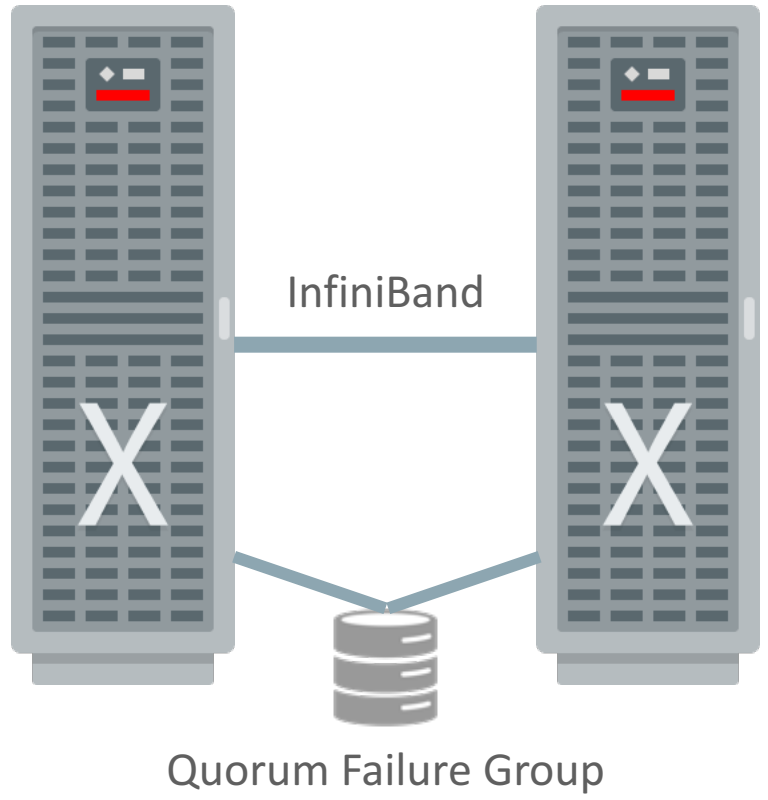
- Smart Scan Offload for Compressed Index Scan
- Smart Scan enhancements for XML
  - Enhancements to XMLExists, XMLCast and XMLQuery
- Smart Scan offload enhancements for LOBs
  - Extended to "LENGTH, SUBSTR, INSTRM CONCAT, LPAD, RPAD, LTRIM, RTRIM, LOWER, UPPER, NLS_LOWER, NLS_UPPER, NVL, REPLACE, REGEXP_INSTR, TO_CHAR"



OLTP: Extended Distance Clusters

NEW IN DB 12.2

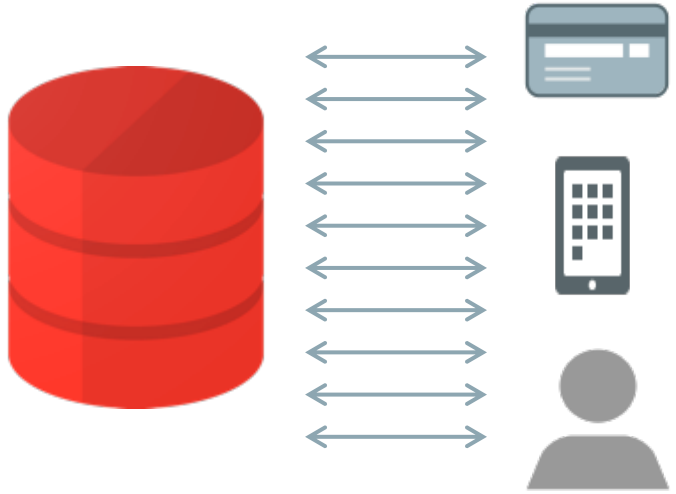
- Data Mirroring between two nearby sites
- InfiniBand connected for high performance
  - Limited to 100m optical cables in 2016 (best for fire cells)
- Implemented using 12.2 ASM Extended Diskgroups
  - Nested failure groups
- Compute nodes at each site read data local to that site
- Data is written 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



OLTP: New Redo Log Write Metrics

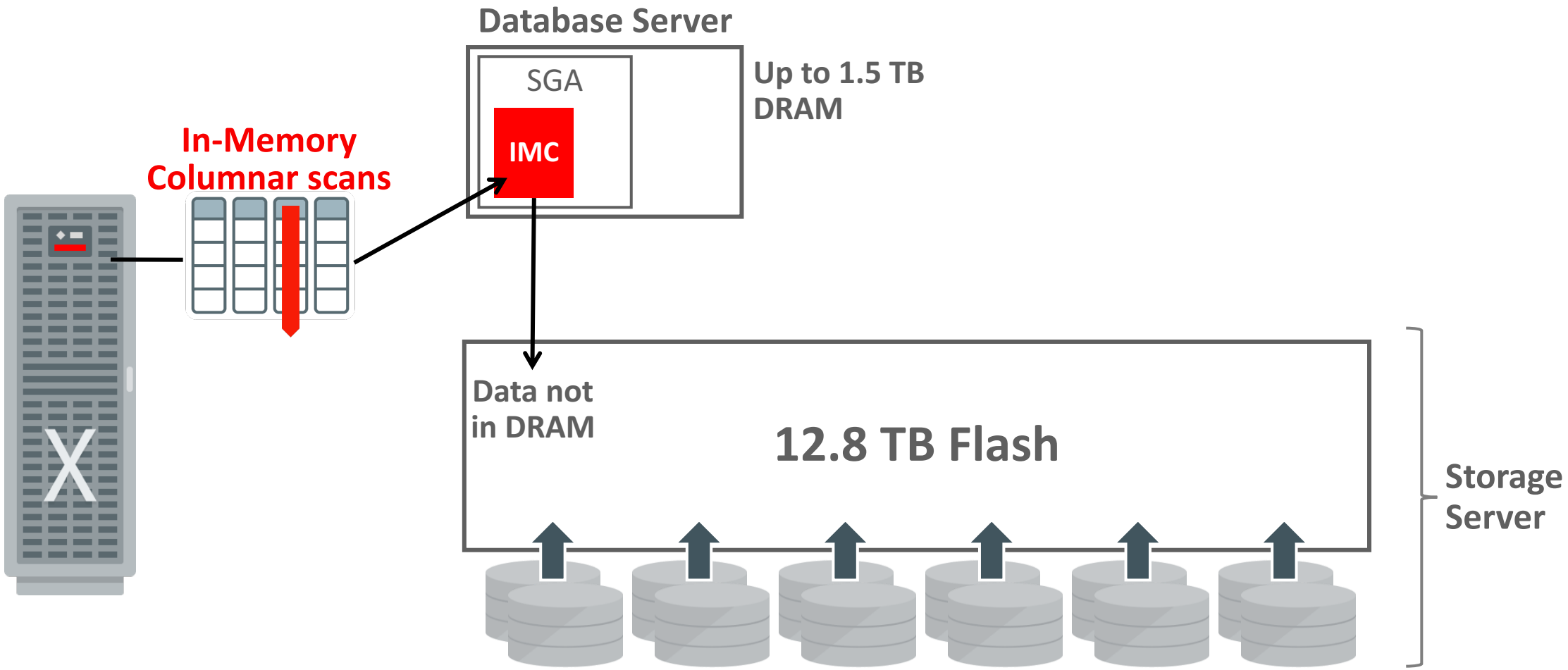
NEW IN DB 12.2

- Redo Log Write response times are very critical to large scale OLTP systems
- New metrics to measure
  - Overall IO latency
  - Networking and other overhead
  - IOs serviced by Flash Log
  - Overall latency per storage server
- Improves visibility of redo log write performance



In-Memory Columnar Formats in DRAM (pre 12.2.1.1.0)

Super-Fast Scans from Memory, but All Queries Complete



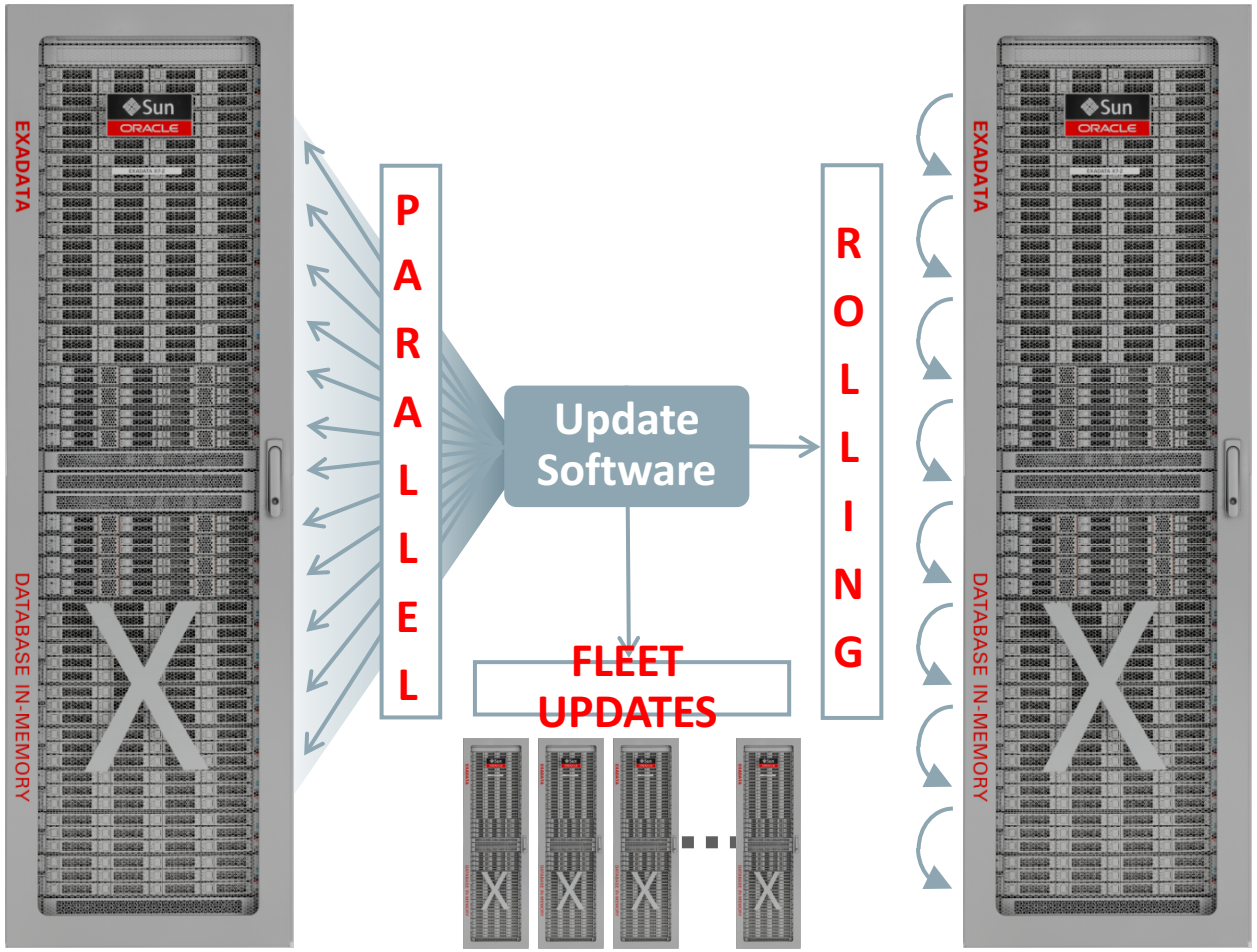
OLTP: End-to-End IO Latency Capping

NEW IN DB 12.2

- Exadata Storage Server software detects and automatically eliminates IO latency outliers on disk and flash media
- On very rare occasions network outliers can deteriorate latency between database and storage servers
- Database 12.2 automatically redirects slow read I/O operations to another Exadata storage server
- Ensures end-to-end low latency for OLTP read I/Os

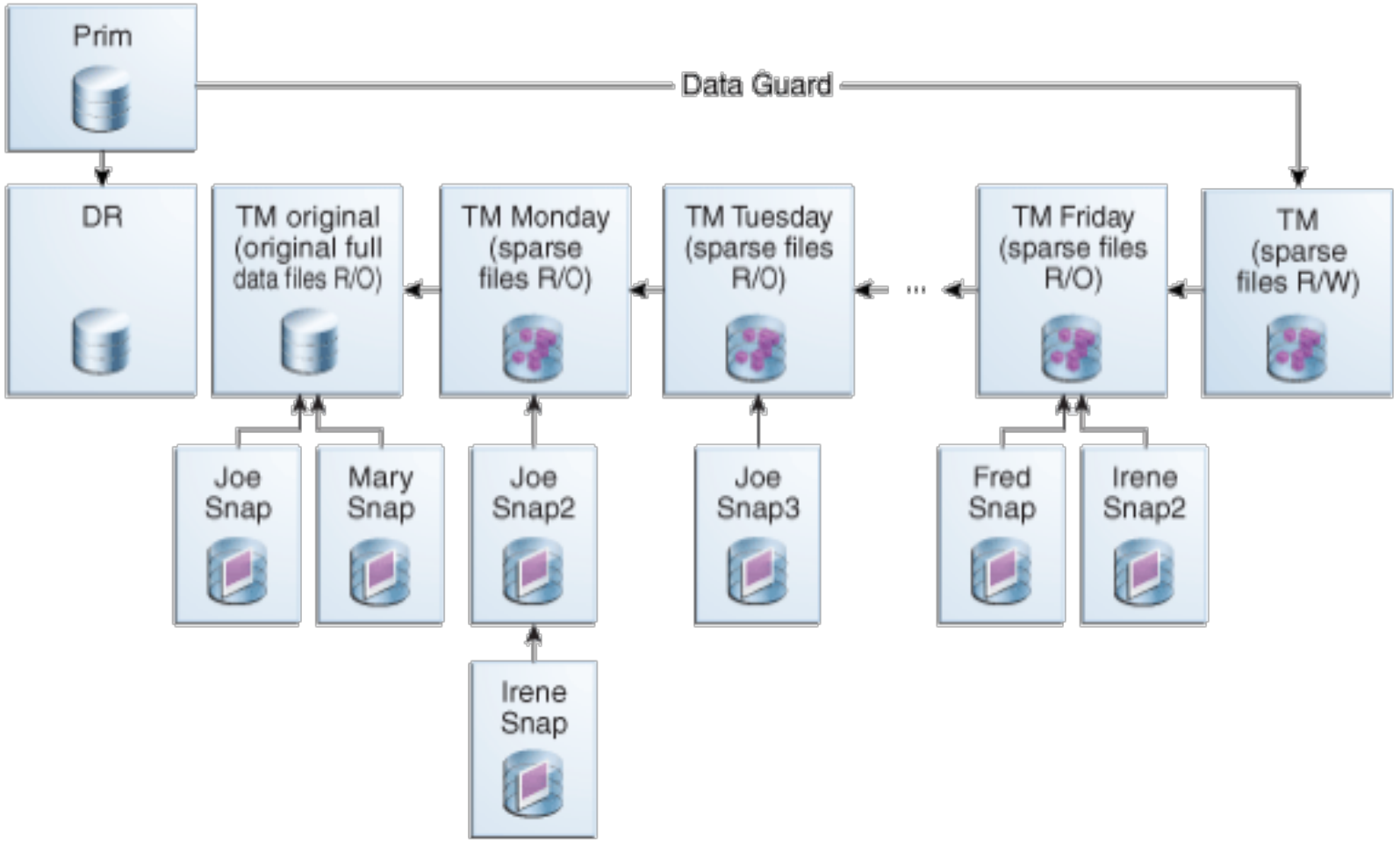
Manageability: Automatic Software Updates at Cloud Scale

- **Single tool updates all Exadata infrastructure software**
  - **600+** software/firmware components per full rack
- **Runs automatically on schedule**
  - Online (rolling)
  - Offline (parallel = fast)
- **Update multiple systems at the same time**
  - Intelligent software manages multiple upgrades as one



Hierarchical Snapshots for Daily Test Masters

Ideal Complement to Agile Development



- Repeat process to create new Exadata Snapshots while keeping prior Exadata Snapshots
- All Sparse Test Masters and Snapshots are sparse sized
- Can have a maximum of 10 Sparse Test Masters