OPENFABRICS ALLIANCE

Infiniband: Enabling Massively Scalable Databases

Tim Shetler, Oracle Product Management
March 15, 2010

But first...

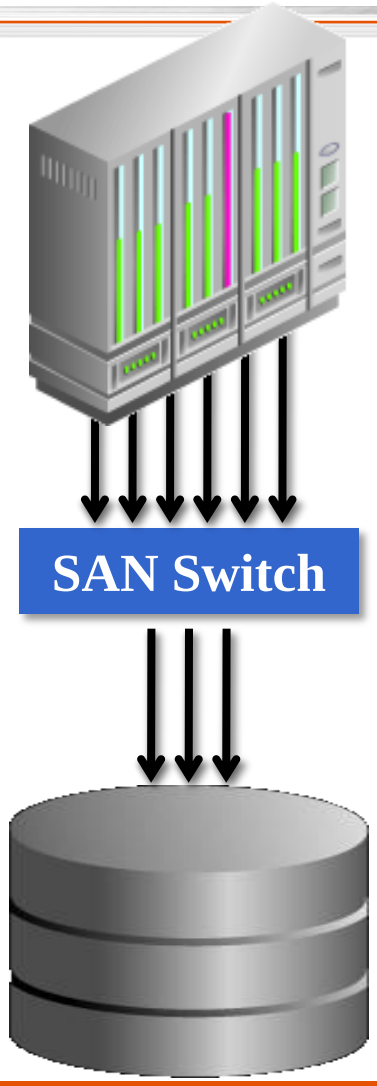
- Place your business card in the fishbowl being passed around for a chance to win an Amazon Kindle!

- The lucky winner will be drawn at the end of this presentation...

Today's Database Trends

- Big data warehouses
  - Multi-terabyte to petabyte
  - Response time in seconds-minutes
- Consolidation
  - Dozens-hundreds of databases
  - Mixed workloads
- Appliances
  - Pre-configured
  - Balanced performance

Traditional DB Configuration



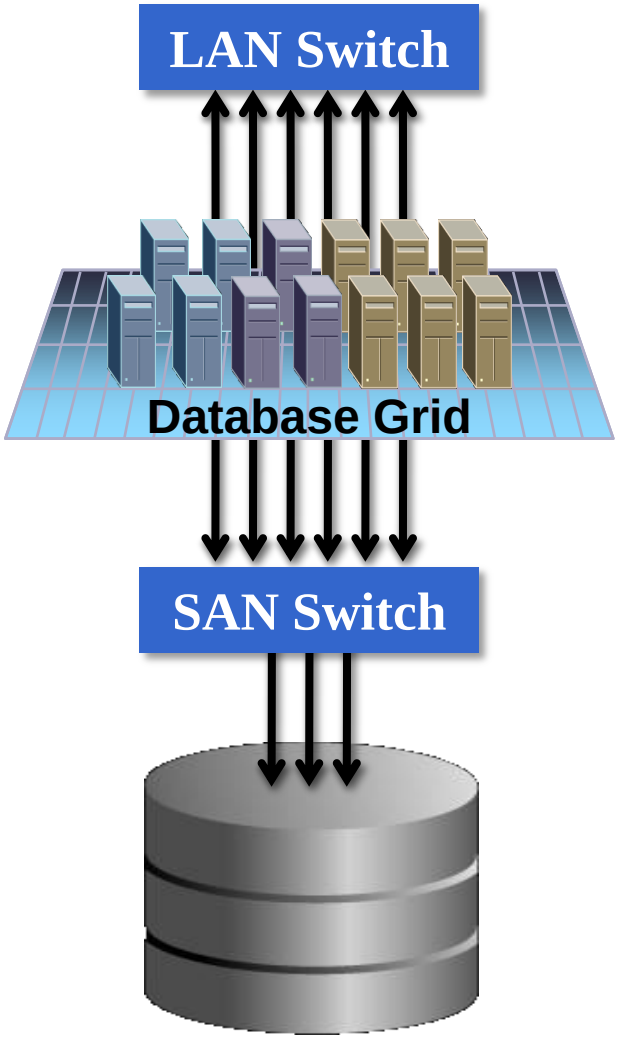
Monolithic SMP Server

- High-Cost Scale-Up
- Limited Scalability

Monolithic Storage Array

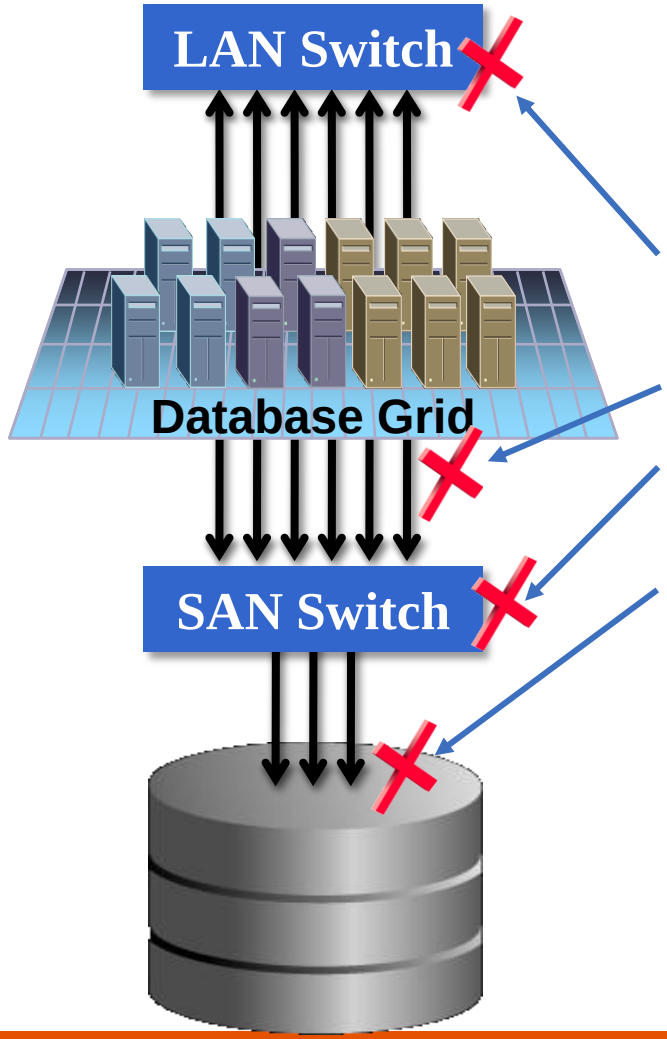
- High-Cost Scale-Up

Grid Scales the Server Tier



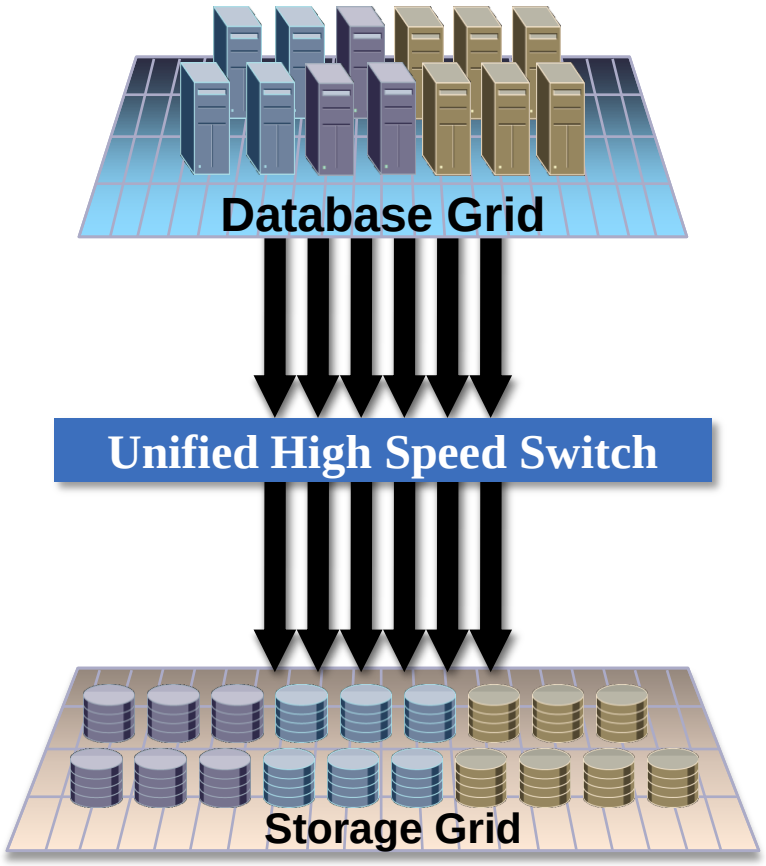
- Real Application Clusters (RAC)
  - Released in 2001 (Oracle 9i)
  - Thousands of production customers
- Scale-out using low cost servers
  - Single system image
  - Highly available architecture
  - Keep pace with latest hardware
- OLTP queries use indexed access
  - One SAN link can serve 50K IOPS
  - Storage arrays with hundreds of disks common

Bottlenecks for Big Data Scans



- Need 10's of Gigabytes per second of I/O
  - Many bottlenecks prevent this today
- LAN switches can't handle load of large joins
- Server nodes need many SAN adapters
- Storage switch cost and SAN complexity increase dramatically
- Large storage arrays cannot deliver bandwidth of hundreds of disks
  - Bottleneck on storage heads and connections to SAN switches
- Result is poor performance for huge data scans (Data Warehouses)

Solving the I/O Bottleneck



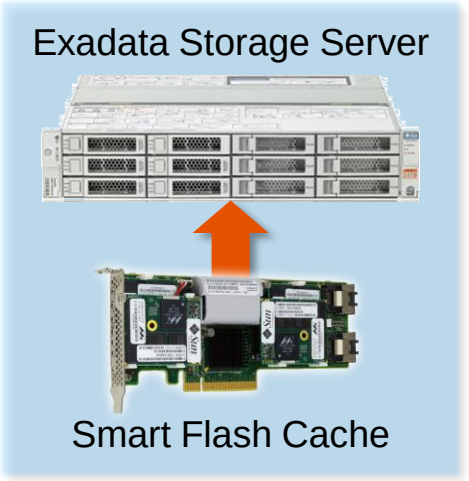
- Bring Grid Architecture to storage
  - Multi-core Intel x86 processors and high volume disks
- Using next generation high speed network
  - 40 Gb Infiniband
  - Unified server and storage network
- Database intelligence in storage tier offloads scans and reduces network traffic
- Wider roads, less traffic
  - Storage and server bandwidth in balance
  - Only relevant data travels the network

Putting it all Together

Sun Oracle Database Machine – the Hardware

* Full-rack configuration

- Oracle Database 11g Server Grid
  - 8 database servers*
  - 64 Intel Xeon cores
- Exadata Storage Server Grid
  - 14 storage servers
  - 112 Intel Xeon cores
  - 100 TB (SAS) or
    336 TB (SATA) raw disk
  - 5.3 TB flash storage
- InfiniBand Network
  - 40 Gb/sec
  - Unified server / storage network
    - Storage to Database
    - RAC interconnect



Putting it all Together

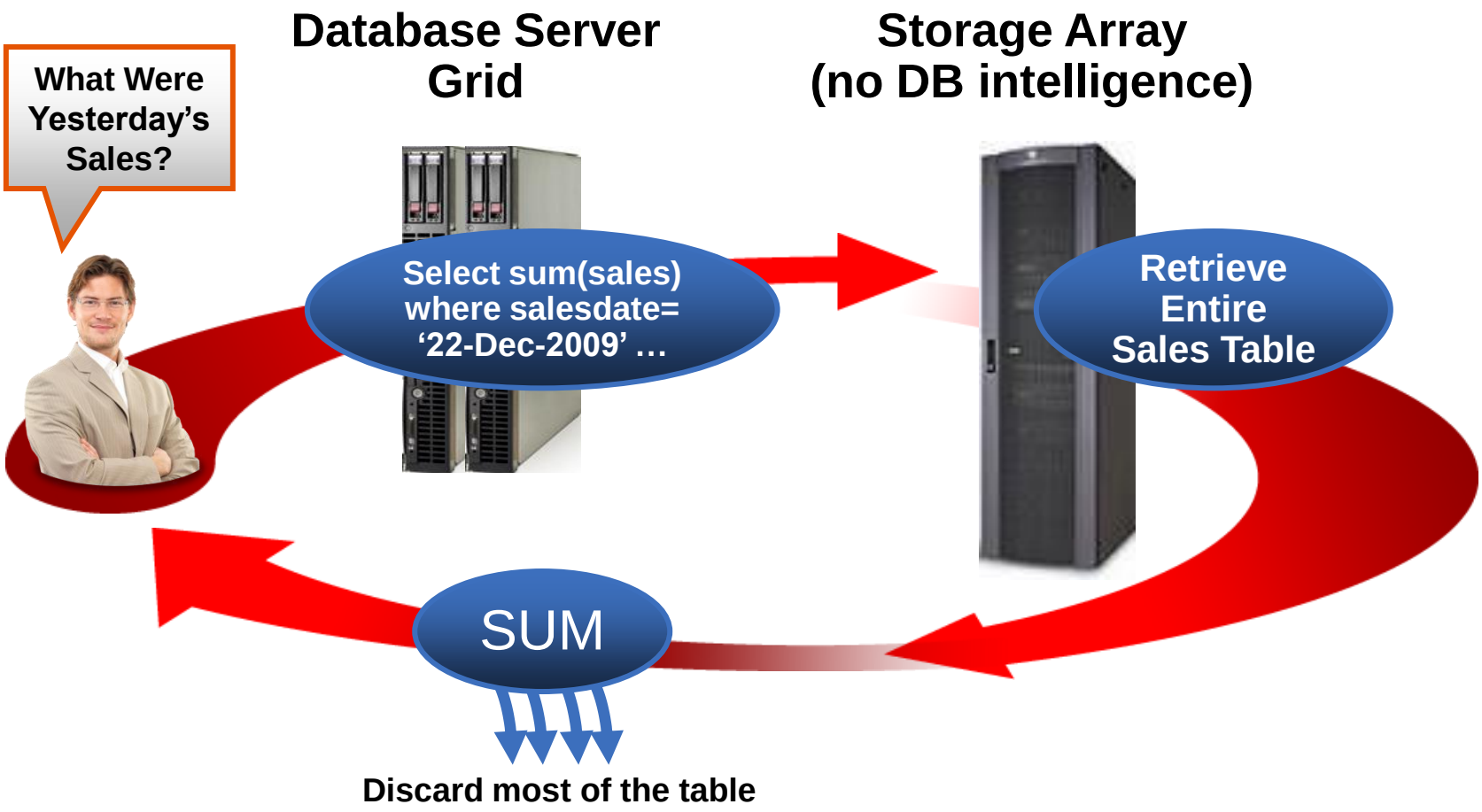
Sun Oracle Database Machine – the Software

Reducing network overhead…

- **Smart Scan –** Query processing within storage
- **Partitioning and Storage Indexes –** I/O elimination
- **Columnar Compression –** I/O and storage reduction
- **Smart Flash Cache –** I/O acceleration
- **Automatic Storage Management** – I/O striping
- **Infiniband with RDS/RDMA –** Latency reduction

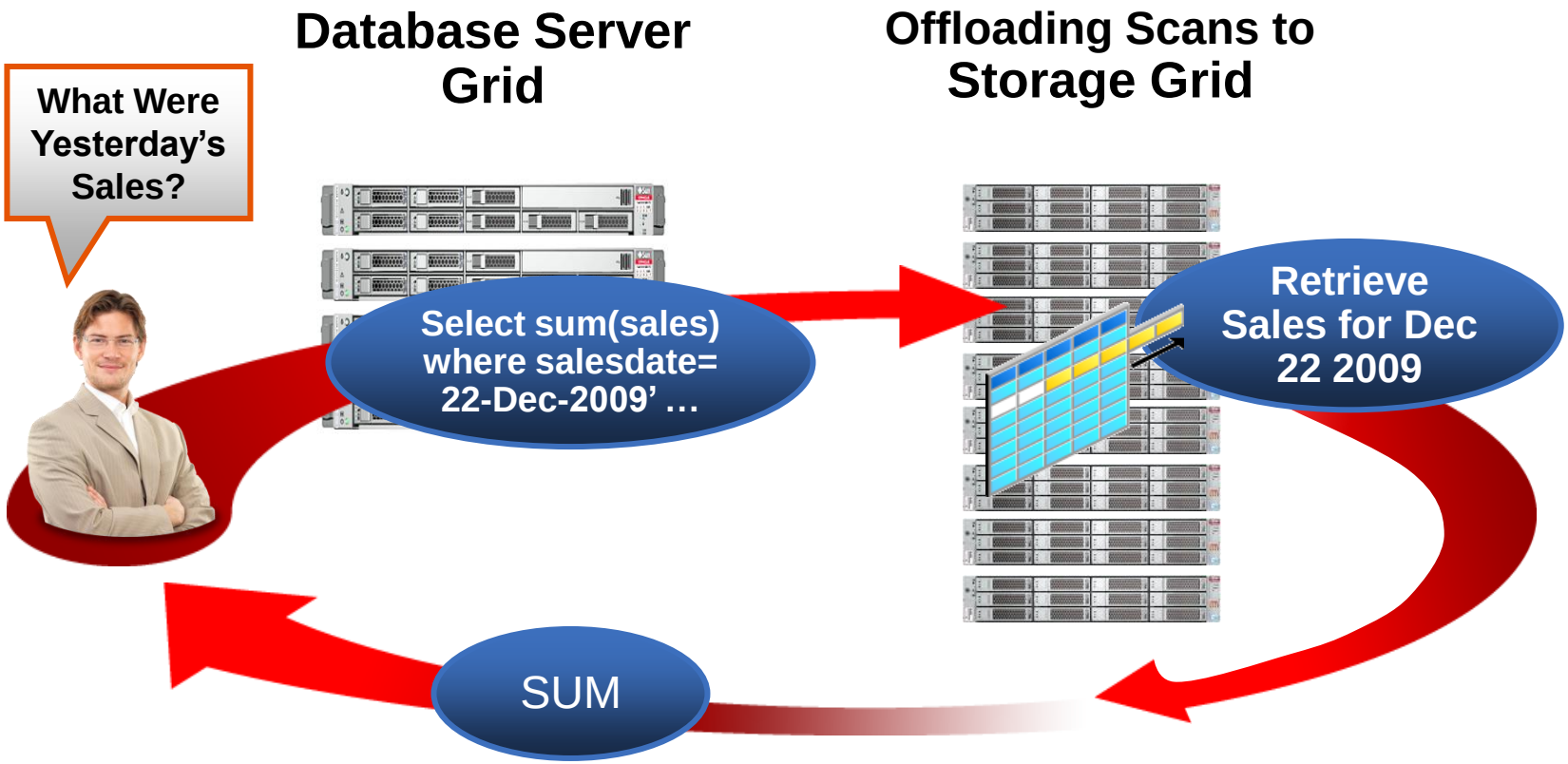
Query Processing – No Offload

The problem with traditional storage

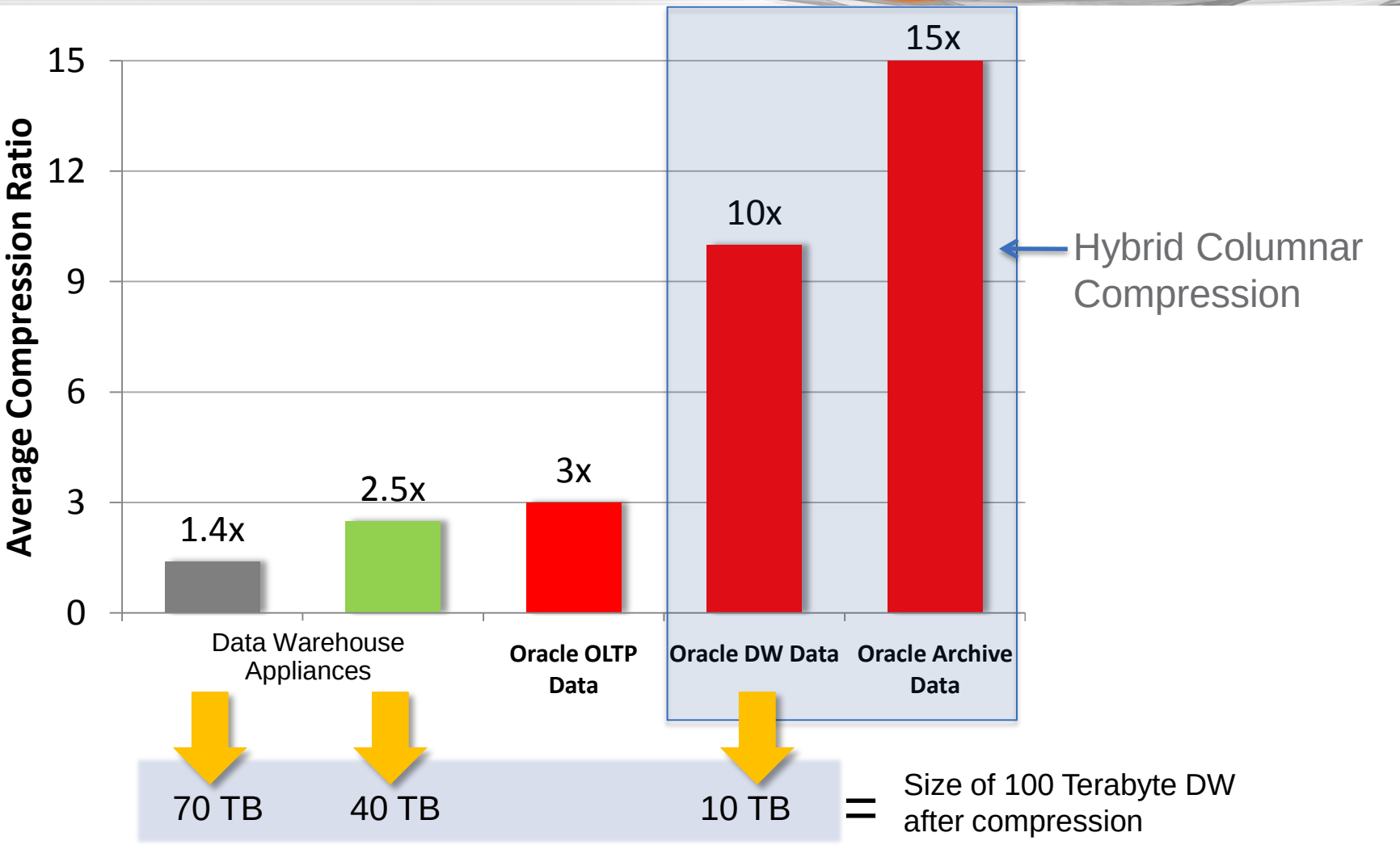


Query Processing – Offloaded

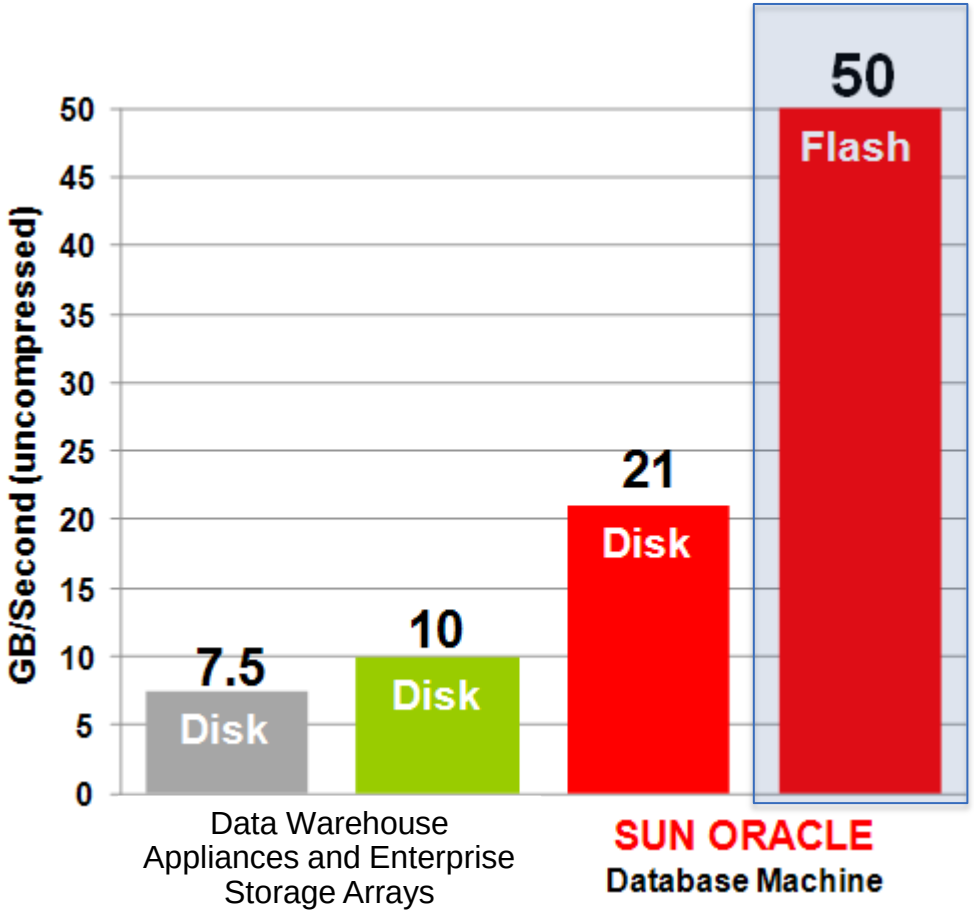
The value of database-aware storage



Columnar Compression



Smart Flash Cache

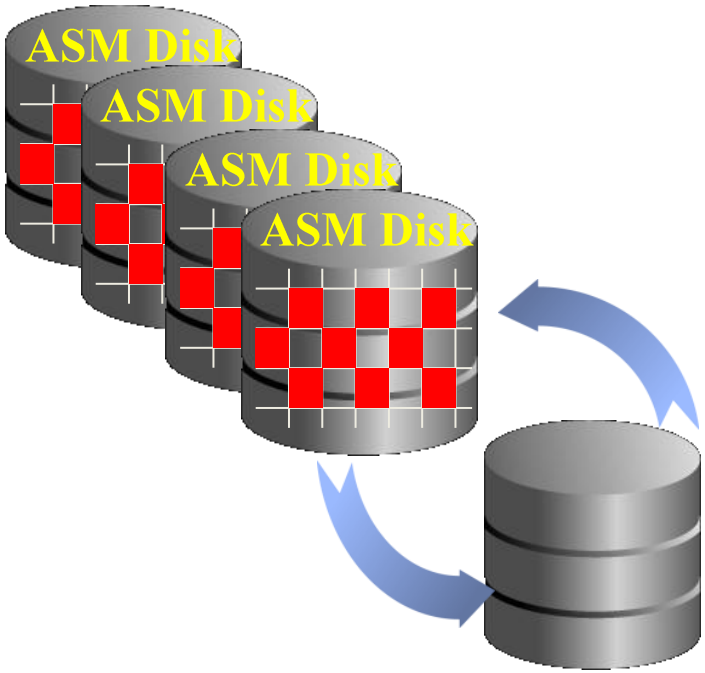


- **I/O Throughput**
  Up to 50 GB per second[1]
- **Random I/O Rate**
  Up to 1 million I/O's per second[2]

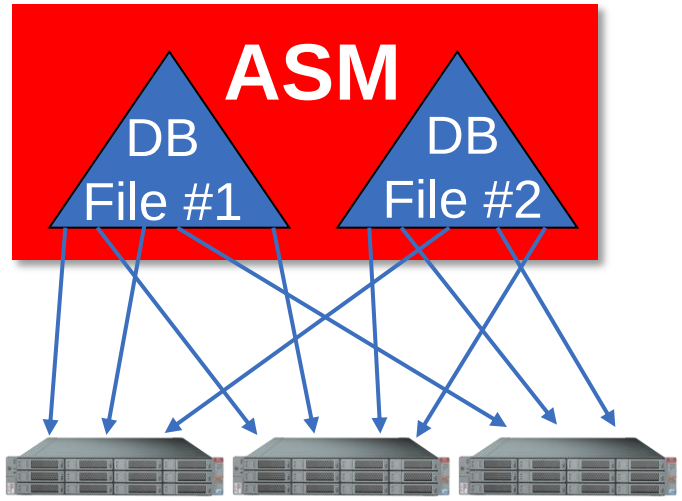
[1] **Uncompressed data from Flash (full rack) – compression increases effective throughput by the compression ratio.**
[2] **Full rack (56 flash cards)**

Automatic Storage Management

- Integrated cluster volume manager
- Flexible data distribution (striping)
- Mirroring
- Automatic data re-balancing

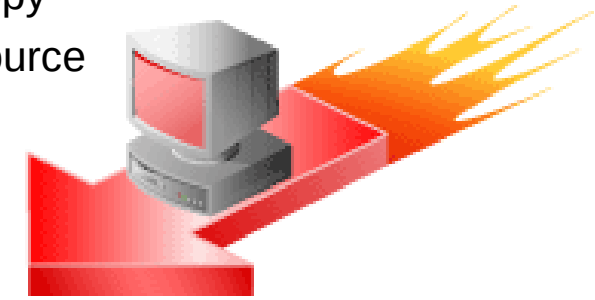


- Manages storage in megabyte allocation units
- Evenly spreads allocation units across all cells and disks in the grid

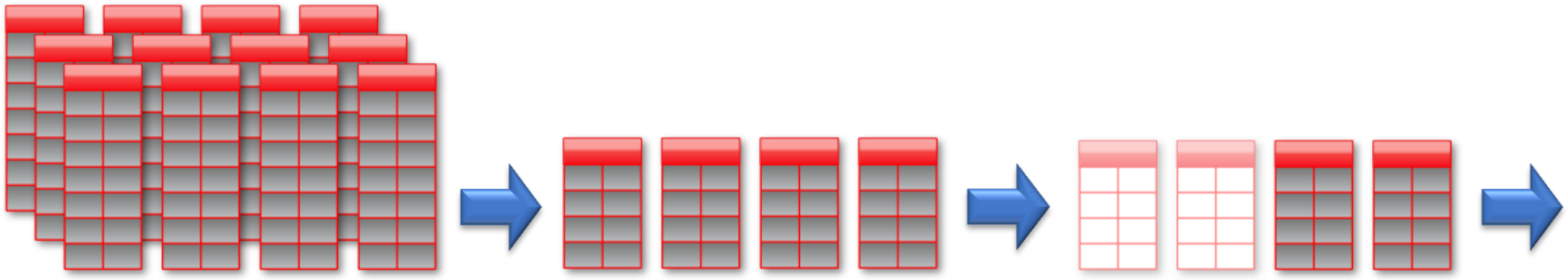


Infiniband Network

- Infiniband is the interconnect
  - Provides highest performance available with the richness of computer network and low overhead of storage network
  - Zero copy and buffer reservation abilities of Storage network
- Infiniband looks like normal Ethernet from a software point of view
  - All IP based tools work transparently – tcp/ip, udp, and everything built on top – http, rsh, ftp, etc.
- Unified Network Fabric
  - Same infiniband network used for grid storage and cluster interconnect
  - Less configuration, lower cost, higher performance
- Uses high performance RDMA Infiniband protocol (RDS V3)
  - Datagram protocol like UDP but reliable and zero copy
  - Implemented by Oracle, available as Linux Open Source
  - Very low CPU overhead



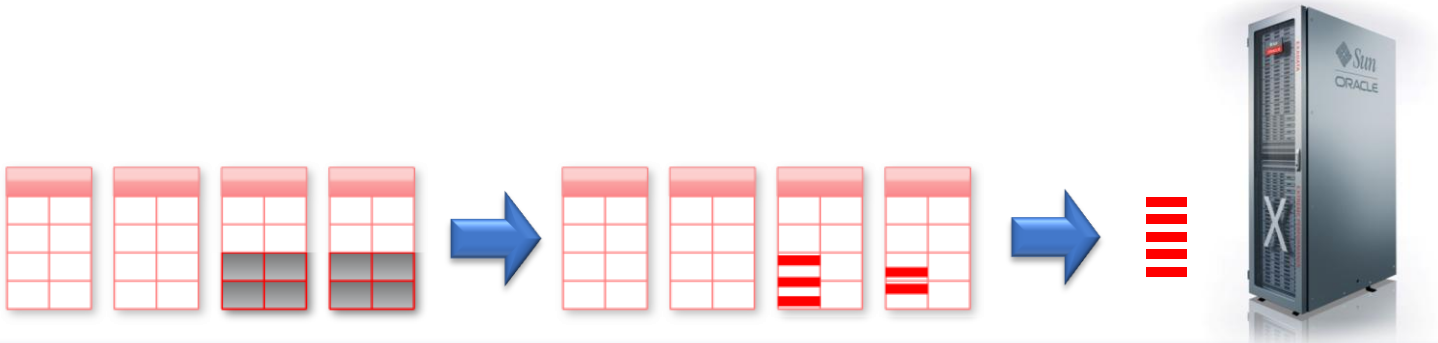
Converting TB to GB



10 TB of user data
(traditionally 10 TB of I/O)

1 TB
(with 10x compression)

100 GB
(with partition pruning)



20 GB
(with storage indexes)

5 GB
(with Smart Scan)

Subsecond "10 TB" scan

No Indexes Required

Consolidation Workloads

- Server rationalization
  - OLTP + OLTP …
  - Data mart + data mart…
  - System life-cycle
    - Production + test + development
- Mixed workload
  - Operational BI
  - Real-time data warehousing
  - Embedded reports, analytics
- Schema integration

Optimizing Mixed Workloads

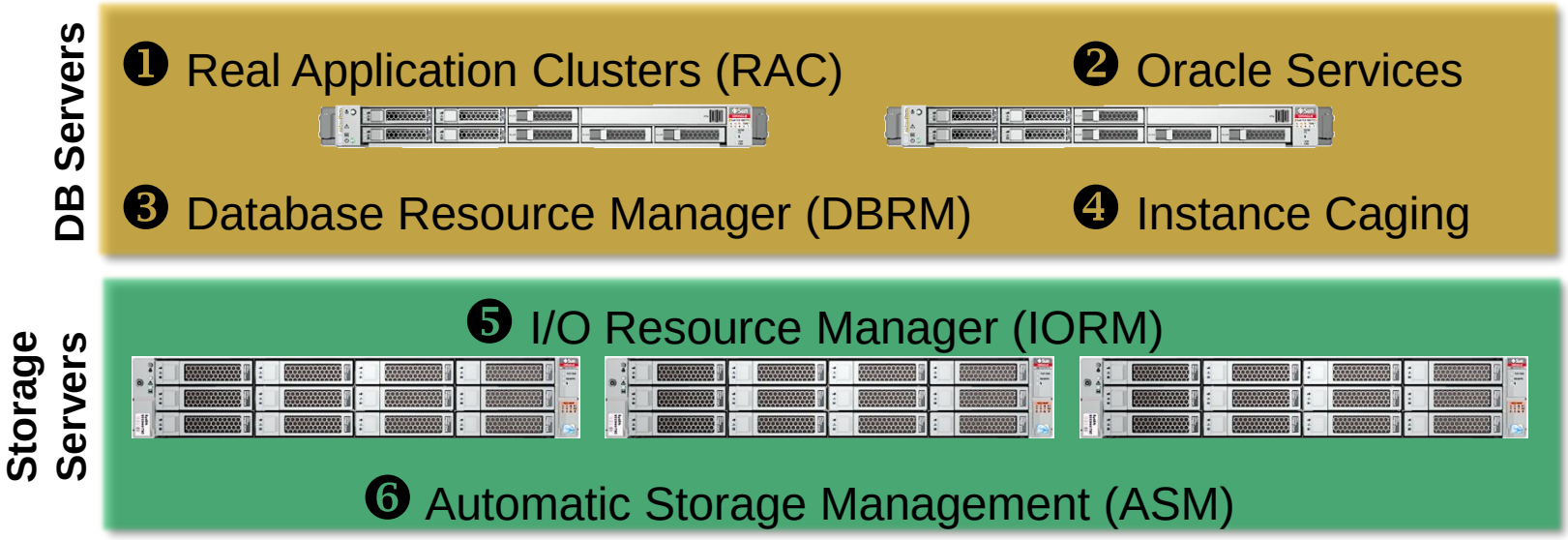
CPU and I/O Sharing

RAC distributes applications across multiple database servers

Oracle Services enables workload placement per application, user group, or workload type

DBRM controls how CPU resources are shared among multiple applications

Instance Caging limits an Oracle instance to a maximum # of CPUs



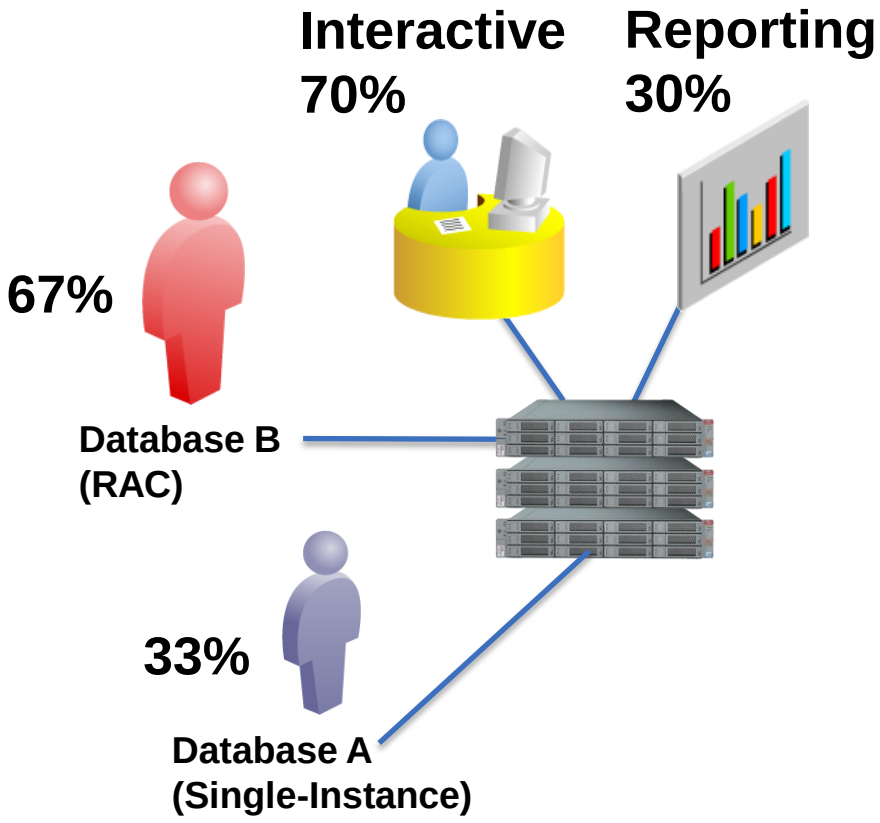
IORM prioritizes I/O requests based on inter- and intra-database priority plans

ASM stripes and mirrors databases across Exadata storage

I/O Resource Management

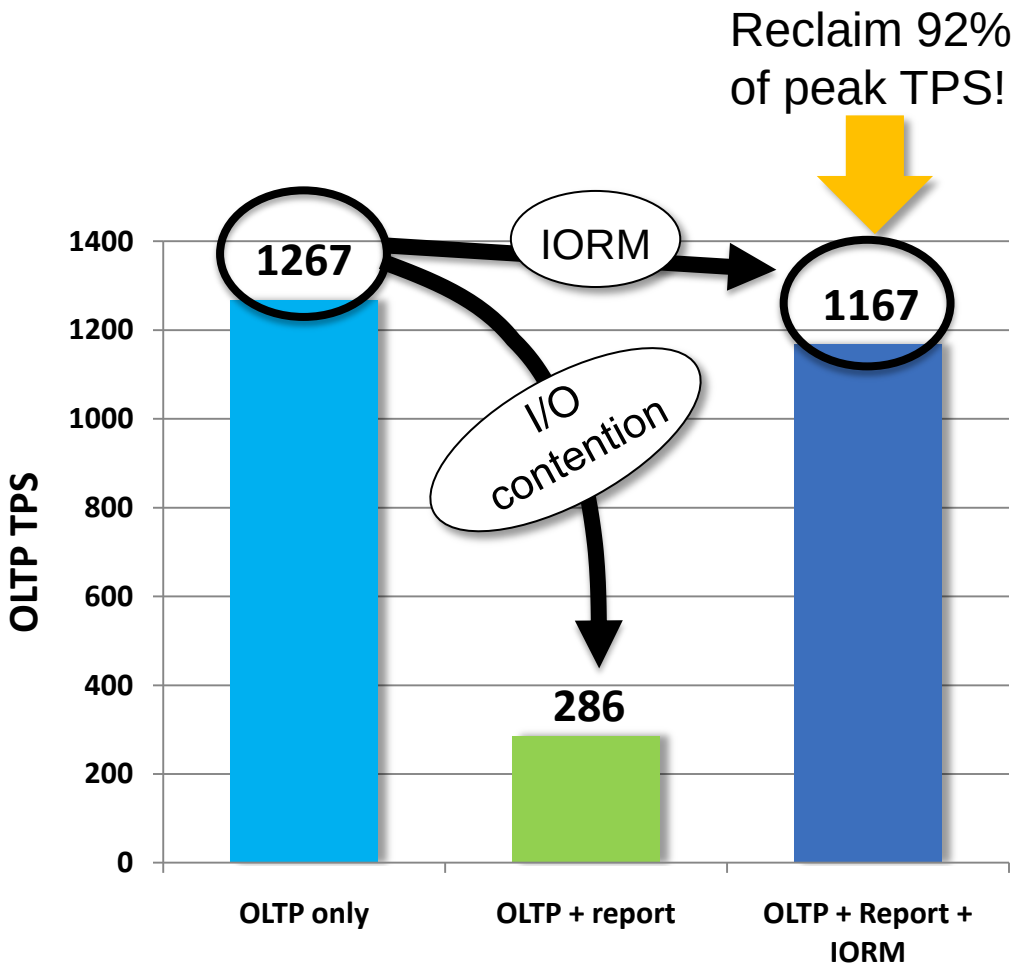
Mixed Workload Environments

- Example Intra-database Plan:
  - Interactive Txns: 70% of I/O bandwidth
  - Reporting Txns: 30% of I/O bandwidth
- Example Inter-database Plan:
  - Database A: 33% of I/O bandwidth
  - Database B: 67% of I/O bandwidth



Controlling a Mixed Workload

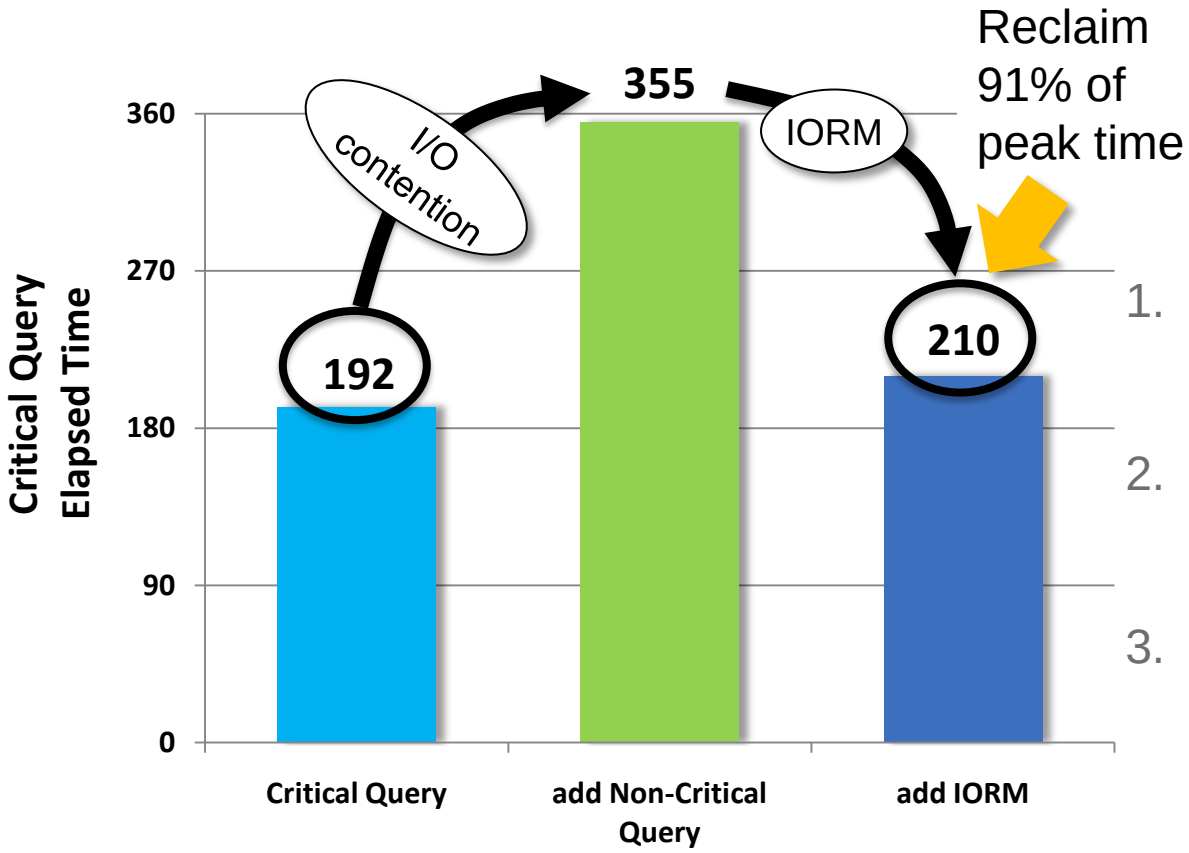
OLTP + Report



1. OLTP app gets 1,267 tps
2. Add Report, tps drops to 286
3. Add IORM, OLTP tps goes back to 1,167

Controlling a Mixed Workload

DSS + DSS



1. Query finishes fast in 192 seconds
2. Add 2nd query, 1st query jumps to 355 seconds
3. Add IORM, 1st query goes back to 210 seconds

Packaging

Sun Oracle Database Machine

Component	Quarter Rack	Half Rack	Full Rack	2-8 Full Racks
Database Servers	2	4	8	16-64
Exadata Storage Servers	3	7	14	28-112
Total Disk Capacity (SAS)	21 TB	50 TB	100 TB	200 – 800 TB
User Data (SAS)	6 TB	14 TB	28 TB	56 – 224 TB
Total Disk Capacity (SATA)	72 TB	168 TB	336 TB	672 – 2,688 TB
User Data (SATA)	21 TB	50 TB	100 TB	200 – 800 TB
I/O Throughput (disks)	4.5 GB/sec	10.5 GB/sec	21 GB/sec	42 - 168 GB/sec
I/O Throughput (flash)	11 GB/sec	25 GB/sec	50 GB/sec	100 - 400 GB/sec
I/O per Second (IOPS)	225,000	500,000	1,000,000	2M – 8M
Racks	1	1	1	2-8

Today's Database Trends

- ✓ Big data warehouses
  - Multi-terabyte to petabyte
  - Response time in seconds-minutes
- ✓ Consolidation
  - Dozens-hundreds of databases
  - Mixed workloads
- ✓ Appliances
  - Pre-configured
  - Balanced performance

Infiniband: A Look Ahead

- The fabric of choice for the foreseeable future
- Future includes…
  - 100 Gb/sec IB
  - PCIe 3.0
  - Incredible x86 Intel/AMD processors
  - Higher performing memory and flash
  - All pushing IB performance/capabilities
- Enterprise requirements
  - Cross-version interoperability
    - Rolling upgrades / heterogeneous clusters
- Keep up the good work!

Now, do you feel lucky...

???


ORACLE®