

Oracle RAC over Mellanox InfiniBand

Most Efficient Solution for Scalable Database

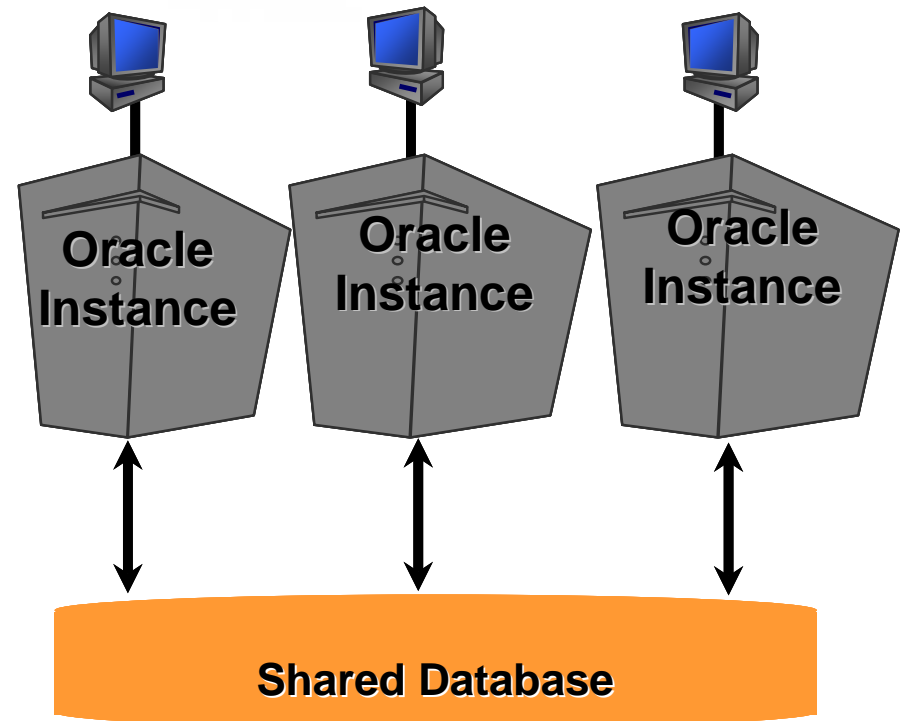


Value Proposition - Oracle Database RAC

→ Oracle Database Real Application Clusters (RAC) provides the ability to build an application platform from multiple systems clustered together

→ Benefits

- Performance
 - Increase performance of a RAC database by adding additional servers to the cluster
- Fault Tolerance
 - A RAC database is constructed from multiple instances. Loss of an instance does not bring down the entire database
- Scalability
 - Scale a RAC database by adding instances to the cluster database



Reliable Datagram Sockets (RDS)



→ What is RDS

- A low overhead, low latency, high bandwidth, ultra reliable, supportable, Inter-Process Communication (IPC) protocol and transport system
- Matches Oracle's existing IPC models for RAC communication
 - Optimized for transfers from 200Bytes to 8MBytes
- Based on Socket API

→ Leverage InfiniBand's built-in high availability and load balance features

- Port failover on the same HCA
- HCA failover on the same system
- Automatic load balancing

→ Open Source on Open Fabric / OFED

- <http://www.openfabrics.org/downloads/OFED/ofed-1.4/OFED-1.4-docs/>

Advantages of RDS over InfiniBand



- Lowering Data Center TCO requires efficient fabrics
 - Oracle RAC 11g will scale for database intensive applications only with the proper high speed protocol and efficient interconnect
- RDS over 10GE
 - 10Gbps not enough to feed multi core Server IO needs
 - Each core may require > 3Gbps
 - Packets can be lost and require retransmit
 - Statistics are not accurate throughput indication
 - Efficiency is much lower than reported
- RDS over InfiniBand
 - The network efficiency is always 100%
 - 40Gbps today
 - Integrated in the Linux kernel
 - More tools will be ported to support RDS, i.e.: netstat, etc.
 - Shows significant real world application performance boost
 - Decision Support System
 - Mixed Batch/OLTP workloads

#1 Price/Performance TPC-H over 11g Benchmark

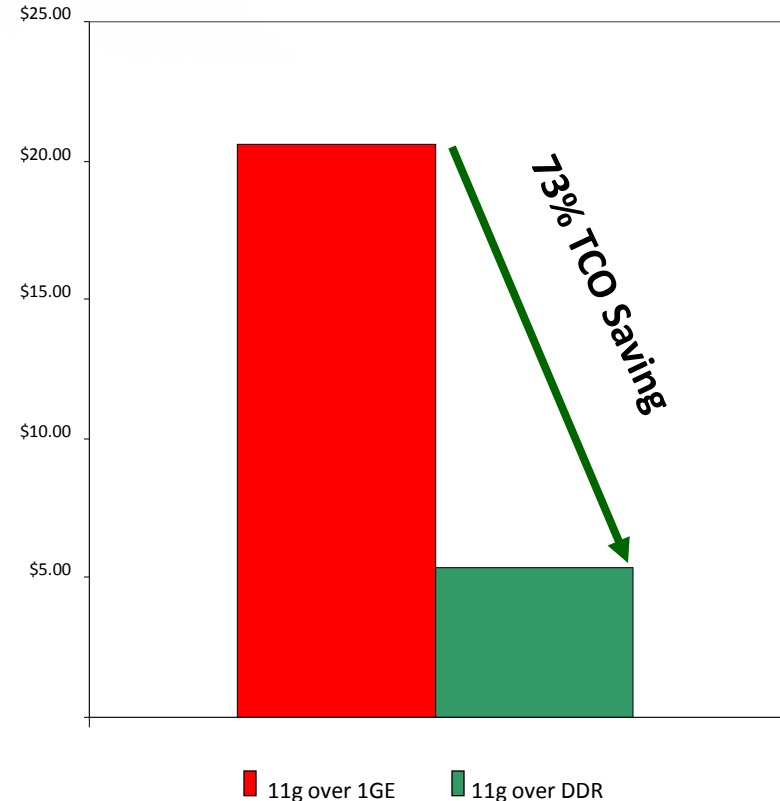


→ 11g over DDR

- Servers: 64 x ProLiant BL460c
 - CPU: 2 x Intel Xeon X5450
 - Quad-Core
- Fabric: Mellanox DDR InfiniBand
- Storage:
 - Native InfiniBand Storage
 - 6 x HP Oracle Exadata

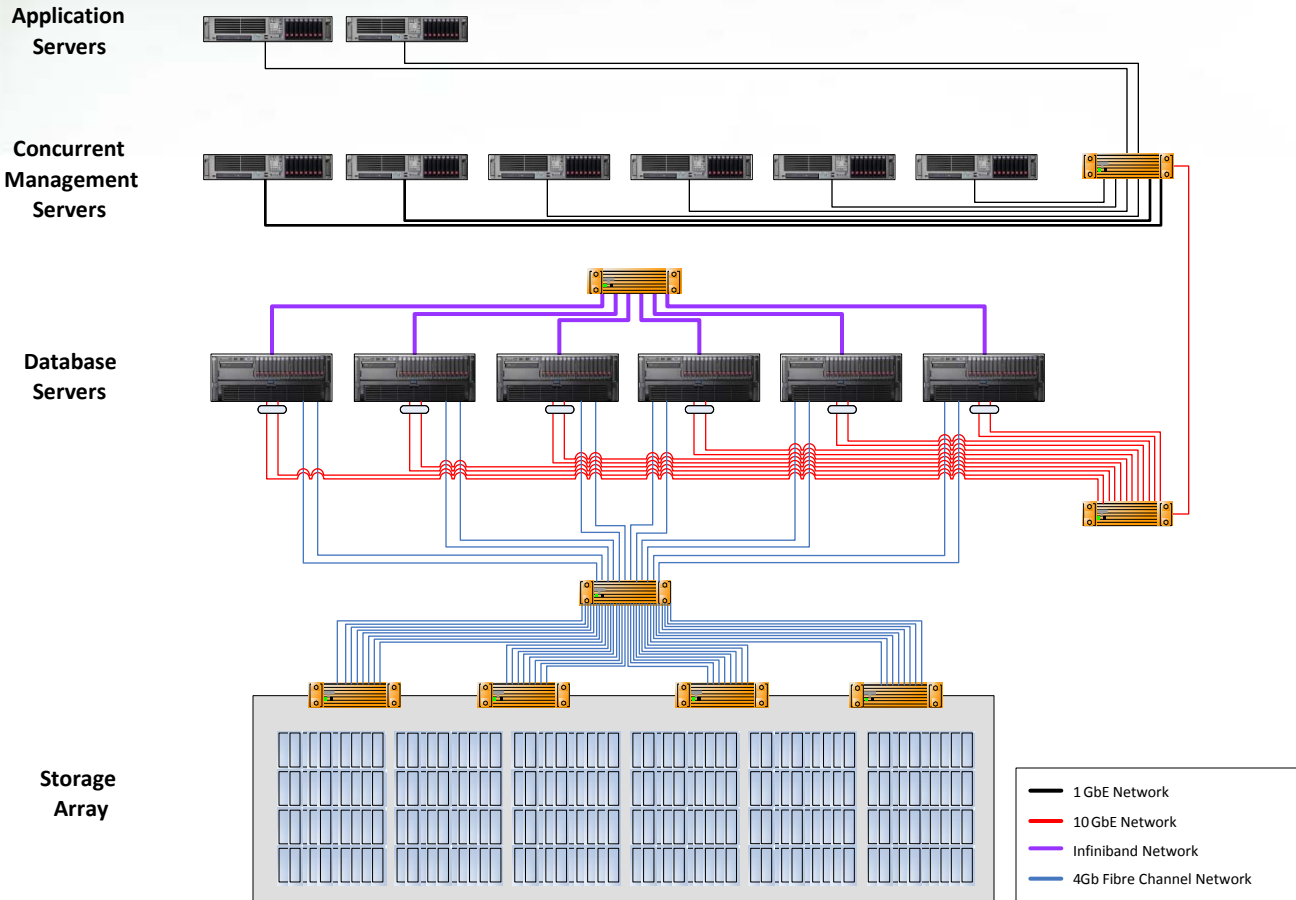
	HP BladeSystem c-Class 128P RAC with HP Oracle Exadata Storage Servers	TPC-H Rev. 2.8.0
		Report Date: June 3, 2009
Total System Cost	Composite Query per Hour Metric	Price / Performance
\$6,320,001USD	1,166,976.6 QphH@1000GB	\$5.42USD \$ / QphH@1000GB

Price / QphH* @1000GB DB



World Record clustered TPC-H Performance and Price/Performance

POC Hardware Configuration



Application Servers

2x HP BL480C
 2 Processors / 8 core X560 3.16GHz
 64GB RAM
 4x 72GB 15K drives
 NIC: HP NC373i 1GB NIC

Concurrent Manager Servers

6x HP BL480C
 2 Processors / 8 core X560 3.16GHz
 64GB RAM
 4x 72GB 15K drives
 NIC: HP NC373i 1GB NIC

Database Servers

6x HP DL580 G5
 4 processors / 24 cores X7460 2.67GHz
 256GB RAM
 8x 72GB 15K drives
 NIC: Intel 10GbE XF SR 2 port PCIe NIC
 Interconnect: Mellanox 4x PCIe Infiniband

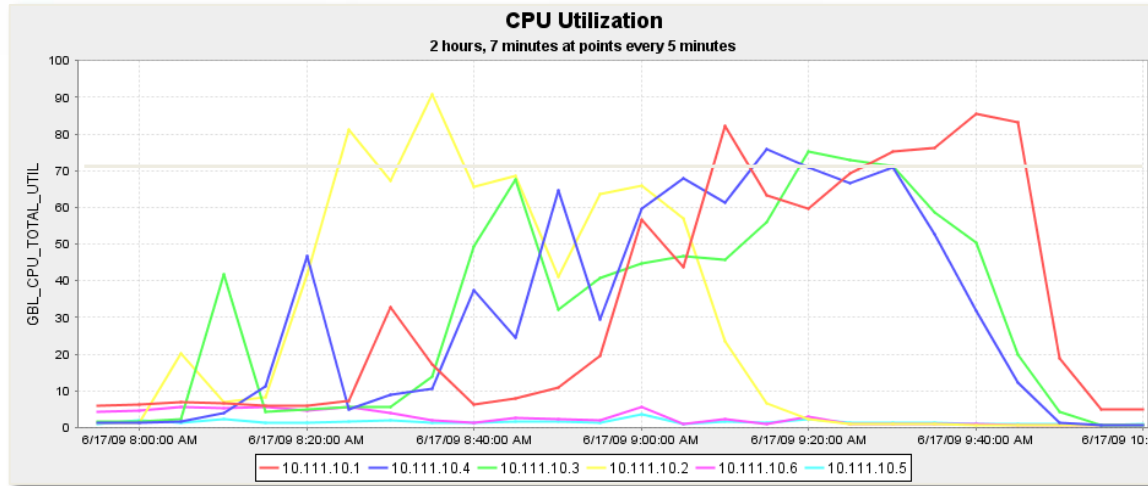
Storage Array

HP XP24000
 64GB cache / 20GB shared memory
 60 Array Groups of 4 spindles
 240 spindles total
 146GB 15K fibre channel disk drives

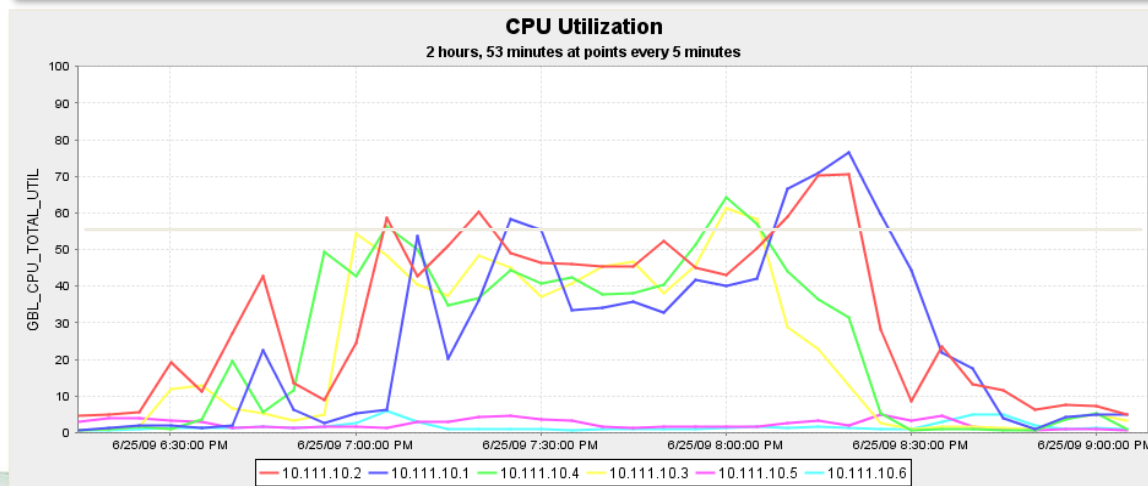
CPU Utilization

- InfiniBand maximize CPU efficiency
 - Enables >20% higher than 10GE

InfiniBand
Interconnect



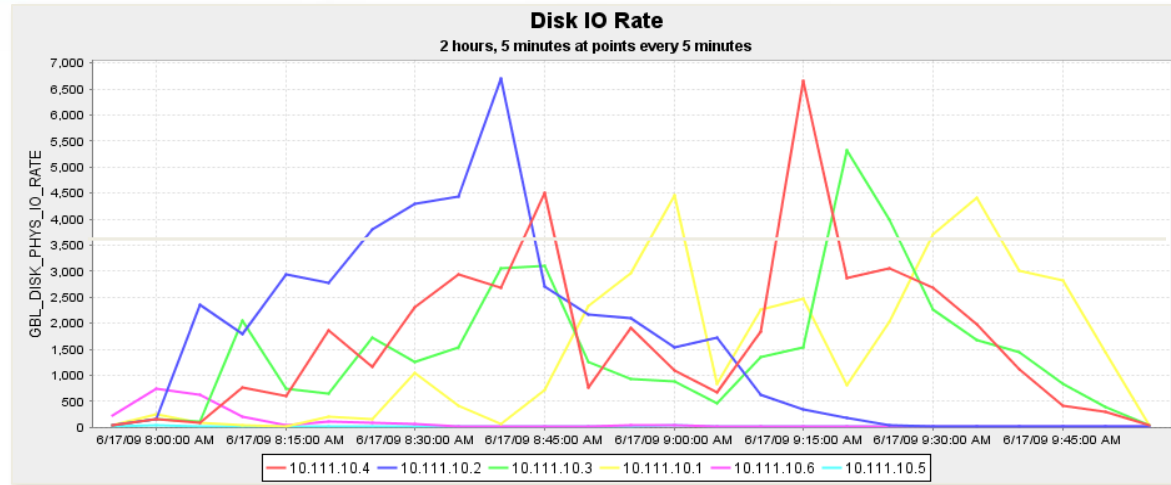
10GigE
Interconnect



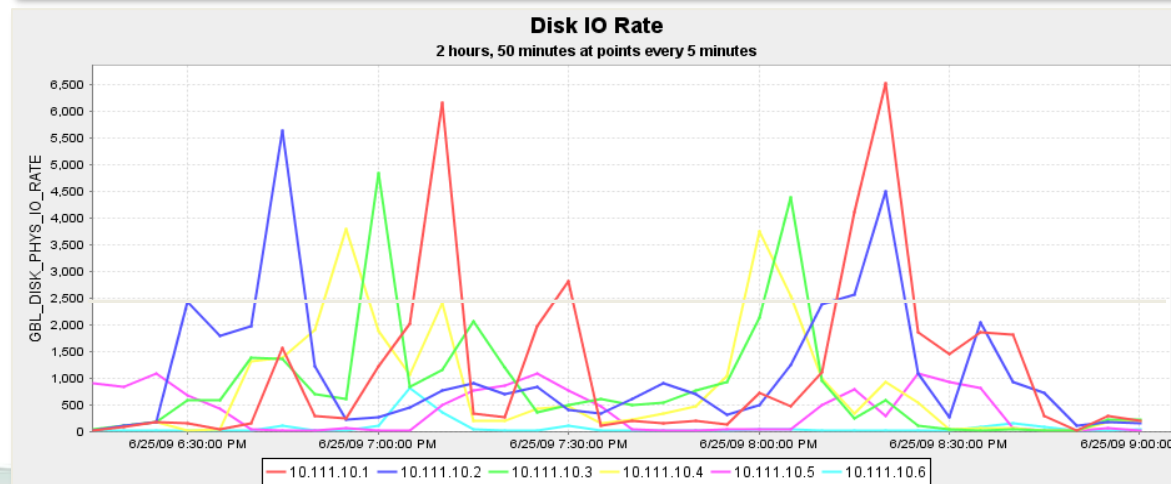
Disk IO Rate

- InfiniBand maximizes Disk utilization
 - Delivers 46% higher IO traffic than 10GE

InfiniBand
Interconnect



10GigE
Interconnect



InfiniBand deliver 63% more TPS vs. 10GE



- TPS Rates for invoice load use case

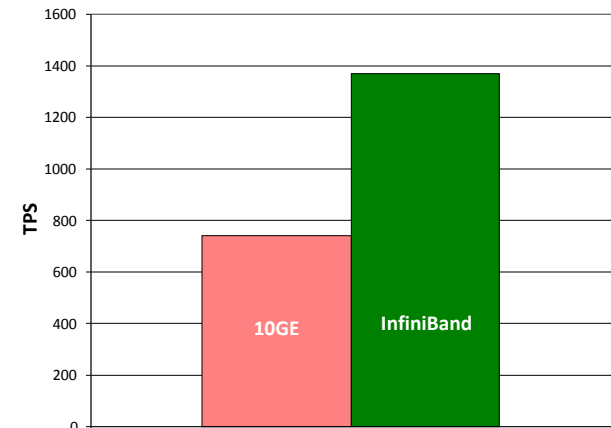
	Activity	Start Time	End Time	Duration	Records	TPS
InfiniBand Interconnect						
1	Invoice Load - Load File	6/17/09 7:48	6/17/09 7:54	0:06:01	9,899,635	27,422.81
2	Invoice Load - Auto Invoice	6/17/09 8:00	6/17/09 9:54	1:54:21	9,899,635	1,442.89
3	Invoice Load - Total	N/A	N/A	2:00:22	9,899,635	1,370.76
10 GigE interconnect						
1	Invoice Load - Load File	6/25/09 17:15	6/25/09 17:20	0:05:21	7,196,171	22,417.98
2	Invoice Load - Auto Invoice	6/25/09 18:22	6/25/09 20:39	2:17:05	7,196,171	874.91
3	Invoice Load - Total	N/A	N/A	2:22:26	7,196,171	842.05

- Work Load

- Nodes 1 through 4: Batch processing
- Node 5: Extra Node not used
- Node 6: EBS Other Activity

- Database size (2 TB)

- ASM
- 5 LUNS @ 400 GB



InfiniBand needs only 4 servers vs. 10 Servers needed by 10GE

TCO Analysis*: \$2.6M Saving



→ Saving over the amortization period

– **\$2,676,278+**

→ Higher Performance at lower cost

→ Data Base Server

– Hardware: \$55,000

– Database 11g: \$95,000 / year

- \$3958 per core per year
- 24 cores per server

– Maintenance / year

- \$1,000

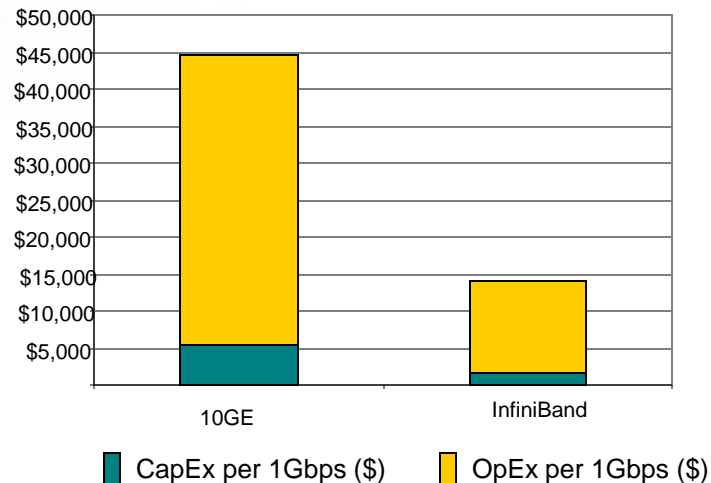
– Power

- 1KWatt/Hour
 - Server: 500 Watt/Hour
 - Cooling: 500 Watt/Hour
 - \$0.2 per KW/Hour

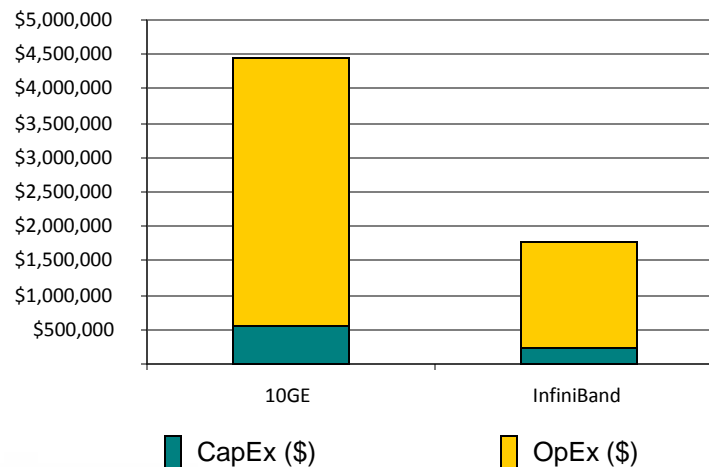
→ Amortization: 4 years

* Running 10 servers for 10GE and 4 Servers for InfiniBand

TCO per 1Gbps over the Amortization Period



TCO over the Amortization Period



Oracle HP Database Appliance



→ Database Machine

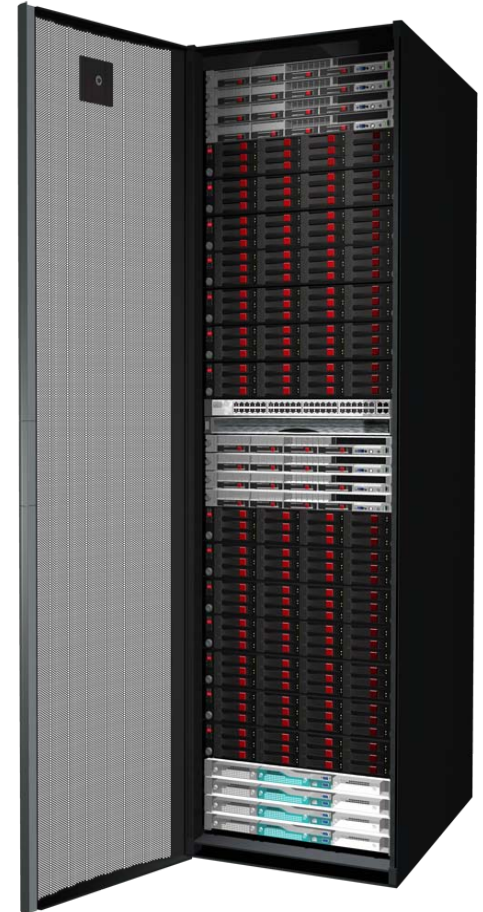
- Standard Rack Servers running Oracle 11g over RDS on EXADATA Storage Servers connected using Mellanox IB DDR products

→ What it solves

- I/O bottleneck between database servers and storage servers
 - Reducing amount of data transferred
 - Increasing size of I/O pipe

→ Result:

- Over 10X Oracle data warehousing query performance over any other solutions



Oracle HP Database Machine

Pre-Configured High Performance Data Warehouse



- 8 DL360 Oracle Database servers
 - 2 quad-core Intel Xeon, 32GB RAM
 - 4 x 146GB Disks
 - Dual-port InfiniBand DDR HCA*
 - Oracle Enterprise Linux & 11g RAC
- 14 DL180 based Exadata Storage Cells
 - 2 quad-core Intel Xeon, 8GB RAM
 - 12 Disks 300GB/1TB SAS/SATA
 - Dual-port InfiniBand DDR HCA*
 - Oracle Enterprise Linux
- 1 Gigabit Ethernet switch
- 4 InfiniBand 24 DDR Ports switches*



* Powered by Mellanox products

What Larry Ellison is Saying



“Exadata shaping up to be our most exciting and successful new product introduction in Oracle’s 30 years history... It can run about 100 times faster in some cases than their standard Oracle environment”

Larry Ellison, Oracle CEO (Q4FY09 earning call)

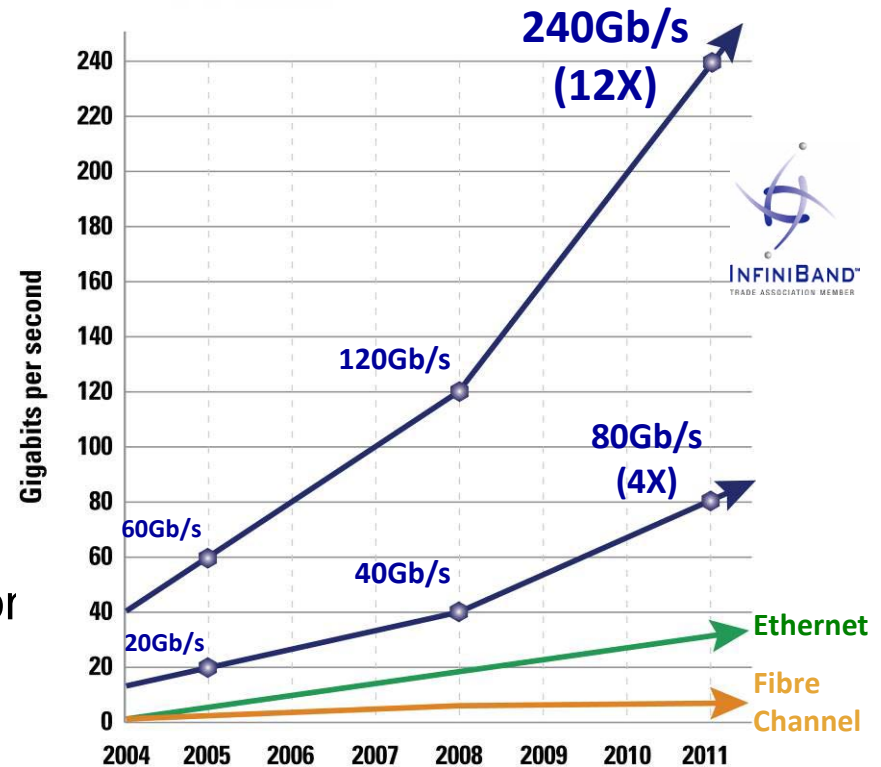


InfiniBand - Technology Leadership



- Industry Standard
 - Hardware, software, cabling, management
 - Design for clustering and storage interconnect
- Price and Performance
 - 40Gb/s node-to-node
 - 120Gb/s switch-to-switch
 - 1us application latency
 - Most aggressive roadmap in the industry
- Reliable with congestion management
- Efficient
 - RDMA and Transport Offload
 - Kernel bypass
 - CPU focuses on application processing
- Scalable for Petascale computing & beyond
- Hardware based QoS
- Virtualization acceleration
- I/O consolidation including storage

The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

Thank You

