

Oracle 9i Real Application Clusters

Marshall Presser
Principal Technologist
Oracle Corporation

Oracle9i Real Application Clusters

Availability
Scalability
Performance
Cost

Oracle9i Real Application Clusters

- Oracle9i Real Application Clusters is designed for today's most demanding deployments
 - Server consolidation means very large user populations
 - Critical e-business requires full time service
 - Rapid growth shortens capacity planning



Oracle9*i* Real Application Clusters

- Major features of Oracle9*i* RAC
 - Scalability with full Cache Fusion architecture
 - Greatest database availability
 - Manageability of a single database
 - Exploitation of technology advances

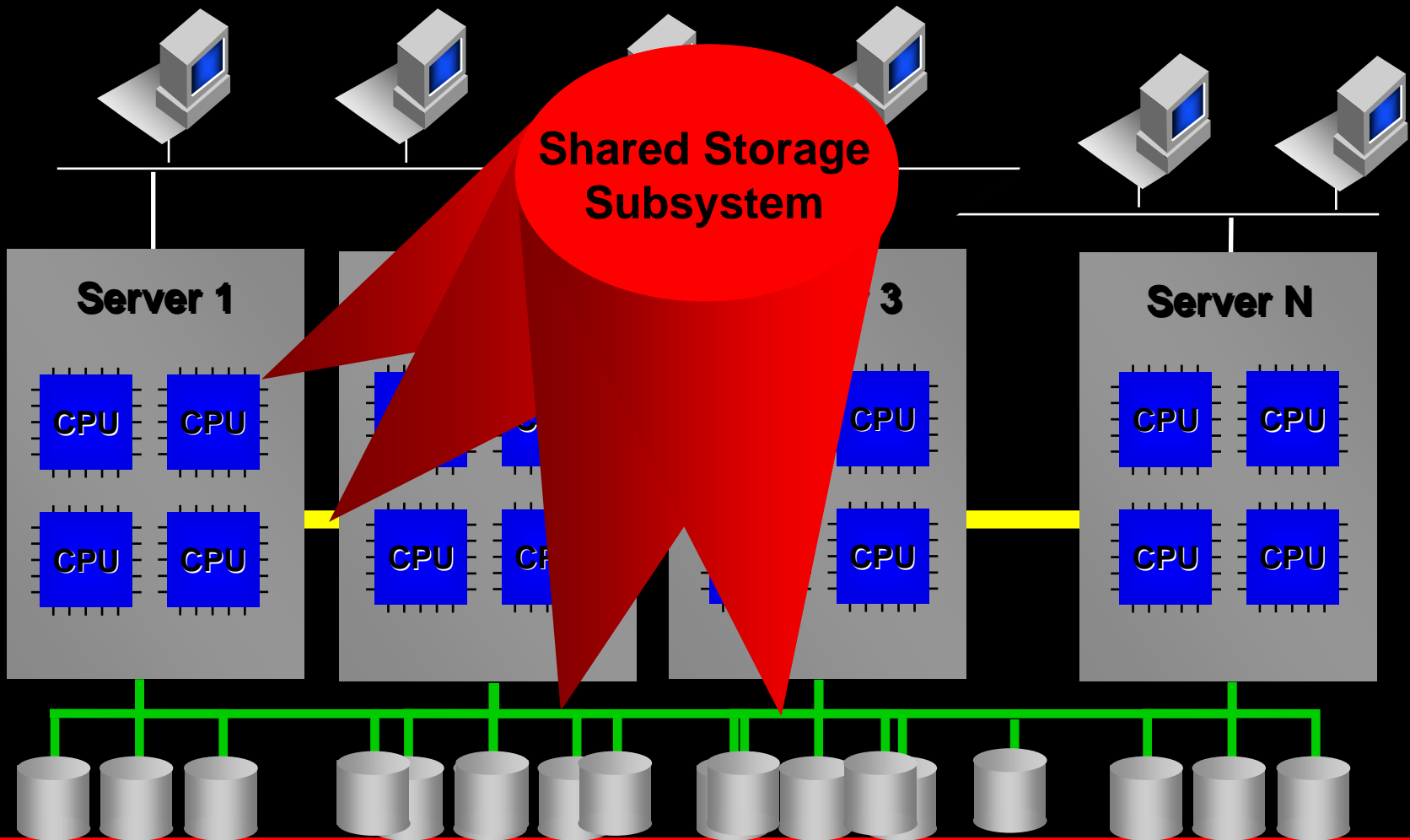


What is a Cluster?

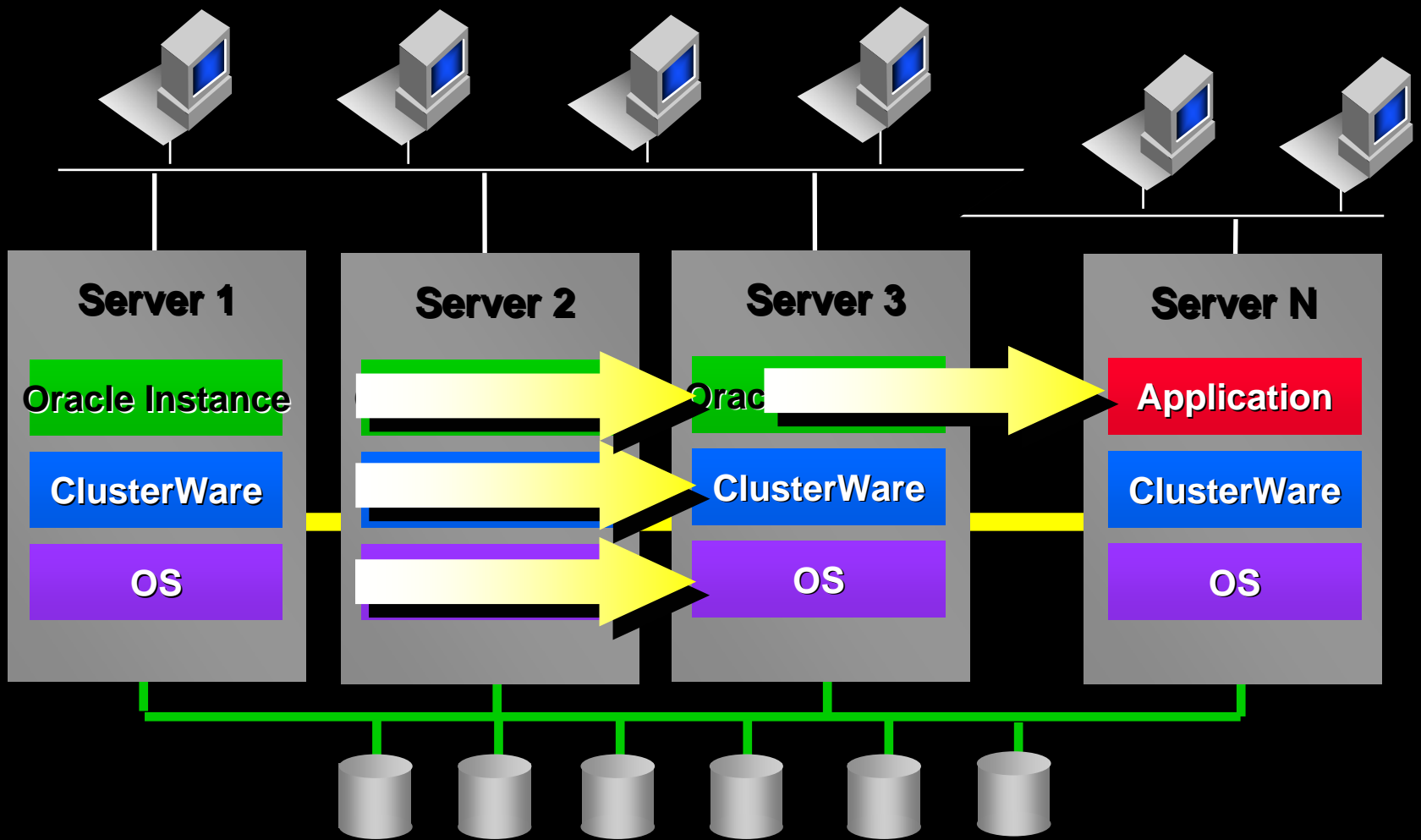
A cluster is a group of independent computers working together as a single system

- **Availability** - Continues running in case of a hardware or software failure
- **Scalability** - New nodes can be added to a cluster to accommodate increased workload
- **Performance** - Workload can be distributed among nodes for optimal performance
- **Cost** – Economics of servers make clusters of small servers more cost effective than large SMP servers

Key Hardware Components

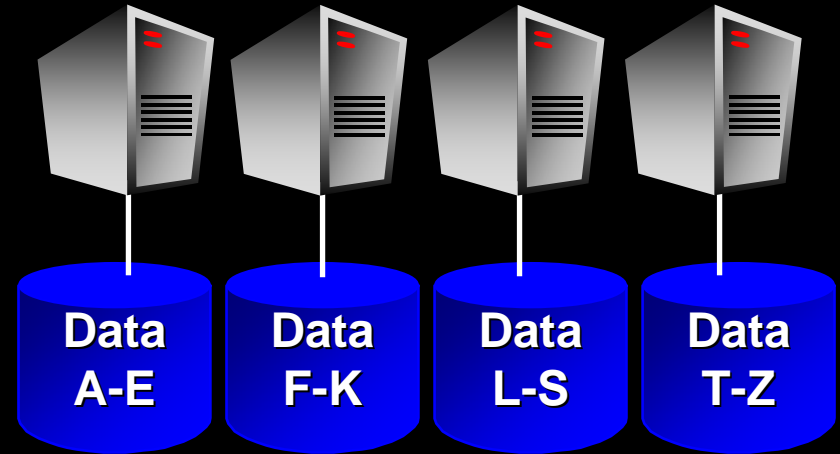
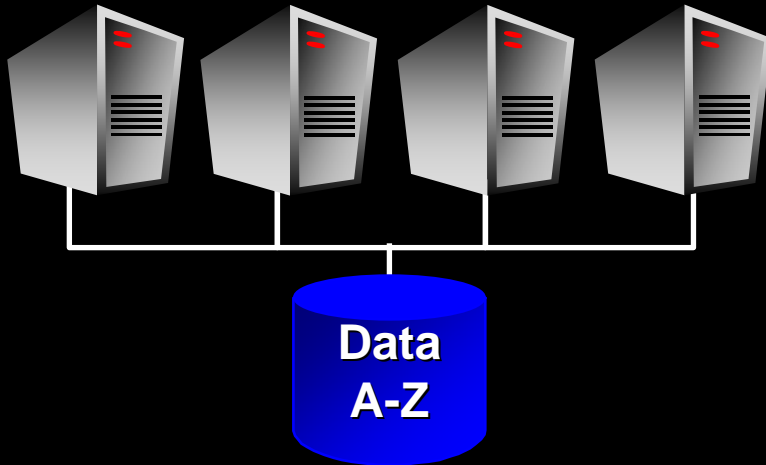


Key Software Components

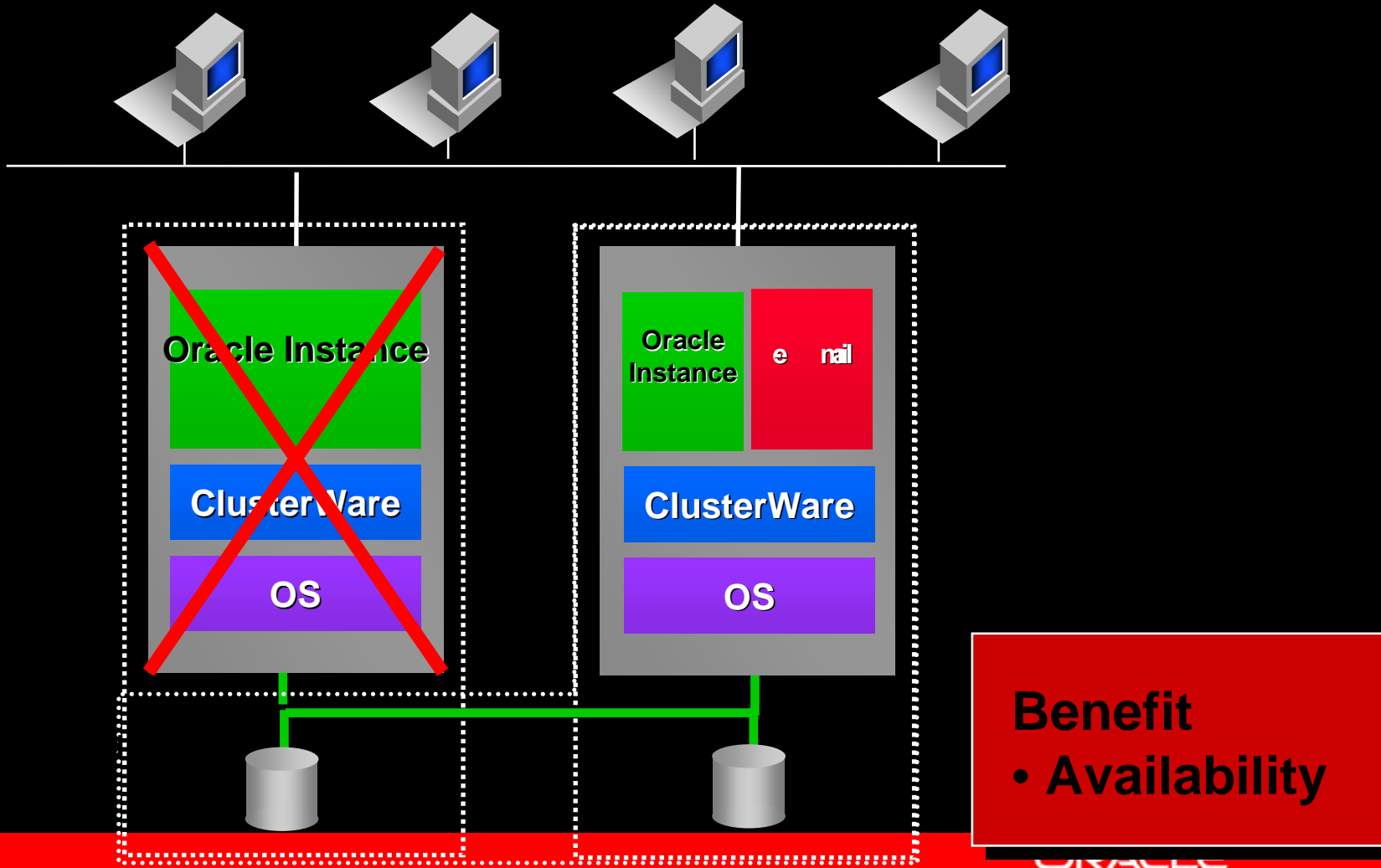


Database Cluster Types

- Shared Cache/Disk
 - Oracle and IBM Mainframes
 - More Reliable As You Add Computers
 - No Data Partitioning Required
- Shared Nothing
 - Microsoft and IBM Unix/NT
 - Less Reliable As You Add Computers
 - Static Data Partitioning



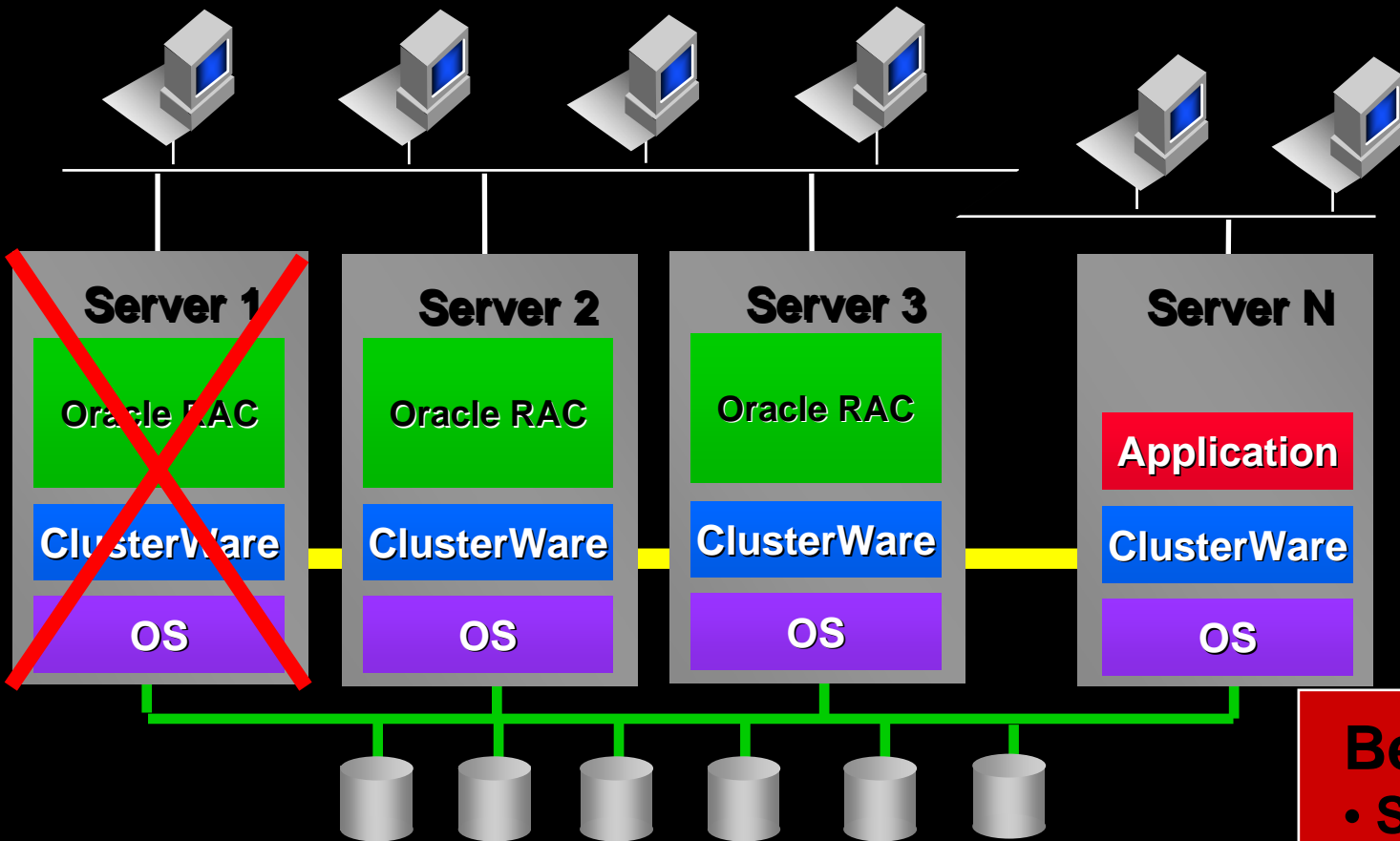
OS Failover Solutions



Benefit
• Availability

ORACLE

Oracle RAC Software Stack



Benefits

- Scalability
- Availability
- Manageability

Comparative Failover Timings

Failover operation	'Cold' Failover	'Hot' Failover
Reconfigures Group Membership	N/A	15 Secs
Reconfigures Distributed Locks	N/A	5 Secs
Failover disk volumes	Up to 20 mins	N/A
Restart Oracle	Up to 5 mins	N/A
Recover Oracle	20 Secs	20 Secs
Total Failover Time	> 25mins	< 60 Secs

On what H/W base and in what year did Oracle first release a multi-instance database system?

- (A) Sun OS - 1993
- (B) HP-UX - 1992
- (C) Windows - 1995
- (D) IBM MVS - 1990
- (E) DEC VMS - 1989

**Any application that runs
on Oracle 9i can run
unmodified on Real
Applications Clusters**

Database Challenges: Scalability

- Scale to Millions of Users
 - Transparently
 - All types of applications
- Scale workloads without limits
- Grow storage easily



Real Application Clusters

Out-of-the-Box Transparent Application Scalability

- In the past clustered databases scaled well for specific types of applications
 - Data Warehouse
 - Parallel-enabled OLTP
- Oracle Real Application Clusters with Cache Fusion is a breakthrough in parallel database technology delivering transparent scalability to all types of applications

Traditional Shared-Disk Clustered Databases

- Maintaining data coherency is a hard problem
 - Need to synchronize updates to shared data
 - The disk is the only medium for data sharing
- Disk I/O latencies appear in the critical path when multiple nodes access shared data
- Disk-based coherency is the main bottleneck to achieving a scalable shared disk cluster
 - Only synthetic fully partitioned workloads scale!

Oracle Real Application Clusters (RAC)

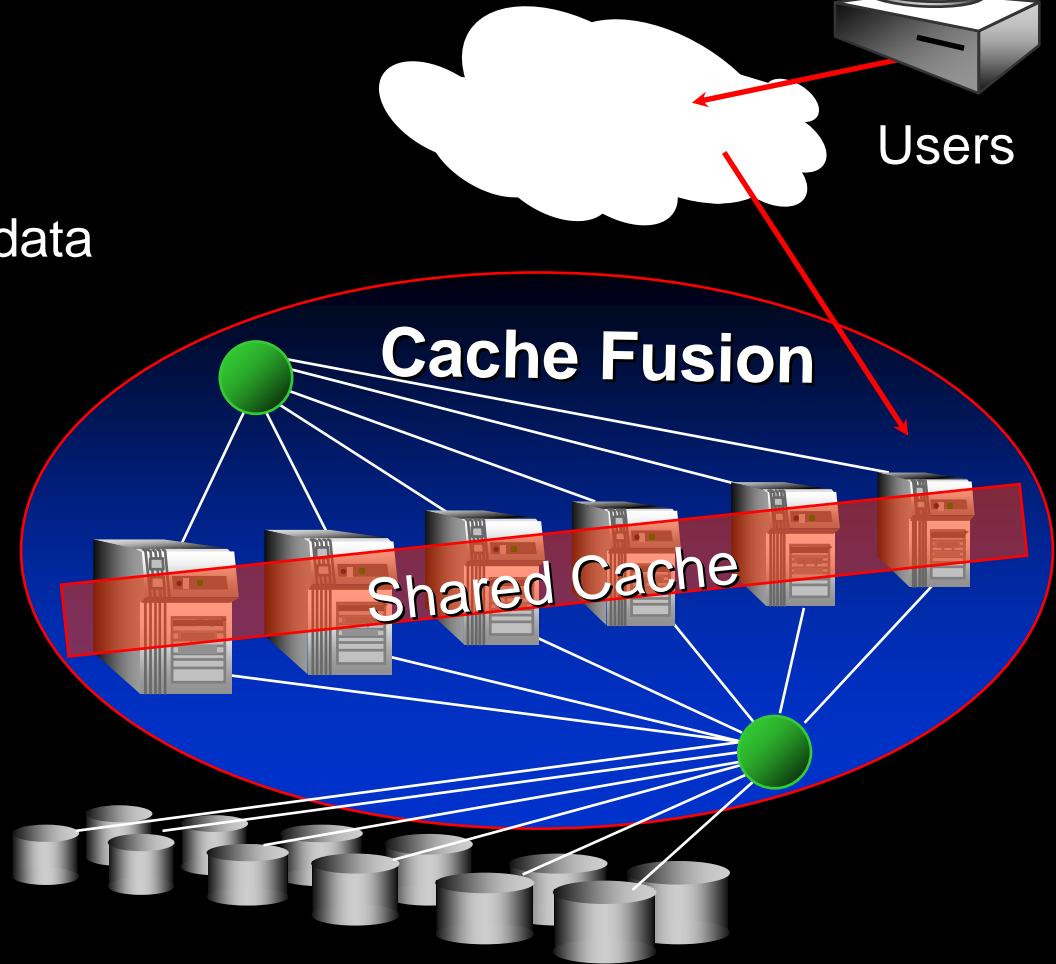
- An application transparent clustered database
 - single node applications run and scale with no changes
- Cluster interconnect fabric replaces the disk as the medium for inter-node data sharing
- Cache Fusion protocol for data sharing results in a scalable cluster for OLTP workloads

Cache Fusion Architecture



Users

- Full Cache Fusion
 - Cache-to-cache data shipping
 - Shared cache eliminates slow I/O
 - Enhanced IPC
- Allows flexible and transparent deployment



What is Cache Fusion?

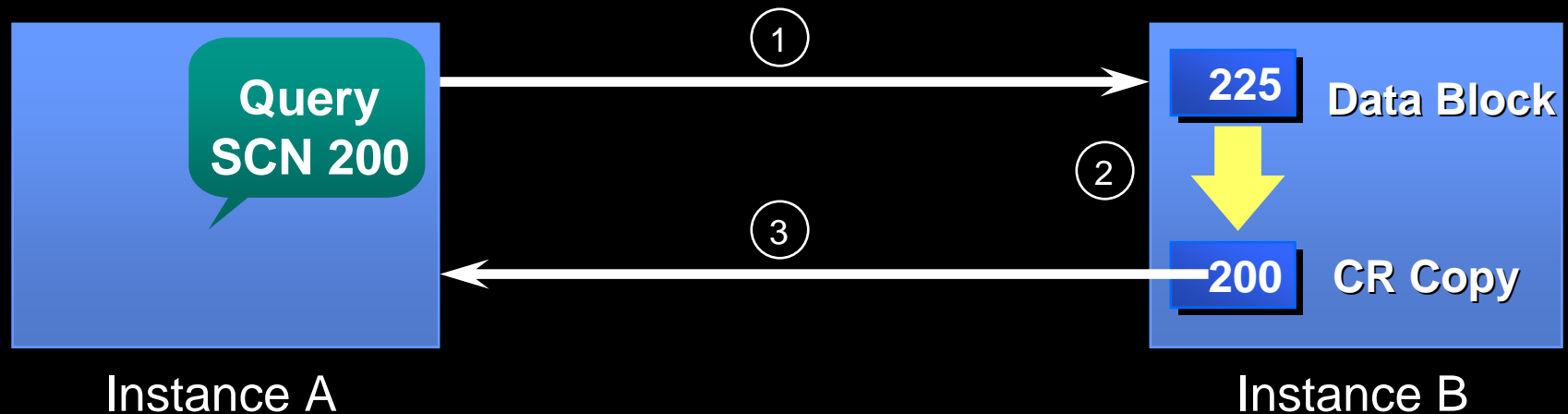
- The underlying technology that enables RAC
- Protocol that allows instances to combine their data caches into a shared global cache
 - Global Cache Service (GCS) coordinates sharing
- Key features are
 - Direct sharing of volatile buffer caches
 - Efficient inter-node messaging framework
 - Fast recovery from node failures using cache and CPU resources from all surviving nodes

Data Sharing Problem

- Read Sharing for Queries
 - query needs to read a data block that is currently in another instance's buffer cache.
- Write Sharing for Updates
 - update needs to modify a data block that is currently in another instance's buffer cache.
- With Cache Fusion, a disk read is performed only if the block is not already in the global shared cache

Cache Fusion Read Sharing

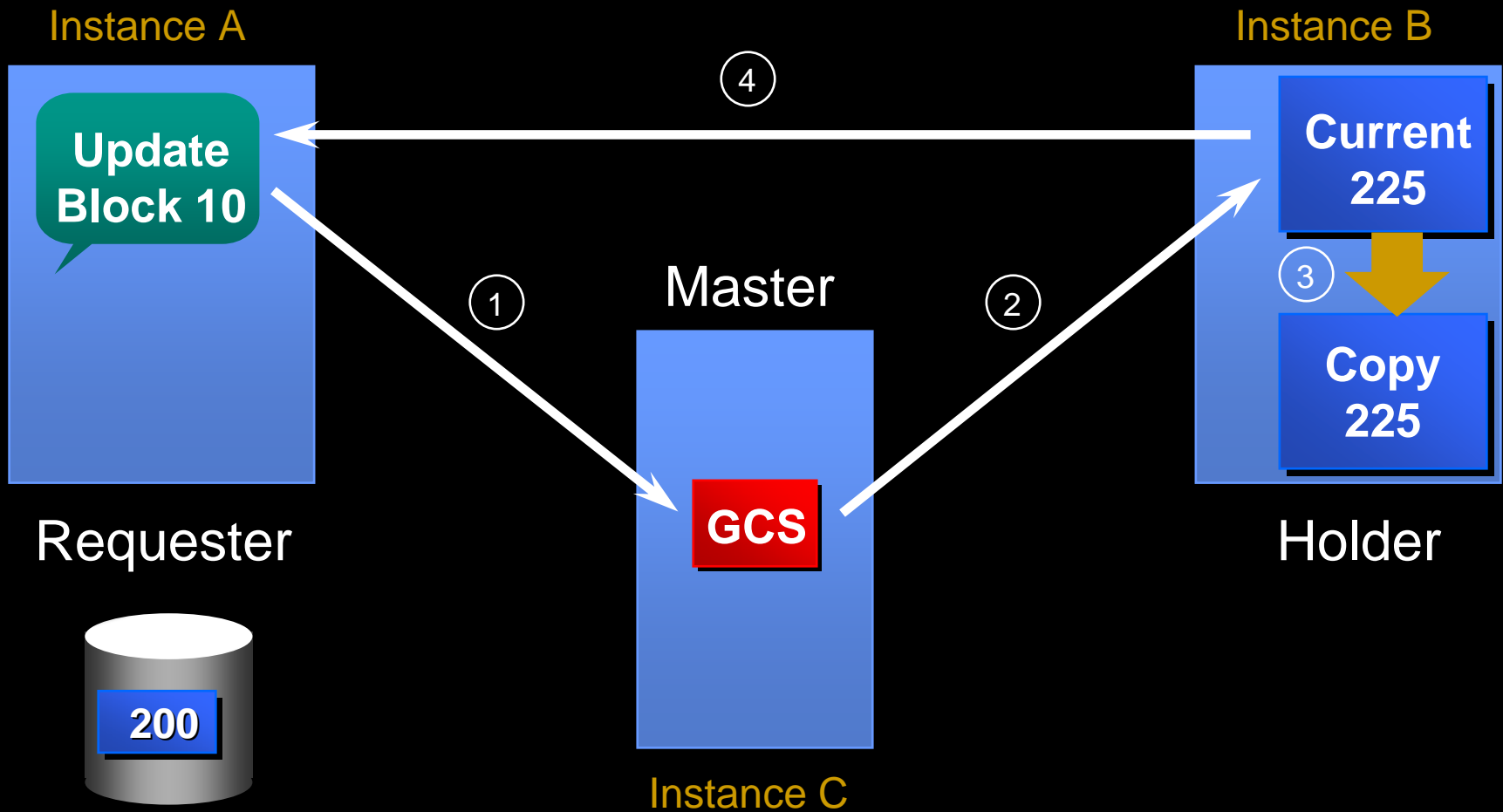
- Uses Oracle's Consistent Read (CR) scheme
 - undo is applied to make a block transactionally consistent to a *System Change Number* (SCN).
 - a CR copy is shipped to the requesting instance



Cache Fusion Write Sharing

- Multiple dirty copies of a data block can exist in the global cache, but only one is *current*
- The current copy can move between instances without first being written to disk
 - Changes are logged if not already on disk
- Non-current dirty copies can directly service queries from any node and instance recovery

Cache Fusion Write Sharing



Marshall's Synchronization Formula

- Total cost of synchronization = number of synchronizations * cost/synchronization
- Number of synchronization is application dependent
- Cost per synchronization is data server dependent
- Goal: drive either term to zero!

Efficient Inter-Node Messaging

- Messaging cost independent of cluster size
 - At most 3 nodes involved in a request
 - requester, holder and master (directory)
 - number of messages to service a request is bounded
- Inter-Node Message Latency
 - exploits high performance interconnect substrates so that on-the-wire message transmission times are minimal
- Frequency of Inter-Node Synchronization
 - adaptive directory migration based on access patterns
 - fast reconfiguration of resources when a node joins/leaves

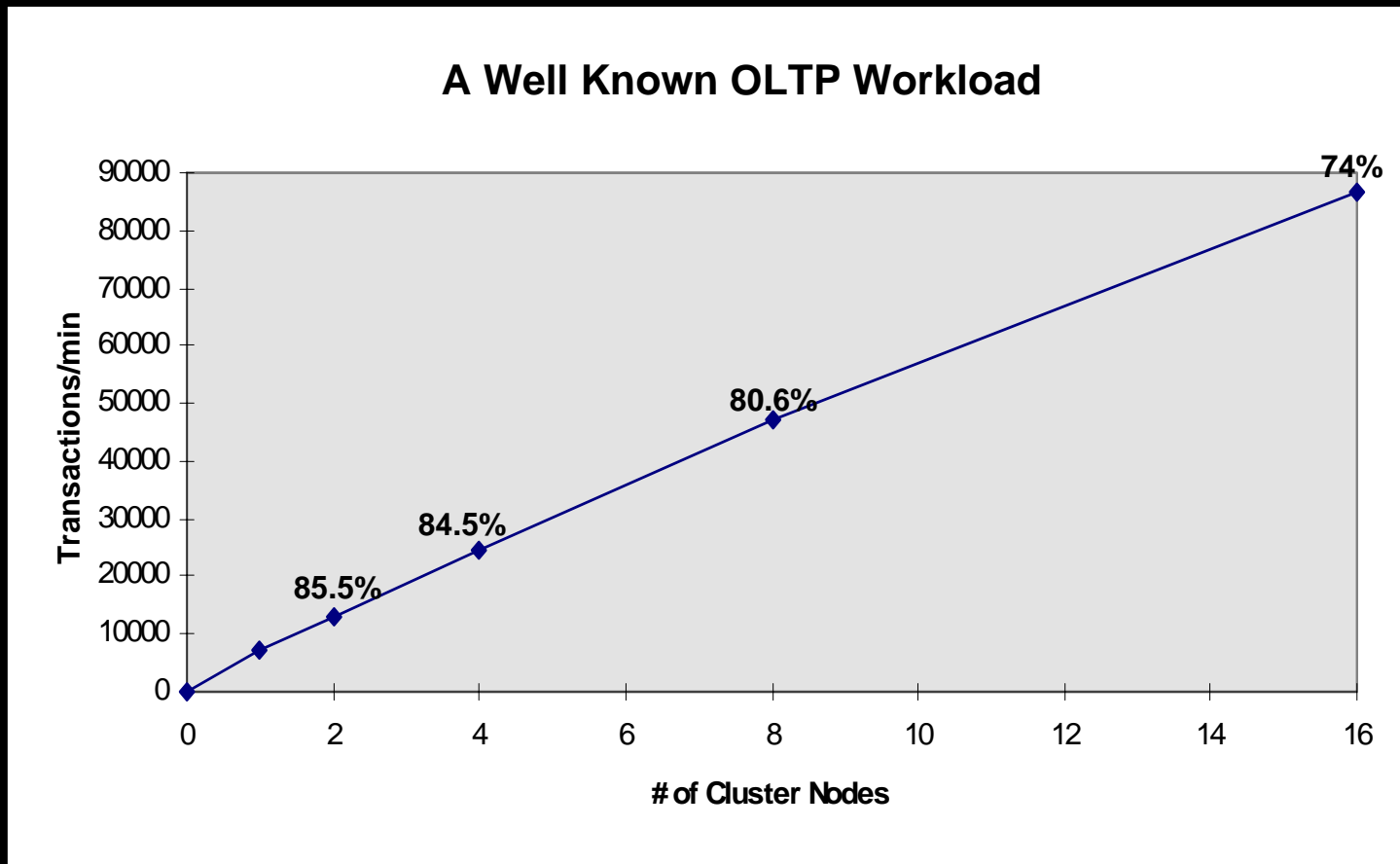
Recovery in a RAC Database

- Survival of one instance guarantees data availability
- Recovery cost is proportional to the number of failures, *not* the total number of RAC nodes
 - cached copies in surviving nodes are used
 - only redo logs from failed instances are applied
- Eliminates disk reads for blocks that are present in a surviving instance's buffer cache.
- Global cache is available after an initial log scan, well before redo application begins.

RAC Support for DSS Workloads

- Cluster-Aware Cost Based Optimizer
 - cluster topology, storage parallelism of an object and disk affinities are considered
 - cost of remote vs local execution
- Function Shipping
 - parallel slaves are sent modified SQL queries
 - reduced messaging cost over data shipping

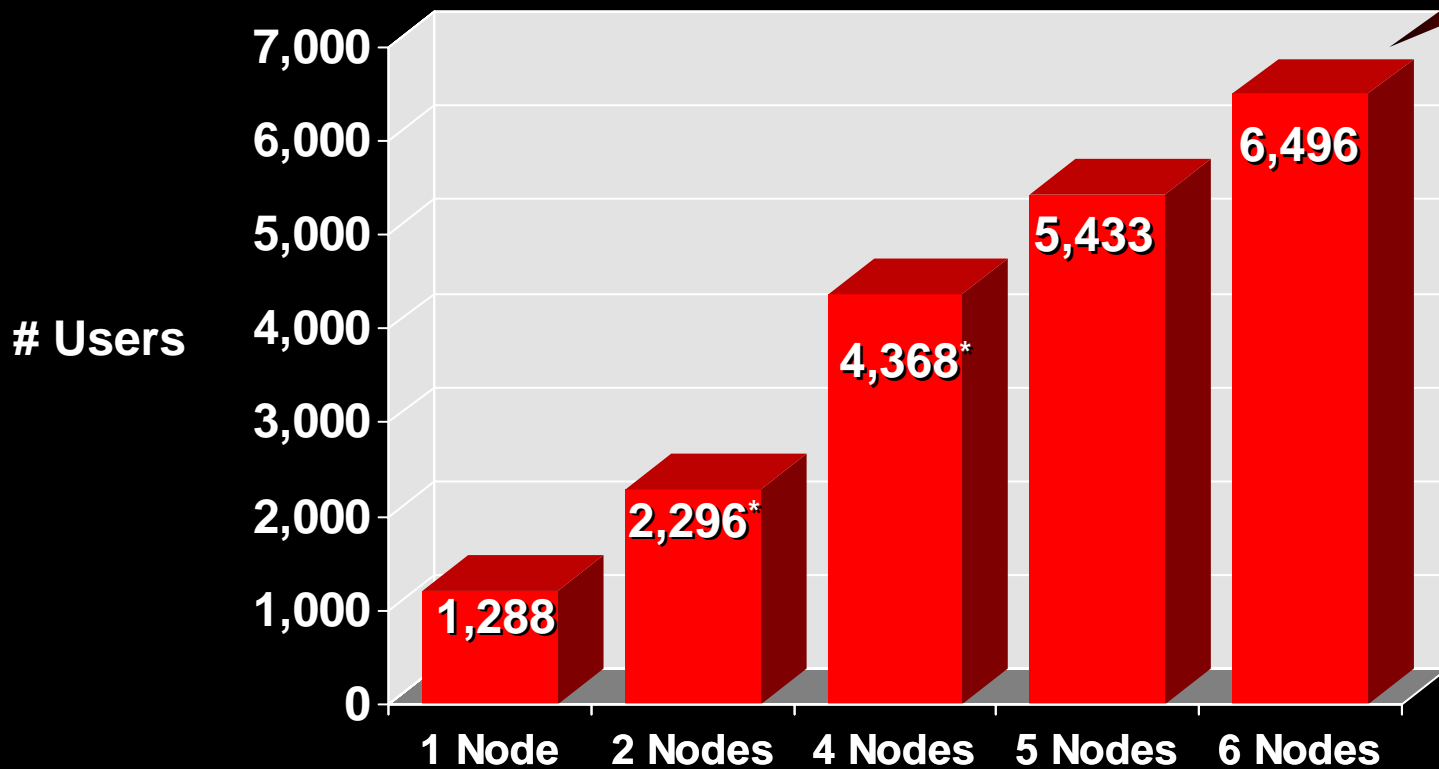
Oracle9i RAC Scalability for OLTP Applications



E-Business Suite Scalability with Oracle9i RAC

Oracle11i E-Business Suite Benchmark

84% Scalability



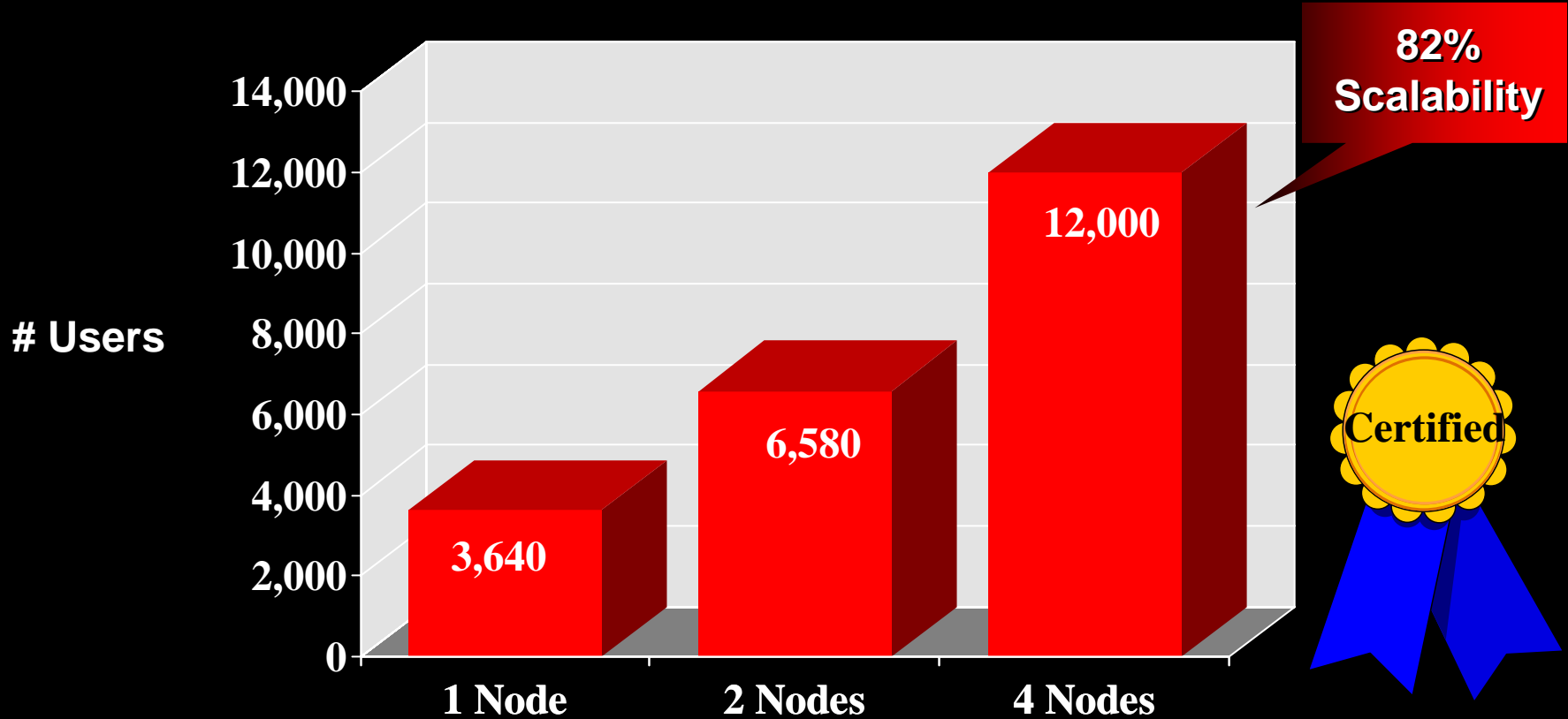
Running on HP Computers

*Audited

ORACLE

SAP Scalability with Oracle9i Real Application Clusters

SD 3-Tier Parallel Benchmark Results Now Official



Running on Compaq AlphaServer Computers

ORACLE

Enhanced IPC

- Global IPC service layer
 - Each requesting server process handles asynch IPC
 - Efficient I/O implementation
- Remote memory operations for direct DMA
 - Exploit modern low latency interfaces
 - Reduced CPU usage
- Intelligent batching of messages
- New V\$ tables for IPC statistics

Lock Simplification and Automation

- Automatic DLM configuration
 - Automatic derivation of DLM configuration
 - No INIT.ORA lock parameters required
 - Improved lock efficiency and memory management

Database Challenges:

Availability

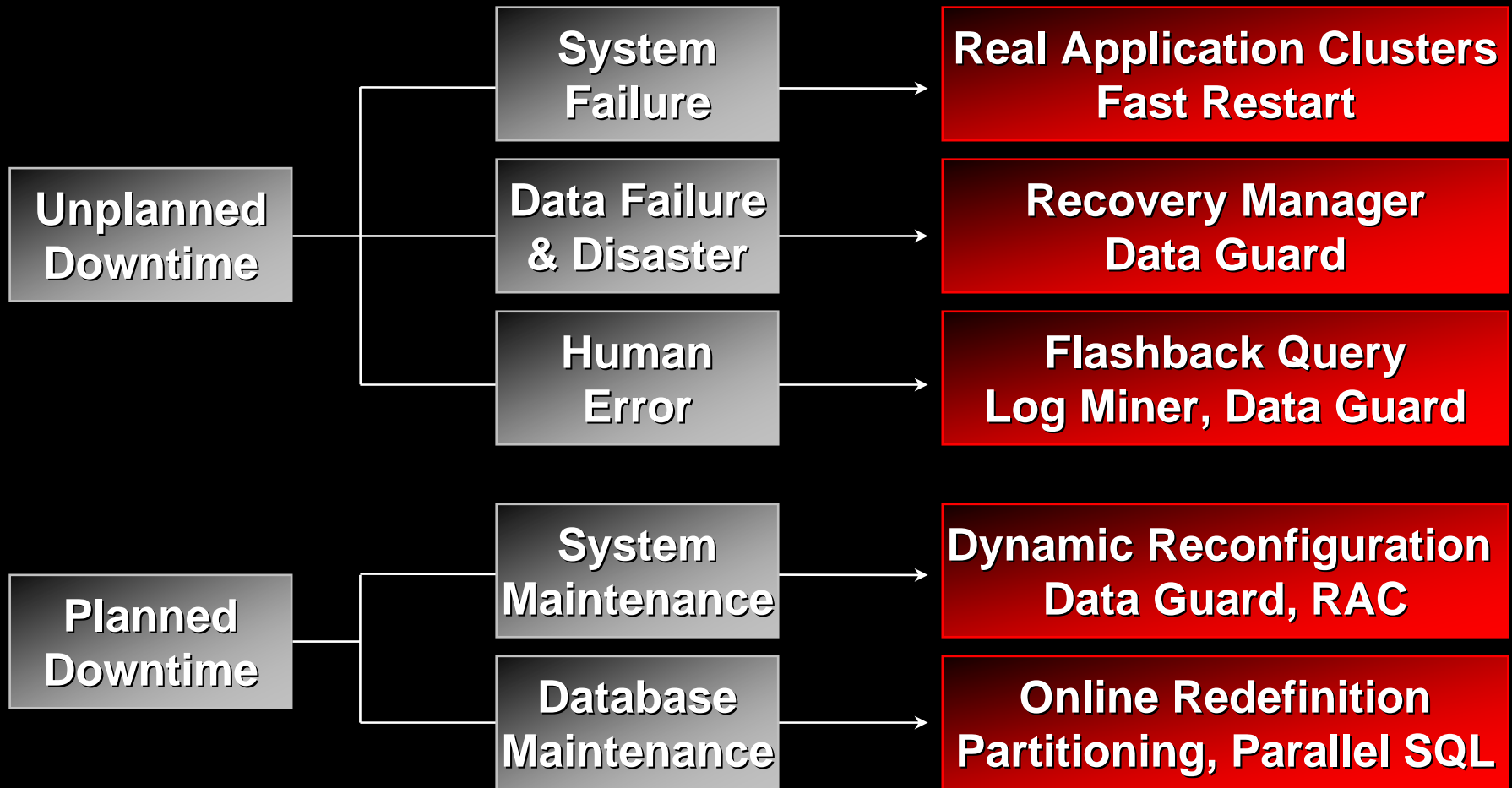
- Be available all the time
 - design for fault tolerance
 - provide fast and reliable fault recovery
 - eliminate maintenance downtime
- Support mission-critical business operations



Real Application Clusters



Oracle9i Handles all Causes of Downtime



The Most Critical HA Issue

- Have a suitable test environment
- Establish HA processes
- Practice HA procedures
- Enforce an HA mentality
- “We can throw all the technology in the world at our customers, but we can’t teach them discipline”

No Single Point of Failure

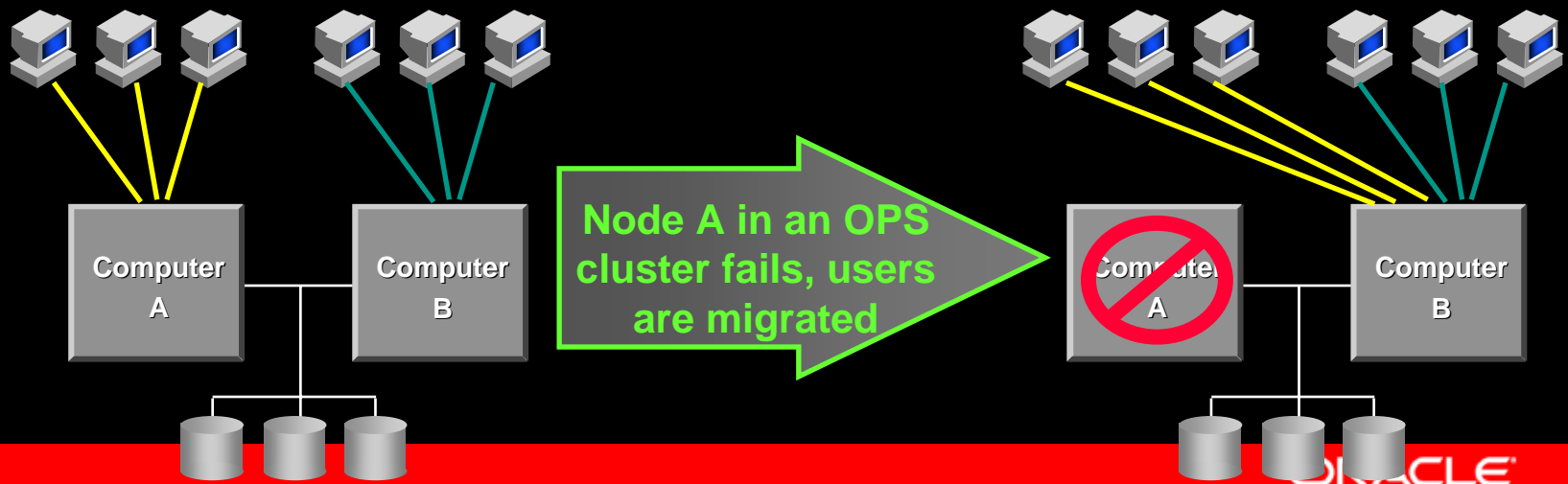
- Real Application Clusters
 - Exploits clusters for very high availability
 - Overcomes the limitations of traditional failover solutions by providing:
 - Concurrent processing
 - Load Balancing
 - Fast time to recovery
- No single point of failure
 - Network, Storage
 - CPU, OS

NEW in Oracle9i: Availability Features

- Reduced time to recovery
 - Concurrent lock reconfiguration and instance (cache) recovery
 - Smarter and more efficient lock reconfiguration
 - Lock replay only for dead masters/locks
 - Batching of reconfiguration messages
 - Deferred/lazy lock remastering
 - Parallel replay processing by multiple LMSs
 - Optimal lock invalidation
 - Optimized special reconfiguration case (l.e. 2->1)
- Fast/reliable detection of node/network failure

High Availability with T.A.F

- Little or no user downtime
- Applications and users are automatically and transparently reconnected to another system
- Queries can continue uninterrupted
- Login context maintained
- DML transactions are rolled back



Database Challenges: **Manageability**

- Create one virtual system to configure and manage
- Single system image for the database integrated with the cluster



Real Application Clusters

Real Application Clusters Manageability

- Single system Image
 - Single Oracle Database
 - One virtual system to configure and manage
 - Single management console
- Cluster-wide monitoring and diagnostics
- Fewer tunable parameters
 - No more GC_FILES_TO_LOCKS !!!

NEW in Oracle9i: Manageability Features

- Improved Single System Image
- Exploit cluster file systems
- First pass diagnostics capability

NEW in Oracle9i: Manageability Features

- Improved tools infrastructure for management
 - Oracle Universal Installer (OUI)
 - Enterprise Manager (OEM)
 - Database Configuration Assistant (DBCA)
 - Net Assistant (NetCA)
 - Recovery Manager (RMAN)
 - Command line interface (SRVCTL)
 - Cluster aware diagnostics (ORADEBUG)

Improved Configuration

- Centralized, persistent configuration storage
 - Eliminates consistency problems with the per node text file-based Parallel Server configuration in prior versions
 - Utilities to migrate previous Parallel Server configurations
- Enhanced DBCA and NetCA functionality
 - Robustness
 - Capability to add and delete instances

Improved Administration

- Dynamic cluster node addition and deletion
 - Add/delete a node in both the system and the database
- Better integration with Oracle Enterprise Manager
 - View and update server side initialization parameter file (SPFILE)
 - Assign private rollback segments to database instances

Improved Administration

- Better integration with OEM (Contd.)
 - Enhanced monitoring capability and events available with OEM and performance packs
 - Cache Fusion statistics
 - Statistics for interconnect block traffic
 - Statistics for the entire database or per instance
 - New EM events associated with new statistics

Improved Diagnosability

- Server side capability
 - First pass analysis on problems
 - Reduces the turnaround time to resolve bugs
 - Enables analysis of intermittent, hard to reproduce problems
 - Reduces need for diagnostic patches

Easy to Use Cluster File system for Windows and Linux

- **Stores all Oracle files (data, logs, executables)**
- **Eliminates need for raw devices**
- **Storage management same as single node**
- **Simplified Backup/restore**
 - **Remove NFS dependency**

Why Oracle RAC for Manageability?

Challenge	Oracle9i	IBM DB2 EEE V7.2	Microsoft SS2000
Single management target	Yes	No	No
Single cluster node limit on Windows 2000	64+	8	24+
Users remain on-line while a node is added	Yes	No	No
Selection of partitioning key unconstrained	Yes	No	No
Disk space usage minimized	Yes	No	No
Downtime to repartition data	None	Days	Days
Users remain on-line if a node fails	Yes	No	No
Users remain on-line during index creation	Yes	No	No
Users on-line during schema change	Yes	No	No

Clusters Are Changing IT Economics

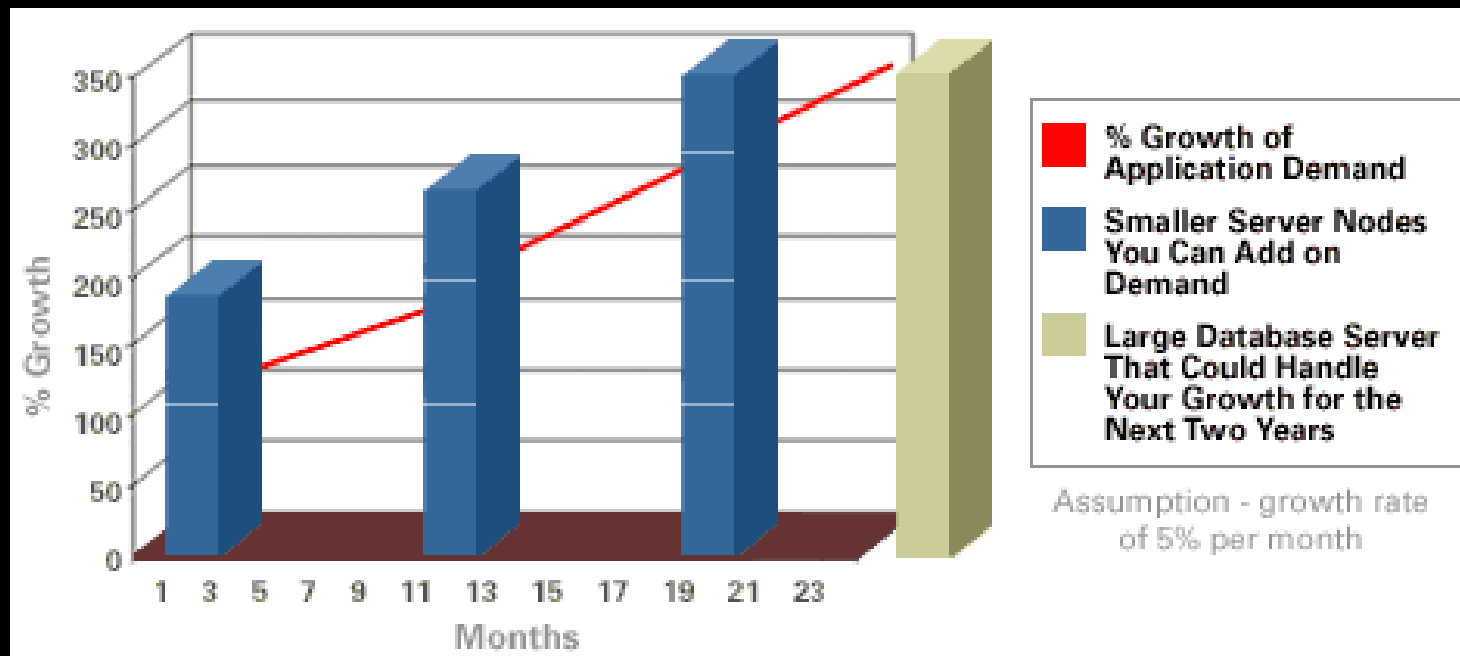
Hardware Costs	Cluster	SMP
Price for 8 Active CPUs	\$96,000	\$234,000
Additional Availability	High	None

Hardware Costs	Cluster	SMP w/ Failover
Price for 8 Active CPUs	\$144,000	\$468,000
Achievable Scalability	> 85%	50%
Additional Availability	High	High

Clusters Are Changing IT Economics

Cluster Configuration	Cost
2 x Sun E10000 (64 CPUs @450MHz, 64Gb)	\$5,200,000
32 x Sun E420R (4 CPUs @450MHz, 4Gb)	\$1,500,000
32 x Compaq DL580 (4 CPUs @700MHz, 4Gb)	\$1,000,000

Capacity on Demand vs Capacity Planning



Oracle9i Delivers Performance and Scalability

- In today's rapidly changing environment:
 - Internet users and transaction volumes grow extremely rapidly
 - Data warehouse systems must support very large data volumes
 - Server consolidation drives the need for large scale systems
- Oracle9i provides the ability to deploy as you grow

Real Application Clusters Scalability and Performance Improvements

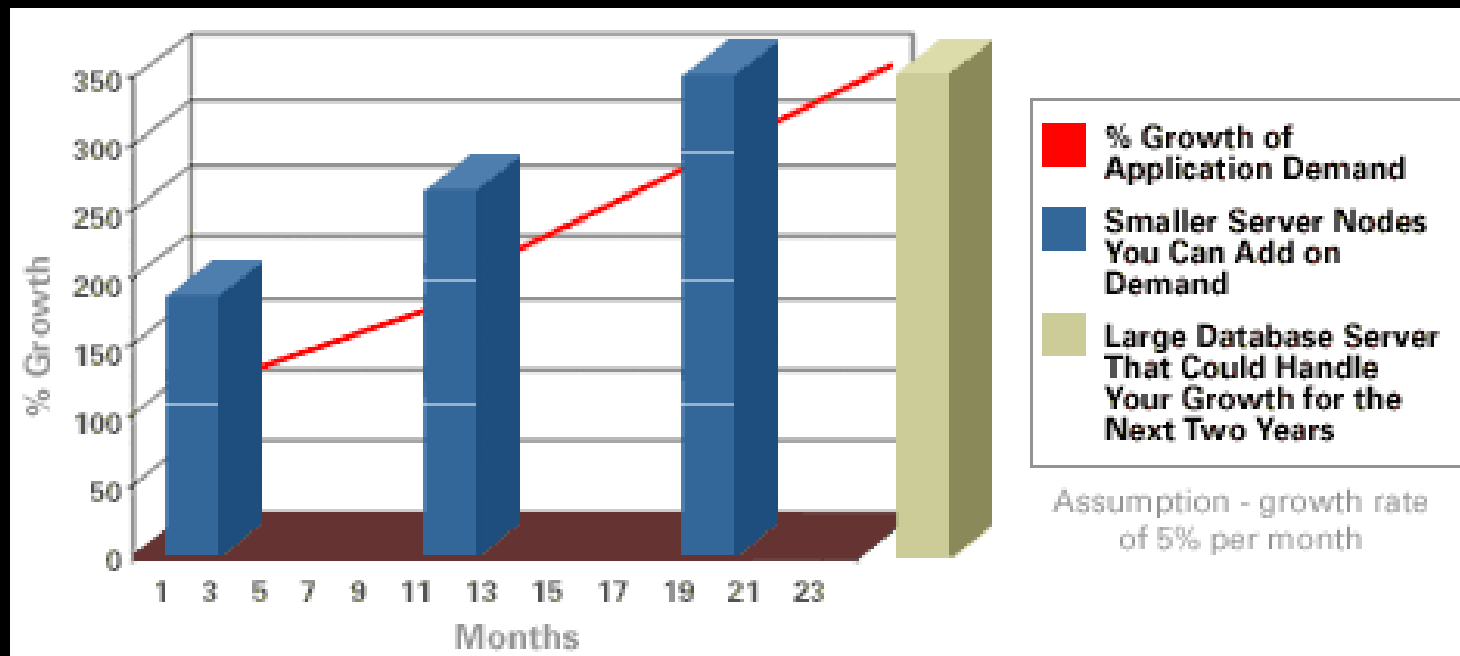
- **Major Reduction in Cluster Messages**
 - **Optimize library cache and row cache latching and multi-node locking**
 - **Direct sends of DLM requests**
 - **Batching of DLM and buffer requests**
 - **Greatly reduce block cleanout**
- **Bitmap segment space management performance improvements remove need to use free list groups**

Clusters Are Changing IT Economics

Hardware Costs	Cluster	SMP
Price for 8 Active CPUs	\$96,000	\$234,000
Additional Availability	High	None

Hardware Costs	Cluster	SMP w/ Failover
Price for 8 Active CPUs	\$144,000	\$468,000
Achievable Scalability	> 85%	50%
Additional Availability	High	High

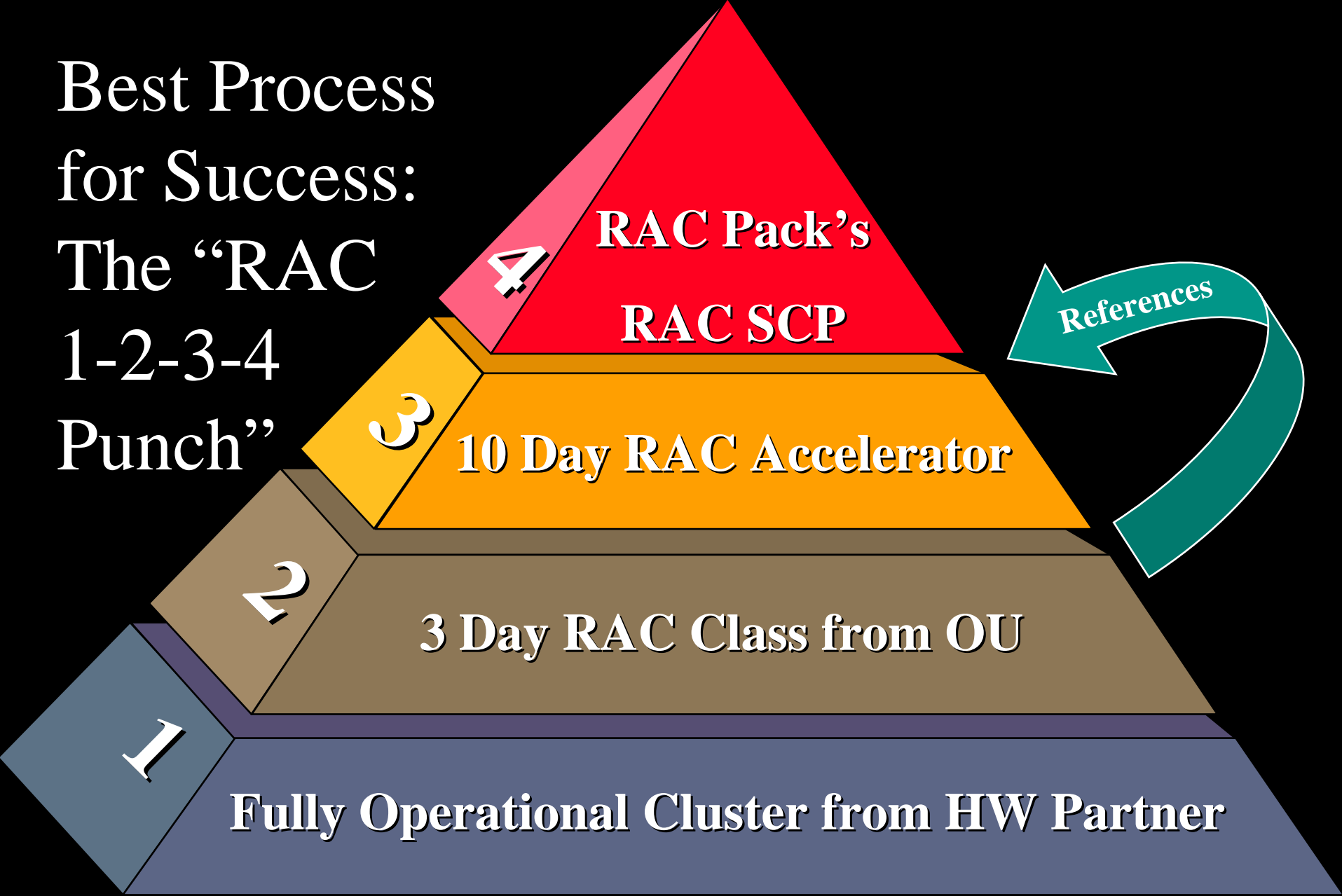
Capacity on Demand vs Capacity Planning



Clusters Are Changing IT Economics

Cluster Configuration	Cost
2 x Sun E10000 (64 CPUs @450MHz, 64Gb)	\$5,200,000
32 x Sun E420R (4 CPUs @450MHz, 4Gb)	\$1,500,000
32 x Compaq DL580 (4 CPUs @700MHz, 4Gb)	\$1,000,000

Best Process
for Success:
The “RAC
1-2-3-4
Punch”



RAC Customers as of May 2003

2,000+ NEW customers

425+ Documented live customers!

124 Live, referenceable customers

- **27 Solution profiles**
- **56 Production press releases**
- **11 Reference forums**
- **6 ROI studies in queue**
- **1 ROI published**

That's all Folks!

ORACLE®

SOFTWARE POWERS THE INTERNET