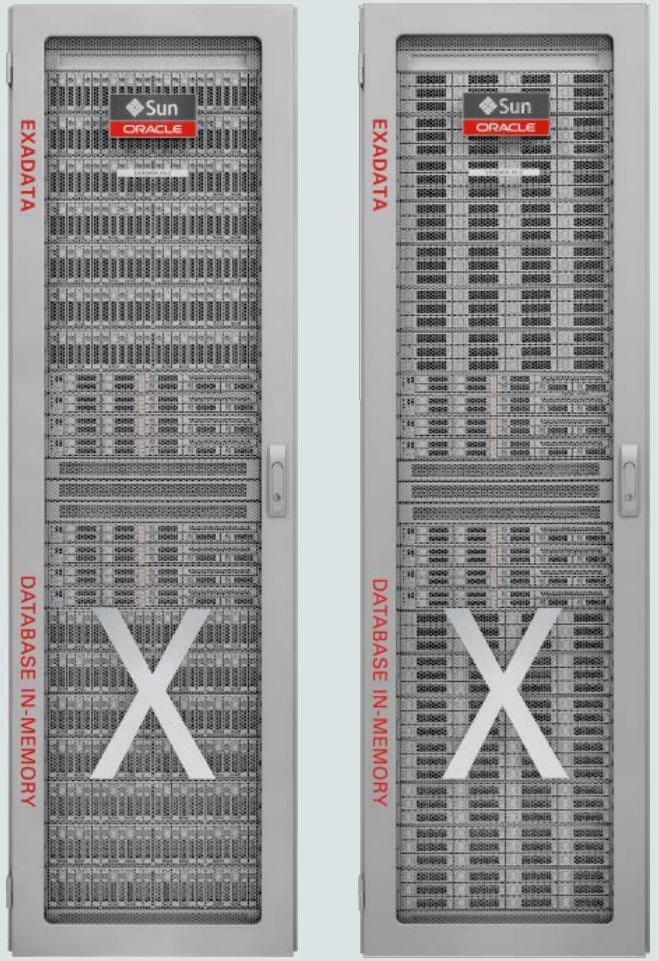
NVMe – A End User Testimonial

Ravi Chari

Technologist,

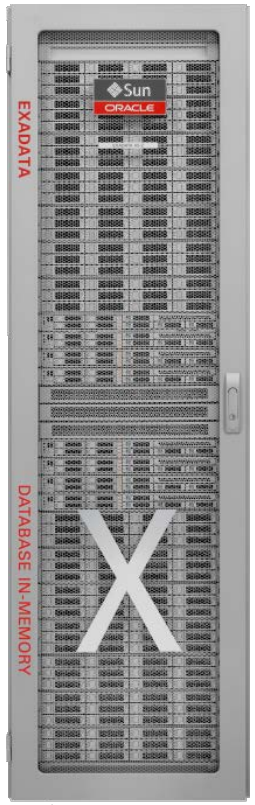
Non Volatile Memory Engineering, Oracle



ORACLE

End User Testimonial?

End user - the person who actually uses a particular product.
Testimonial- a public tribute to someone and to their achievements



Truth in Advertising!

- **Oracle is a Member of the Board of Directors of NVMExpress.**
- **Oracle is a very active contributor to NVMExpress Eco System (Especially Solaris and Oracle Enterprise Linux)**
- **Oracle introduced its first NVMExpress Devices in Exadata X5 Systems in December 2014.**

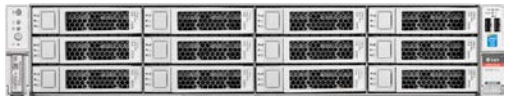
ORACLE

Exadata X5 Storage Servers

Extreme Flash Storage Server

All-Flash

High-Capacity Storage Server



Disk + Flash Cache

State-of-the-art **NVMe PCIe** flash
Consistently Low Response Times
Optimized InfiniBand I/O Protocols

Exadata Storage Server Software

Smart Scan (SQL Offload)
Smart Flash Cache
I/O Resource Management
Hybrid Columnar Compression

Performance	Extreme Flash	High-Capacity
Analytic Scans	263 GB/s	140 GB/s
OLTP Reads (8K)	4.14 M IOPS	4.14 M IOPS
OLTP Writes (8K)	4.14 M IOPS	2.69 M IOPS
Flash Latency	0.25 ms @ 2M IOPS	0.25 ms @ 1M IOPS

Capacity	Extreme Flash	High-Capacity
Cores (for SQL offload)	16	16
Disk (per server)	-	48 TB
Flash (per server)	12.8 TB	6.4 TB
Disk (full rack)*	-	672 TB
Flash (full rack)*	179.2 TB	89.6 TB

* Full Rack : 8 DB servers, 14 storage servers

ORACLE / SUN Evolution to NVM Express

Year: 2011

SUN FLASH ACCELERATOR F20 PCIe CARD

KEY FEATURES

- Over 100K IOPS Performance
- Over 1,000 MB/s Bandwidth
- 96GB user capacity
- Embedded Flash controllers for high performance and compatibility
- Highly reliable, high endurance Sun FlashFire technology
- Compact Low-profile PCIe form factor to fit most servers

Year: 2012

SUN FLASH ACCELERATOR F40 PCIe CARD

KEY FEATURES

- 400 GB capacity
- Up to 149K IOPS (8K) performance
- Over 2.0 GB/s throughput
- 95 microsecond write latency
- Embedded Flash controllers for greater performance , compatibility and low CPU burden
- Advanced write endurance
- Proactive monitoring features
- Low-profile PCIe form factor

ORACLE / SUN Evolution to NVM Express

Year: 2013

SUN FLASH ACCELERATOR F80 PCIe CARD

KEY FEATURES

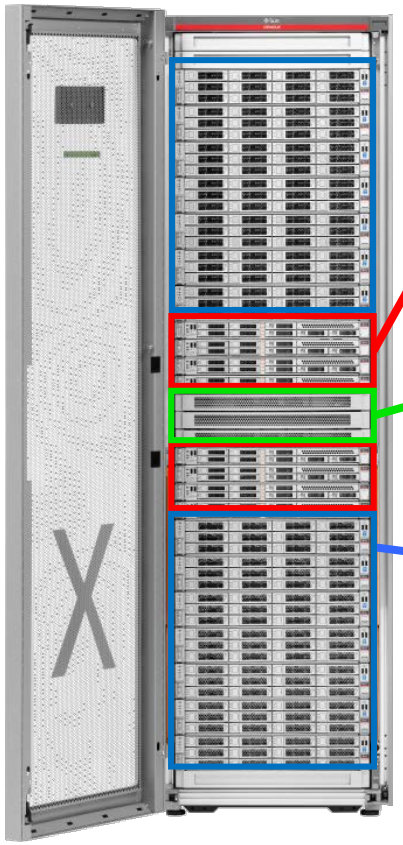
- 800 GB capacity
- 155K random IOPS (8K), 2.1 GB/sec throughput performance
- 84 microsecond write latency (8K transfer size)
- Advanced write endurance and proactive health monitoring
- Optimized with Oracle's systems and software.
- Compatible with Oracle's Database Smart Flash Cache and Advanced Compression

Year 2014

NVMExpress Based Cards & U.2 SSDs!

- Standardizes register set, feature set, and command set where there were proprietary PCIe solutions before
- Designed to scale for next generation NVM, agnostic to NVM type used
- Streamlined & simple command set (13 required commands)
- All parameters for 4KB command in single 64B command
- Supports deep queues (64K commands per queue, up to 64K queues)

Exadata X5-2 Product Components

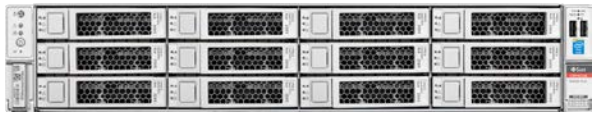


- **Scale-Out Database Servers**
  - **Two 18-core x86 Processors (36 cores)**
  - Oracle Linux 6
  - Oracle Database Enterprise Edition
  - Oracle VM (optional)
  - Oracle Database options (optional)
- **Fastest Internal Fabric**
  - 40 Gb/s InfiniBand
  - Ethernet External Connectivity
- **Scale-Out Intelligent Storage**
  - **High-Capacity Storage Server**
  - **Extreme Flash Storage Server**
  - **Exadata Storage Server Software**

X5-2 Database Server

36 cores per server
256 – 768 GB DRAM

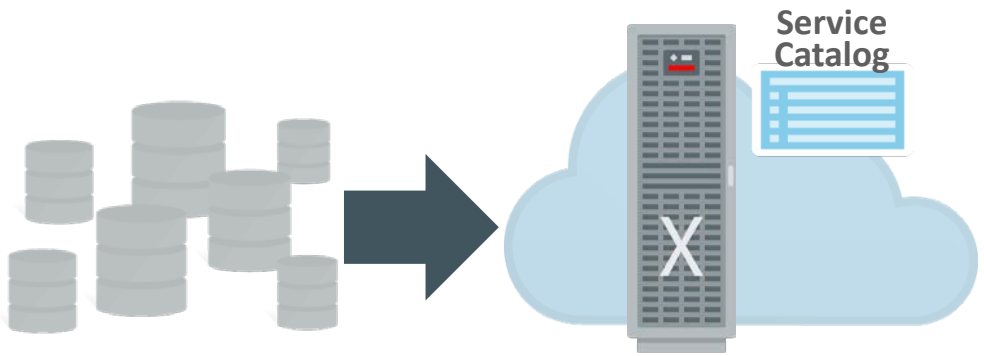
High-Capacity Storage Server



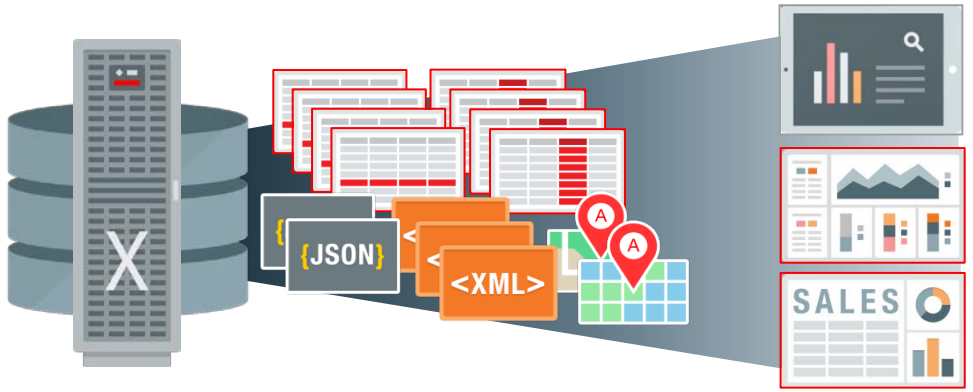
Extreme Flash Storage Server

Exadata Use Cases

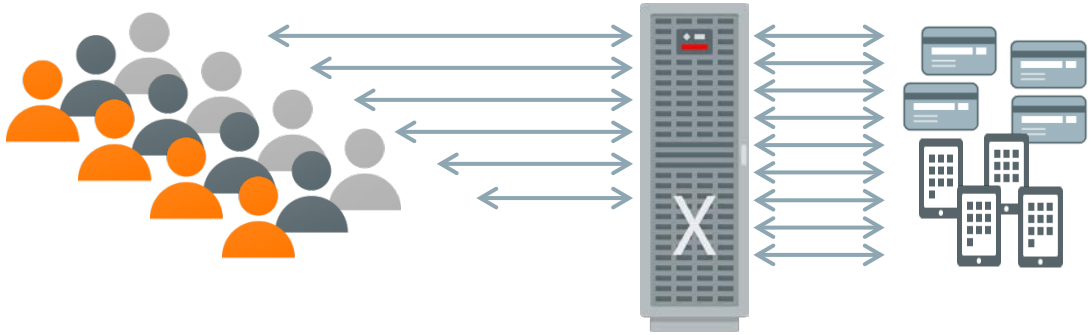
- **DATABASE CONSOLIDATION / DBaaS**



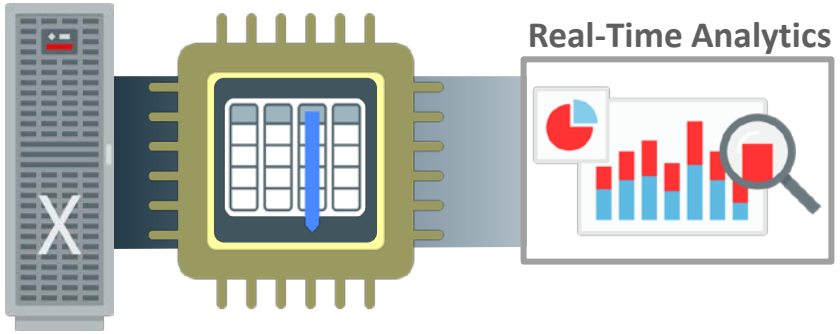
- **DATA WAREHOUSING**



- **ONLINE TRANSACTION PROCESSING**



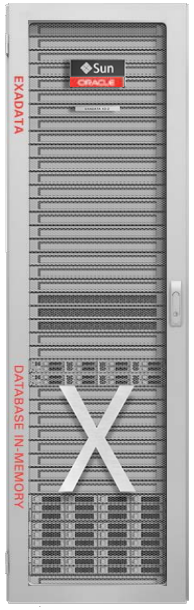
- **IN-MEMORY DATABASE**



Exadata Elastic Configurations

Optimize Exadata for any Workload

Qtr Rack

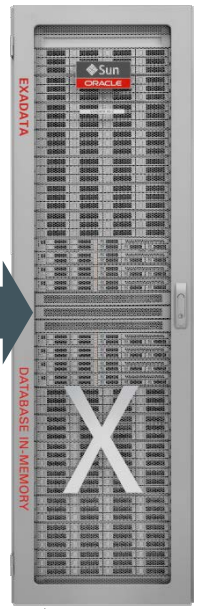


Database Server

Extreme Flash Storage

High-Capacity Storage

Full Rack

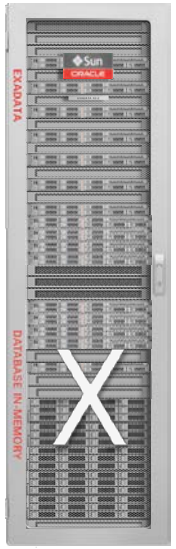


Start with
2 Database Servers
3 Storage Servers

Add Servers
Any Kind
Any Quantity

Configuration Examples

DB In-Memory Machine



15 DB Servers
5 Storage Servers

576 DB Cores
13.3 TB RAM
192 TB Disk

Extreme Flash OLTP Machine

11 DB Servers
11 Storage Servers

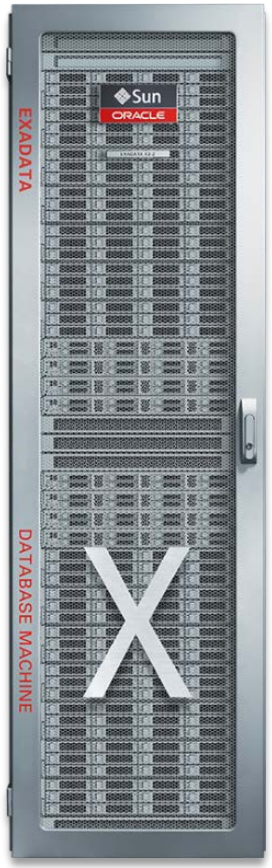
396 DB Cores
8 TB RAM
140 TB Flash

Data Warehousing Machine

8 DB Servers
14 Storage Servers

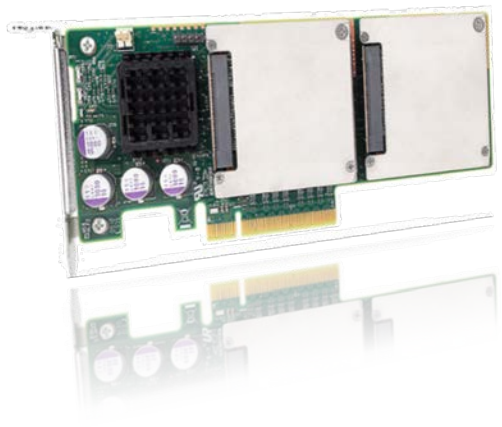
512 Cores
90 TB Flash Cache
672 TB Storage

Oracle's Flash Architecture



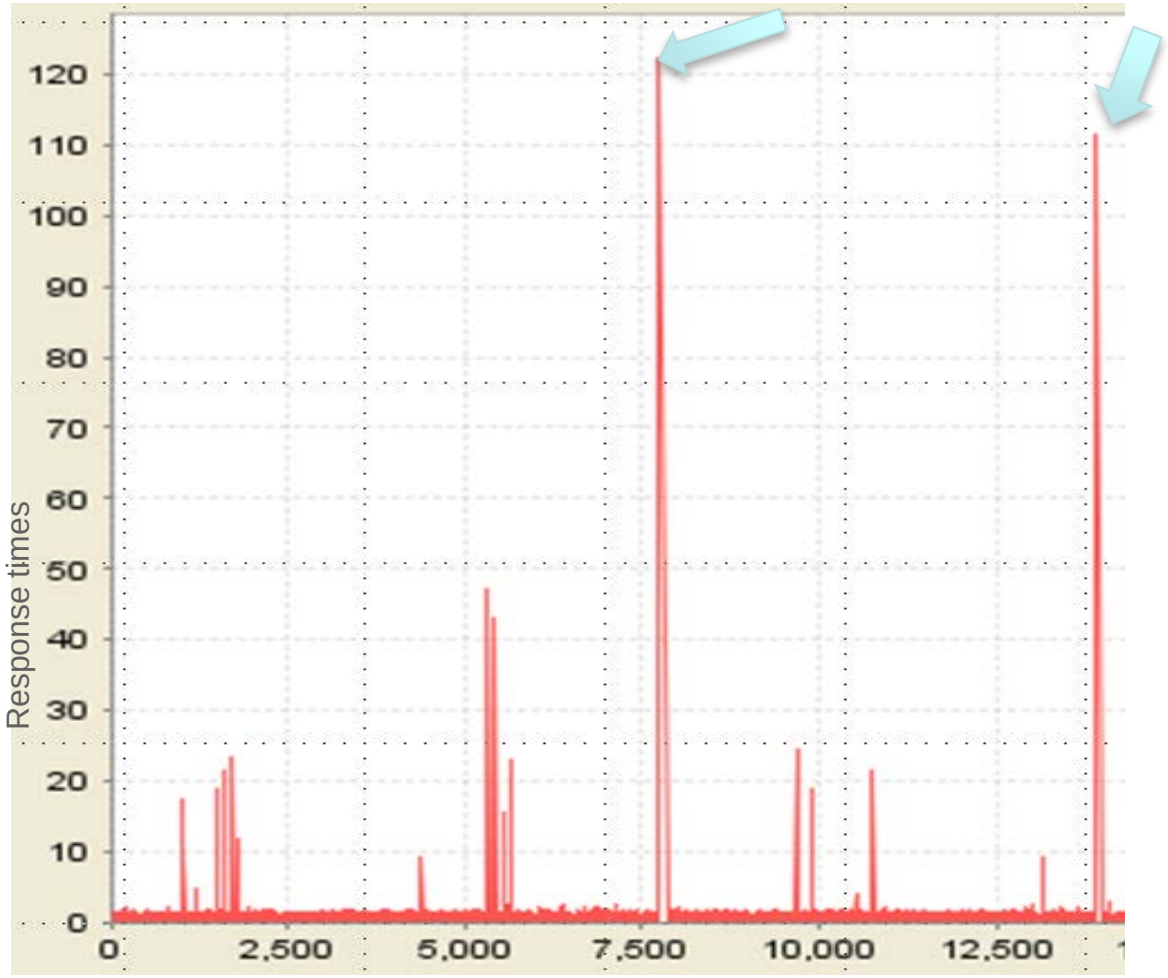
- Scale out architecture
  - adds flash capacity and performance by adding storage servers
  - adds networking and CPU needed to process flash in one unit
- Database Aware Storage
  - Metadata about IO present on the cell
- Flash on the Storage Server enables sharing
  - A block on disk is stored in only one 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
- Write-back flash cache
  - Caches writes from the database not just reads
- RAC-aware from day one

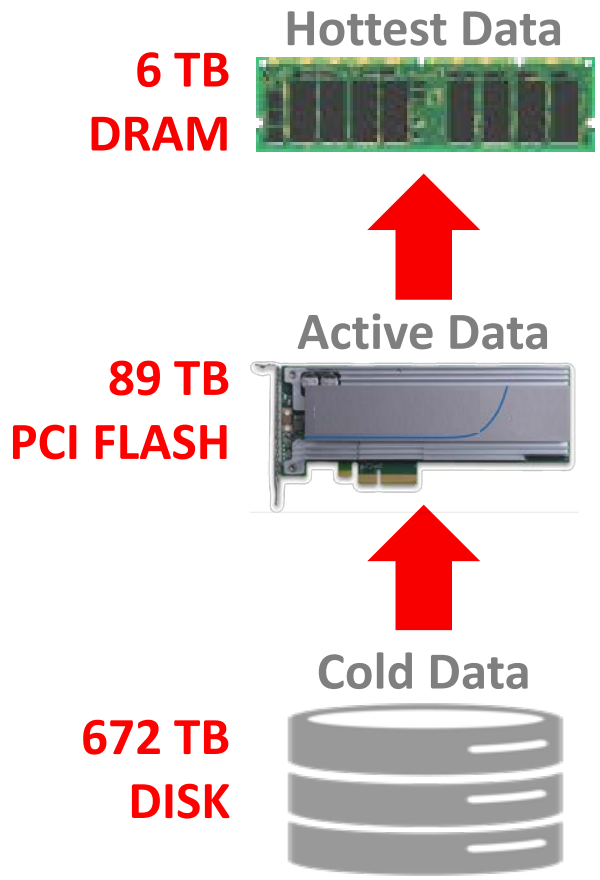
Flash And Database Logs



- Flash has very good *average* write latency
- Greatly improves user transaction response time
- Flash occasional outliers, one or two orders of magnitude slower
- OLTP workloads dislike such large variations
- *Oracle's Approach:* Write to Flash and the DRAM cache in the disk controller simultaneously to even out the impact of outliers
  - the first to complete "wins" so that outliers are avoided (on either medium)

Most Cost Effective Database Storage

- **Exadata software transparently gives best of memory, flash, disk**
  - **Cost and Capacity** of SAS Disk Storage
  - **I/Os** of Scale-Out PCI Flash
  - **Speed** of In-Memory DB
- **Hybrid Columnar Compression (HCC)**
  - **Industry best data compression (10x average) for analytics & archive**
  - Data remains compressed in flash, memory, backups, standbys



Per standard DB Machine full rack
8 DB, 14 HC storage servers

Comparison to Old system

Metric	Exadata ODS	Monolithic Hardware ODS	Comparison
Single Block Reads	1.5 ms	3.8 ms	> 2x
Log File Synch Waits	.85 ms	5.7 ms	> 6x

Note: The Exadata ODS is over twice the workload as the previous version. In addition, the Exadata system is shared with several databases, while the Monolithic Hardware was dedicated.

General Comments On Latency & IOPs

What are the alternatives to NVM Express in Enterprise Use Cases?

What are the implications of New Non-Flash Non Volatile Memories?

Will new NVMs require something completely different ?

Integrated Cloud

Applications & Platform Services


ORACLE®