

ORACLE®

Session id: 36777

User-mode I/O in Oracle 10G with ODM and DAFS

Jeff Silberman

Systems Architect

Network Appliance

Margaret Susairaj

MTS Server Technologies

Oracle Corp



Agenda

- The Transportation Revolution
- Concepts: RDMA, DAT, DAPL, DAFS
- RDMA and Oracle 10G
- The DAFS API: User-mode I/O and OS bypass
- ODM : The File I/O API for Oracle 10G
- Oracle 10G RAC and InfiniBand
- Performance
- Summary, Q&A

The Transportation Revolution

- “dumb” networks vs. reliable data movers
- Data copies vs. RDMA
- Ethernet vs. InfiniBand
- Kernel mode I/O vs. User-mode I/O
- Unix I/O vs. ODM

Concepts

- Remote Direct Memory Access (RDMA)
- Direct Access Transports (DAT)
- Direct Access Provider Library (DAPL)
- Direct Access File System (DAFS)

RDMA

- Memory to memory access over a network
- Requires both intelligent transports and intelligent network interface cards (NICs)
- Cannot be done over “standard” Gigabit Ethernet
- Operations defined with respect to the server
- Examples:
 - FC/VI, GbE/VI, DAPL/IB

Direct Access Transports (DAT)

- Both RDMA read and RDMA write operations supported
- Multiple concurrent virtual connections
- Asynchronous I/O
- Direct Data Placement
- Kernel Bypass

DAT is transport agnostic

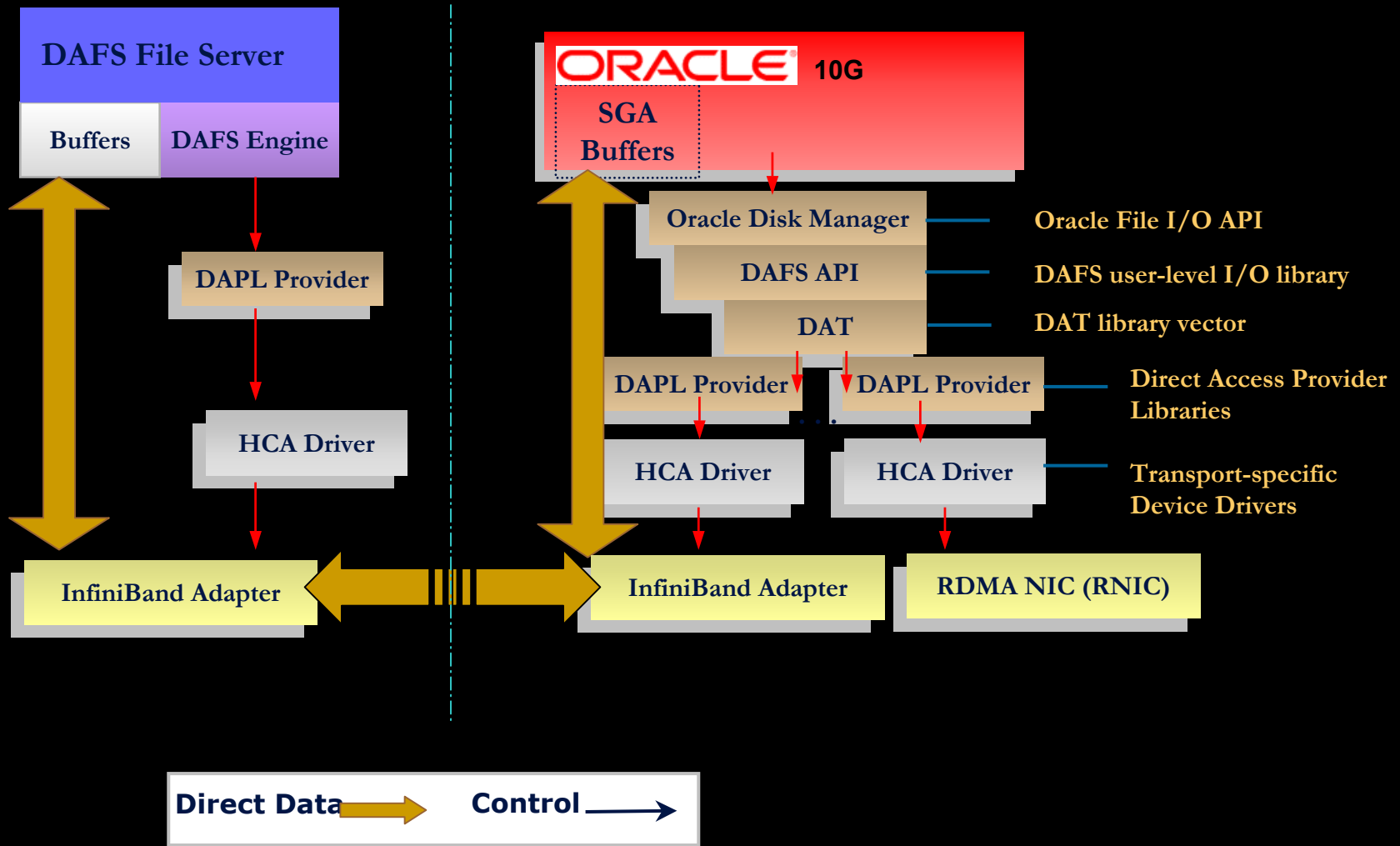
Direct Access Provider Library (DAPL)

- Standards-based API for DAT
 - Over 40 companies including both Oracle and IBM
- Designed to facilitate higher-level RDMA protocols
 - Examples: DAFS, Oracle RAC
- DAPL “providers” are typically the NIC providers
- A portable API for RDMA transports

Direct Access File System (DAFS)

- DAFS is a remote file access protocol
- DAFS derives heavily from NFSv4
- Target is local data-center file sharing
- Ideal cluster file system for RAC
- Rich set of Oracle-inspired semantics
- Will always perform better than TOE's
 - Zero touch, zero data copy

Oracle 10G and RDMA



Oracle 10G and RDMA

- Low latency
- High Bandwidth
- Memory to memory transfer
- Minimal CPU intervention
- User-mode I/O
 - ✓ Data block transfers for cache fusion
 - ✓ Storage I/O requests
 - ✓ Lock request messages

DAFS API: User-Mode I/O

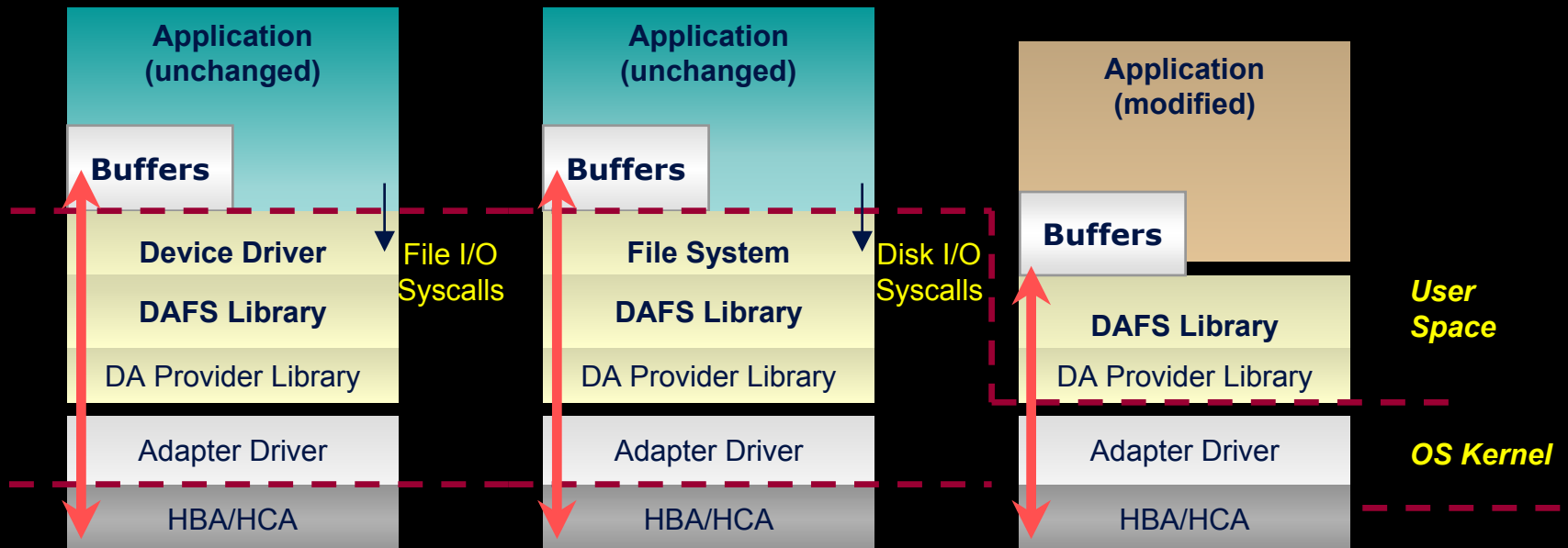
- Memory Registration
- Asynchronous I/O
- Security / Authentication
- I/O Fencing
- I/O Completion Groups
- Multi-path I/O

DAFS Implementation Models

Kernel File System

Raw Device Driver

User Library



Application Transparency

Performance

Oracle Disk Manager (ODM)

- The File I/O API for Oracle
- Performance of Raw Disk with the Manageability of Files

Oracle Disk Manager (ODM)

Problem	Solution
No consistent standard I/O interfaces. I/O interfaces vary with each operating system variant.	The ODM API semantics are invariant across all OS platforms including Windows
No standard asynchronous I/O model for regular files. Asynchronous I/O, if it was provided, relied on special kernel-based device drivers.	ODM supports both synchronous and asynchronous I/O for any regular files in an ODM file system
No standard for batching I/O requests within a single I/O call.	The <code>odm_io()</code> function provides batch I/O capability, which minimizes the number of system calls and kernel traps
Excess system resources consumed when each process in an Oracle instance must open each datafile in the instance	ODM provides shared file identifiers. A given file-id can be used by any process in the instance, thereby reducing the number of opens, instance wide.

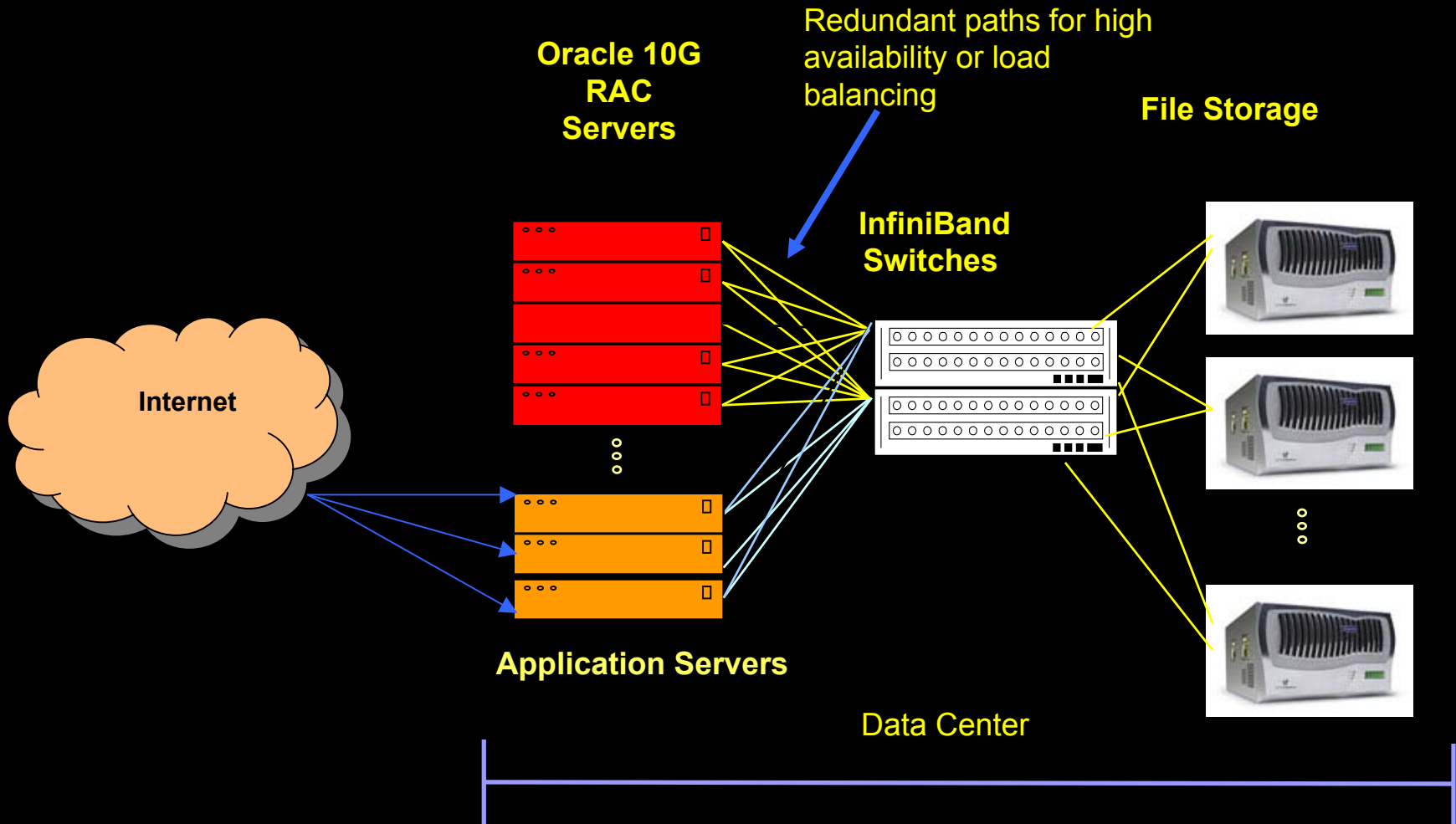
ODM Advanced File Semantics

- Open with 'share' key
- Files not visible until file is initialized
- Files cannot be deleted if open references exist

ODM version 2

- Zero data copy
 - Zero touch of data, from storage to SGA
- Memory registration
- Asynchronous I/O
 - Polling more, waiting less ...
- Non-shared file ids
 - Same semantics as with Unix file descriptors
- Portability
 - DAFS API semantics are invariant across platforms

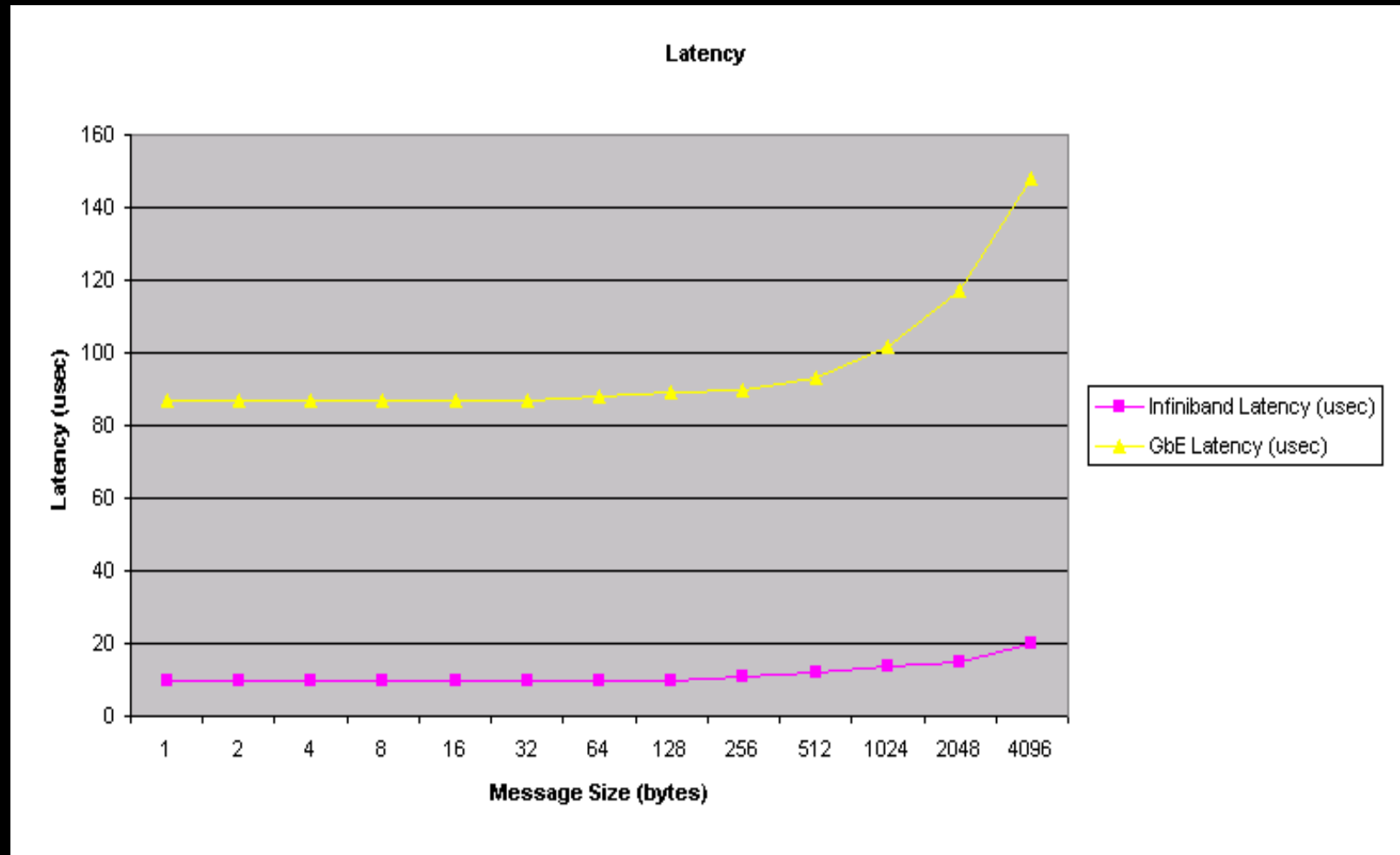
Oracle 10G RAC



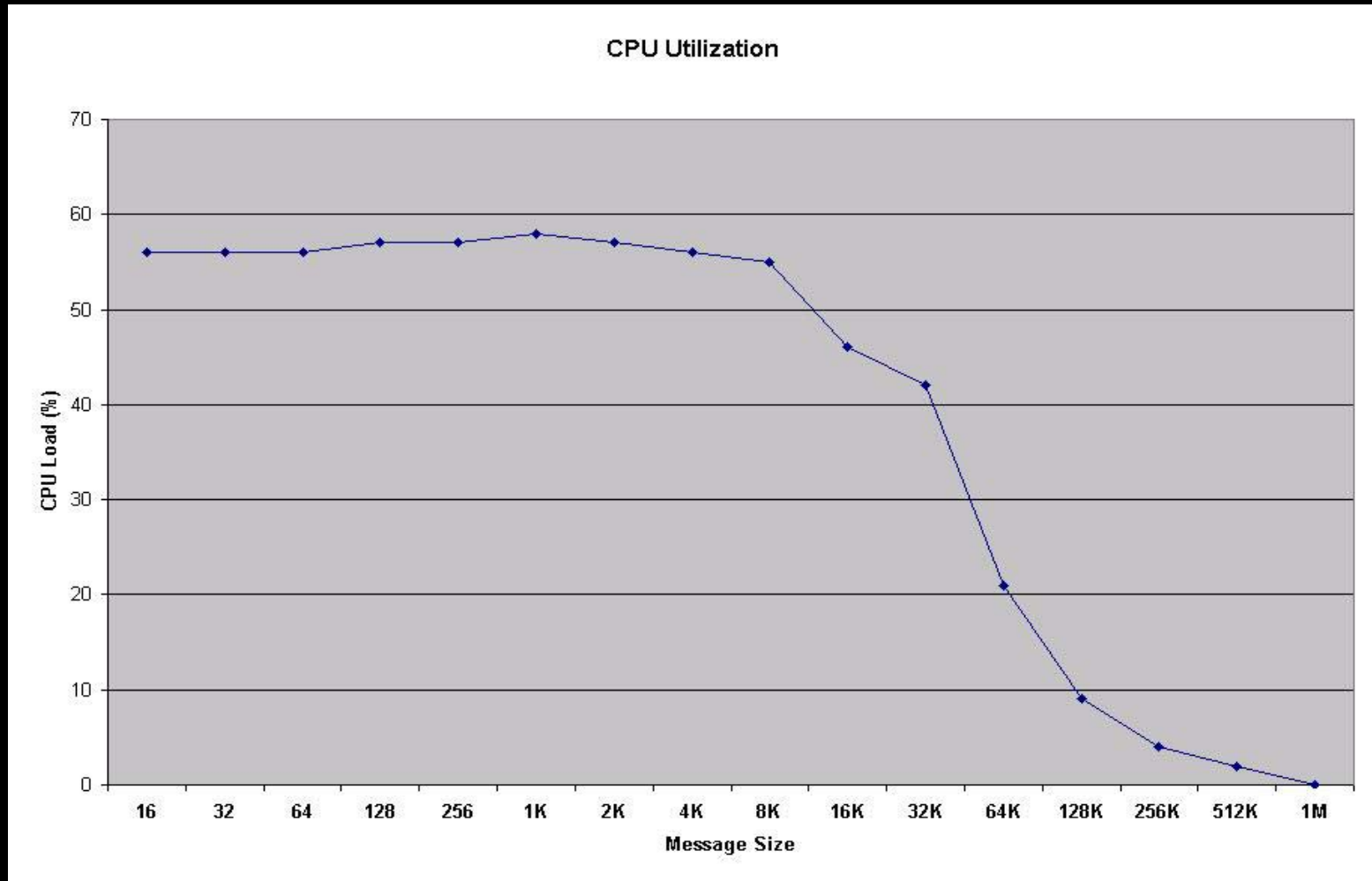
Performance

- Thanks to Ariel Cohen from Topspin Communications
- One client / one server
 - 1.8 GHz Xeon CPU
 - 133 MHz PCI-X bus
 - 4x IB HCA (10 Gbs)
 - Gigabit Ethernet w/ checksum offload support
 - Jumbo frame size of 9000
 - RedHat Linux 7.3

Performance



Performance



NFS and RDMA

Network Appliance and Sun Microsystems Collaborate on Next- Generation System Area Network Technology



[Printable Page](#)

**Aim to Make RDMA File Access Near Wire Speed on 10Gb/s
Transports**

Sunnyvale, Calif. and Mountain View, Calif. - June 24, 2003 -- Network Appliance, Inc. (NASDAQ: NTAP) and Sun Microsystems, Inc. (NASDAQ: SUNW) announced today that the companies have entered a wide-ranging technical collaboration focused on delivering the highest possible throughput from 10Gb/s transports thereby accelerating data exchange between computers of the future.

Evolution and Revolution

- Hungry apps and database must look elsewhere for extra CPU power
 - OS bypass for I/O
- High performance transports are here today
 - InfiniBand offers 10Gbs w/ 10 usec latency
- Unix and Windows do not provide user-level I/O
 - The DAFS API does
- Oracle 10G RAC w/ a single pipe
 - Both RAC/IP and user-level file I/O over one IB pipe

“Please keep your seatbelts fastened ...”

Next Steps

High Availability Sessions from Oracle

Tuesday in Moscone Room 304

11:00 AM

***How Oracle Database 10g
Revolutionizes Availability and
Enables the Grid***

3:30 PM

***Oracle Recovery Manager (RMAN)
10g: Reloaded***

5:00 PM

***Proven Techniques for Maximizing
Availability***

Wednesday in Moscone Room 304

8:30 AM

***Oracle Database 10g - RMAN and ATA
Storage in Action***

11:00 AM

***Oracle Data Guard: Maximum Data
Protection at Minimum Cost***

1:00 PM

***Oracle Database 10g Time
Navigation: Human-Error Correction***

4:30 PM

***Data Guard SQL Apply: Back to the
Future***

ORACLE

For More Info On Oracle HA Go To <http://otn.oracle.com/deploy/availability/>

Next Steps

High Availability Sessions from Oracle

Thursday

8:30 AM in Moscone Room 304

***Oracle Database 10g Data
Warehouse Backup and Recovery:
Automatic, Simple, Reliable***

8:30 AM in Moscone Room 104

***Building RAC Clusters over
InfiniBand***

**Database HA Demos All Four Days
In The Oracle Demo Campground**

Real Application Clusters

Data Guard

Database Backup & Recovery

Flashback Recovery

***LogMiner, Online Redefinition, and
Cross Platform Transportable
Tablespaces***

ORACLE

For More Info On Oracle HA Go To <http://otn.oracle.com/deploy/availability/>

**Reminder –
please complete the
OracleWorld online session
survey**

Thank you.

A large, stylized logo in the background consisting of a grey 'Q', a red ampersand '&', and a grey 'A'. The text 'QUESTIONS' and 'ANSWERS' is overlaid on this logo.

QUESTIONS
ANSWERS