

# Exadata: Delivering Memory Performance with Shared Flash

Setting New Standards for Database Performance



September 18–22, 2016  
San Francisco

**Kothanda Umamageswaran**  
Vice President, Exadata Development

**Gurmeet Goindi**  
Technical Product Strategist, Exadata

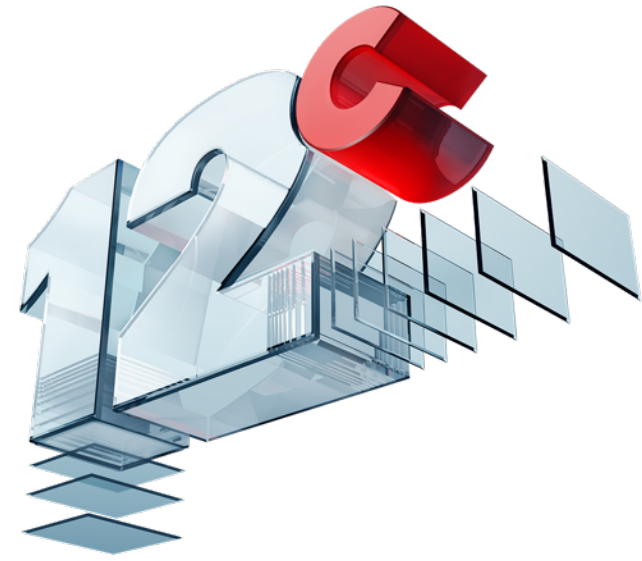


# Safe Harbor Statement

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.

# Announcing Oracle Database 12c Release 2 on Oracle Cloud

- Available now
  - Exadata Express Cloud Service
- Coming soon
  - Database Cloud Services
  - Exadata Cloud Machine

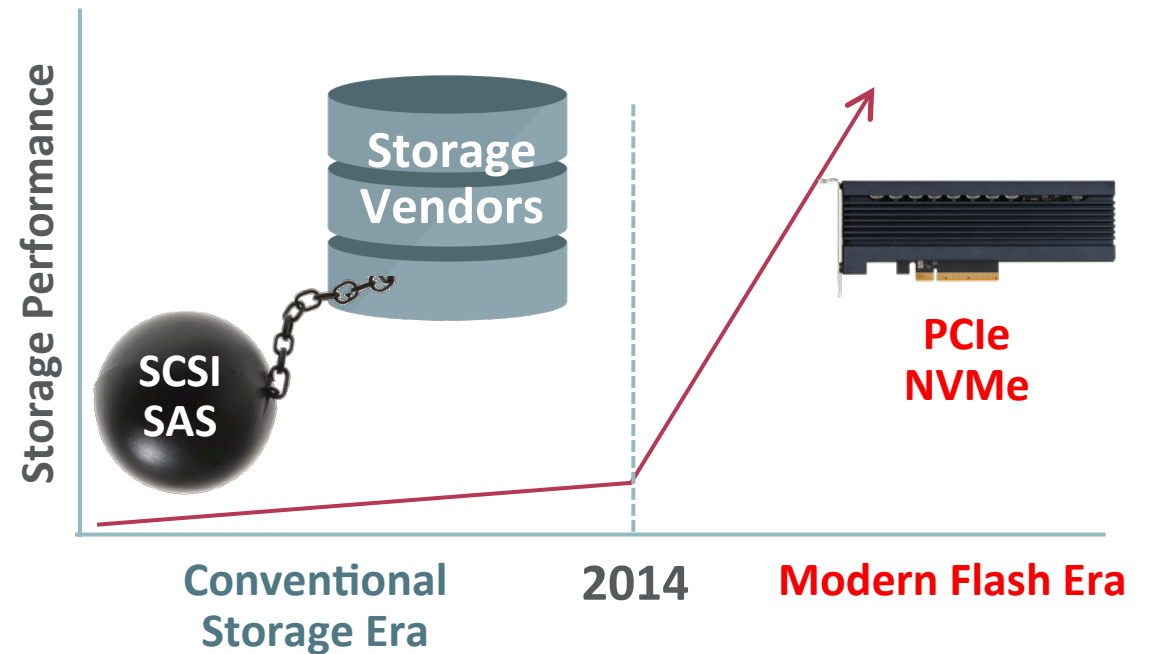


Oracle is presenting features for Oracle Database 12c Release 2 on Oracle Cloud. We will announce availability of the On-Prem release sometime after Open World.

# Did You Miss the Storage Revolution?

## Good Chance Your Storage Vendor Did Too

- Incumbent storage vendors have decades old investment in legacy protocols keeping them from adopting new technologies
- PCIe Flash with NVMe interface is a new interface that realizes full flash potential
- PCIe/NVMe storage architectures are *orders of magnitude faster* than what you probably use today
- Available now with Oracle Exadata storage



# Solid State Media is Very Different Than Spinning Disk

- Compared to Spinning Disk, Flash
  - Is many orders of magnitude faster
  - Has many orders of magnitude higher bandwidth
  - Has extremely low latency
  - Has wearing issues as it ages, but technology is catching up
  - Is expensive, but the price gap is shrinking
- Every storage vendor has some flash based solution for your Database

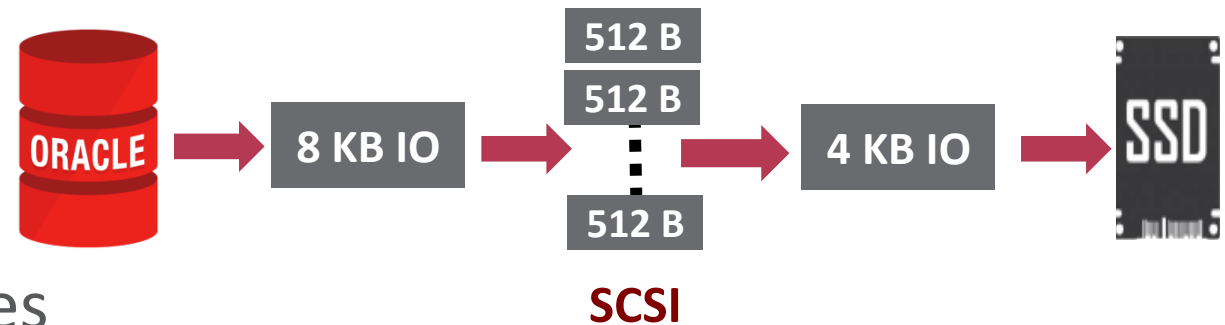
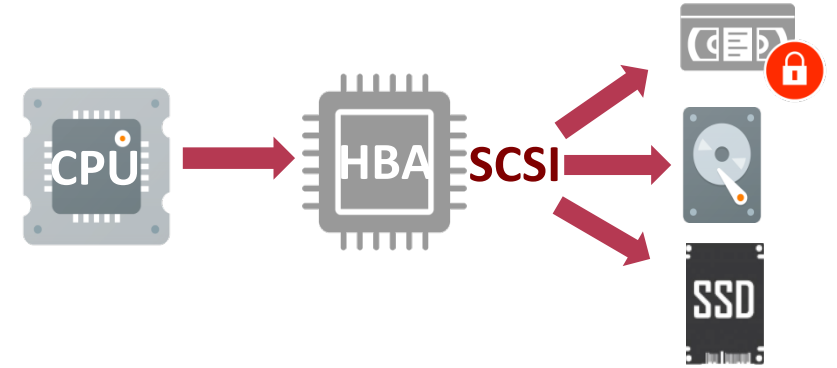


***Q: Will my database realize the full benefit of flash technology ?***

***A: It will depend on how fast you can move the data from the flash to the database***

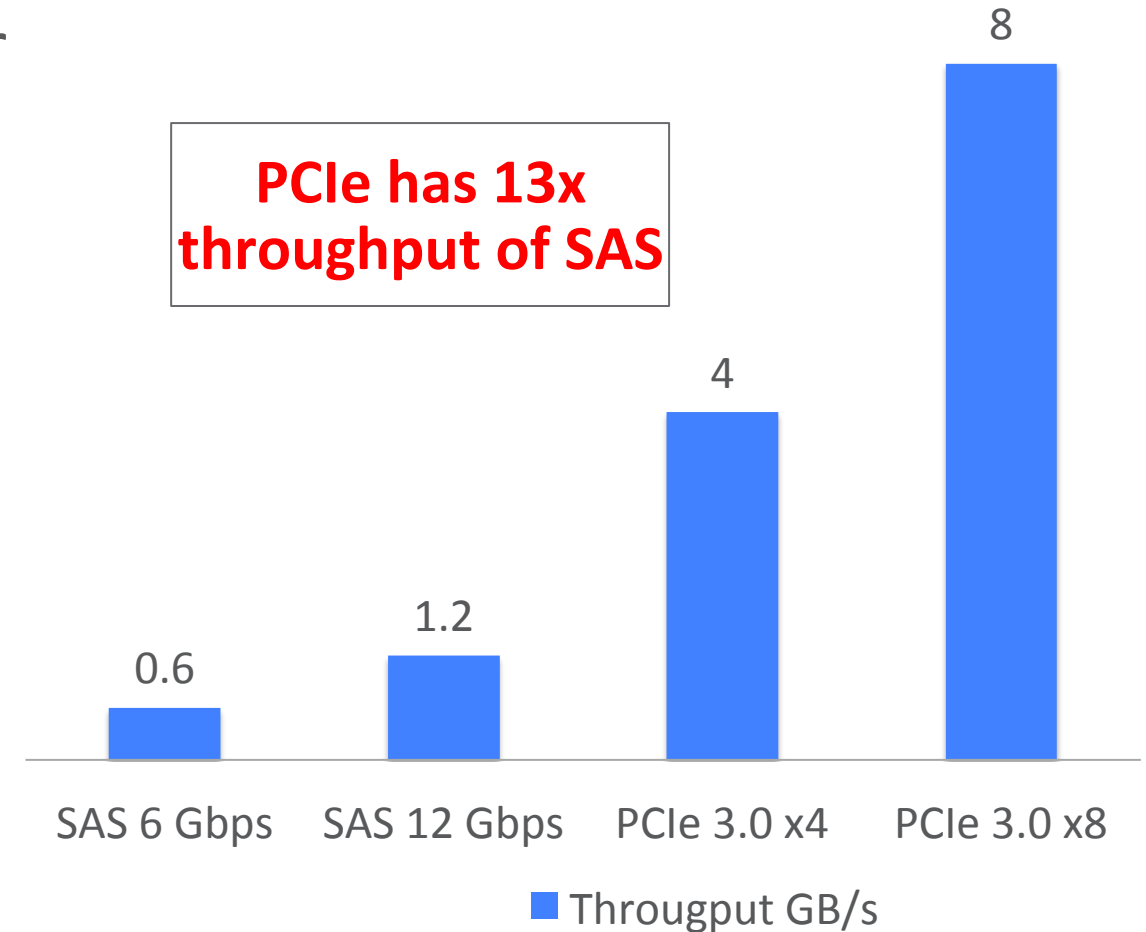
# SCSI Access Model

- SCSI was designed for tapes and HDDs
- HDDs are **sequential** whereas Flash devices are **massively parallel**
- Traditional IO stack is optimized for spinning media
  - 512 Byte block size transfers
  - Flash and databases do 4KB/8KB IOs
- Using legacy interfaces like **SCSI** **fundamentally bottlenecks** flash drives



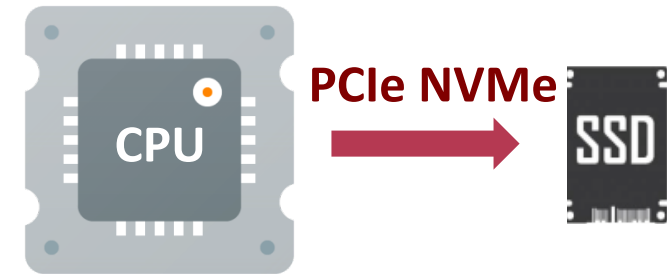
# PCI Express Vs SAS Connectivity

- PCI Express is orders of magnitude faster than SAS, and is getting faster
- PCI Express has the same characteristics as Flash
  - High Throughput
  - Low Latency
- Using legacy interconnects like **SAS** **fundamentally bottlenecks** flash drives



# PCI Express Flash with NVMe Interface

- **Non Volatile Memory Express** is a brand new ground up interface designed for flash
- NVMe is inherently **parallel**
- NVMe provides native atomic **IO size affinity** for databases
- NVMe IO stack massively **reduces** CPU utilization and latency



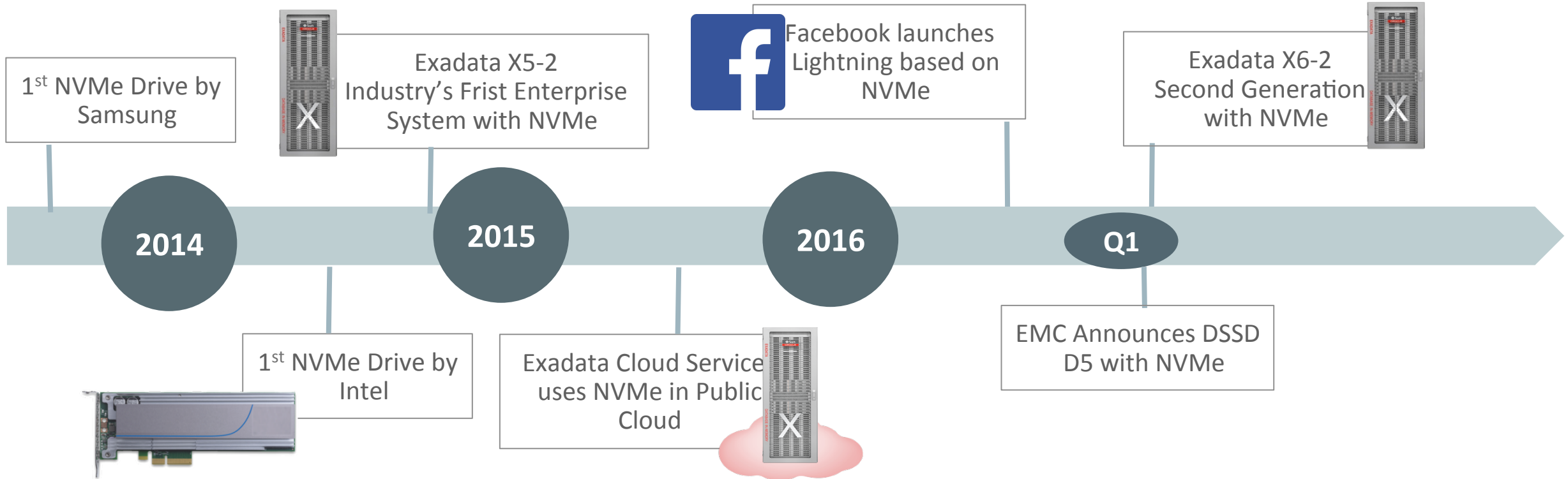
**NVMe is 2.2x  
Faster than SCSI**

***PCI Express Flash with NVMe Interface is the right choice for your Database***

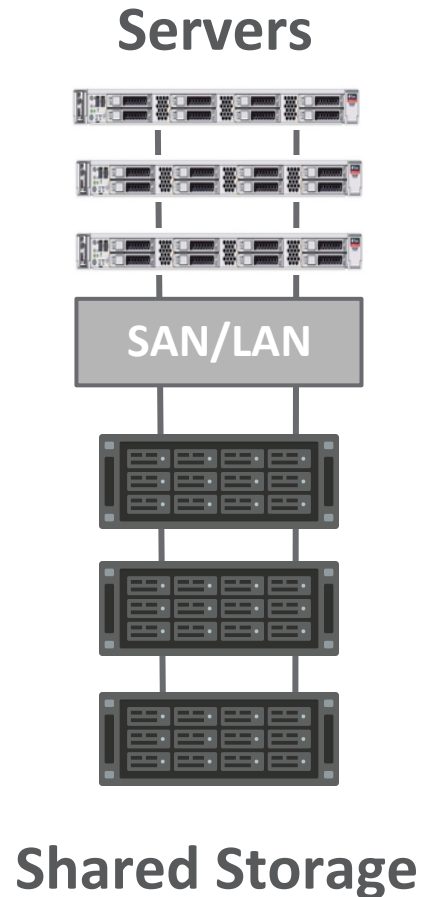


# Exadata is Leading NVMe Adoption

*Thousands of Exadata systems shipped with NVMe Flash since 2014*



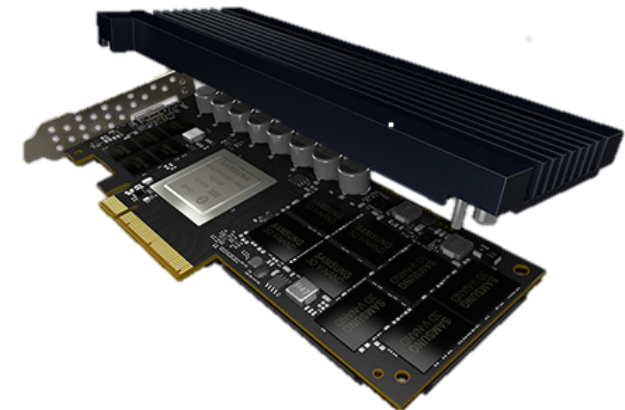
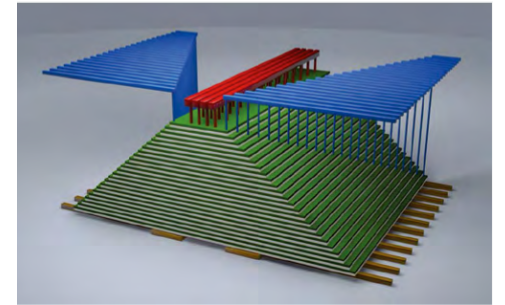
# Shared Storage Has Many Advantages over Local Storage



- Much better **space utilization**
- Much better **security, management, reliability**
- Enables DB **consolidation, DB high availability, RAC scale-out**
- **Shares storage performance**
  - Aggregate performance of shared storage can be dynamically used by any server that needs it

# New Exadata X6 Super-Capacity and Performance Flash

- 3D V-NAND **3.2TB/card** (2X previous card capacity)
  - **48 layer NAND**
  - No tradeoffs - faster writes, lower power, higher endurance
- Latest, most modern interface – NVMe (introduced in X5)
- Fastest flash card on market by wide margin
  - Only flash card on market with PCI 8-lane scale bandwidth **~ 5.4GB/sec**
  - Highest IOs per second
  - Lowest outliers – 99.995% write IOs complete within 250us



# NVMe PCI-e Flash Disrupts the Storage Array Model

New improvements are causing **100X bottlenecks** across shared storage stack



Latest PCIe Flash  
**5.4 GB/sec**

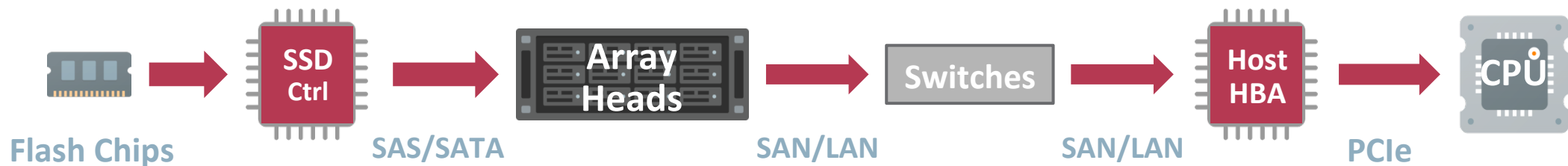


SAN Link = 40Gb  
**5 GB/sec**  
**Less than 1 Flash card**

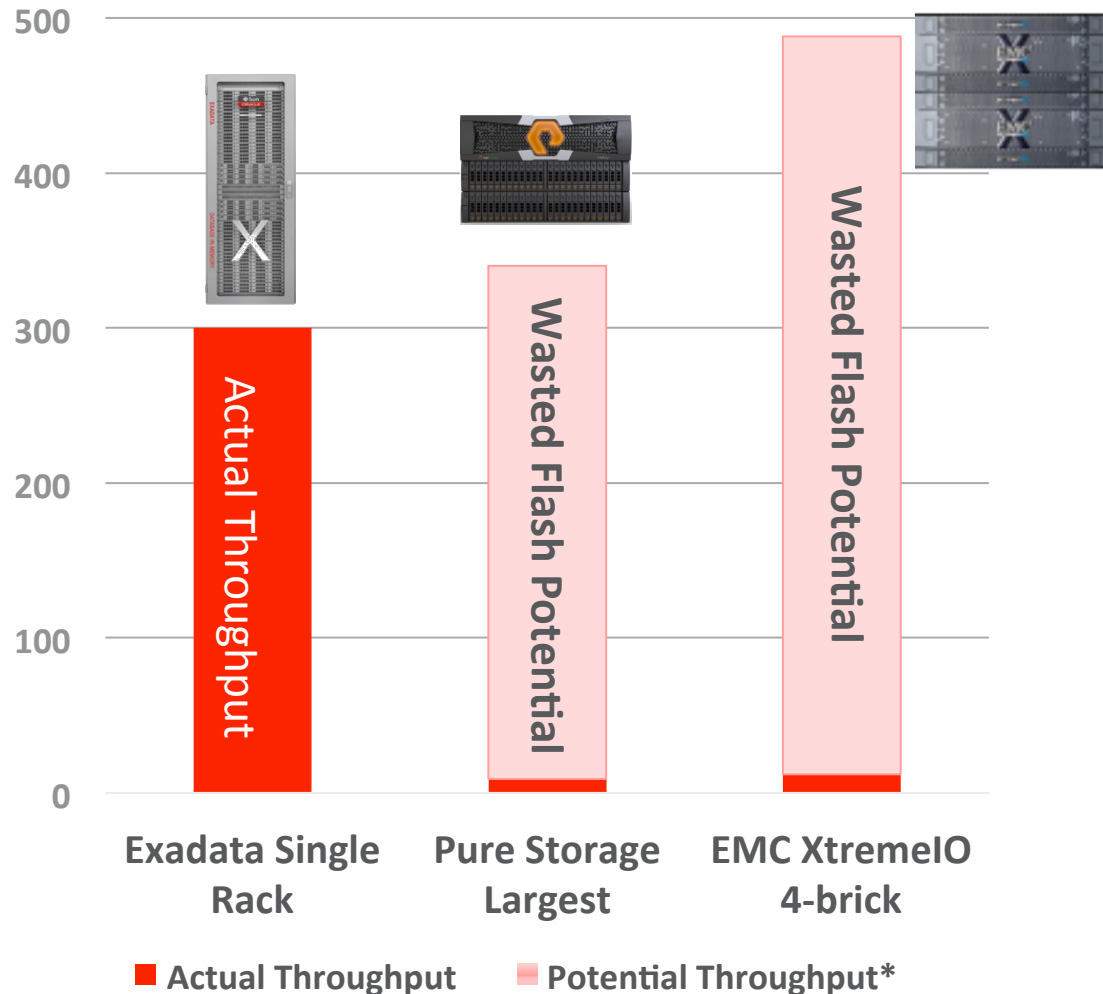


Leading All Flash Array  
**24 GB/sec**  
**Less than 5 Flash card**

All-Flash Storage Array IO Path: many steps, each adds **latency** and creates **bottlenecks**

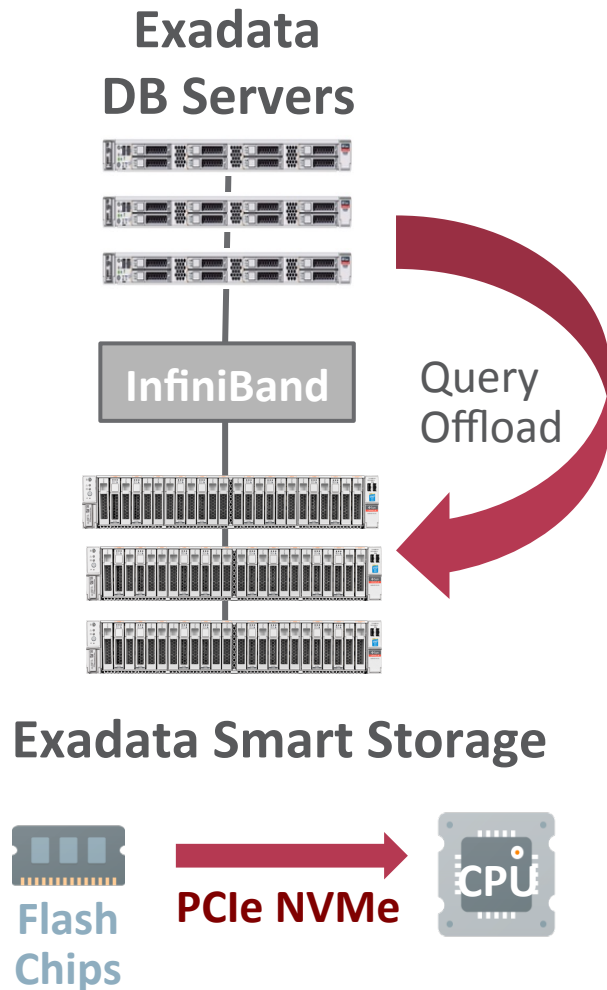


# Only Exadata Achieves Full Performance of Shared Flash



- **Leading All-Flash Storage Arrays achieve under 3% of potential flash throughput**
  - Pure Storage – 132 MB/sec per flash drive
  - EMC XtremIO – 120 MB/sec per flash drive
  - Spinning disk level throughput!
  - AND can't scale-out for higher performance
  - AND can't share even this slow performance due to bottleneck at server inputs
- **Exadata X6 achieves full flash throughput**
  - 5400 MB/sec per drive
- **Exadata also achieves much faster OLTP IOs**
  - 5.6 Million IOPs, 250us latency even at 2.4M IOs

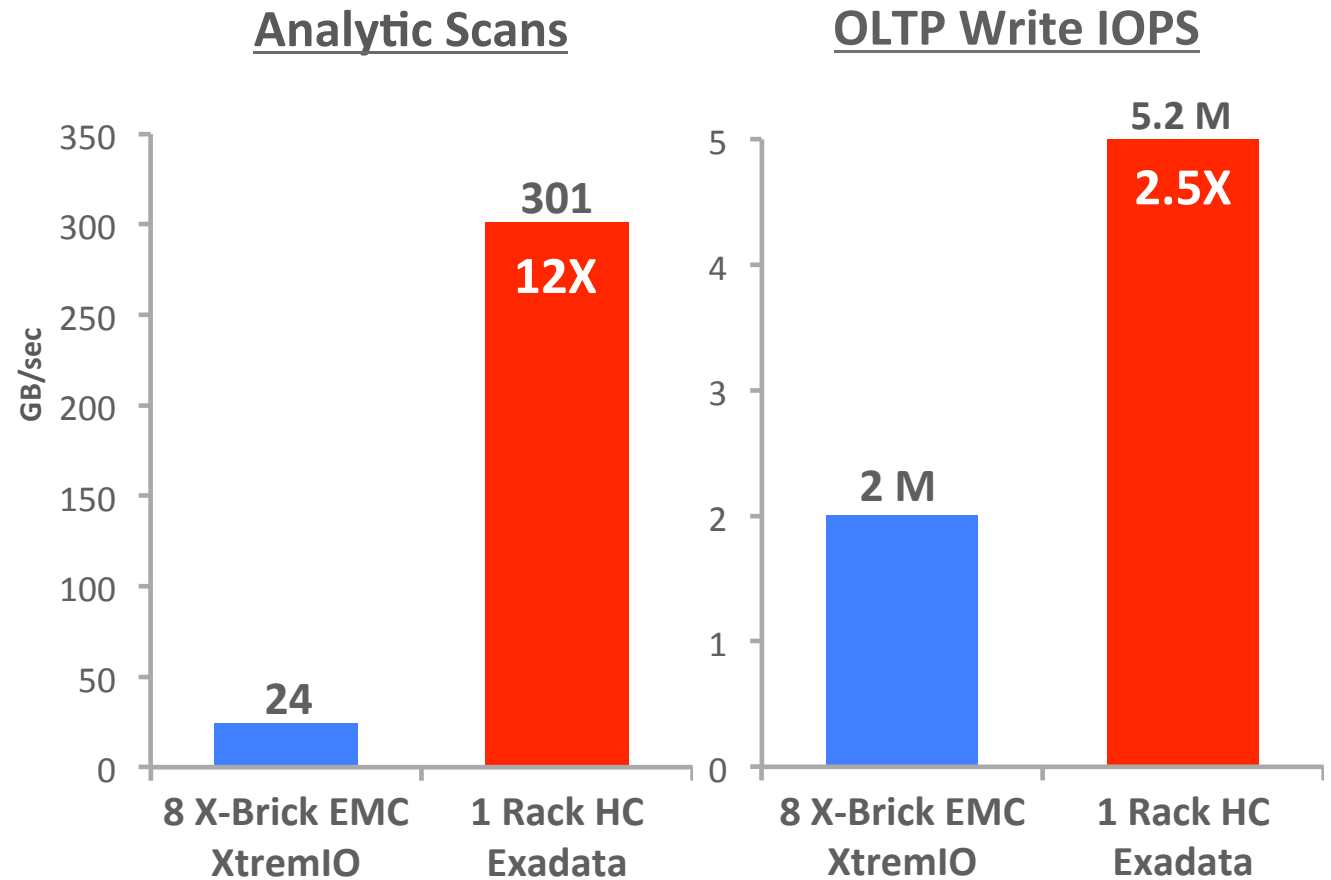
# Exadata Achieves Memory Performance with Shared Flash



- Exadata X6 delivers **300GB/sec flash bandwidth** to any server
  - Approaches 800GB/sec aggregate **DRAM** bandwidth of DB servers
- **Must move compute to data to achieve full flash potential**
  - Requires owning full stack, can't be solved in storage alone
- **Fundamentally, Storage Arrays can share flash capacity but not flash performance**
  - Even with next gen scale-out, PCIe networks, or NVMe over fabric
  - E.g. new EMC DSSD has **3-6 times slower** throughput than Exadata X6
- **Shared storage with memory level bandwidth** is a paradigm change in the industry
  - Get near DRAM throughput, with the capacity of shared flash

# Exadata X6 I/O is Much Faster than All-Flash EMC

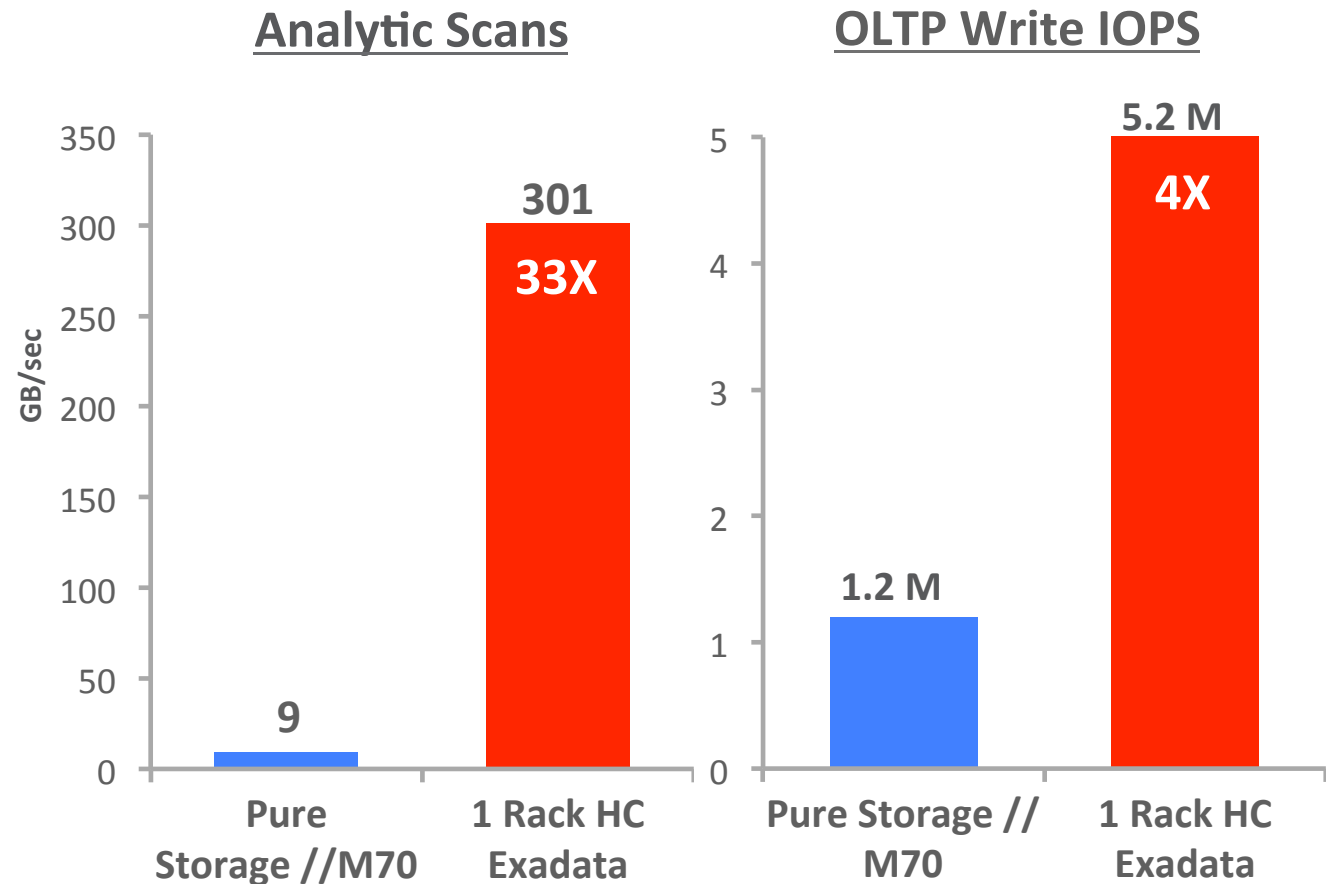
- One **High Capacity** Exadata beats the fastest EMC XtremIO **all-flash** array in every performance metric
  - **12X more throughput**
  - **2.5X more IOPS**
  - **2X faster latency**



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

# Exadata X6 I/O is Much Faster than All-Flash Pure Storage

- One **High Capacity** Exadata beats the fastest Pure Storage **all-flash** array in every performance metric
  - **33X more throughput**
  - **4X more IOPS**
  - **4X faster latency**



Pure Storage does not scale higher - Exadata scales by adding racks



# Getting Memory performance with Shared Flash using Smart Software

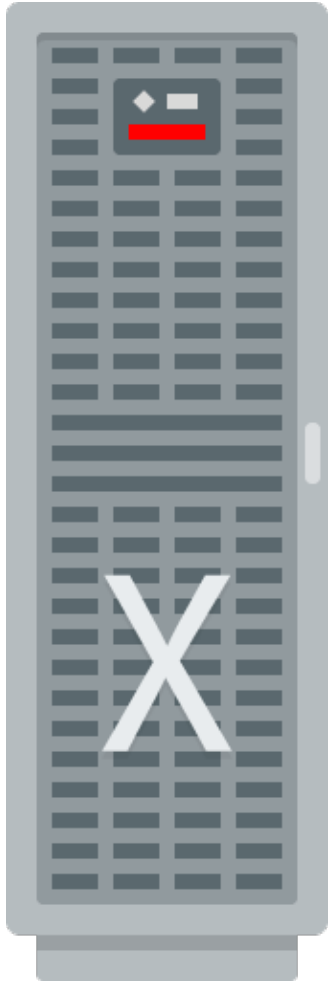
# Oracle's Infrastructure Innovations in Flash



- Oracle Exadata V2: First to bring flash storage to the database market
- Oracle Exadata X3: Doubled flash capacity
- Oracle Exadata X4: 100GB/s throughput scans in a single rack
- Oracle Exadata X5: **Lowest latency NVMe** and increases scans to **263GB/s**
- Oracle Exadata X5: **Hot-pluggable NVMe server** for the database
- Oracle Linux: First Linux vendor with production NVMe drivers
- Oracle Exadata X6: Highest throughput over **350GB/s** and lowest latency

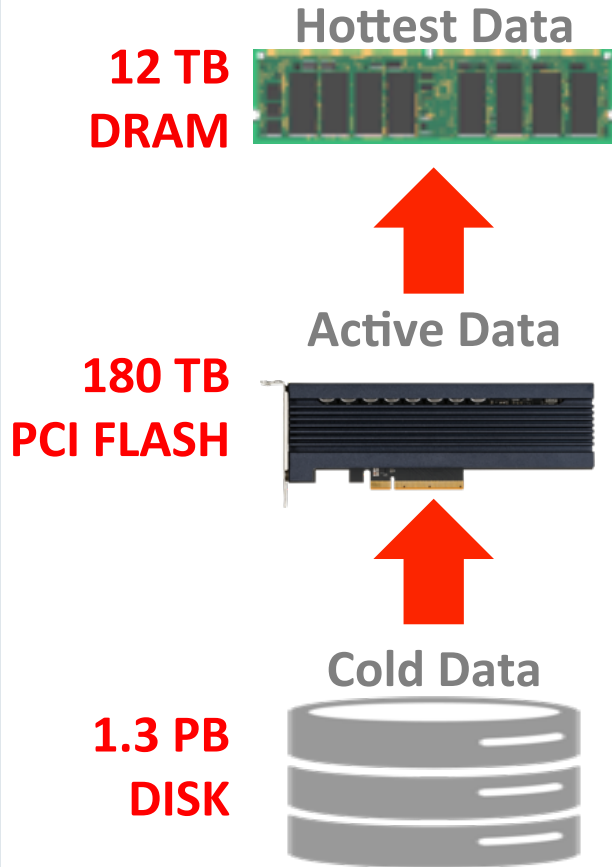


# Oracle's Software Innovations in Flash



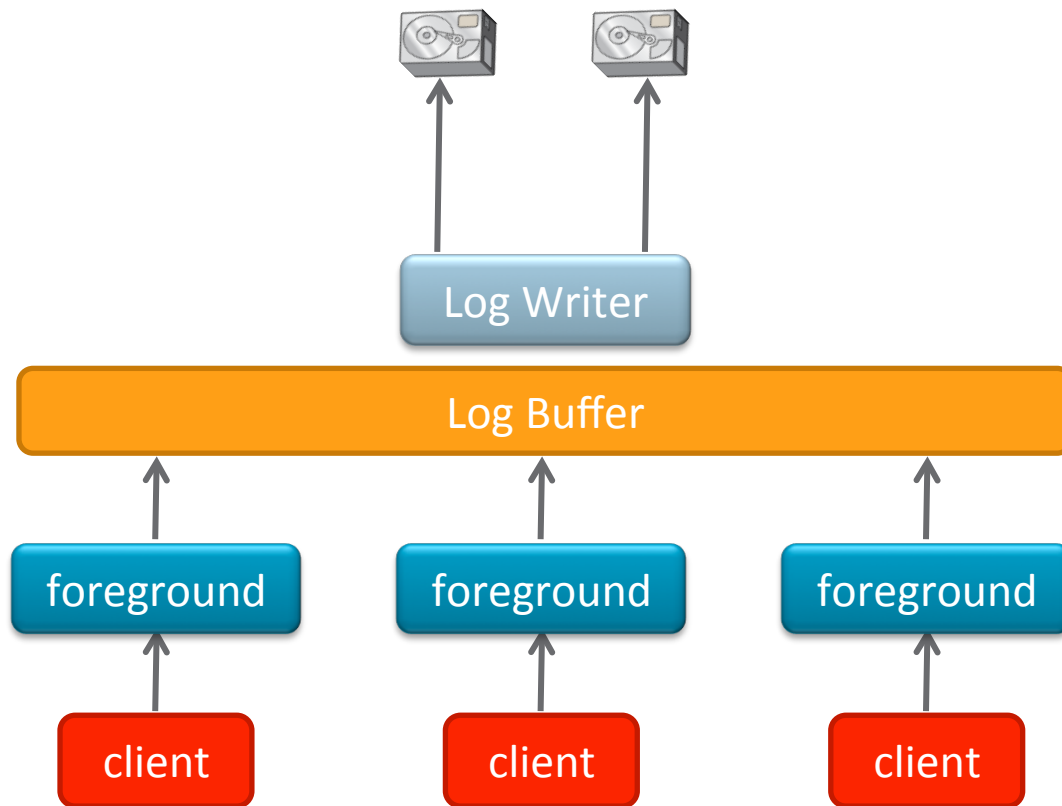
- Exadata Smart Flash Cache
- Exadata Smart Flash Log
- Exadata Smart Flash Cache Scan Awareness
- Exadata Smart File Initialization
- Exadata Smart Columnar Flash Cache
- Exadata Smart Flash Cache Space Resource Management
- **Upcoming:** Exadata Smart In Memory Formats in Flash
- **Upcoming:** Smart write burst and temp IO in Flash Cache

# Exadata Smart Flash Cache



- Understands different types of I/Os from database
  - Skips caching I/Os to backups, data pump I/O, archive logs, tablespace formatting
  - Caches Control File Reads and Writes, file headers, data and index blocks
  - **Enables more space for relevant user data**
- Immediately adapts to changing workloads
- Write-back flash cache
  - Caches writes from the database not just reads
- **Doesn't need to mirror in flash for read intensive workloads**
  - Flash arrays store both mirror copies always in flash increasing your cost
- **Smart Scans can run at the throughput of flash drives**
  - Flash arrays need lots of servers with lots of processes and still cannot match Smart Scan throughput of single query
- Provides performance of flash at cost of disk

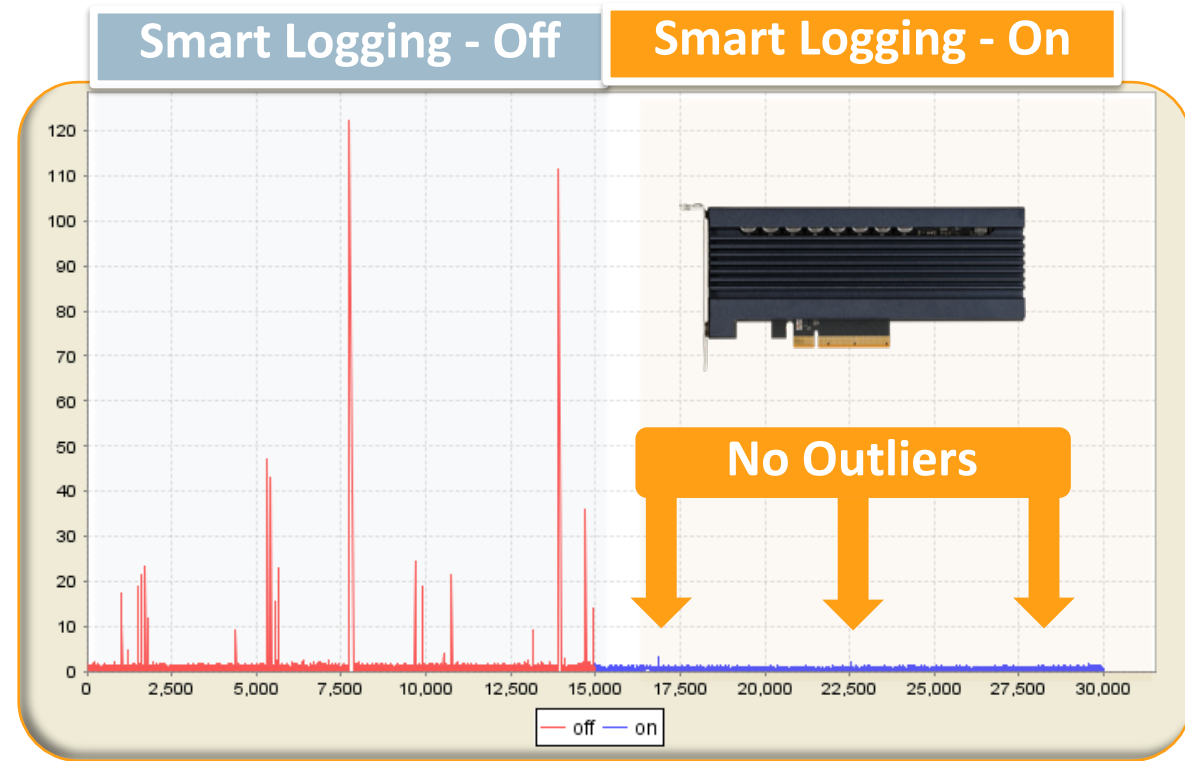
# Exadata Smart Flash Log



- Outliers in log IO slow down lots of clients
- Outliers from any one copy of mirror slow down all the foregrounds
  - Database wait time goes up by  $\#foregrounds * Stall\ time$
  - Backlog doesn't clear immediately like an accident on the freeway and increases "log file sync" waits
- Performance critical algorithms like space management and index splits are sensitive to log write latency
- Legacy storage IO cannot differentiate redo log IO from others
- UPS protected cache in traditional storage seems to work initially until the cache is overwhelmed by other writes
  - Measure log file latency with full backup or a data load running

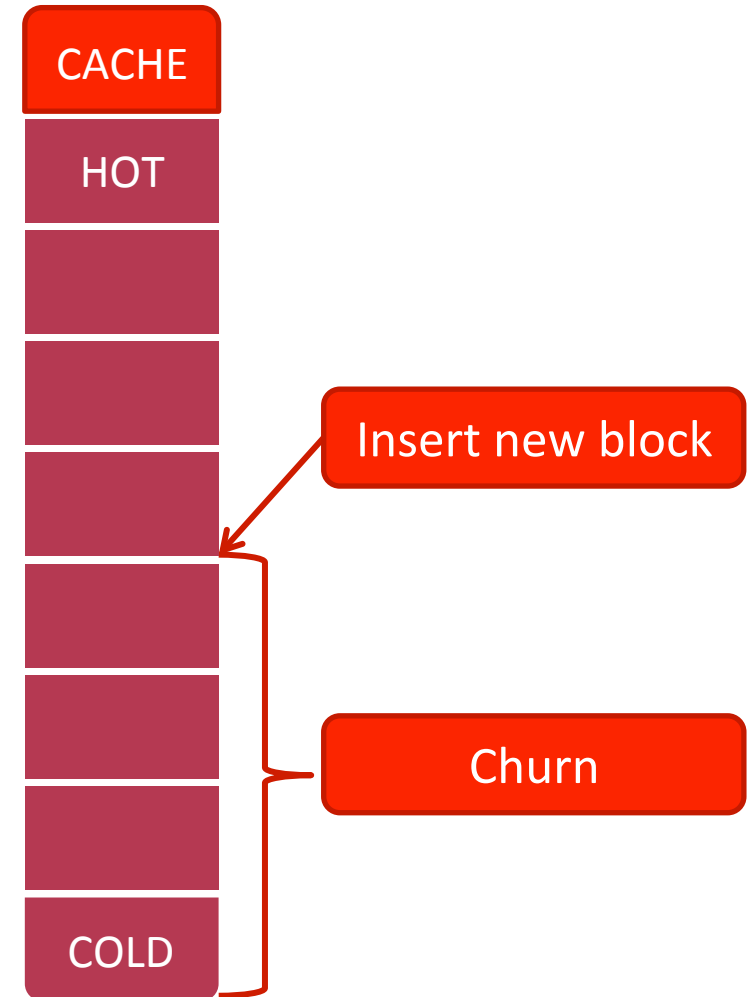
# Exadata Smart Flash Log

- Smart Flash Log uses flash as a parallel write cache to disk controller cache
- Whichever write completes first wins (disk or flash)
- Reduces response time and outliers
  - “log file parallel write” histogram improves
  - Greatly improves “log file sync”
- Uses almost no flash capacity (< 0.1%)
- Network resource management provides priority for redo log I/Os across the network
- **OLTP workloads transparently accelerated and provide predictable response times**



# Exadata Smart Flash Cache Scan Awareness

- On a traditional cache, if you scan dataset larger than cache size
  - Blocks 0,1,2,3 brought into cache, cache is full
  - Scanning Blocks 20,21,22,23 replaces 0,1,2,3 in cache
- Repeat the same scan
  - Block 0,1, 2, 3 will replace blocks 20,21,22,23
  - Block 20,21,22,23 will again replace block 0,1,2,3
- Traditional caches churn with no actual benefit
- Some implementations call the insertion of new block in the middle scan resistant



# Exadata Smart Flash Cache Scan Awareness

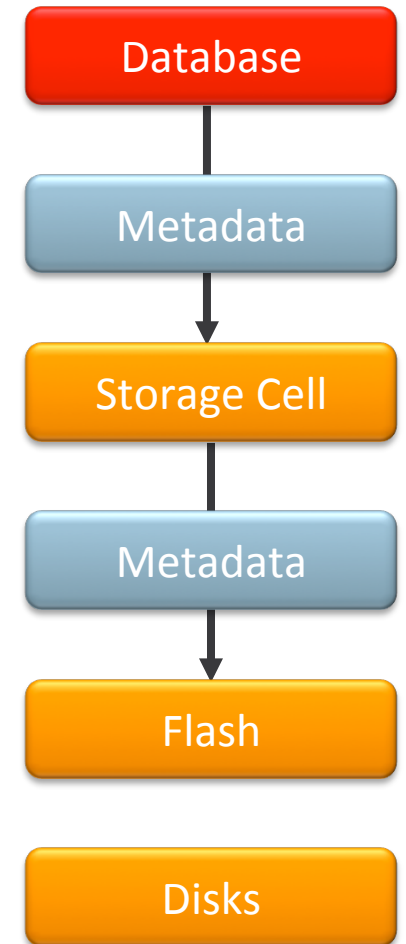
- Exadata Smart Flash Cache is scan resistant
  - Ability to bring subset of the data into cache and not churn
  - OLTP and DW scan blocks can co-exist
- Nested scans bring in repeated accesses
  - Repeat, For each item in large table, scan small table
  - Smart enough to pull the small table into flash since it is accessed repeatedly even though the size of large table alone is larger than flash cache
- No need to set “KEEP” attribute in data warehouses
- Scans automatically use flash for extreme performance
- Scans won't blow out the cache providing predictable OLTP performance





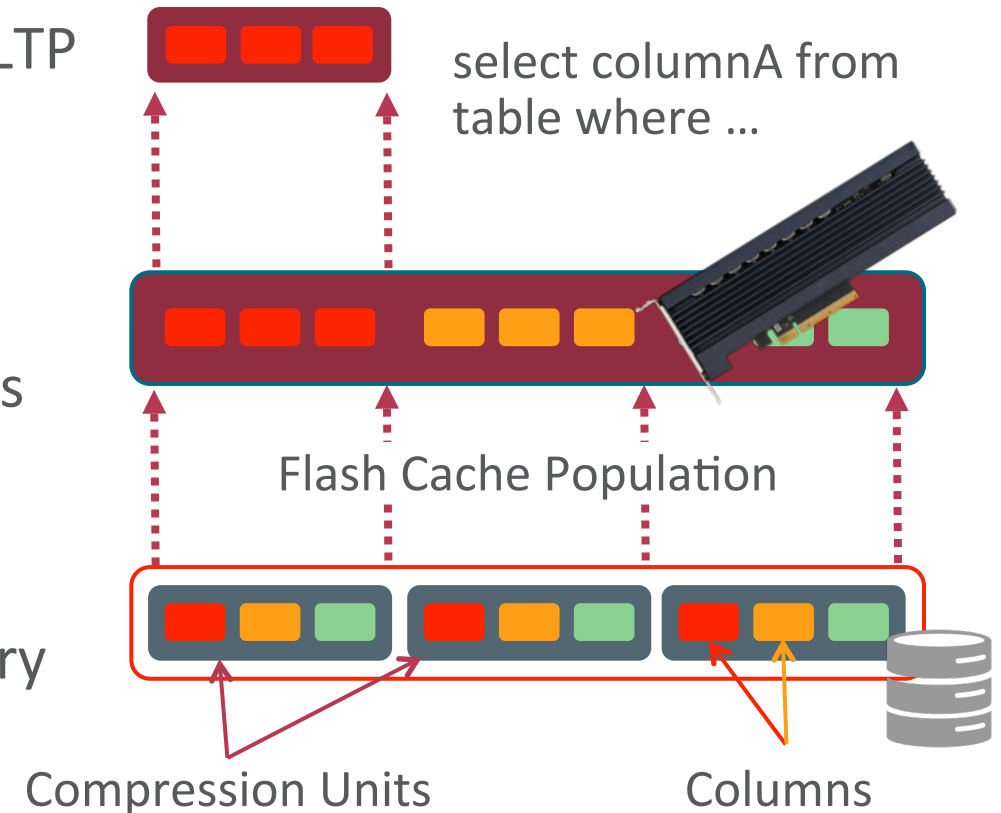
# Exadata Smart File Initialization

- Combine the benefits of Smart Initialization and Writeback Flash Cache
  - Write **file creation meta-data** to writeback flash cache
  - Tiny amount of flash space used to cache large portions of initialized data on disk
  - Initialization I/Os to disk deferred or not performed if data loaded
- Create tablespace, file extensions, autoextend show benefit
- Redo log initialization included in Exadata 12.1.1.1.0
- **File creation sped up by over 10x**



# Exadata Smart Columnar Flash Cache

- Hybrid Columnar Compression balances need for OLTP and Analytics
- As CPUs get faster want even faster scans
- Smart Flash Cache automatically transforms blocks from hybrid columnar to pure columnar for analytics during flashcache population
- Dual format representation for single row lookups
- Only selected columns read from flash during a query
- **Up to 5x query speedup**



# Smart Flash Cache Space Resource Management



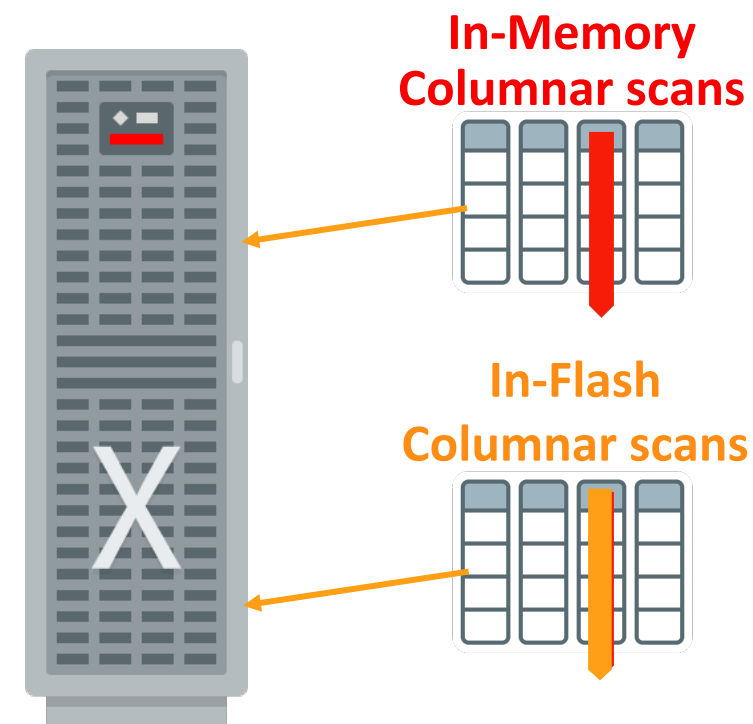
- Flash Cache is a shared resource
- Database as a Service creates need for efficient resource sharing
- Specify minimum (flashCacheMin) and maximum (flashCacheLimit) sizes, or fixed allocations (flashCacheSize), a database can use in the flash cache

```
ALTER IORMPLAN -  
  
  dbplan=( (name=sales, flashCacheSize=100G), -  
           (name=finance, flashCacheLimit=100G, flashCacheMin=20G), -  
           (name=schain, flashCacheSize=200G) )
```

- Container database resource specified at the storage
- Pluggable database container resource limits expressed as percentages in the container database
- Database and Pluggable database I/O resource management is unique to Exadata
- Predictable performance for database queries – no more noisy neighbor

# 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

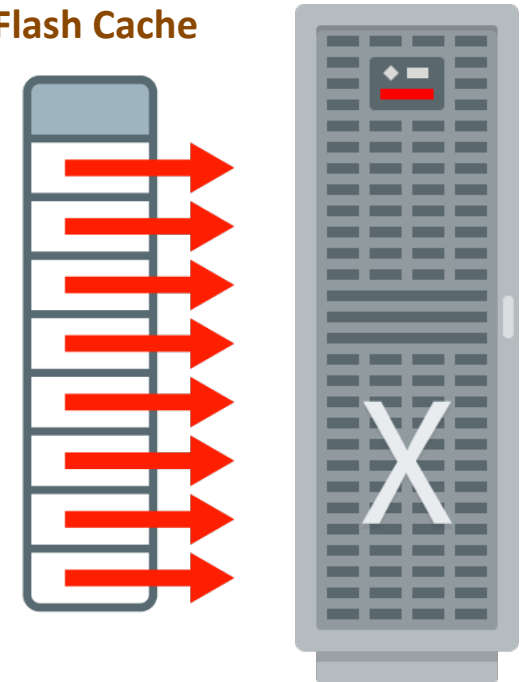


*Upcoming release of Exadata Software*

# Upcoming: Smart write bursts and temp IO in flash cache

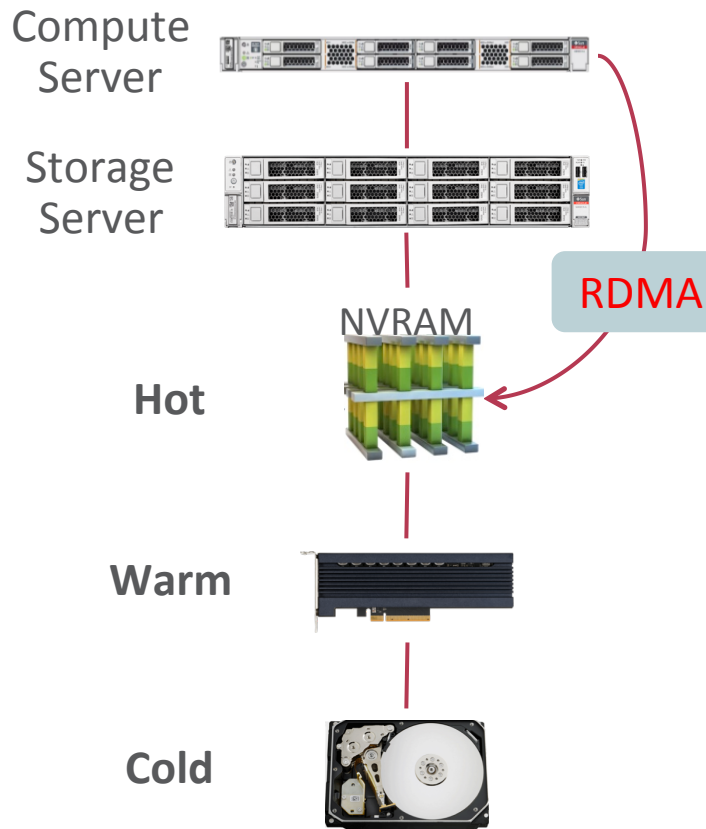
- Write throughput of four flash cards has become greater than the write throughput of 12-disks
- When database write throughput exceeds the throughput of disks, smart flash cache intelligently caches writes
  - Schema changes during application upgrades rewrite entire tables in some packaged applications
  - Large database consolidations can have write bursts at the same time
- 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 to prioritize OLTP data and does not remove hot OLTP lines from the cache
- Smart flash wear management for large writes
- Much faster scans and disk writes

## Write Bursts and Temp IO in Flash Cache



*Upcoming release of Exadata Software*

# Preview: Non-volatile Memory Tier in Exadata Storage



- Exadata Storage Servers will add a non-volatile memory (NVRAM) cache in front of Flash memory
  - Similar to current Flash cache in front of disk
  - RDMA direct access to NVRAM gives **20x lower latency** than Flash
- NVRAM used as a cache effectively increases its capacity by 10x
- Expensive NVRAM shared across servers for lower cost
- NVRAM mirrored across storage servers for fault-tolerance

# Exadata Smart Flash Benefits

- Smart Flash Cache is database aware
- Smart Flash Logging avoids redo log outliers
- Smart Flash Cache Scan provides subset scanning and is table scan resistant
- Smart File Initialization creates a file by writing meta-data to flash cache
- Smart Columnar Flash Cache extends columnar benefit to storage
- Smart Flash Cache Space Resource Management provides granular control
- **Upcoming:** Smart Flash cache with in memory formats enables massive capacity for vector processing
- **Upcoming:** Smart write burst and temp IO in Flash Cache

# Integrated Cloud

## Applications & Platform Services



ORACLE®