Update: InfiniBand for Oracle RAC Clusters

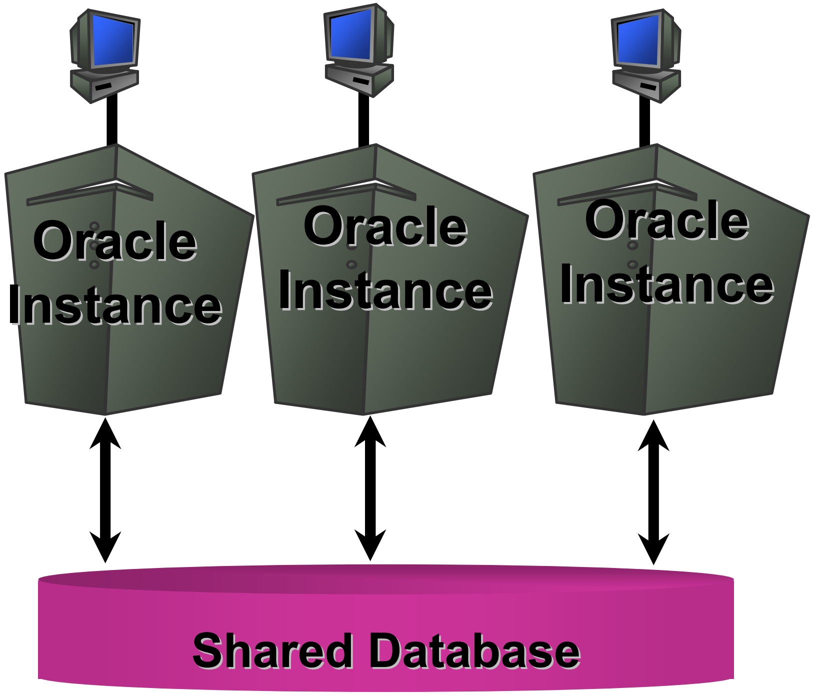
Paul Tsien, Oracle

Agenda

- Oracle RAC Overview
- Reliable Datagram Sockets
- Server Vendor Validation
- RAC on InfiniBand ~ Customer Experience
- What’s next?

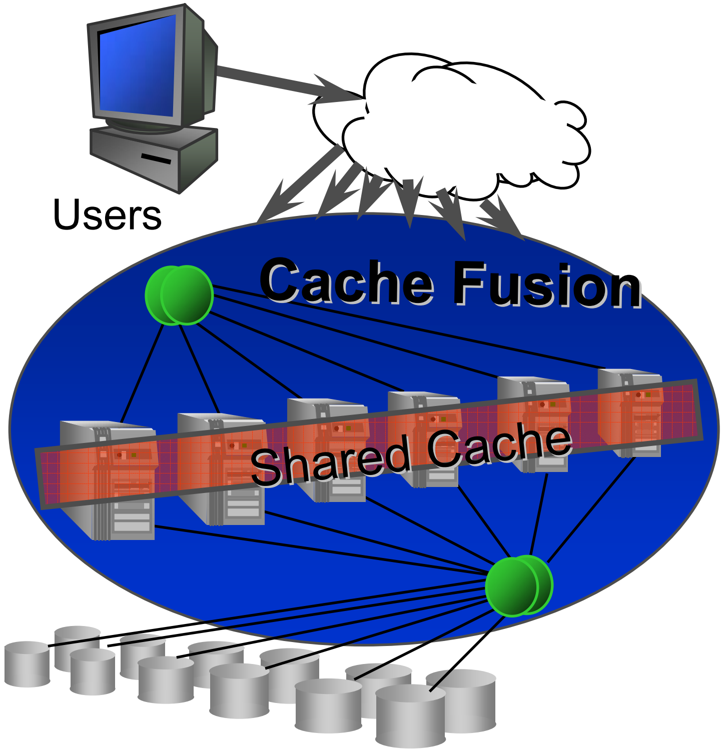
Oracle RAC

Oracle Real Application Clusters



- Oracle Real Application Clusters (RAC) provides the ability to build an application platform from multiple systems that are clustered together
- Allows applications to become
  - Highly scalable
  - Highly available
- Chosen to avoid a single node failure, causing application downtime
  - Eliminates a node as single point of failure

Real Application Clusters

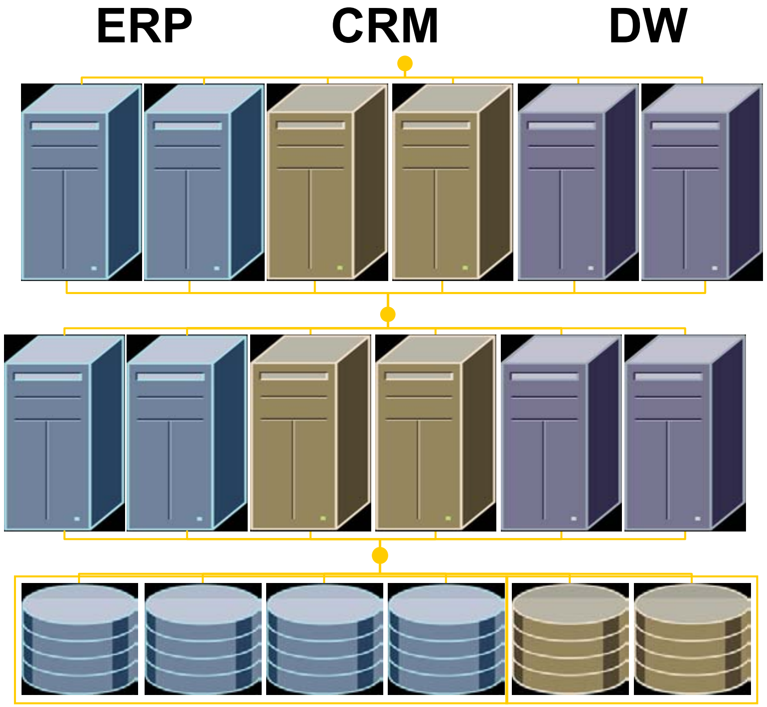


- World’s best Scalability with Cache Fusion
  - Cache-to-cache data shipping
  - Scales off-the-shelf applications with no changes
- World’s best Availability with Fast-Start Fault Recovery
  - Node failure is transparent to applications
  - Recovers from node failure in 17 seconds - workload independent
  - Pre-warmed cache speeds restart
  - Easily add and delete nodes

The Ultimate Parallel Architecture

Real Applications in a Real Grid

- Existing Apps
  - Financials, MFG, HR and CRM
  - Collaboration Suite
  - In house developed
  - DSS
  - ISV Apps
- Easy Migration
- Improve Utilization



Oracle RAC IPC

- RAC IPC
  - Thousands of processes
  - 200K+ associations (not connections)
  - 64 nodes
- Oracle IPC Usage
  - New grid aware applications will significantly increase IPC utilization
    - Approach database I/O rates
    - Very large messages

Reliable Datagram Sockets

Vision Statement

- A low overhead, low latency, high bandwidth, ultra reliable, supportable, IPC protocol and transport system
  - Which matches Oracle's existing IPC models for RAC communication
  - Optimized for transfers from 200 bytes to 8 MB

Goal and Objective

- Support for a reliable datagram IPC
  - Based on Socket API
  - Minimal code change / testing for Oracle
  - Runs over InfiniBand, 10 Gig Ethernet, and iWARP
  - 6 month validation / certification for RAC

Goal and Objective

- Leverage InfiniBand’s built-in availability and load balance features
  - Port failover on the same HCA
  - HCA failover on the same system
  - Automatic load balancing

Reliable Datagram IPC

- UDP – Oracle adds reliable delivery via user mode wire protocol engine
  - Two sockets per process, thousands of messages on wire
  - Slow sends times (windowing,acks,retrans)
  - Holds together but degenerates under CPU load
  - Well tested !

RDS IPC over InfiniBand

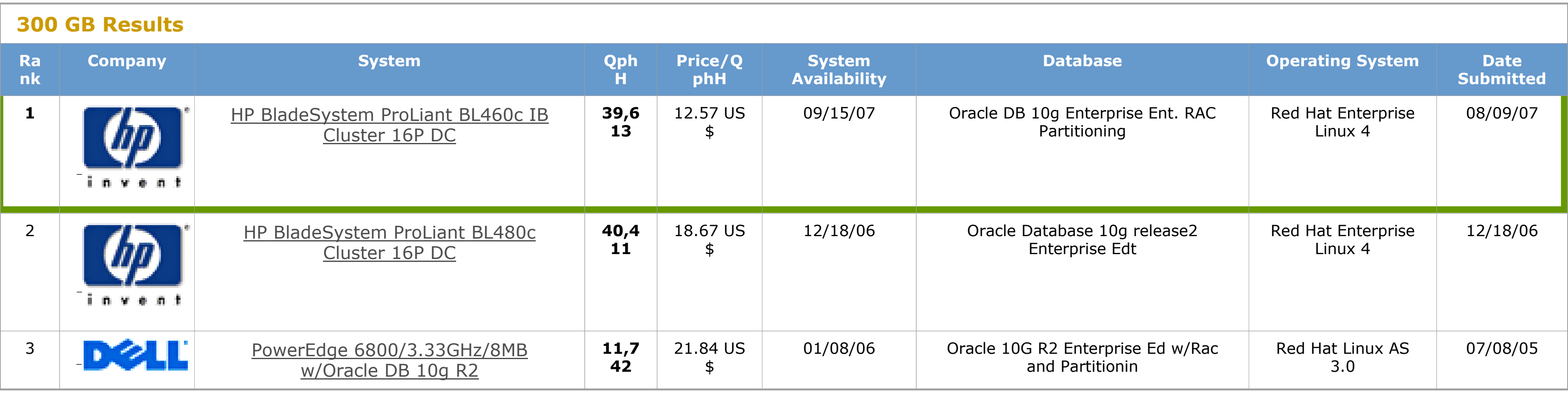
- RD – Reliable Datagram IPC over IB
  - Minimal Oracle code change
  - Stable code and easily passed all Oracle regression tests
  - Supports fail-over across and within HCAs
- Oracle internal interconnect tests show…
  - 50% less CPU than IP over IB, UDP
  - ½ latency of UDP (no user-mode acks)
  - 50% faster cache to cache Oracle block throughput

RDS IPC over IB

- Uses IB reliable connection (RC)
- Node to Node level connection
  - User mode sockets share small pool of node to node RCs
  - Formed either dynamically at send or at system startup

300GB TPC-H – Oracle RAC Clusters

Cluster entries sorted by price/performance…

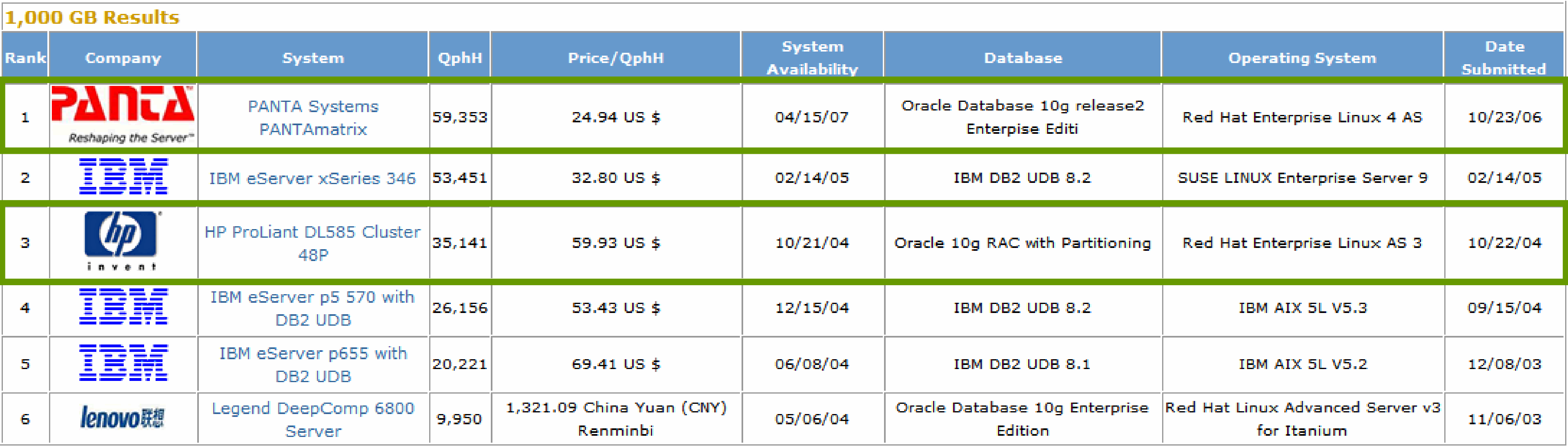


300 GB Results

Rank	Company	System	QphH	Price/QphH	System Availability	Database	Operating System	Date Submitted
1	hp invent	HP BladeSystem ProLiant BL460c IB Cluster 16P DC	**39,613**	12.57 US $	09/15/07	Oracle DB 10g Enterprise Ent. RAC Partitioning	Red Hat Enterprise Linux 4	08/09/07
2	hp invent	HP BladeSystem ProLiant BL480c Cluster 16P DC	**40,411**	18.67 US $	12/18/06	Oracle Database 10g release2 Enterprise Edt	Red Hat Enterprise Linux 4	12/18/06
3	DELL	PowerEdge 6800/3.33GHz/8MB w/Oracle DB 10g R2	**11,742**	21.84 US $	01/08/06	Oracle 10G R2 Enterprise Ed w/Rac and Partitionin	Red Hat Linux AS 3.0	07/08/05

1 TB TPC-H – Oracle RAC Clusters

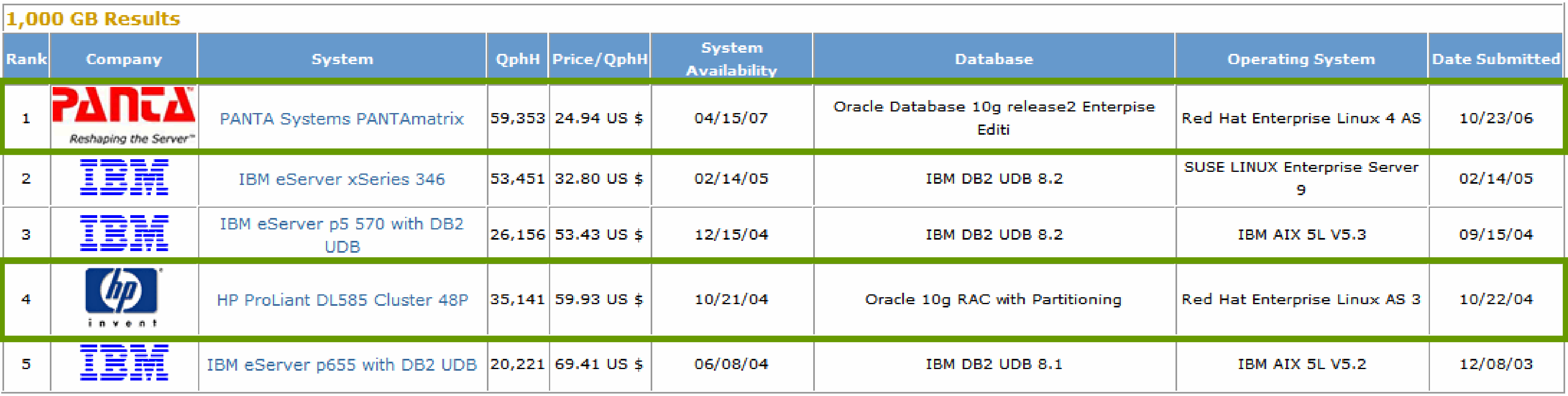
Cluster entries sorted by performance…



1,000 GB Results

Rank	Company	System	QphH	Price/QphH	System Availability	Database	Operating System	Date Submitted
1	PANTA Reshaping the Server™	PANTA Systems PANTAmatrix	59,353	24.94 US $	04/15/07	Oracle Database 10g release2 Enterpise Editi	Red Hat Enterprise Linux 4 AS	10/23/06
2	IBM	IBM eServer xSeries 346	53,451	32.80 US $	02/14/05	IBM DB2 UDB 8.2	SUSE LINUX Enterprise Server 9	02/14/05
3	hp invent	HP ProLiant DL585 Cluster 48P	35,141	59.93 US $	10/21/04	Oracle 10g RAC with Partitioning	Red Hat Enterprise Linux AS 3	10/22/04
4	IBM	IBM eServer p5 570 with DB2 UDB	26,156	53.43 US $	12/15/04	IBM DB2 UDB 8.2	IBM AIX 5L V5.3	09/15/04
5	IBM	IBM eServer p655 with DB2 UDB	20,221	69.41 US $	06/08/04	IBM DB2 UDB 8.1	IBM AIX 5L V5.2	12/08/03
6	lenovo联想	Legend DeepComp 6800 Server	9,950	1,321.09 China Yuan (CNY) Renminbi	05/06/04	Oracle Database 10g Enterprise Edition	Red Hat Linux Advanced Server v3 for Itanium	11/06/03

Cluster entries sorted by price/performance…



1,000 GB Results

Rank	Company	System	QphH	Price/QphH	System Availability	Database	Operating System	Date Submitted
1	PANTA Reshaping the Server™	PANTA Systems PANTAmatrix	59,353	24.94 US $	04/15/07	Oracle Database 10g release2 Enterpise Editi	Red Hat Enterprise Linux 4 AS	10/23/06
2	IBM	IBM eServer xSeries 346	53,451	32.80 US $	02/14/05	IBM DB2 UDB 8.2	SUSE LINUX Enterprise Server 9	02/14/05
3	IBM	IBM eServer p5 570 with DB2 UDB	26,156	53.43 US $	12/15/04	IBM DB2 UDB 8.2	IBM AIX 5L V5.3	09/15/04
4	hp invent	HP ProLiant DL585 Cluster 48P	35,141	59.93 US $	10/21/04	Oracle 10g RAC with Partitioning	Red Hat Enterprise Linux AS 3	10/22/04
5	IBM	IBM eServer p655 with DB2 UDB	20,221	69.41 US $	06/08/04	IBM DB2 UDB 8.1	IBM AIX 5L V5.2	12/08/03

Server Vendor Validation of RAC on InfiniBand

HP ORACLE CTC
http://www.hporaclectc.com/

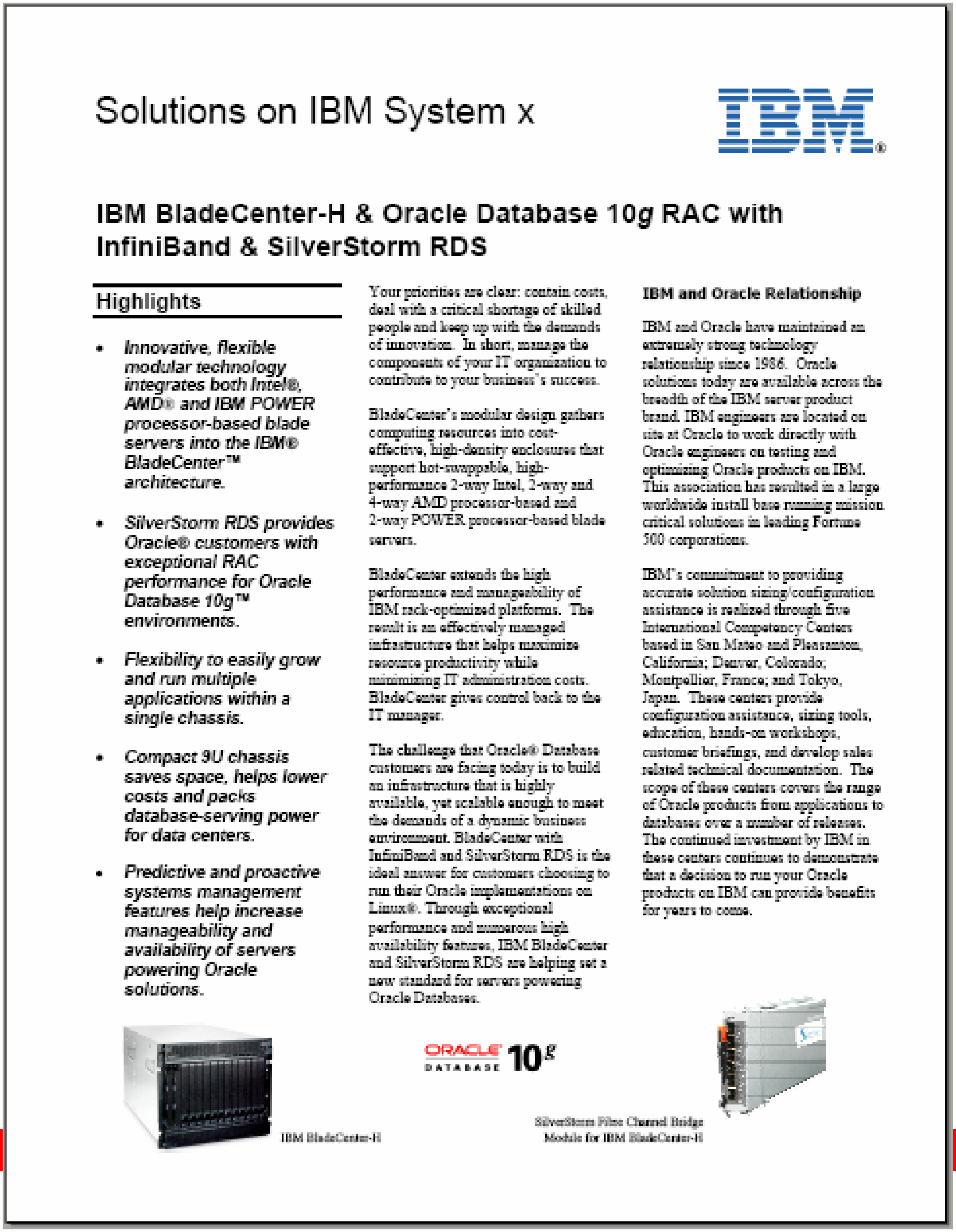
Oracle RAC10g on Linux With InfiniBand

Yann Allandit
Christine Perrin
Date: 22-August-06

System x

&

BladeCenter



Solutions on IBM System x

IBM BladeCenter-H & Oracle Database 10*g* RAC with InfiniBand & SilverStorm RDS

Highlights

- *Innovative, flexible modular technology integrates both Intel®, AMD® and IBM POWER processor-based blade servers into the IBM® BladeCenter™ architecture.*
- *SilverStorm RDS provides Oracle® customers with exceptional RAC performance for Oracle Database 10g™ environments.*
- *Flexibility to easily grow and run multiple applications within a single chassis.*
- *Compact 9U chassis saves space, helps lower costs and packs database-serving power for data centers.*
- *Predictive and proactive systems management features help increase manageability and availability of servers powering Oracle solutions.*

Your priorities are clear: contain costs, deal with a critical shortage of skilled people and keep up with the demands of innovation. In short, manage the components of your IT organization to contribute to your business's success.

BladeCenter's modular design gathers computing resources into cost-effective, high-density enclosures that support hot-swappable, high-performance 2-way Intel, 2-way and 4-way AMD processor-based and 2-way POWER processor-based blade servers.

BladeCenter extends the high performance and manageability of IBM rack-optimized platforms. The result is an effectively managed infrastructure that helps maximize resource productivity while minimizing IT administration costs. BladeCenter gives control back to the IT manager.

The challenge that Oracle® Database customers are facing today is to build an infrastructure that is highly available, yet scalable enough to meet the demands of a dynamic business environment. BladeCenter with InfiniBand and SilverStorm RDS is the ideal answer for customers choosing to run their Oracle implementations on Linux®. Through exceptional performance and numerous high availability features, IBM BladeCenter and SilverStorm RDS are helping set a new standard for servers powering Oracle Databases.

IBM and Oracle Relationship

IBM and Oracle have maintained an extremely strong technology relationship since 1986. Oracle solutions today are available across the breadth of the IBM server product brand. IBM engineers are located on site at Oracle to work directly with Oracle engineers on testing and optimizing Oracle products on IBM. This association has resulted in a large worldwide install base running mission critical solutions in leading Fortune 500 corporations.

IBM's commitment to providing accurate solution sizing/configuration assistance is realized through five International Competency Centers based in San Mateo and Pleasanton, California; Denver, Colorado; Montpellier, France; and Tokyo, Japan. These centers provide configuration assistance, sizing tools, education, hands-on workshops, customer briefings, and develop sales related technical documentation. The scope of these centers covers the range of Oracle products from applications to databases over a number of releases. The continued investment by IBM in these centers continues to demonstrate that a decision to run your Oracle products on IBM can provide benefits for years to come.

IBM BladeCenter-H

SilverStorm Fibre Channel Bridge Module for IBM BladeCenter-H

DELL POWER SOLUTIONS - MAY 2007

POWER SOLUTIONS

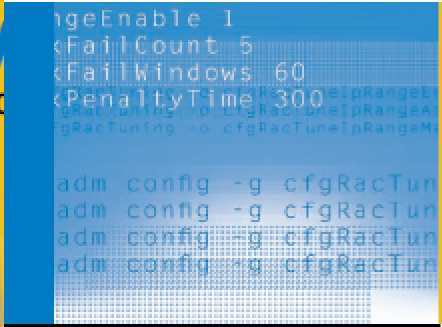
TRANSFORMING YOUR IT FRAMEWORK INTO A SCALABLE ENTERPRISE

MAY 2007 • $4.95

Access Everywhere
Business Any

How Microsoft Exchange Server 20

Access Everywhere. Business Anywhere. How Microsoft Exchange Server 2007 Makes It Hap



Using Reliable Datagram Sockets Over InfiniBand for Oracle Database 10*g* Clusters

BY ZAFAR MAHMOOD
ANTHONY FERNANDEZ
GUNNAR K. GUNNARSSON

Reliable Datagram Sockets (RDS) Over InfiniBand can provide a horizontally scalable, high-performance alternative to traditional vertical scaling for enterprises using Oracle® Database 10*g* and Oracle Real Application Clusters (RAC). This article discusses the advantages of using RDS Over InfiniBand to build scalable, high-performance Oracle RAC clusters with cost-effective, industry-standard Dell™ and QLogic components.

Related Categories:

Clustering

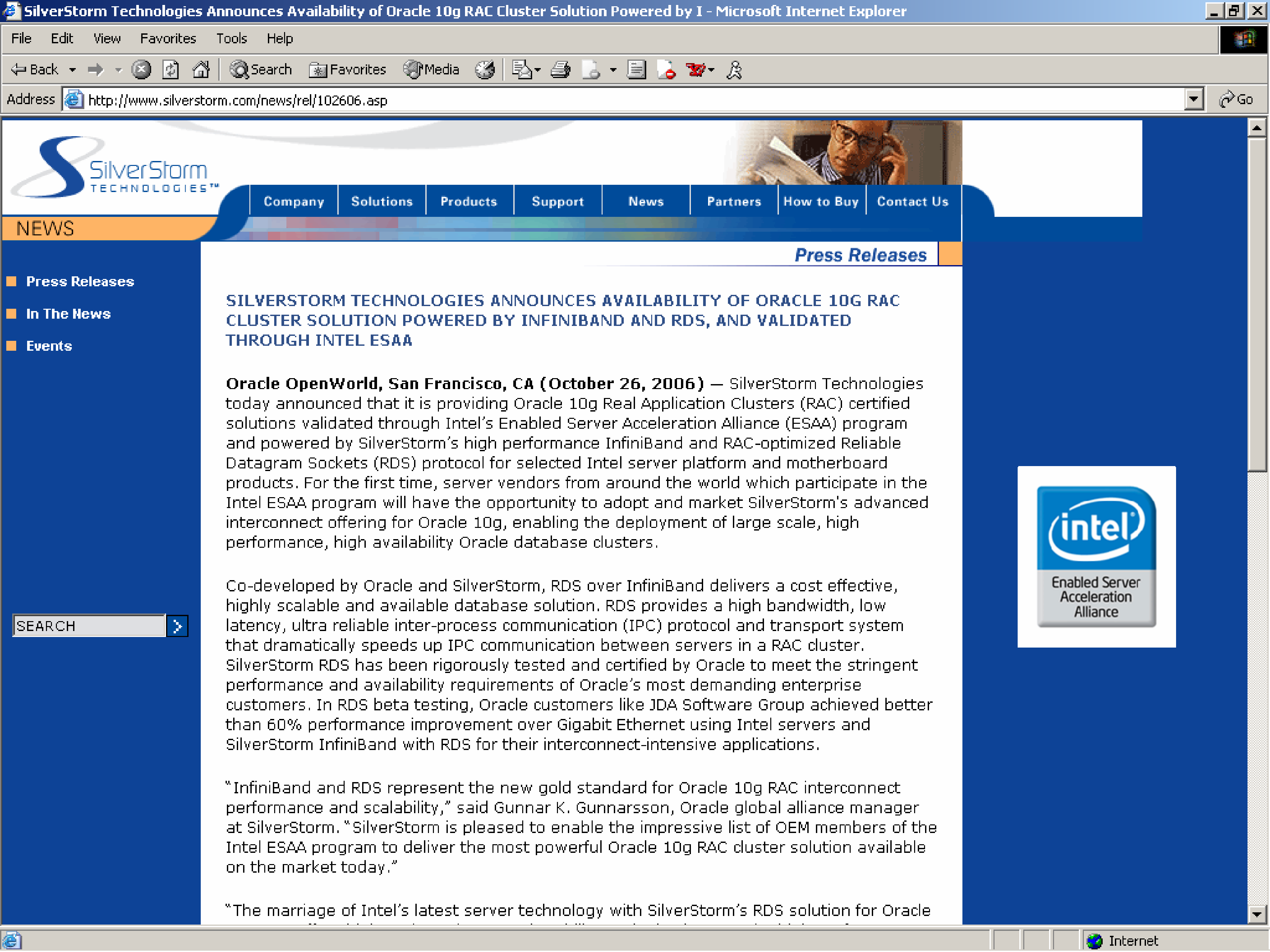
Database

Dell/EMC storage

InfiniBand

In the past, large database systems were often synonymous with costly mainframes. Today, however, grid computing with commodity servers can provide many advantages for large databases, including cost-effectiveness, scalability, high performance and availability, network consolidation, and simple installation and management. Two key technologies enabling this type of system for large databases are Oracle Real Application Clusters (RAC) and industry-standard grid components with efficient interconnects that provide high throughput and low latency for data traffic between components. Oracle Database 10*g* and Oracle

Although support problems and the lack of a standard protocol has historically made implementing InfiniBand for clusters of this size a challenge, Oracle Database 10*g* Release 2 (R2) and the 10.2.0.3 patch set support a cluster interconnect protocol developed by Oracle and QLogic specifically for Oracle RAC called Reliable Datagram Sockets (RDS), which is agnostic to underlying Remote Direct Memory Access (RDMA)–capable devices and simplifies implementation. This protocol can work over either an RDMA-capable Ethernet network interface card (NIC) or an InfiniBand host channel adapter (HCA).



SilverStorm Technologies Announces Availability of Oracle 10g RAC Cluster Solution Powered by I - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.silverstorm.com/news/rel/102606.asp Go
SilverStorm TECHNOLOGIES™
Company Solutions Products Support News Partners How to Buy Contact Us
NEWS
Press Releases
In The News
Events
SEARCH
intel Enabled Server Acceleration Alliance
Internet

Press Releases

SILVERSTORM TECHNOLOGIES ANNOUNCES AVAILABILITY OF ORACLE 10G RAC CLUSTER SOLUTION POWERED BY INFINIBAND AND RDS, AND VALIDATED THROUGH INTEL ESAA

Oracle OpenWorld, San Francisco, CA (October 26, 2006) — SilverStorm Technologies today announced that it is providing Oracle 10g Real Application Clusters (RAC) certified solutions validated through Intel's Enabled Server Acceleration Alliance (ESAA) program and powered by SilverStorm's high performance InfiniBand and RAC-optimized Reliable Datagram Sockets (RDS) protocol for selected Intel server platform and motherboard products. For the first time, server vendors from around the world which participate in the Intel ESAA program will have the opportunity to adopt and market SilverStorm's advanced interconnect offering for Oracle 10g, enabling the deployment of large scale, high performance, high availability Oracle database clusters.

Co-developed by Oracle and SilverStorm, RDS over InfiniBand delivers a cost effective, highly scalable and available database solution. RDS provides a high bandwidth, low latency, ultra reliable inter-process communication (IPC) protocol and transport system that dramatically speeds up IPC communication between servers in a RAC cluster. SilverStorm RDS has been rigorously tested and certified by Oracle to meet the stringent performance and availability requirements of Oracle's most demanding enterprise customers. In RDS beta testing, Oracle customers like JDA Software Group achieved better than 60% performance improvement over Gigabit Ethernet using Intel servers and SilverStorm InfiniBand with RDS for their interconnect-intensive applications.

"InfiniBand and RDS represent the new gold standard for Oracle 10g RAC interconnect performance and scalability," said Gunnar K. Gunnarsson, Oracle global alliance manager at SilverStorm. "SilverStorm is pleased to enable the impressive list of OEM members of the Intel ESAA program to deliver the most powerful Oracle 10g RAC cluster solution available on the market today."

"The marriage of Intel's latest server technology with SilverStorm's RDS solution for Oracle

Customer Experience

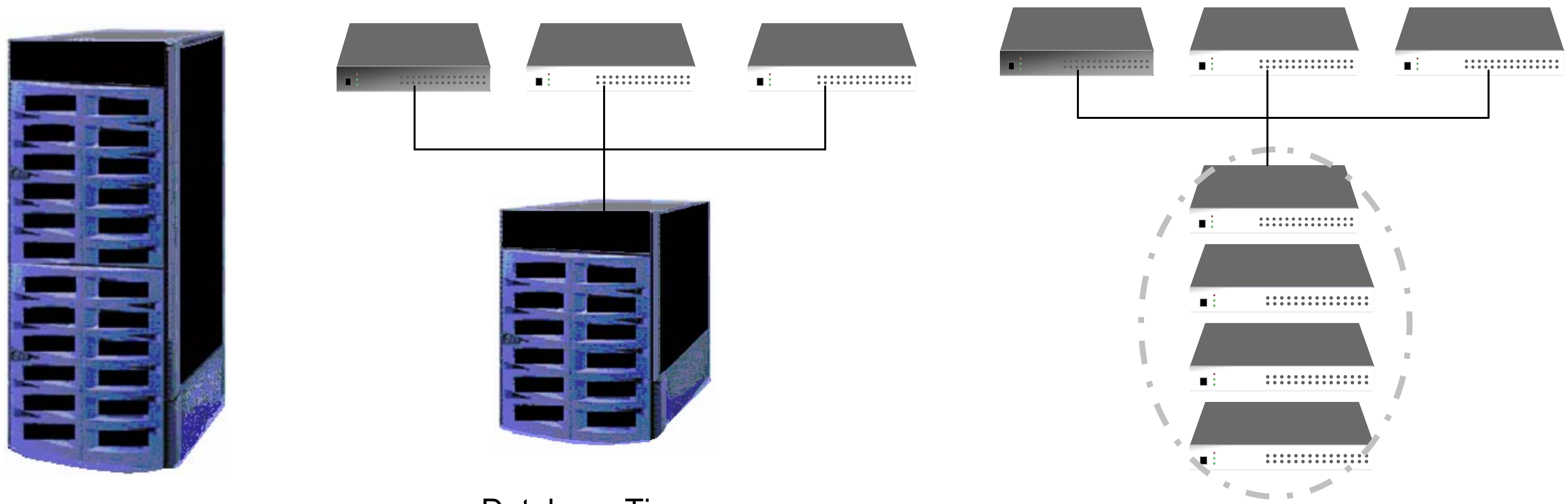
Customer Requirements

- Improve application performance (throughput and latency)
- Maintain data availability
- Lower TCO through commodity hardware and improve performance/scalability
- Want to implement Grid and Utility computing

Results

- RDS/IB shows significant real world application performance gains for certain workloads: DSS and mixed Batch/OLTP workloads
  - Throughput and latency
- Customers are interested in unified fabric for cost and manageability reasons
  - Reservation/QoS

JDA Software ~ Shifting Deployment Paradigm


Application Tier on Commodity servers
Application Tier on Commodity servers
Application and Database on Same SMP Server
Database Tier
Database Tier on Commodity Servers

Monolithic SMP

- Application
- Database

Mixed Configuration

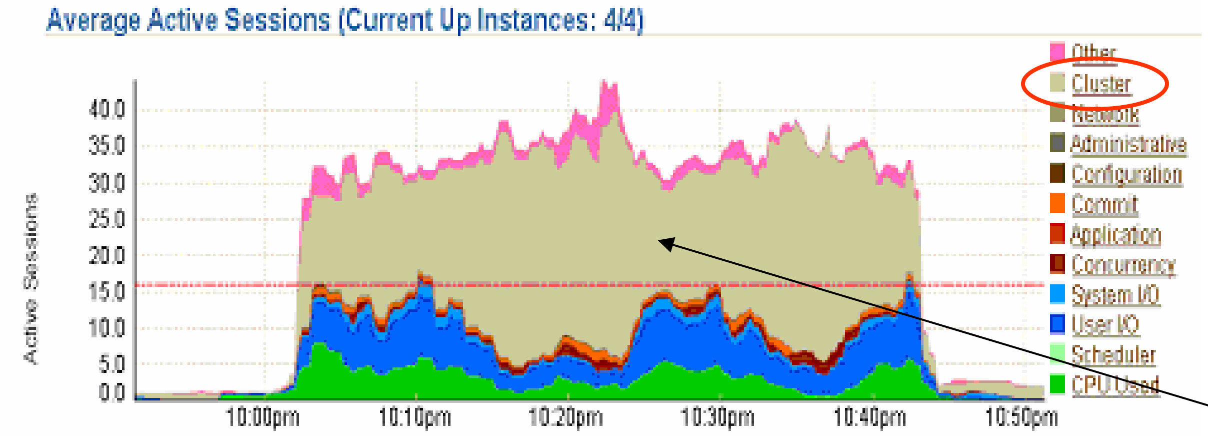
- Commodity Application Servers
- SMP Database Servers

Grid Computing

- All Commodity Servers

Past → Present → Future

JDA Software ~ RDS Performance



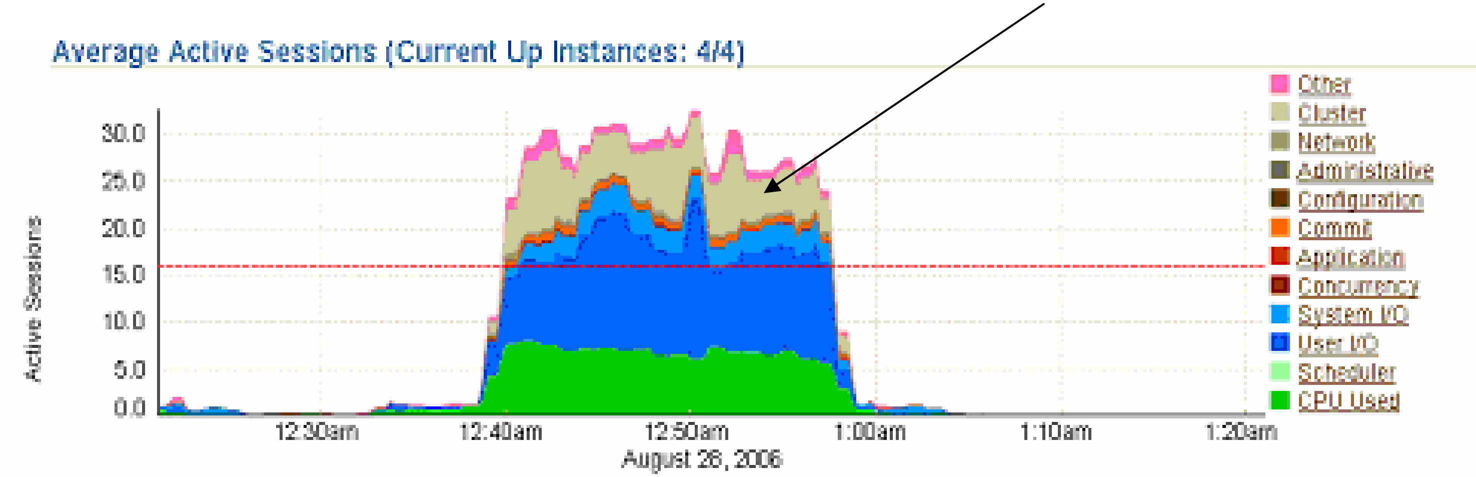
66 minutes

IB/RDS is the highest performance interconnect available for Oracle 10*g* RAC

- **< ½ latency of GE or IPoIB**
- **4X+ bandwidth of GE**
- **< ½ CPU utilization of GE or IPoIB**

Customer feedback…

Significant RAC cluster interconnect contention eliminated with IB/RDS



25 minutes

OEM data courtesy of
JDA Software Group

Oracle RAC InfiniBand Deployments

- Oracle Redwood Shores (multiple)
- Oracle Atlanta & Reston ETCs
- Oracle Austin Data Center
- Key partners, integrators & enterprise customers

Oracle ETC Reston

JDA® REAL DEMAND CHAIN RESULTS.

KELLY SERVICES
Celebrating 60 Years of Excellence 1946-2006

Oracle ETC Atlanta

AcuityBrands. Lighting
LITHONIA LIGHTING®

UF UNIVERSITY of FLORIDA
The Foundation for The Gator Nation

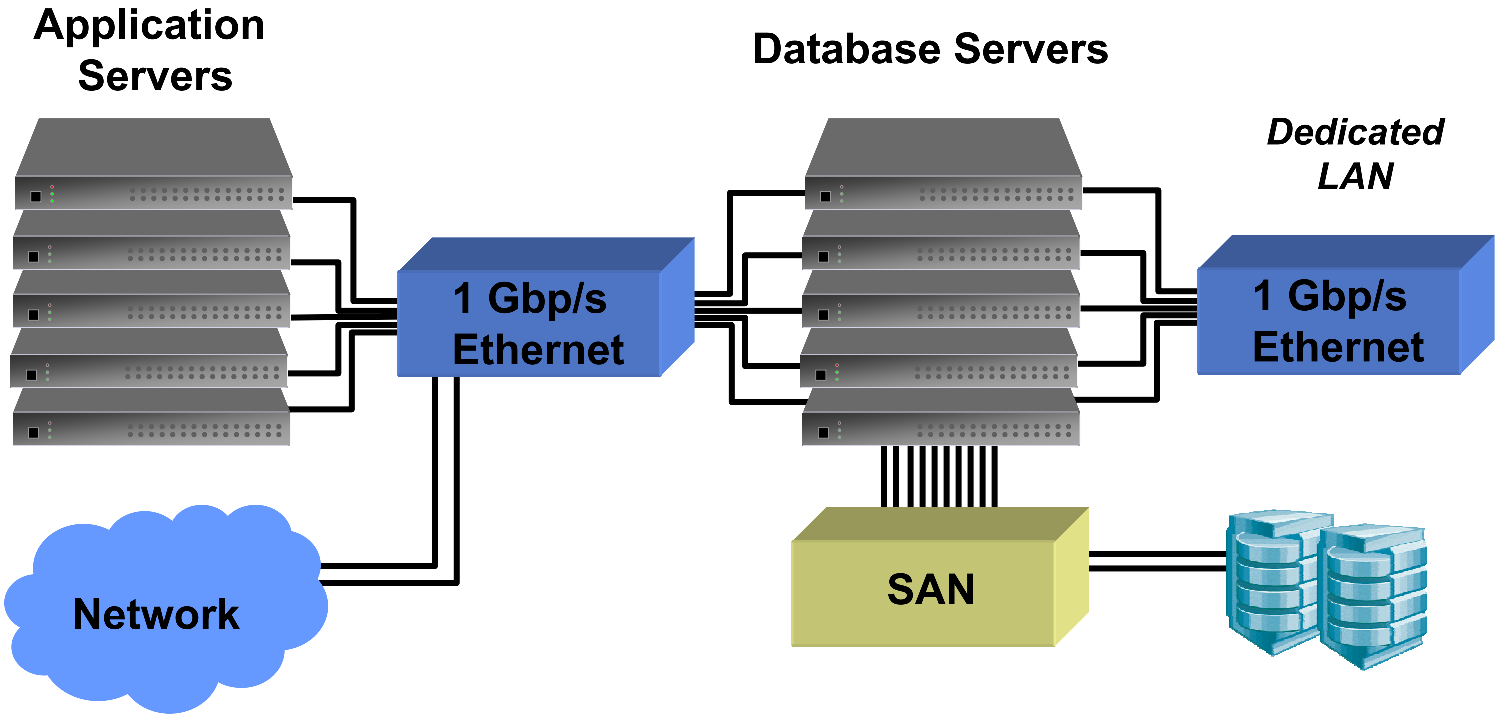
RDS Status

- Oracle 10*g* Release 2 supports SilverStorm/QLogic RDS
  - Excellent performance and stability
  - Compatible with Oracle 11*g*
- Open Source RDS
  - Oracle and partners are testing/certifying Open Source RDS (OFED 1.2) on InfiniBand with Oracle 11*g*
  - Next version of the Open Source RDS (zcopy support) spec is available
- All tier one Unix system vendors are developing/testing RDS

What’s next?

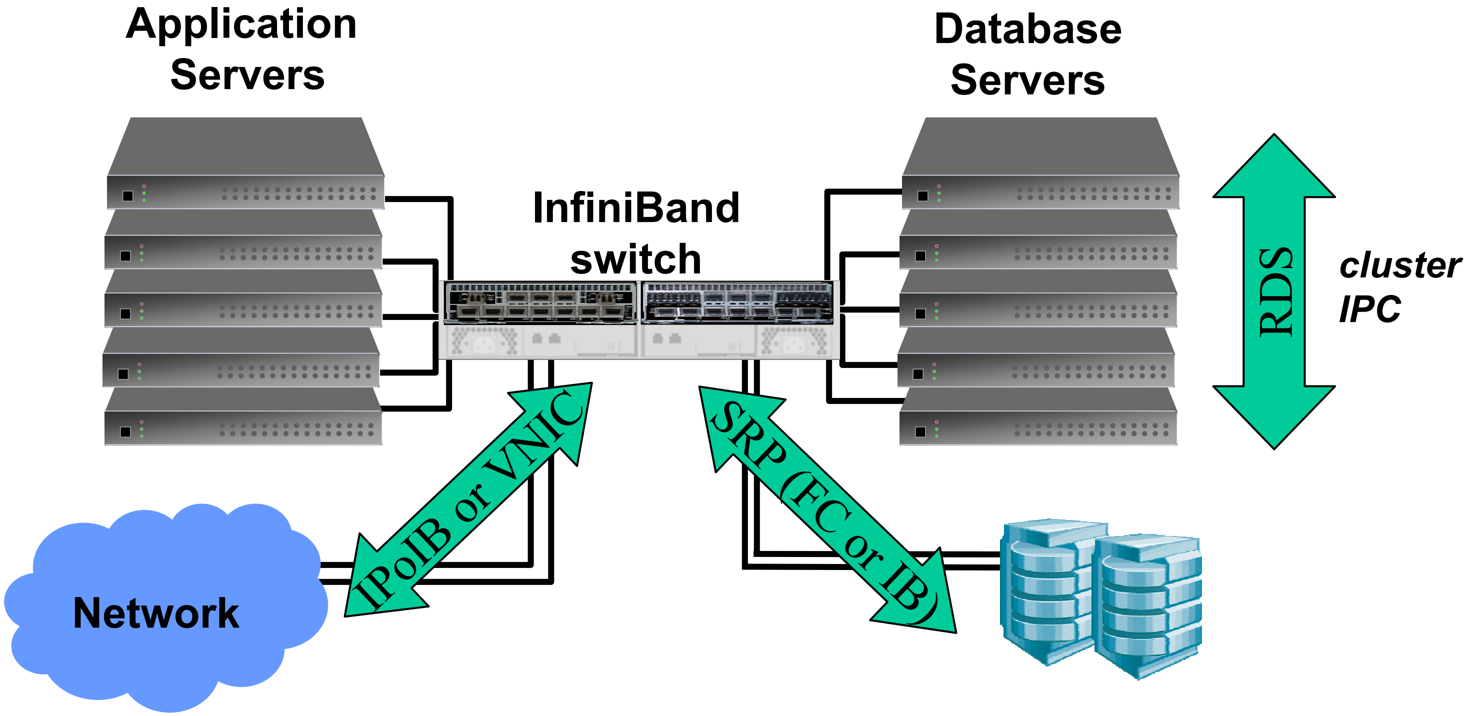
Database Clustering and I/O Deployment

Traditional FC & Dual GE Network Topology



Network Consolidation for RAC

Unified, Secure, Reduced Cost InfiniBand Fabric


Application Servers
Database Servers
InfiniBand switch
RDS
cluster IPC
IPoIB or VNIC
SRP (FC or IB)
Network