Oracle Exadata: The World’s Fastest Database Machine
Exadata Database Machine Architecture

Ron Weiss, Exadata Product Management

Exadata Database Machine

Best Platform to Run the Oracle Database

- Best Machine for **Data Warehousing**
- Best Machine for **OLTP**
- Best Machine for **Database Consolidation**

Hardware and Software

Engineered to Work Together

Exadata Hardware Architecture

Scaleable Grid of industry standard servers for compute and storage

- Eliminates long-standing tradeoff between Scalability, Availability, Cost

Database Grid

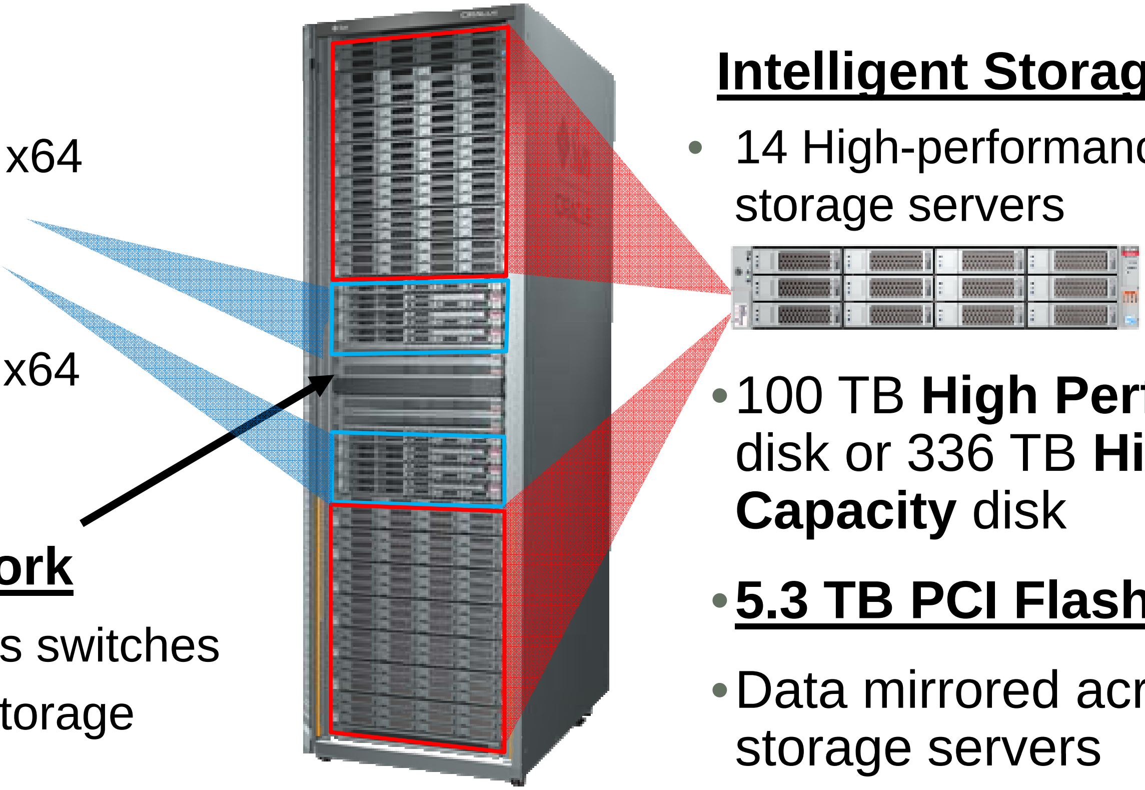
- 8 Dual-processor x64 database servers

or

- 2 Eight-processor x64 database servers

InfiniBand Network

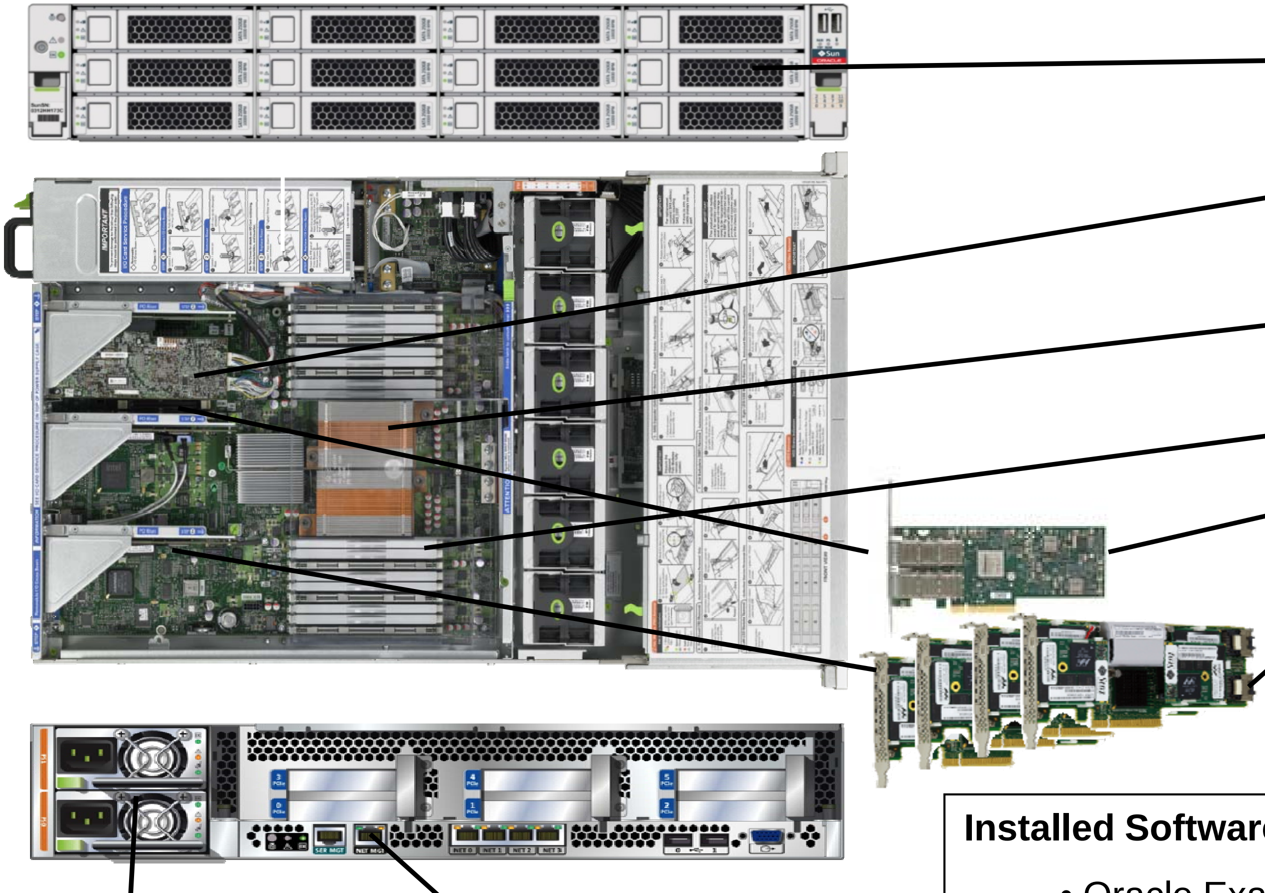
- Redundant 40Gb/s switches
- Unified server & storage network



Intelligent Storage Grid

- 14 High-performance low-cost storage servers
- 100 TB **High Performance** disk or 336 TB **High Capacity** disk
- **5.3 TB PCI Flash**
- Data mirrored across storage servers

Exadata Storage Server Hardware (Sun Fire X4270 M2)



12 x 3.5” Disk Drives
(600GB SAS or 2TB SAS)

Disk Controller SAS2 HBA with 512M battery backed cache

2 six-core Intel® Xeon® L5640 Processors

24 GB DRAM (6 x 4GB)

Dual-port InfiniBand QDR (40Gb/s) HCA

4 x 96GB Flash PCIe Cards

Dual-redundant, hot-swappable power supplies

ILOM

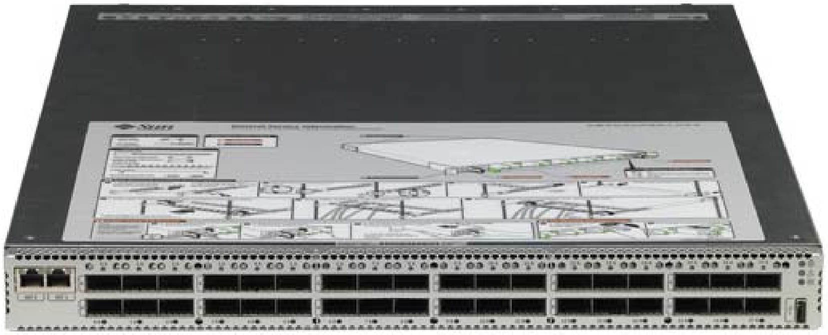
Installed Software:

- Oracle Exadata Storage Server Software
- Oracle Linux

Flash in the Exadata Storage Server

- Flash vs Disk tradeoff
  - 10X-100X better performance but 10X more expensive
- Exadata goal is get performance of Flash but at the price point of disk
- 4 x 96GB Sun F20 Flash Accelerator PCIe Cards in each storage server
  - 384 GB of Flash per Exadata Storage Server
- Choice of PCIe form factor over SSD for performance reasons
  - No disk controller bottleneck

InfiniBand Network



- Unified InfiniBand Network
  - Storage Network
  - RAC Interconnect
  - External Connectivity (optional)
- High Performance, Low Latency Network
  - 80 Gb/s bandwidth per link (40 Gb/s each direction)
  - SAN-like efficiency (Zero copy, buffer reservation)
  - Simple manageability like IP network
- Protocols
  - Zero-copy Zero-loss Datagram Protocol (ZDP RDSv3)
    - Linux Open Source, Low CPU overhead (Transfer 3 GB/s with 2% CPU usage)
  - Internet Protocol over InfiniBand (IPoIB) for external connectivity
    - Looks like normal Ethernet to host software (tcp/ip, udp, http, ssh,…)

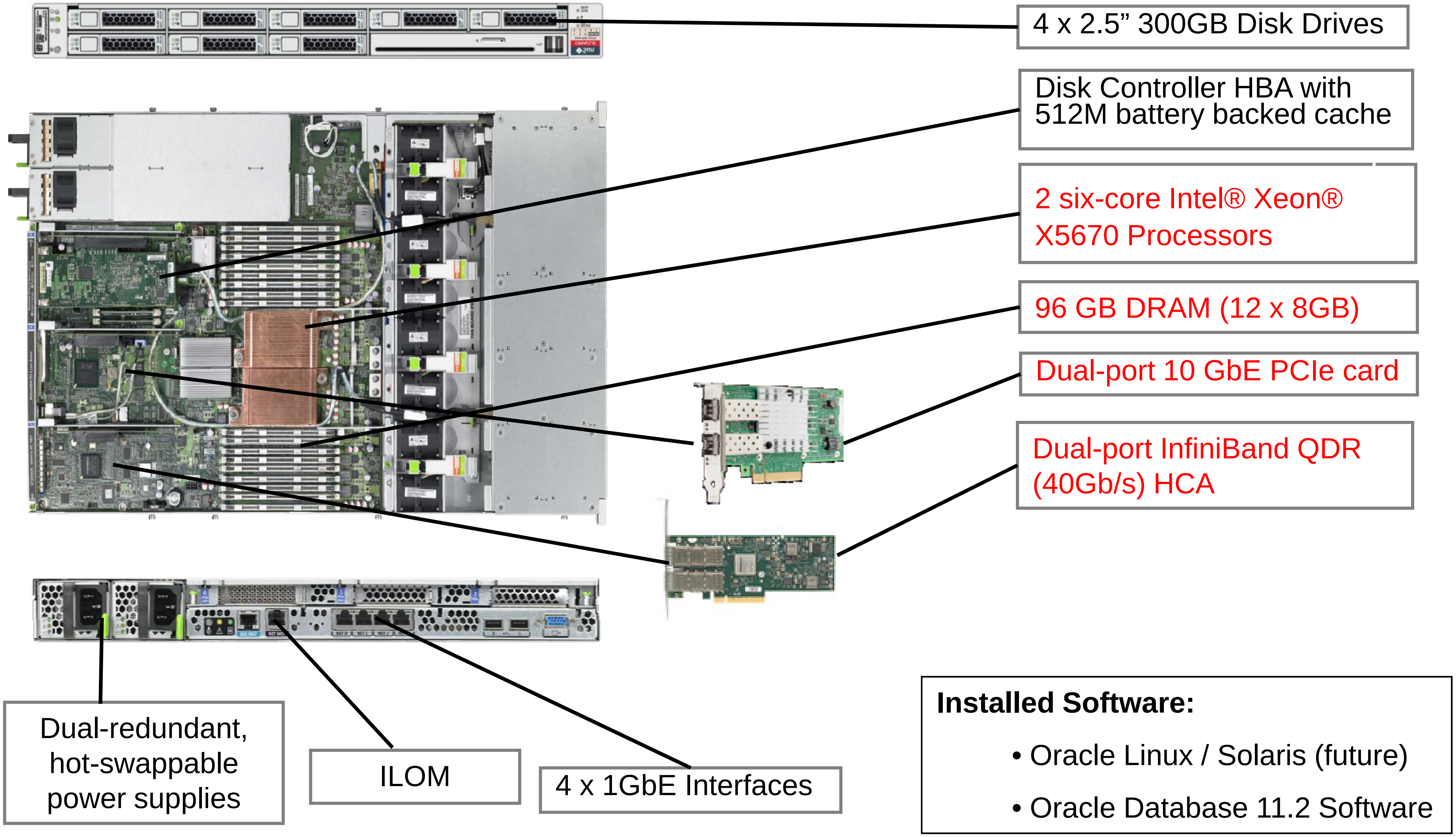
InfiniBand Network

- Uses Sun Datacenter 36-port Managed QDR (40Gb/s) InfiniBand switches
  - Runs subnet manager and automatically discovers network topology
  - Only one subnet manager active at a time
  - 2 “leaf” switches to connect individual server IB ports
  - 1 “spine” switch in Full Rack and Half Rack for scaling out to additional Racks
- Database Server and Exadata Servers
  - Each server has Dual-port QDR (40Gb/s) IB HCA
  - Active-Passive Bonding – Assign Single IP address
    - Performance is limited by PCIe bus, so active-active not needed
  - Connect one port from the HCA to one leaf switch and the other port to the second leaf switch for redundancy

Database Machine Models

- X2-2 and X2-8 - Two types of Database Machine models
  - Difference is the number and size of the database servers
- X2-2 uses smaller two-socket X4170 M2 servers
  - 6 cores per socket
- X2-8 uses larger eight-socket X4800 server
  - 8 cores per socket

X2-2 Database Servers (Sun Fire X4170 M2)



4 x 2.5” 300GB Disk Drives

Disk Controller HBA with 512M battery backed cache

2 six-core Intel® Xeon® X5670 Processors

96 GB DRAM (12 x 8GB)

Dual-port 10 GbE PCIe card

Dual-port InfiniBand QDR (40Gb/s) HCA

Dual-redundant, hot-swappable power supplies

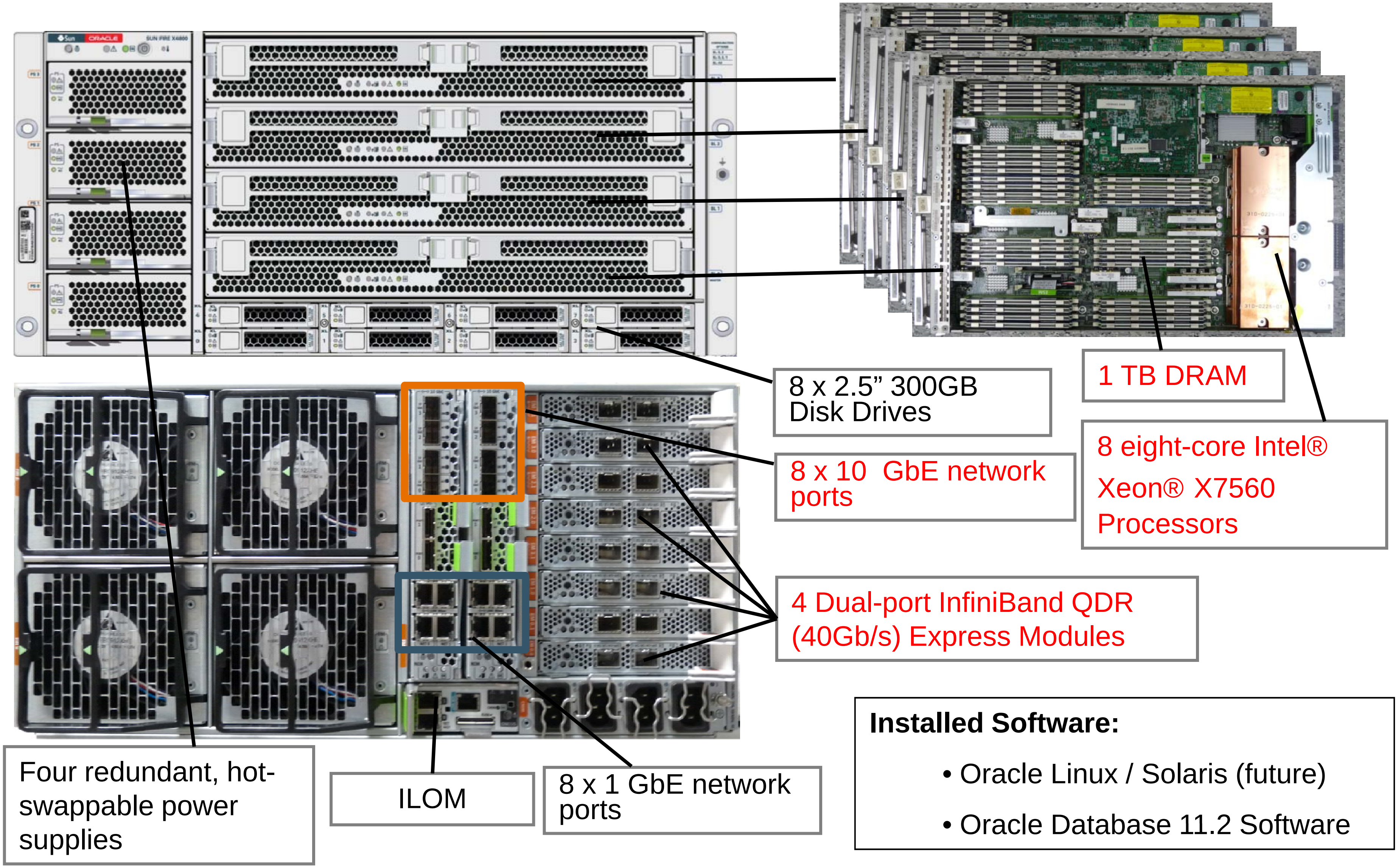
ILOM

4 x 1GbE Interfaces

Installed Software:

- Oracle Linux / Solaris (future)
- Oracle Database 11.2 Software

X2-8 Database Server (Sun Fire X4800)



8 x 2.5" 300GB Disk Drives

1 TB DRAM

8 x 10 GbE network ports

8 eight-core Intel® Xeon® X7560 Processors

4 Dual-port InfiniBand QDR (40Gb/s) Express Modules

Four redundant, hot-swappable power supplies

ILOM

8 x 1 GbE network ports

Installed Software:

- Oracle Linux / Solaris (future)
- Oracle Database 11.2 Software

Complete Family Of Database Machines

For OLTP, Data Warehousing & Consolidated Workloads

Oracle Exadata X2-2

Quarter Rack

Half Rack

Full Rack

Oracle Exadata X2-8

Full Rack

Exadata Database Machine X2-8 Full Rack

Extreme Performance for Consolidation, Large OLTP and DW

- 2 x 64 Eight-processor Database servers (Sun Fire 4800)
  - High Core, High Memory Database Servers
  - 128 CPU cores (64 per server)
  - 2 TB (1 TB per server)
  - 10 GigE connectivity to Data Center
    - 16 x 10GbE ports (8 per server)
- 14 Exadata Storage Servers X2-2
  - All with High Performance 600GB SAS disks

  OR
  - All with High Capacity 2 TB SAS disks
- 3 Sun Datacenter InfiniBand Switch 36
  - 36-port Managed QDR (40Gb/s) switch
- 1 “Admin” Cisco Ethernet switch
- Redundant Power Distributions Units (PDUs)

Add more racks for additional scalability

Exadata Database Machine X2-2 Full Rack

Pre-Configured for Extreme Performance

- 8 x 64 Dual-procesor Database Servers (Sun Fire X4170 M2)
  - 96 cores (12 per server)
  - 768 GB memory (96GB per server)
  - 10 GigE connectivity to Data Center
    - 16 x 10GbE ports (2 per server)
- 14 Exadata Storage Servers X2-2
  - All with High Performance 600GB SAS disks

  OR
  - All with High Capacity 2 TB SAS disks
- 3 Sun Datacenter InfiniBand Switch 36
  - 36-port Managed QDR (40Gb/s) switch
- 1 “Admin” Cisco Ethernet switch
- Keyboard, Video, Mouse (KVM) hardware
- Redundant Power Distributions Units (PDUs)

Add more racks for additional scalability

Exadata Database Machine X2-2 Half Rack

Pre-Configured for Extreme Performance

- 4 x 64 Dual-procesor Database Servers (Sun Fire X4170 M2)
  - 48 cores (12 per server)
  - 384 GB memory (96GB per server)
  - 10 GigE connectivity to Data Center
    - 8 x 10GbE ports (2 per server)
- 7 Exadata Storage Servers X2-2
  - All with High Performance 600GB SAS disks

  OR
  - All with High Capacity 2 TB SAS disks
- 3 Sun Datacenter InfiniBand Switch 36
  - 36-port Managed QDR (40Gb/s) switch
- 1 “Admin” Cisco Ethernet switch
- Keyboard, Video, Mouse (KVM) hardware
- Redundant Power Distributions Units (PDUs)

Can Upgrade to a Full Rack

Exadata Database Machine X2-2 Quarter Rack

Pre-Configured for Extreme Performance

- 2 x 64 Dual-procesor Database Servers (Sun Fire X4170 M2)
  - 24 cores (12 per server)
  - 192 GB memory (96GB per server)
  - 10 GigE connectivity to Data Center
    - 4 x 10GbE ports (2 per server)
- 3 Exadata Storage Servers X2-2
  - All with High Performance 600GB SAS disks

  OR
  - All with High Capacity 2 TB SAS disks
- 2 Sun Datacenter InfiniBand Switch 36
  - 36-port Managed QDR (40Gb/s) switch
- 1 “Admin” Cisco Ethernet switch
- Keyboard, Video, Mouse (KVM) hardware
- Redundant Power Distributions Units (PDUs)

Can Upgrade to an Half Rack

Scale to 8 Racks by Just Adding Cables

Full Bandwidth and Redundancy

Half and Full racks can be connected

- Eight X2-2 Full Racks
  - 768 CPU cores and 6.1 TB memory for database processing
  - 1,344 CPU cores for storage processing
  - 42.4 TB Flash Storage
  - 800 TB or 2,688 TB Raw Disk Storage

- Eight X2-8 Racks
  - 1,024 CPU cores and 16 TB of memory for database processing
  - 1,344 CPU cores for storage processing
  - 42.4 TB Flash Storage
  - 800 TB or 2,688 TB Raw Disk Storage

X2-2 and X2-8 Full Rack

	X2-8 Full Rack	X2-2 Full Rack
Database Servers	**2**	**8**
Cores (Total)	128 (2.26 GHz)	96 (2.93 GHz)
Memory (Total)	2048 GB	768 GB
1 GbE Ports (Total)	16	32
10 GbE Ports(Total)	16	16
InfiniBand Switches	**3**	
Exadata Storage Servers	**14**	
Flash (Total)	5.3 TB	
Raw Storage (Total)	100 TB or 336 TB	
Raw Disk Data Bandwidth	**25 GB/s***	
Raw Flash Data Bandwidth	**50 GB/s**	
Flash IOPS (8k Reads)	**1,000,000**	

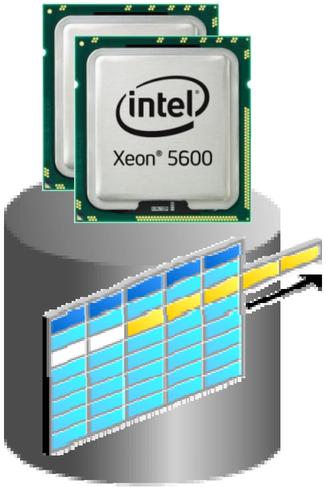
* Using High Performance 15K RPM disks

ORACLE

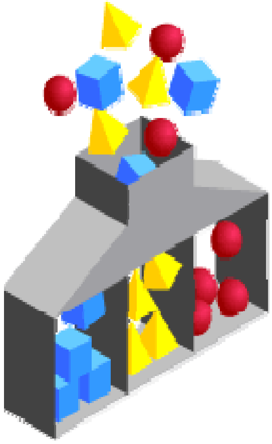
Database Server Operating System Choices

- Two Operating System Choices on the database servers
  - Oracle Linux
  - Solaris 11 Express (x86) – Coming Soon
- Customers choose their preferred database server OS at installation time
  - No pricing difference
  - No performance difference
  - Choice driven by familiarity and expertise with the OS

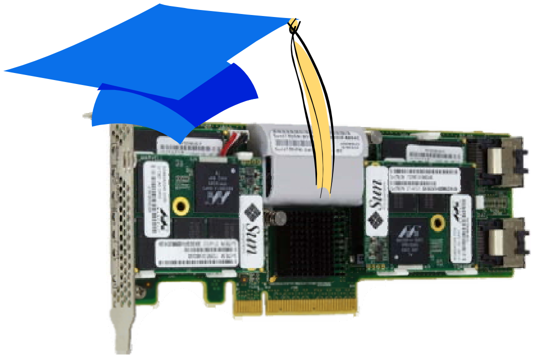
Keys to Speed and Cost Advantage



Exadata
Intelligent Storage
Grid



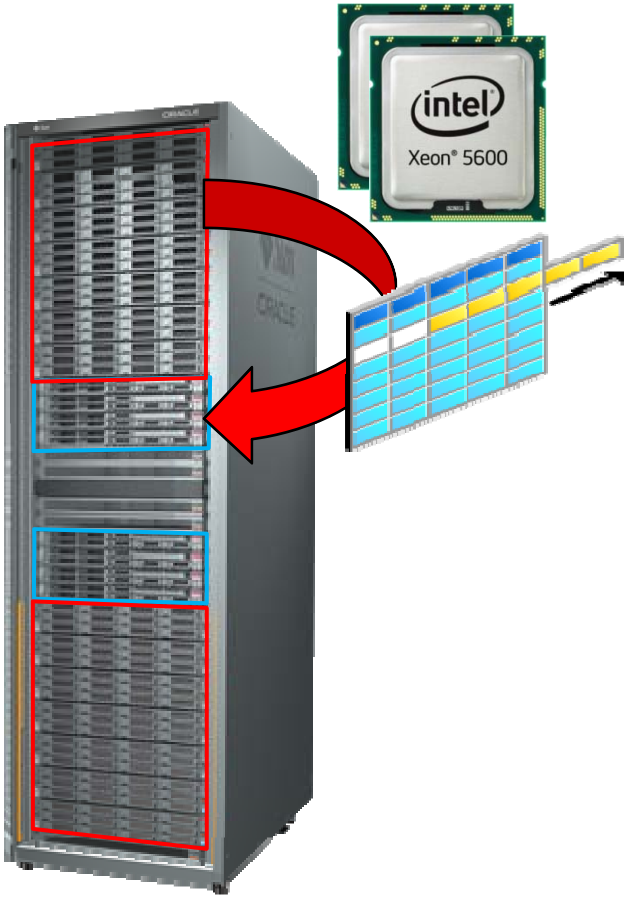
Exadata Hybrid
Columnar
Compression



Exadata Smart
Flash Cache

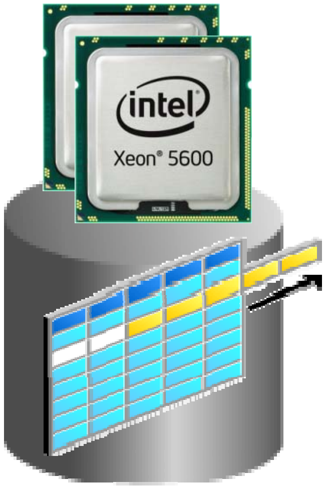
Exadata Intelligent Storage Grid

Most Scalable Data Processing



- **Data Intensive processing runs in Exadata Storage Grid**
  - Filter rows and columns as data streams from disks (112 Intel Cores)
- Example: How much product X sold last quarter
  - Exadata Storage Reads 10TB from disk
  - Exadata Storage Filters rows by Product & Date
  - Sends 100GB of matching data to DB Servers
- Scale-out storage parallelizes execution and removes bottlenecks

Exadata Intelligent Storage



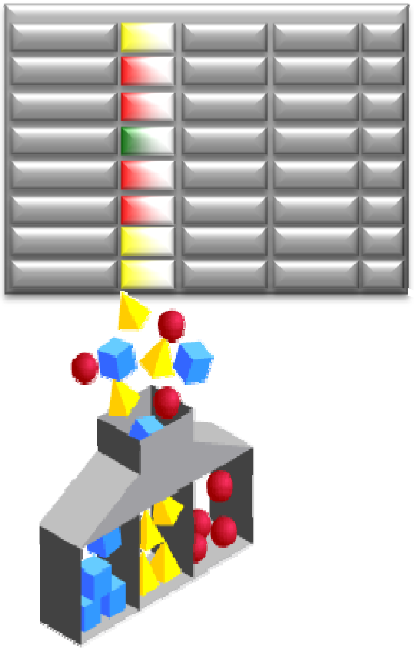
Exadata Intelligent Storage Grid

- Exadata storage servers also run more complex operations in storage
  - **Join filtering**
  - **Incremental backup filtering**
  - **I/O prioritization**
  - **Storage Indexing**
  - **Database level security**
  - **Offloaded scans on encrypted data**
  - **Data Mining Model Scoring**
  - **Smart File Creation**

- 10x reduction in data sent to DB servers is common

Exadata Hybrid Columnar Compression

Highest Capacity, Lowest Cost



- Data is organized and compressed by column
  - Dramatically better compression
- Speed Optimized **Query Mode** for Data Warehousing
  - **10X compression typical**
  - **Runs faster because of Exadata offload!**
- Space Optimized **Archival Mode** for infrequently accessed data
  - **15X to 50X compression typical**

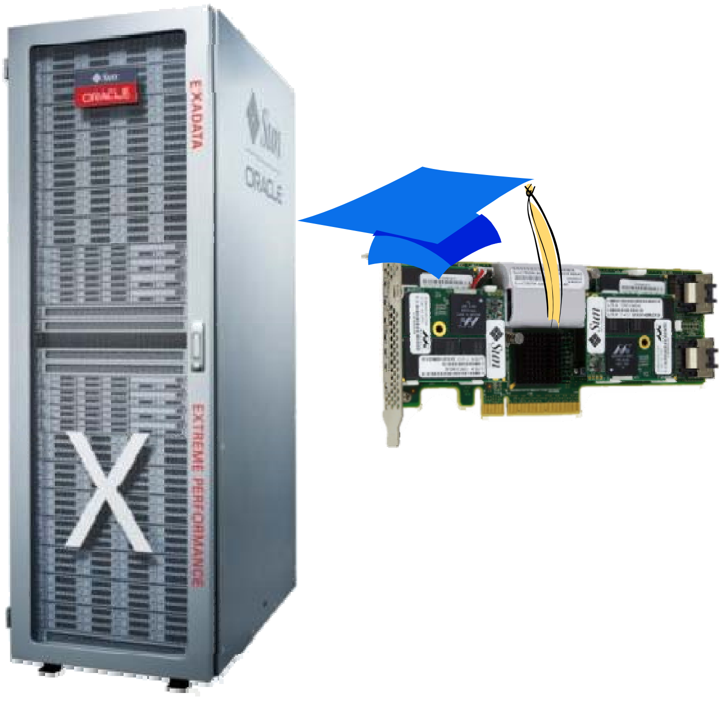
Faster and Simpler

Backup, DR, Caching, Reorg, Clone

Benefits Multiply

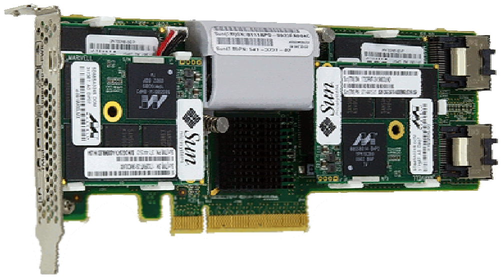
Exadata Smart Flash Cache

Extreme Performance OLTP



- Exadata has **5 TB** of flash
  - **56 Flash PCI cards avoid disk controller bottlenecks**
- **Intelligently manages flash**
  - Smart Flash Cache holds hot data
  - **Gives speed of flash, cost of disk**
- Exadata flash cache achieves:
  - Over **1 million IO/sec from SQL** (8K)
  - Sub-millisecond response times
  - **50 GB/sec query throughput**

Smart Flash Cache



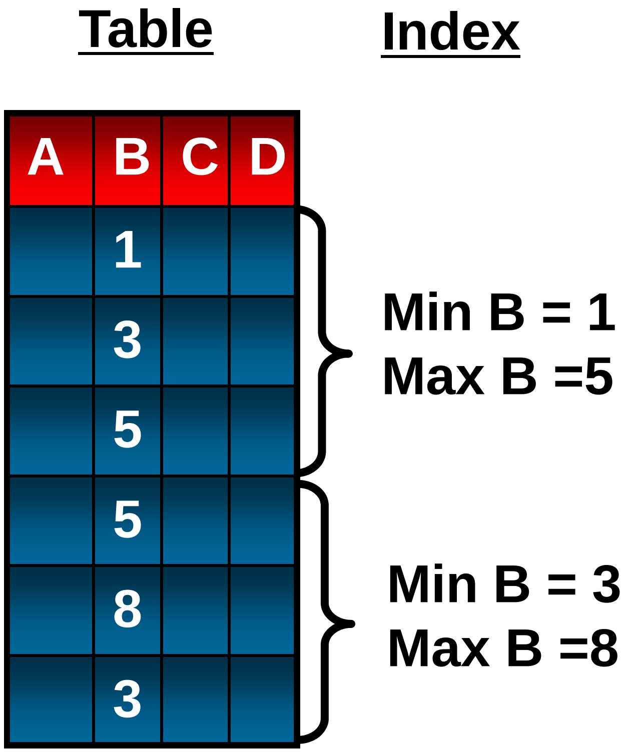
- Understands different types of I/Os from database
  - Skips caching I/Os to mirror copies
  - Skips caching backups
  - Skips caching data pump I/O
  - Skips caching tablespace formatting
  - Resistant to table scans
  - Control File Reads and Writes are cached
  - File header reads and writes are cached
  - Data Blocks and Index blocks are cached

Smart Flash Cache Keep Directive

- DBA can enforce that an object is kept in flash cache
  - `ALTER TABLE calldetail STORAGE (CELL_FLASH_CACHE KEEP)`
- Can be set like other storage clause values
  - At table level, partition level, during creation time etc.
- Table scans on objects marked with cell_flash_cache keep run through the flash cache
  - Disk bandwidth full rack – 25GB/s
  - Flash bandwidth full rack – 50GB/s

Exadata Storage Index

Transparent I/O Elimination with No Overhead



Table

A	B	C	D
	1		
	3		
	5		
	5		
	8		
	3		

Index

Min B = 1
Max B =5

Min B = 3
Max B =8

- Exadata Storage Indexes maintain summary information about table data in memory
  - Store MIN and MAX values of columns
  - Typically one index entry for every MB of disk
- Eliminates disk I/Os if MIN and MAX can never match “where” clause of a query
- Completely **automatic and transparent**

Select * from Table where B<2 - Only first set of rows can match

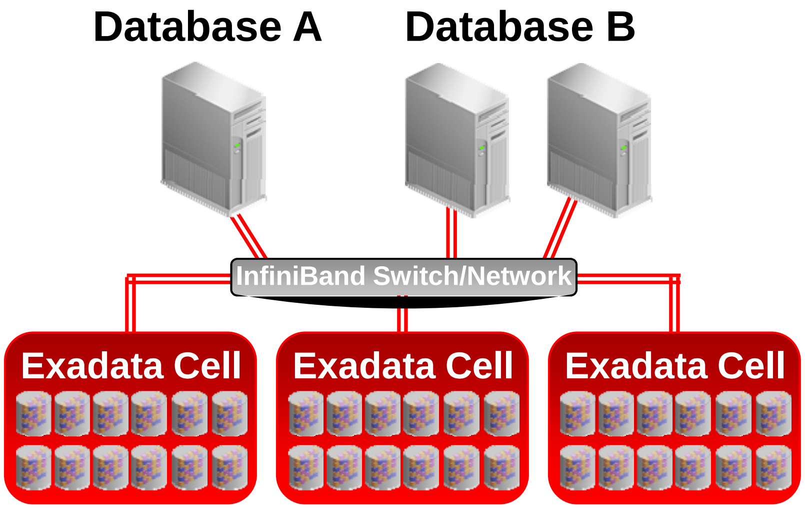
Most Secure Database Machine

- Moves decryption from software to hardware
  - Over 5x faster
  - Leverages AES-NI compliant hardware
- Near zero overhead for fully encrypted database
  - Queries decrypt data at hundreds of Gigabytes/second
- DB2, Teradata and Netezza do not have database managed encryption
  - Must write into every application module

Exadata I/O Resource Management

Mixed Workloads and Multi-Database Environment

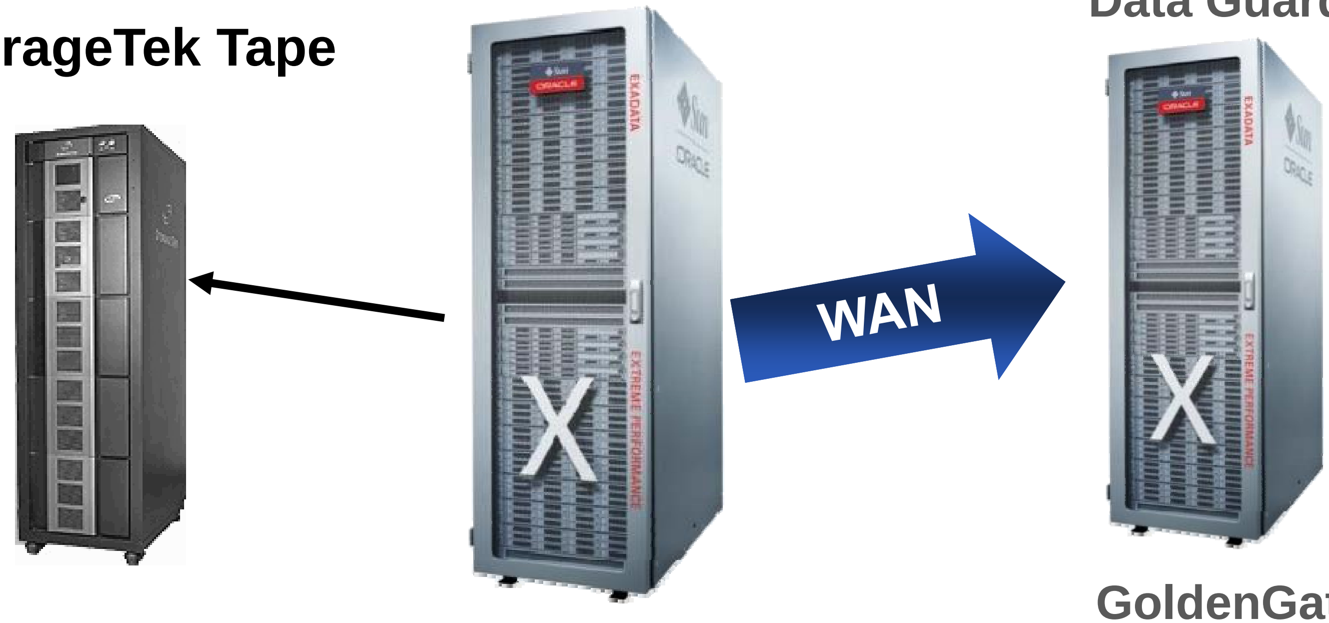
- Ensure different databases are allocated the correct relative amount of I/O bandwidth
  - Database A: 33% I/O resources
  - Database B: 67% I/O resources
- Ensure different users and tasks within a database are allocated the correct relative amount of I/O bandwidth
  - Database A:
    - Reporting: 60% of I/O resources
    - ETL: 40% of I/O resources
  - Database B:
    - Interactive: 30% of I/O resources
    - Batch: 70% of I/O resources



Best Machine for Database Consolidation

- Exadata serves as farm/cloud for databases
  - Large memory enables many databases to be consolidated
  - Extreme performance for complex workloads that mix OLTP, DW, batch, reporting
  - I/O and CPU resource management isolates workloads

Best and Fastest HA



StorageTek Tape

WAN

Active Data Guard

GoldenGate Replication

- Full backup
  - 20 TB/hour disk-to-disk in Exadata
  - 8 TB/hour Exadata to tape backup
    - Tape drive limited
- Incremental backup is 10x faster

- Real-Time Active Replica
- Data Guard keeps up with 5TB/hour compressed loads

Exadata Summary

- Best for OLTP
  - Smart Flash Cache
  - 1 Million I/Os per Second
- Best for Warehousing
  - Intelligent Scale-Out storage
    - 10x faster queries
  - 10x Data Compression
- Best for Consolidation
  - Terabytes of Memory
  - Mix OLTP, DW, batch, reporting in single machine

- Complete Ready-to-Run System
- Full database encryption with near zero overhead
- Runs all Oracle Applications unchanged

ORACLE

Q&A