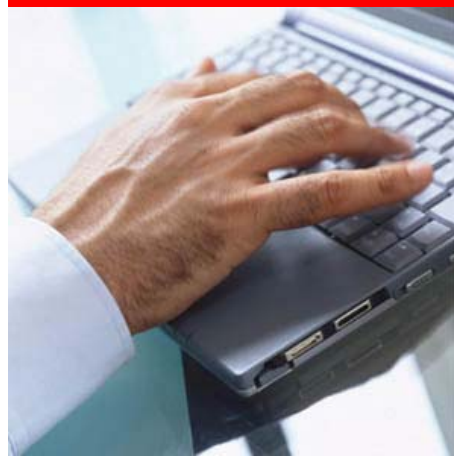

ORACLE
OPEN
WORLD

ORACLE

Next-Generation Interconnect Protocol: Reliable Datagram Sockets (RDS) and InfiniBand

Paul Tsien, Oracle
William Song, JDA Software Group, Inc. (formerly Manugistics, Inc.)

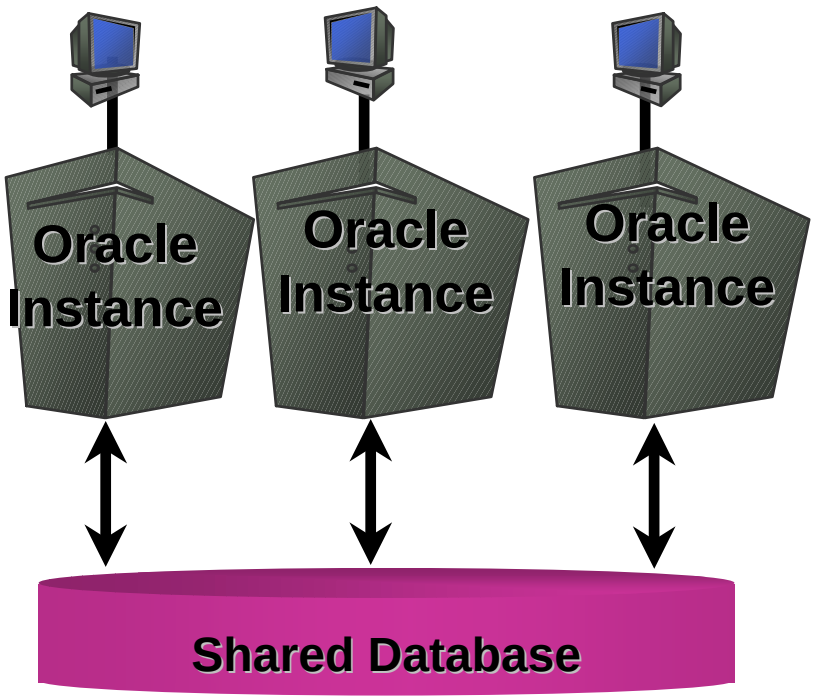
Agenda



- Oracle RAC 10*g*
- What is RDS (Reliable Datagram Sockets)?
- Open Source RDS for Linux
- Beta Customer Experience
- JDA’s Oracle RDS Project

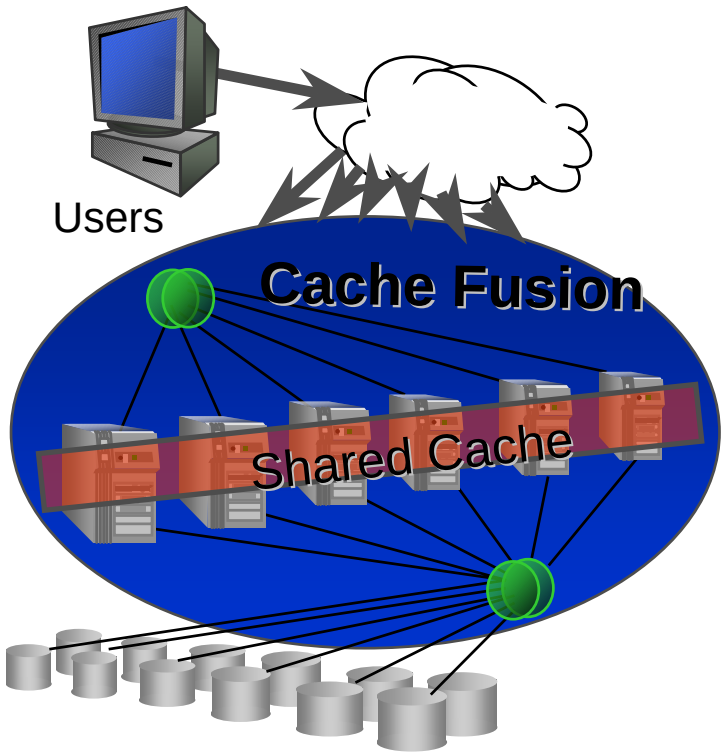
Oracle RAC 10*g*

Oracle RAC 10*g*



- Oracle Real Application Clusters (RAC) 10*g* provides the ability to build an application platform from multiple systems that are clustered together
- Allows applications to become
  - Highly scaleable
  - Highly available
- Chosen to avoid a single node failure, causing application downtime
  - Eliminates a node as single point of failure

Real Application Clusters

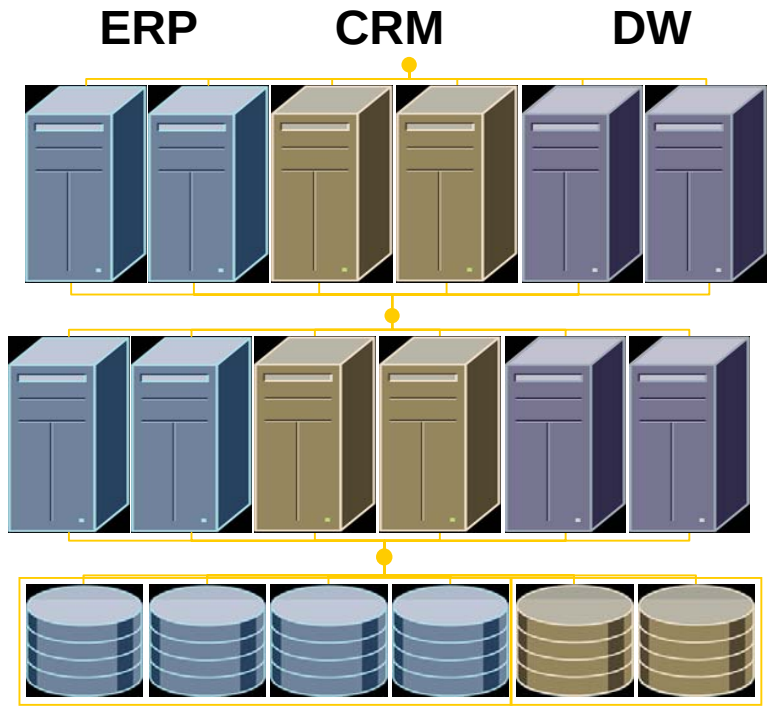


- **World's best Scalability with Cache Fusion**
  - **Cache-to-cache data shipping**
  - **Scales off-the-shelf applications with no changes**
- **World's best Availability with Fast-Start Fault Recovery**
  - **Node failure is transparent to applications**
  - **Recovers from node failure in 17 seconds - 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

What is RDS (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 !

Available Options

- uDAPL / itAPI – not supporting
- IP over IB – high CPU overhead
- SDP – connection oriented
- We want to take our existing well tested UDP module, shut off most of it to run over an O/S provided RD IPC

RDS IPC over InfiniBand

- RD – Reliable Datagram IPC over IB co-developed by Oracle and SilverStorm Technologies
  - Minimal Oracle code change
  - Stable code and easily passed all Oracle regression tests
  - Supports fail-over across and within HCAs
- Oracle internal interconnect test tool shows
  - 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

Open Source RDS

- SilverStorm RDS contributed to OpenFabrics (Industry Consortium)
- Oracle is building interconnect-agnostic Open Source RDS for Linux

  http://oss.oracle.com/projects/rds/
- Oracle will support RDS on Linux
- Oracle RDS will be pulled into OFED
- Oracle RDS will support InfiniBand, 10 Gig Ethernet, and iWARP

RDS Status

- Oracle support for SilverStorm RDS GA in 10.2.0.3
  - RDS beta testing completed, excellent performance and stability
- Open Source RDS
  - Oracle is developing/testing Open Source RDS on InfiniBand
- All tier one Unix system vendors are developing/testing RDS

Beta 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 Group, Inc.

Application Test Participants:

JDA Software Group, Inc.
SilverStorm Technologies
Oracle Corporation
Intel Corporation

Overview

- Collaborative Test Effort (JDA, Oracle, SilverStorm, and Intel)
- Why consider Oracle Real Application Clusters (RAC) 10*g*?
- RDS InfiniBand and Oracle RAC 10*g* Scalability
- JDA Grid Computing Architecture and Applications
- Test Results

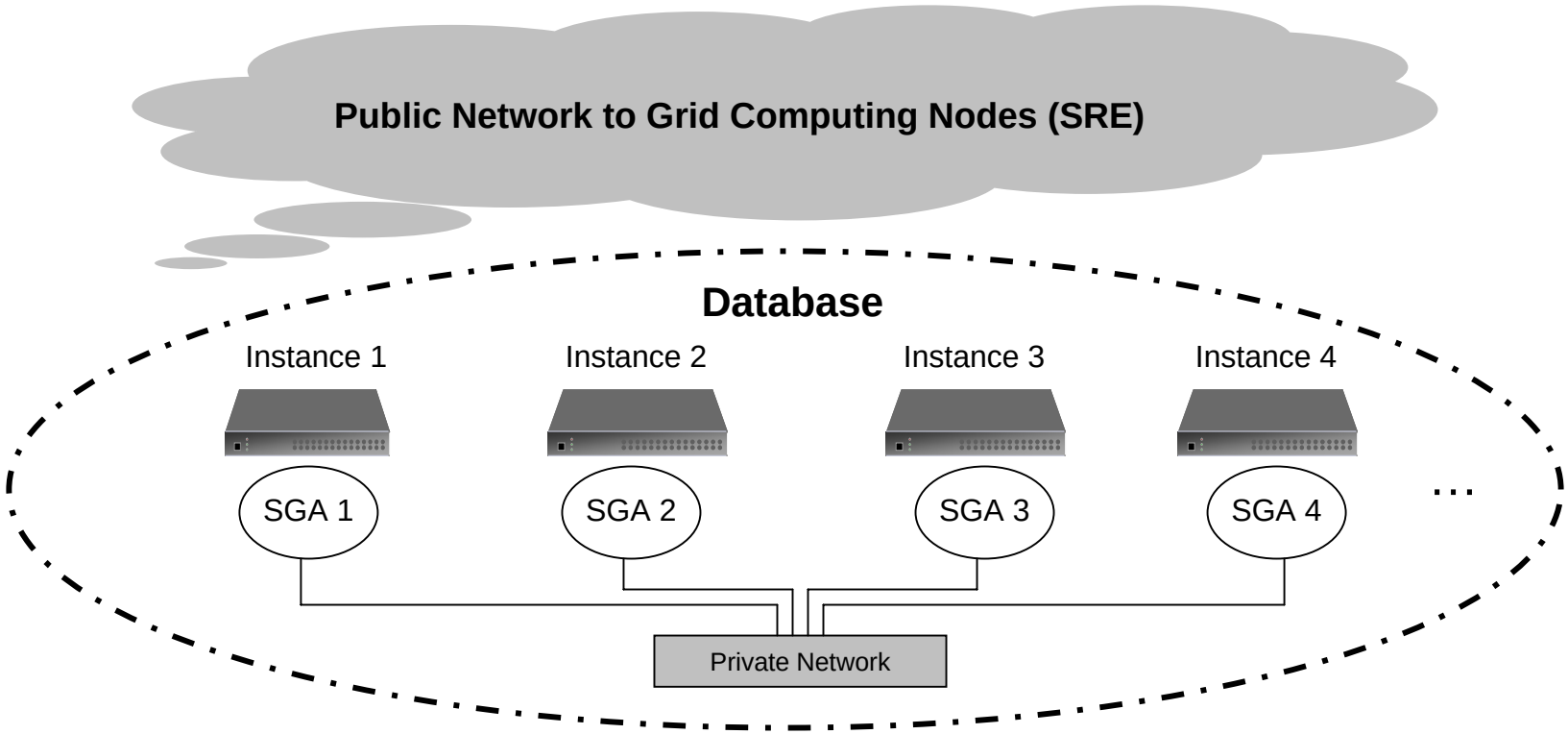
Collaborative Effort and Participation

- Collaborate in Oracle's Early Access Program to test and benchmark Reliable Datagram Sockets (RDS) over InfiniBand.
- Participate in SilverStorm's RDS Beta Program and validate InfiniBand network consolidation of RAC interconnect and SAN storage connectivity.
- Intel contributed all commodity servers allowing us to execute test plan.
- **Validate performance on real world JDA's Strategic Supply and Demand Management (SSDM) applications.**

Why JDA Applications?

- JDA's Strategic Supply and Demand Management (SSDM) applications are **rigorous, intense, and demanding** especially at the database tier, solving very large-scale planning, scheduling, and revenue optimization problems – Enterprise DSS.
- We employ a **Grid Computing Architecture** at the application tier, while using Oracle as the data store for client input data and algorithm solution output.
- We enable our application **scalability and performance** by regulating the number of grid computing nodes running across a network of distributed commodity servers.

What is Oracle Real Application Clusters (RAC) 10*g* Database?



- Multiple Instances
- One Database
- SGA database memory of all instances aggregated and appears as one single database to applications through Cache Fusion.

Why Use Oracle Real Application Clusters (RAC) 10*g* Database?

- Performance

  Increase performance of a RAC database by adding additional servers to the cluster.

- Fault Tolerance

  A RAC database is made up of multiple instance. While performance may degrade, loss of an instance does not bring down the entire database.

- Scalability

  Scale a RAC database by adding instances to the cluster database.

RDS InfiniBand

- Oracle RAC 10*g* will scale for **database intensive** applications **only** with the proper high speed protocol and private interconnect.
- Reliable Datagram Sockets (RDS) co-developed by Oracle Corporation and SilverStorm Technologies.

Industry Trends

- Changing server landscape due to market pressures from Intel and AMD. Increasing CPU performance on 2 CPU and 4 CPU Intel EM64T/AMD-64 servers are outpacing the CPU performance of larger 8 CPU – 32 CPU SMP servers.
- Application vendors are embracing Clustering, Grid, and Utility computing.
- Companies looking to lower TCO through commodity hardware without sacrificing performance or scalability.
- Clients are requesting Oracle RAC database.

JDA Grid Computing Architecture

- Originally name Service Request Environment (SRE)
- SRE framework is written in PL/SQL – wrapped and resides inside the database
- SRE Computing Nodes are written in Java

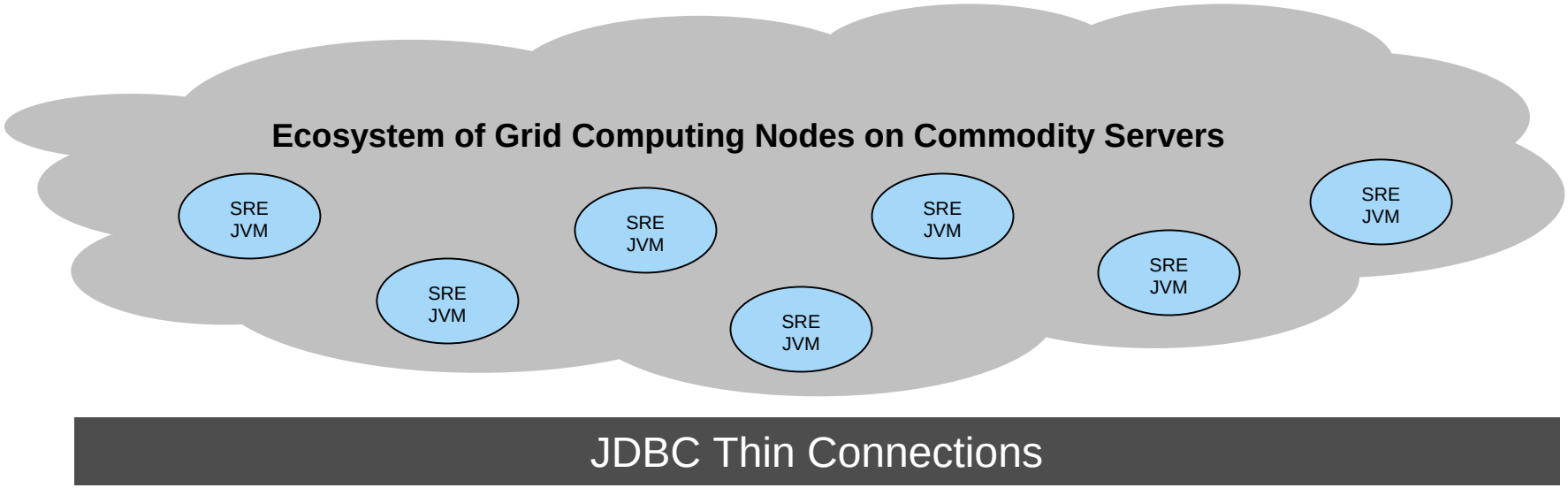
  Autonomous, no single master node,
  self-sustaining, kill failed nodes, spawn new nodes

  Multithreaded multiple concurrent database connections

- The database **is** the reliable persistent communication layer, media, and channel for all grid computing nodes.

  Leverage all the advantages of Oracle's database technology –performance, fault tolerance and scalability

JDA Grid Computing Architecture



JDA Grid Computing Applications

- We have advance and mature technology for solving large problems by dividing it up into smaller actionable jobs that can be resolved concurrently by Grid Computing Node Pools running on a number of distributed **commodity servers**.
- This is our **core technology** and what differentiates us from others in being able to scale and solve very large Supply Chain Planning, Scheduling, and Revenue Optimization problems.
- We've already **commoditized** the Application Tier.

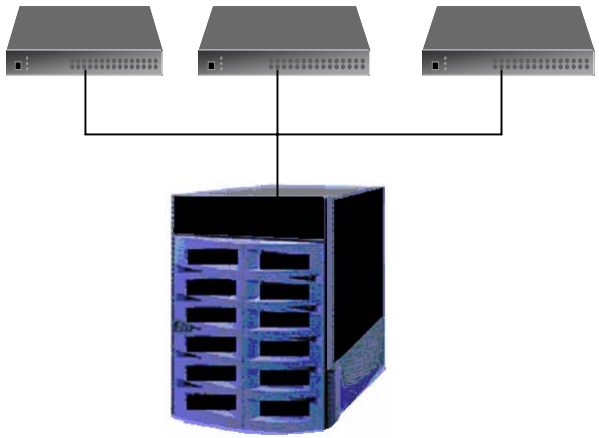
JDA Reasons for Testing RAC

- **Lower capital cost** of hardware by as much as 80% at the database tier
- Remove the **barrier to entry** by reducing the cost of the initial implementation
- Provide **Incremental scalability** by allowing RAC instances to be added to the cluster without losing value in the initial investment of servers
- **Complete our Grid Computing Architecture** by bringing it to the database tier
- **Reduce the total cost of implementation** making deals easier to close without diluting our sales margins

Shifting Trend in Deployment Paradigm

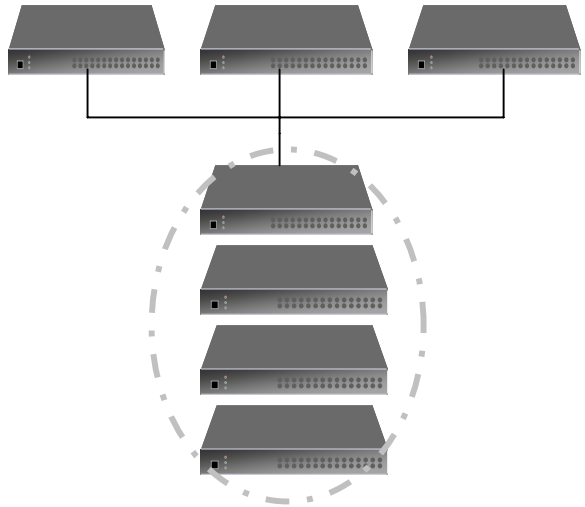
Application and Database on Same SMP Server

Application Tier on Commodity servers



Database Tier

Application Tier on Commodity servers



Database Tier on Commodity Servers

Monolithic SMP
- Application
- Database

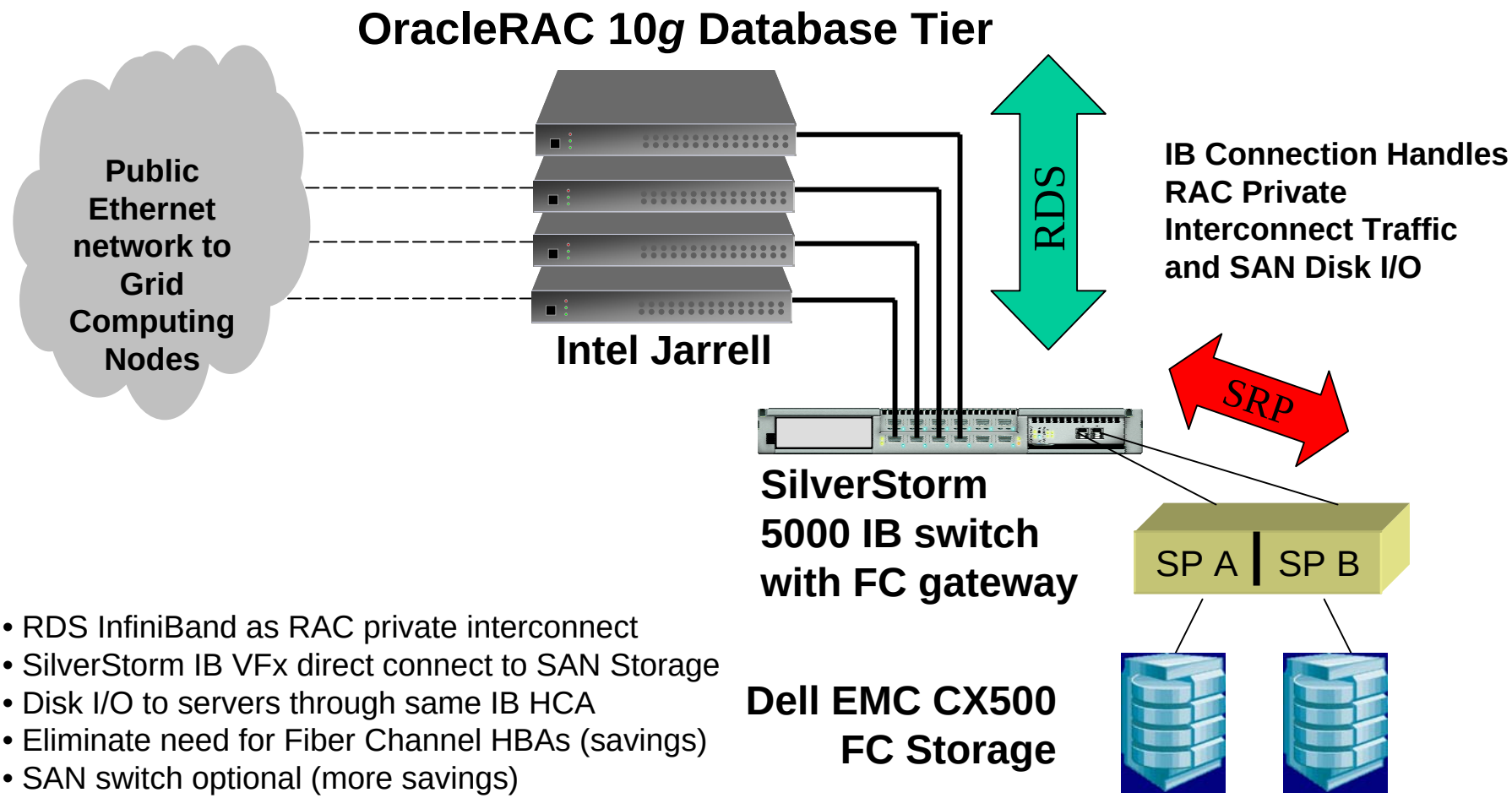
Mixed Configuration
- Commodity Application Servers
- SMP Database Servers

Grid Computing
- All Commodity Servers

Past → Present → Future

JDA REAL DEMAND CHAIN RESULTS.

InfiniBand Test Configuration



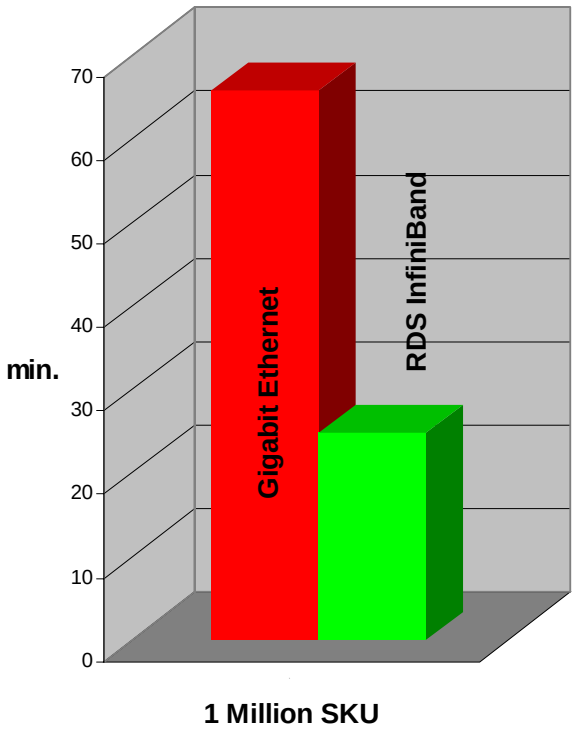
- RDS InfiniBand as RAC private interconnect
- SilverStorm IB VFx direct connect to SAN Storage
- Disk I/O to servers through same IB HCA
- Eliminate need for Fiber Channel HBAs (savings)
- SAN switch optional (more savings)

Performance Measure

- A common measure for performance in the Supply Chain Industry is the number of Stock Keep Units (SKU) planned or schedule over time – [SKU/hr].

- Using JDA's **Fulfillment Planning** application, run 1 million SKU through the plan process.

RDS InfiniBand vs Gigabit Ethernet



Time to Plan 1 Million SKU

- 66 min. on Gigabit Ethernet
- 25 min. on RDS InfiniBand

62% Improvement on SilverStorm InfiniBand with RDS

JDA REAL DEMAND CHAIN RESULTS.

ORACLE

Gigabit Ethernet vs InfiniBand Costs

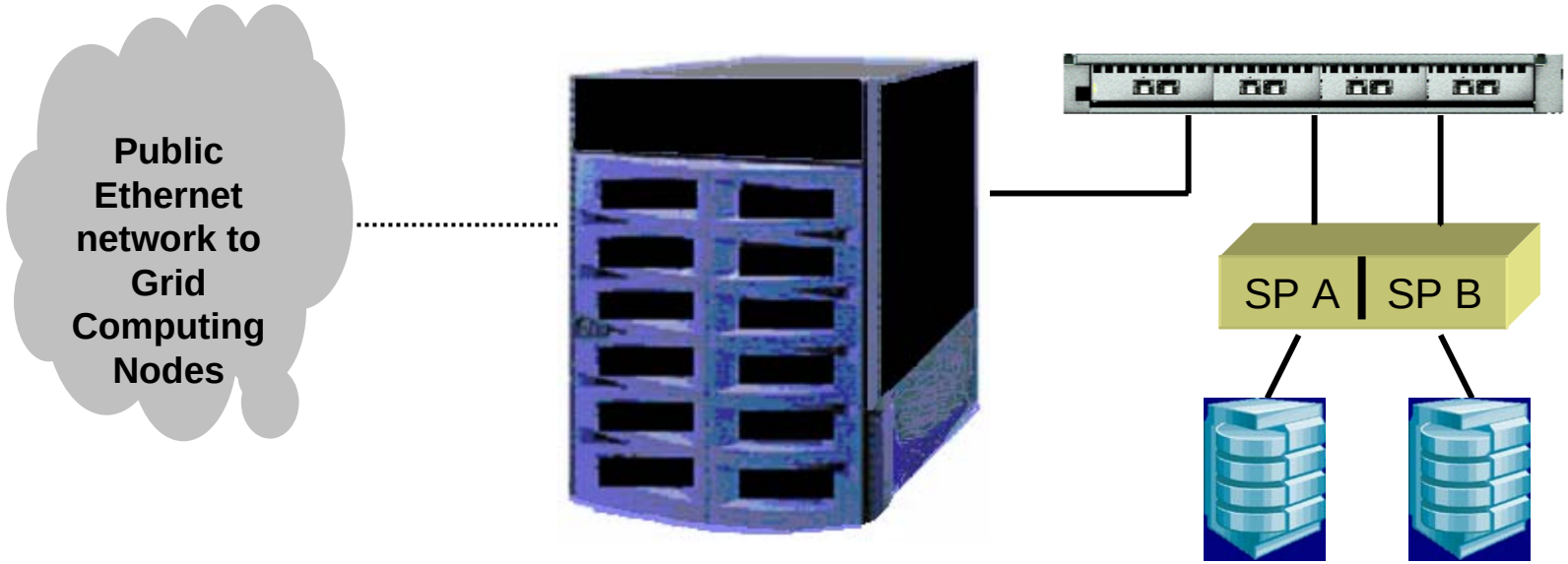
- Ethernet network & Fiber Channel SAN $19,721
  - Popular brand of enterprise class GE switch
  - Popular brand of enterprise class FC switch & HBAs

- InfiniBand unified fabric for RAC $12,825
  - SilverStorm 5000 multi-protocol InfiniBand switch with FC gateway
  - InfiniBand HCAs

35% cost reduction with SilverStorm InfiniBand network consolidation

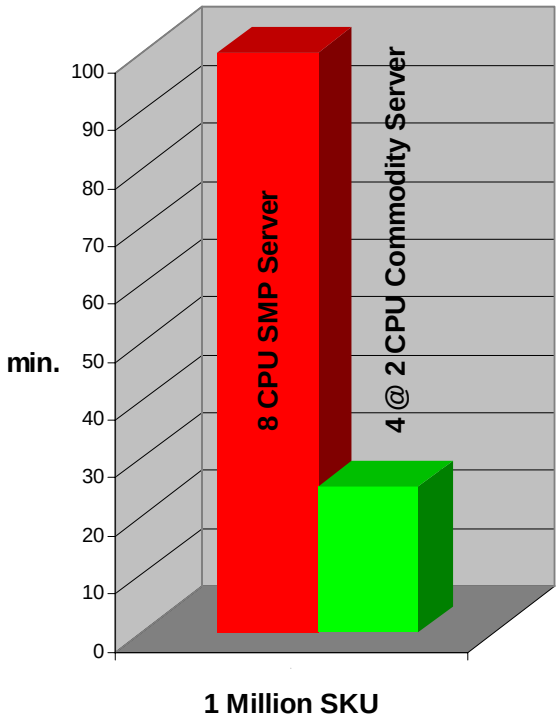
JDA REAL DEMAND CHAIN RESULTS.

SMP Test Configuration



8 CPU SMP Server

SMP vs Commodity Performance



Time to Plan 1 Million SKU

- **100 min**. on 8 CPU SMP Server
- **25 min.** on 4 Commodity Server RAC with RDS InfiniBand

75% Performance Improvement on Intel Commodity RAC Servers with RDS InfiniBand

JDA REAL DEMAND CHAIN RESULTS.

ORACLE

SMP vs Commodity Cost

$120,000

8 CPU SMP Server

$20,000

4 @ 2 CPU **Intel EM64T**
1 @ SilverStorm RDS-enabled InfiniBand Switch

83% Cost Reduction in hardware for Intel Commodity RAC Servers and SilverStorm RDS InfiniBand vs SMP Server

JDA REAL DEMAND CHAIN RESULTS.

ORACLE

Price/Performance of SMP vs Commodity

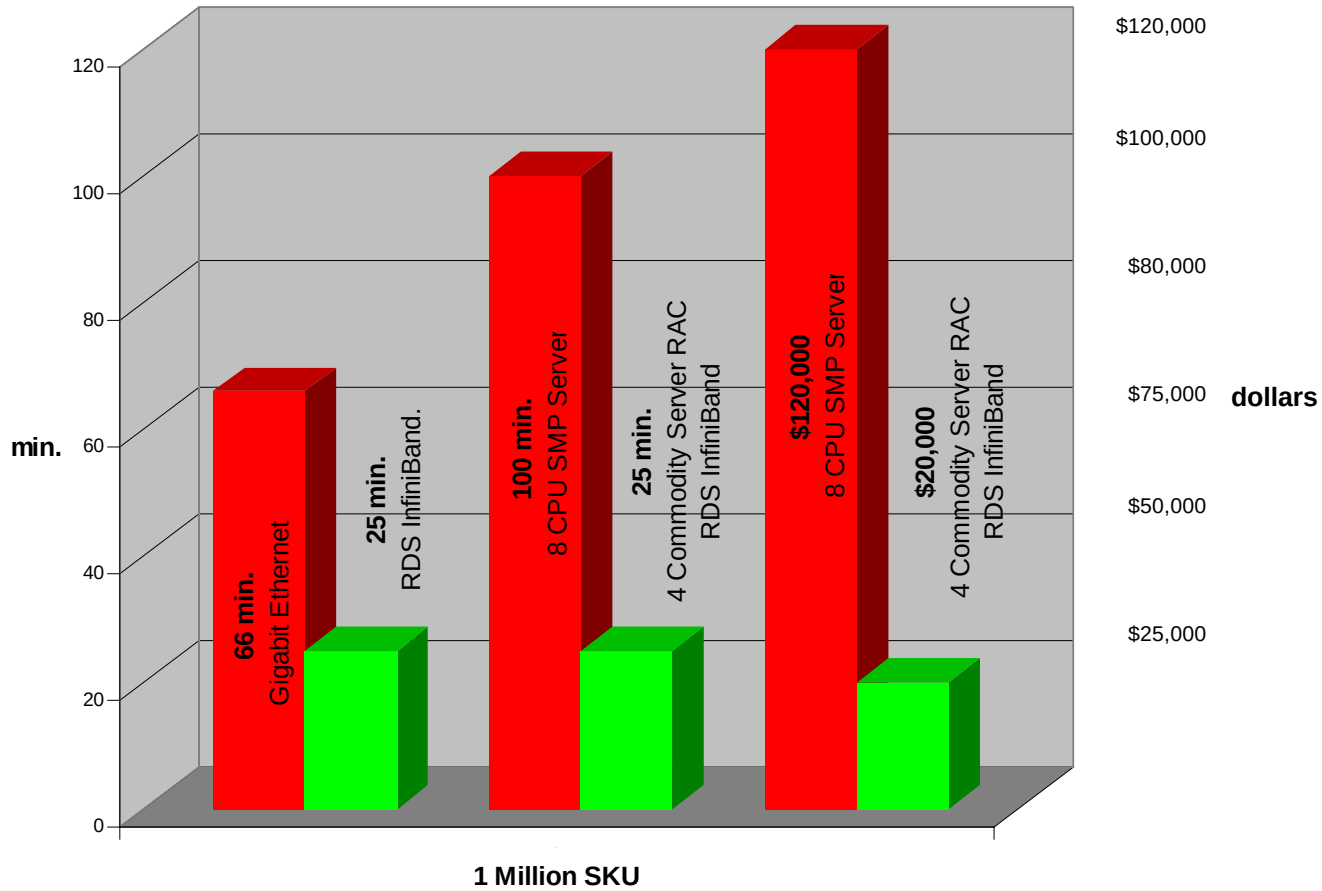
- 8 CPU SMP
  - Dollar cost to process 1M SKU = **0.200 [$ hr/SKU]***
  - 1M SKU/100min = 600,000 SKU/hr
  - $120,000
- 4 @ 2 CPU Commodity RAC
  - Dollar cost to process 1M SKU = **0.008 [$ hr/SKU]***
  - 1M SKU/25min = 2,400,000 SKU/hr
  - $20,000

*(*Similar to $/TCP-H as applied to Strategic* Supply and Demand Management *Industry)*

96% Price/Performance Improvement on Intel Commodity RAC Servers with RDS InfiniBand

JDA REAL DEMAND CHAIN RESULTS.

Speed, Performance, and Lower Cost



RDS InfiniBand and Oracle10g RAC on Commodity Servers

JDA REAL DEMAND CHAIN RESULTS.

ORACLE

Summary of Improvements

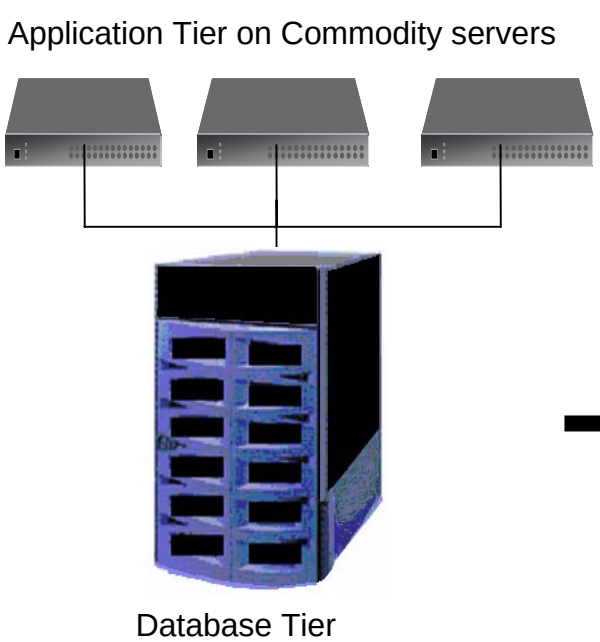
62% Speed Improvement on SilverStorm InfiniBand with RDS vs Gigabit Ethernet

75% Performance Improvement on Intel Commodity RAC Servers with SilverStorm RDS InfiniBand vs SMP Server

83% Cost Reduction in Hardware for Intel Commodity RAC Servers with SilverStorm RDS InfiniBand vs SMP Server

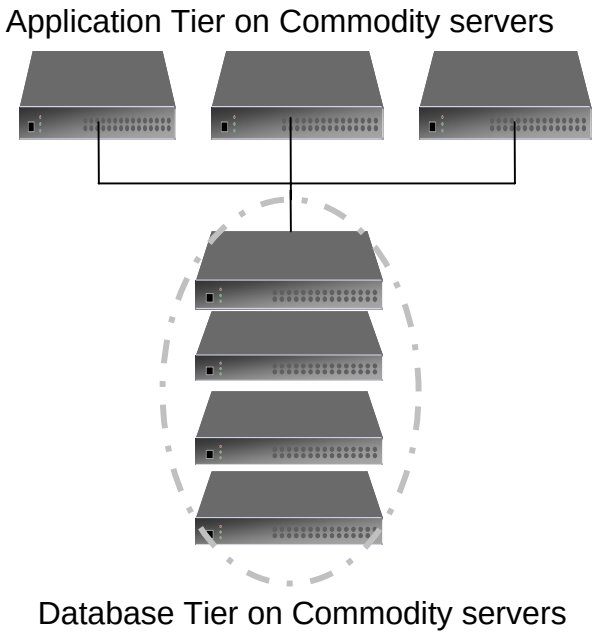
96% Price/Performance Improvement on Intel Commodity RAC Servers with SilverStorm RDS InfiniBand

Complete Grid Computing Architecture



Mixed Configuration

- Commodity Application servers
- SMP Database servers



Complete Grid Computing Solution

- All Commodity servers

JDA Completes Grid Computing Architecture Solution

JDA REAL DEMAND CHAIN RESULTS.

ORACLE

Next Steps For JDA

- Continue Oracle10g RAC Database Test Effort.
  - Replicate success with Oracle10g RAC and RDS InfiniBand across JDA application suite
- Benchmark
  - Benchmark performance results on higher volume and larger scale
  - Determine limits of performance and scalability of Oracle10g RAC and RDS InfiniBand
- Can we scale to high water mark of **1 Million [SKU/hr/CPU]** on Oracle10g RAC with RDS InfiniBand?

For More Information…

- Visit SilverStorm booth 2733 for more information and to see a demo of Oracle 10g RAC with RDS