Linux - a solution for today's enterprise-class computing

Sanhita Sarkar

Oracle Corporation

Sanhita.Sarkar@oracle.com

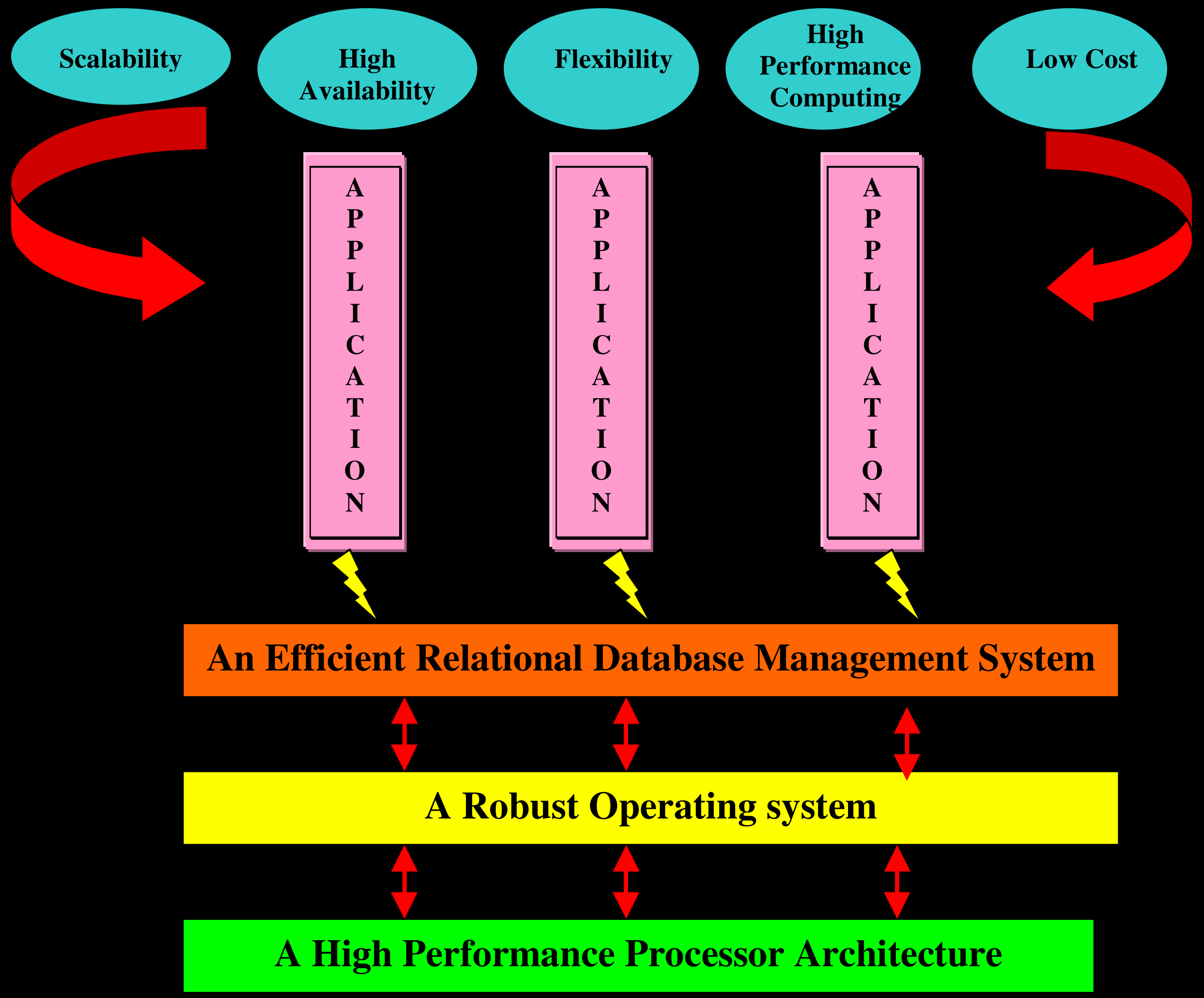
Topics for Discussion

- The challenges of today's enterprises
- The concept of EPIC and evolution of IA-64 architecture
- The Itanium2 Chip Design
- Key enhancements in IA-64 Linux Kernel – Design and Implementation
- Optimizing the Compiler for Itanium2
- Oracle RAC 10g Database on Linux meets all challenges of today's growing enterprises
- Oracle RAC 10g and today's On-Demand concept - Virtualization, Dynamic Provisioning and clustering technology is the enterprise flexibility solution
- HP/Oracle Performance Benchmarks

The challenges of today’s enterprises

- A Processor architecture providing excellent performance for running business applications
- An Operating system with robust, scalable and efficient memory, process and I/O management
- An efficient RDBMS with features capable of exploiting the processor architecture and the operating system kernel
- Scalability of applications as business grows
- High Availability for mission-critical business applications
- Flexibility requirements from OS, Database and hardware for dynamic growth of enterprises
- **Overall goal:** Enterprise business applications always running at peak performance at the lowest cost

Enterprise Challenges



Scalability

High Availability

Flexibility

High Performance Computing

Low Cost

APPLICATION

APPLICATION

APPLICATION

An Efficient Relational Database Management System

A Robust Operating system

A High Performance Processor Architecture

IA-64 : A High Performance Fault-Tolerant Architecture for Enterprise-Class Computing

Solves the challenges of reliable and fast computing

Evolution of IA-64 architecture

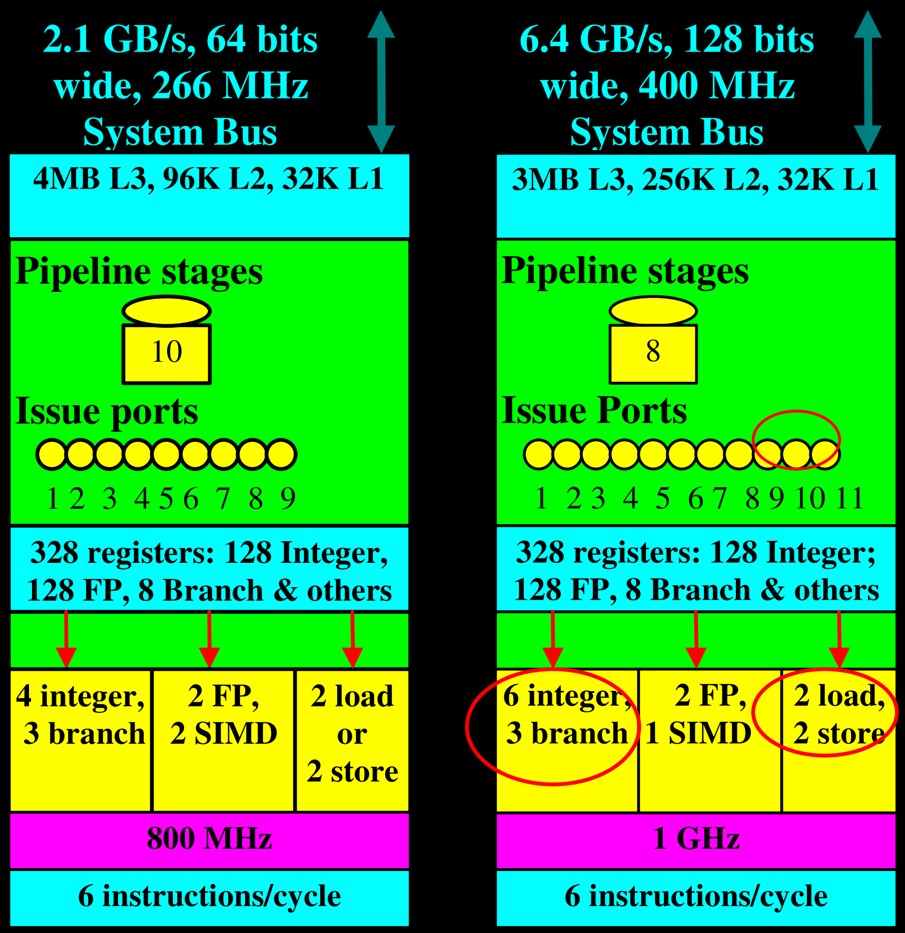
- CISC and RISC architectures – Limitations
- The EPIC concept
- Evolution of IA-64 architecture based on Itanium
- Requirement of efficient compilers running on Itanium
- Predication, Control and Data Speculation, Software Pipelining, Data Prefetching
- A fully predicated architecture providing Fault Tolerance and Reliability

Itanium2 Chip Design

- Three-level Cache Organization
- Memory hierarchy
  - Efficient use of the three cache levels
  - Latencies of L1, L2 and L3 caches

- Large number of registers and larger register width
- Large number of floating point units
- Large Cache, prefetching, elimination of branch predictions
- HP Scalable Processor Chipset zx1
  - Provides a low-cost, low-latency, high-bandwidth connection between the processor, memory and I/O.
  - Allows effective transfer rate of 400M transfers/sec to 6.4 GB/s
  - zx1 can support up to 4 processors off the McKinley bus and a variety of I/O configurations
  - Provides an efficient clustering solution

Itanium vs Itanium 2
Chip Architecture

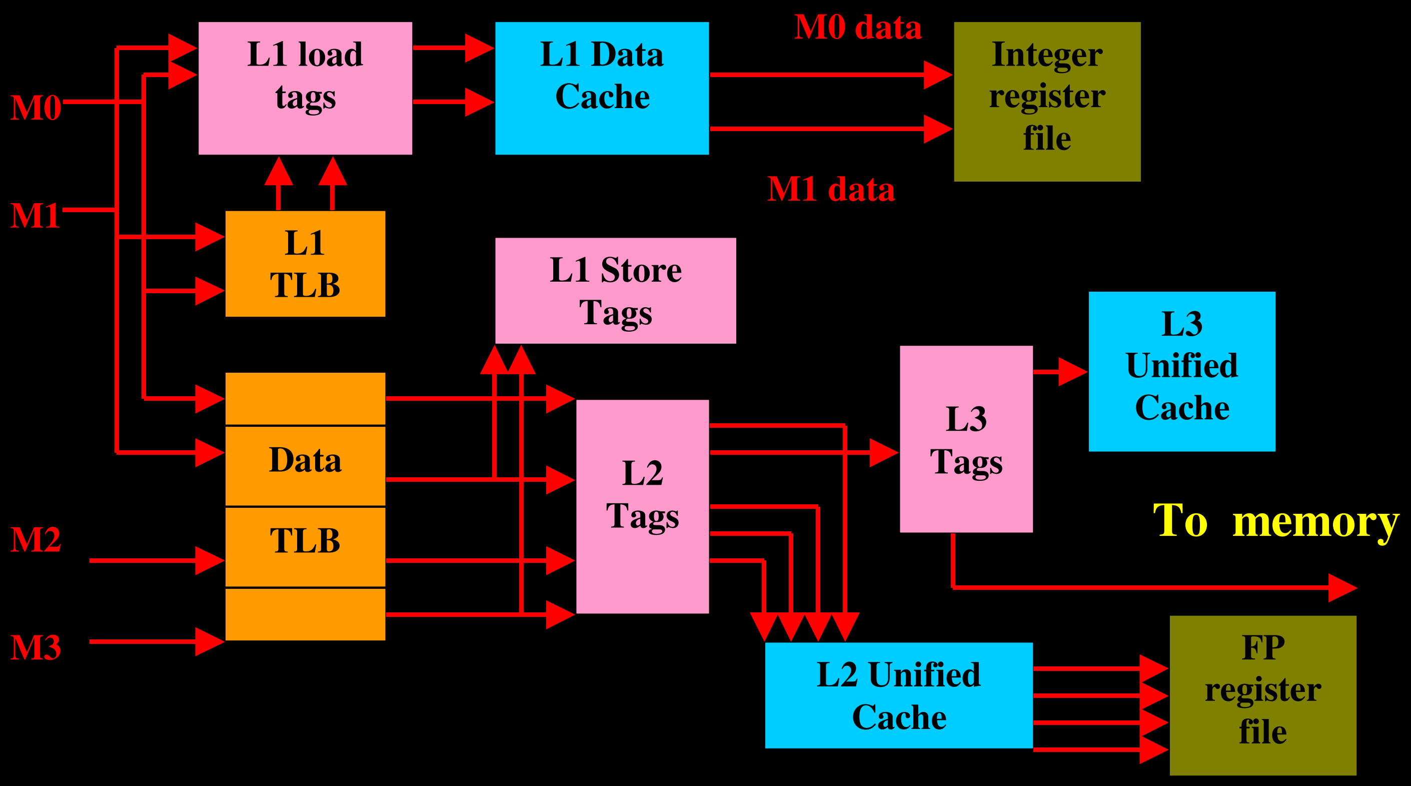


Itanium processor	Itanium2 processor	
2.1 GB/s, 64 bits wide, 266 MHz System Bus	6.4 GB/s, 128 bits wide, 400 MHz System Bus	3x increase System bus bandwidth
4MB L3, 96K L2, 32K L1	3MB L3, 256K L2, 32K L1	Large cache, Reduced latencv
Pipeline stages 10	Pipeline stages 8	
Issue ports 1 2 3 4 5 6 7 8 9	Issue Ports 1 2 3 4 5 6 7 8 9 10 11	Additional Issue Ports
328 registers: 128 Integer, 128 FP, 8 Branch & others	328 registers: 128 Integer; 128 FP, 8 Branch & others	
4 integer, 3 branch	6 integer, 3 branch	Additional execution units
2 FP, 2 SIMD	2 FP, 1 SIMD	
2 load or 2 store	2 load, 2 store	
800 MHz	1 GHz	Increased core frequency
6 instructions/cycle	6 instructions/cycle	

Latest Itanium2 has 6MB L3, 1.5GHz throughput, 50% higher application performance

HP WORLD 2004
Solutions and Technology Conference & Expo

Memory Hierarchy : 3-level Cache Organization



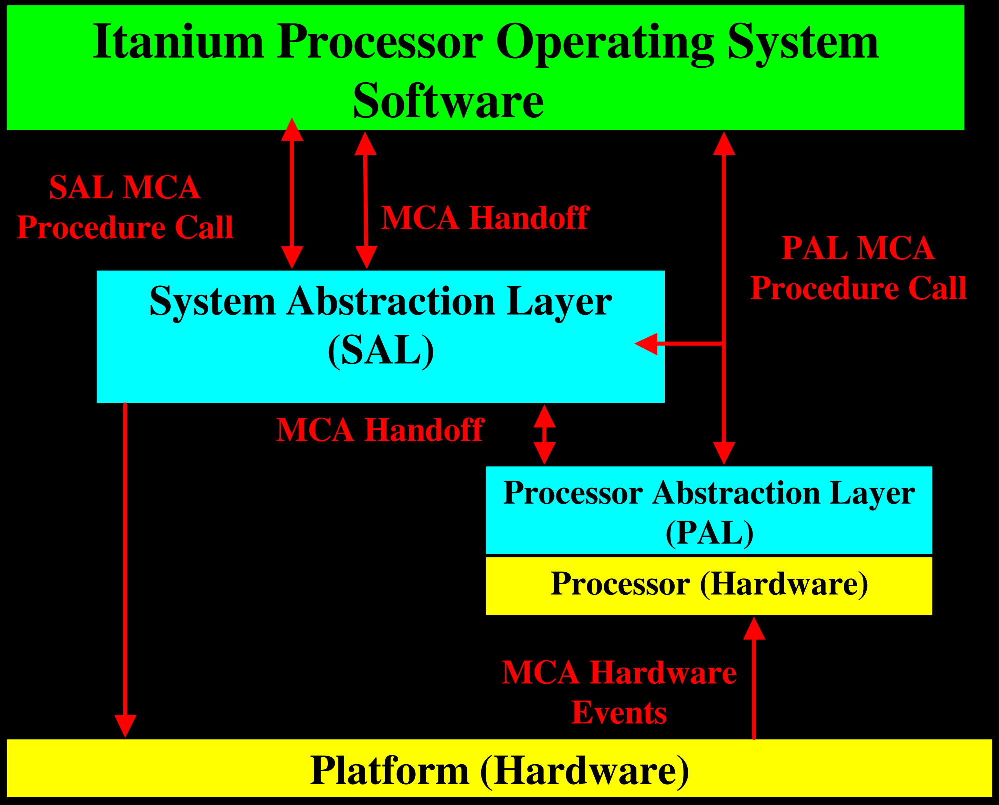
Itanium2: The cache Levels and their Effective use

- The three-tiered cache levels provide a balanced trade-off between speed and size
- Both data and instructions need to be in the smallest and fastest cache.
- Prefetching instructions and data in the proper cache at the proper time enhances performance
  - Too much prefetching is bad – it will cause “hot” cache lines to be flushed from the cache.
  - Organize the data structures so that data that is accessed together is also grouped within the same cache line.
- Segregate integer and floating point variables so they don’t fall in the same cache line

Itanium2: Performance Monitoring Unit (PMU)

- Interface to PMU in Itanium2 consists of specialized registers programmed to capture occurrences of certain architectural events – *pmd* and *pmc*
- PMU has advanced monitoring features like Event Address Registers allowing to pinpoint performance problems (>= 300 events)
- Looking at cycle accounting shows how the processor is spending its time in the code
  - **Unstalled execution** – useful work (+50% -> well tuned)
  - **Data access** – cycles spent waiting for data to be retrieved from memory hierarchy (< 40% -> well tuned)
  - **Instruction access** – cycles spent waiting for instructions to be retrieved from memory hierarchy (< 5% -> well tuned)
  - **Branch miss-predicts** – cycles spent recovering from miss-predicted branches (< 5% -> well tuned)
  - **Register Stack engine (RSE)** – cycles spent spilling and filling registers during procedure calls (< 5% -> well tuned)
  - **Scoreboard** – cycles spent waiting for registers/functional units (< 5% -> well tuned)

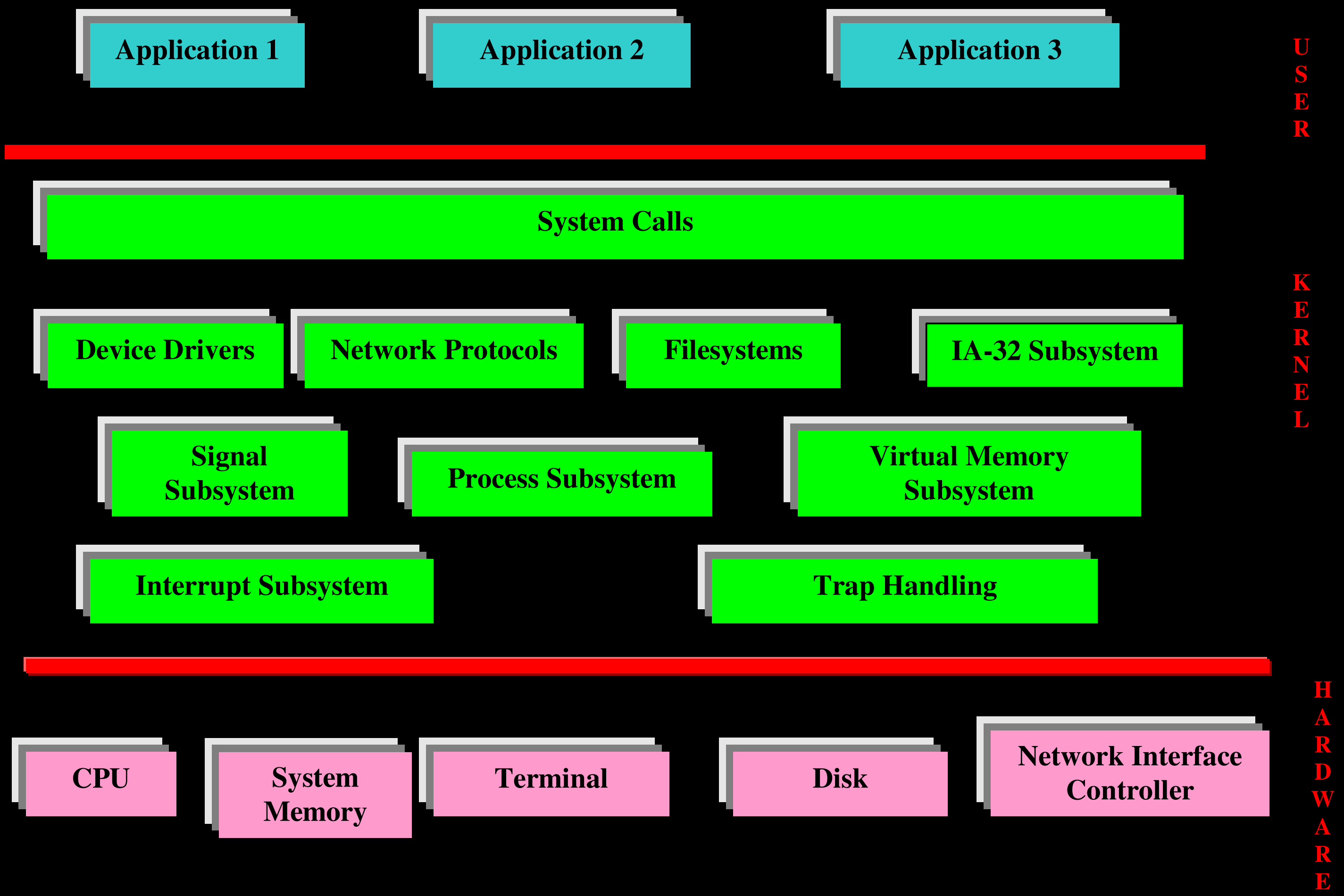
Error Handling by PAL and SAL on Itanium 2



Linux – An operating system for today’s enterprises

The IA-64 Linux Kernel meets the High Performance Computing challenges of enterprise applications

The IA-64 Linux Kernel - Overview



Application 1

Application 2

Application 3

USER

System Calls

Device Drivers

Network Protocols

Filesystems

IA-32 Subsystem

KERNEL

Signal Subsystem

Process Subsystem

Virtual Memory Subsystem

Interrupt Subsystem

Trap Handling

CPU

System Memory

Terminal

Disk

Network Interface Controller

HARDWARE

IA-64 Linux Kernel - Enhancements

- Enhanced Linux I/O Subsystem
  - Asynchronous I/O
    - Allows processes to submit I/O requests without waiting for I/O completions
    - Non-blocking semantics of AIO allow Oracle database writer to quickly write dirty buffers from database buffer cache to disk upon notification of I/O completion
    - Allows Oracle processes to issue multiple I/O requests with a single system call, rather than a large number of distinct I/O requests
  - Global Lock elimination
    - In earlier Linux releases, I/O requests were queued one at a time while holding a global lock (io_request_lock), used for the entire device block subsystem
    - Fine grained locking now allows a separate lock for each individual block device
    - I/O requests now queued holding a lock specific to the request queue
    - Scalable concurrent I/O queuing scheme allows better I/O throughput under heavy Oracle database load
  - Variable Block I/O Sizes
    - In earlier Linux releases, I/Os were broken up into several blocks with max size of 4K
    - Raw I/O requests of 512-byte units can now be submitted as one single request with variable block sizes to a max of 16K.
    - Improves performance and scalability of I/O operations

Enhancements Contd…..

- **Virtual memory: Huge TLB Pages**
  - Few TLB entries on a processor results in high TLB miss rate for Oracle applications accessing large amounts of memory
  - Required each TLB mapping very large to reduce processor's TLB
  - The feature, Huge TLB pages, allows applications to benefit from large pages
  - With large page support, a processor now deals with more memory for each page table entry -> page table entries well cached -> TLB miss rate decreases significantly
  - A single page table entry addressing vast amounts of memory -> lower number of page table entries -> reduces page table size -> reduces memory usage by page tables
  - Eliminates swapping physical pages ->entire buffer cache locked in physical memory -> enhanced system performance

Enhancements Contd…..

- Enhanced Process Scheduler
  - Doubles the number of useable processors
  - Supports up to 16 logical CPUs (or 8 hyperthreaded CPU pairs)
  - Support for a timed semaphore call (semtimedop) allows a large timeout value for operations previously using a timer function
  - Oracle applications benefit in performance
  - High user applications would see the most benefit from this feature

Optimized Compiler Instructions

- IA-64 Compiler exploits the features of the IPF architecture in the best possible way

  - Speculation : control and data
  - Predication
  - Pipelining

- **Control Speculation:** **executes instructions prior to conditions**

```
int a,b;                        ld.s       t1 = [p] ;
extern int *p;                  add        b = t1,2
extern int global;      cmp.ne.unc p1,p0 = condition,0 ;;
if(condition) {             (p1) chk.s      b, L2
a = global;                 L1:
b = *p + 2;                     ...
}                           L2: ld         t1 = [p] ;;
                                add        b = t1,2
                                br         L1
```

Optimized Compiler Instructions

- **Data Speculation:** executes load prior to store instructions

```
int a,b;

extern int *p;
extern int *q;
*p = a;
b = *q + 2;
```

```
        ld.a     t1 = [q];;

        add      b = t1,2
        st       [p] = a
        chk.a    b, L2
L1:
...
L2:     ld       t1 = [q];;
        add      b = t1,2
        br       L1
```

Optimized Compiler Instructions

- **Data Predication:** control dependency to data dependency

```
          if (a == 0) {
              x = 5;
          } else {
              x = *p;
          }
```

```
Using Branches                  Using data predication
cmp.ne.unc p1,p0 = a,0 ;;     cmp.ne.unc  p1,p2 = a,0 ;;
(p1) br    L1 ;;              (p2)  mov         x = 5
     mov   x = 5              (p1)  ld          x = [p]
     br    L2 ;;
L1:  ld    x = [p]
L2:
```

- High-level compiler optimizations: +O3, +O4, etc
  - **Helps inlining, faster numerical codes, faster code for math library functions, etc**

The Scalability and High Availability Challenges - Oracle RAC 10g is a solution

Scalability and High Availability Challenges

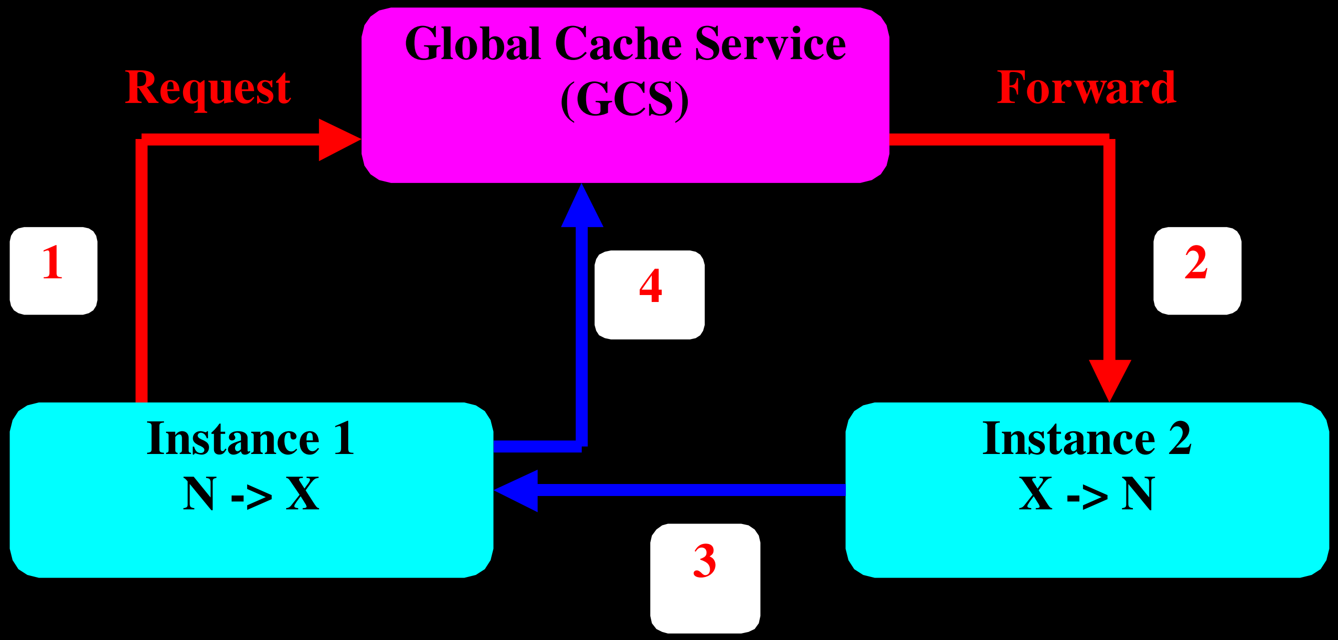
- Traditional IT infrastructures are not making the best use of available computing resources.
- Applications are independently constructed, custom configured, and sized only for the peak load.
- Trade-off between scalability for peak and lower investment to avoid idle capacity.
- As business expands, the cost of incrementally adding capacity to high end SMP is enormous – tens to hundreds of thousands of dollars, if not more.
  - Avoid adding capacity, trying to barely get by at processing peaks.
  - Avoid too much idle capacity sitting around most of the time.
- **In other words, the bigger the box, the greater the incremental cost of ‘scaling up’**
- Enterprise runs mission-critical applications. Downtime not a possibility. Disaster Recovery and Data protection – an absolute requirement.

Oracle RAC 10g: Scalability and High Availability

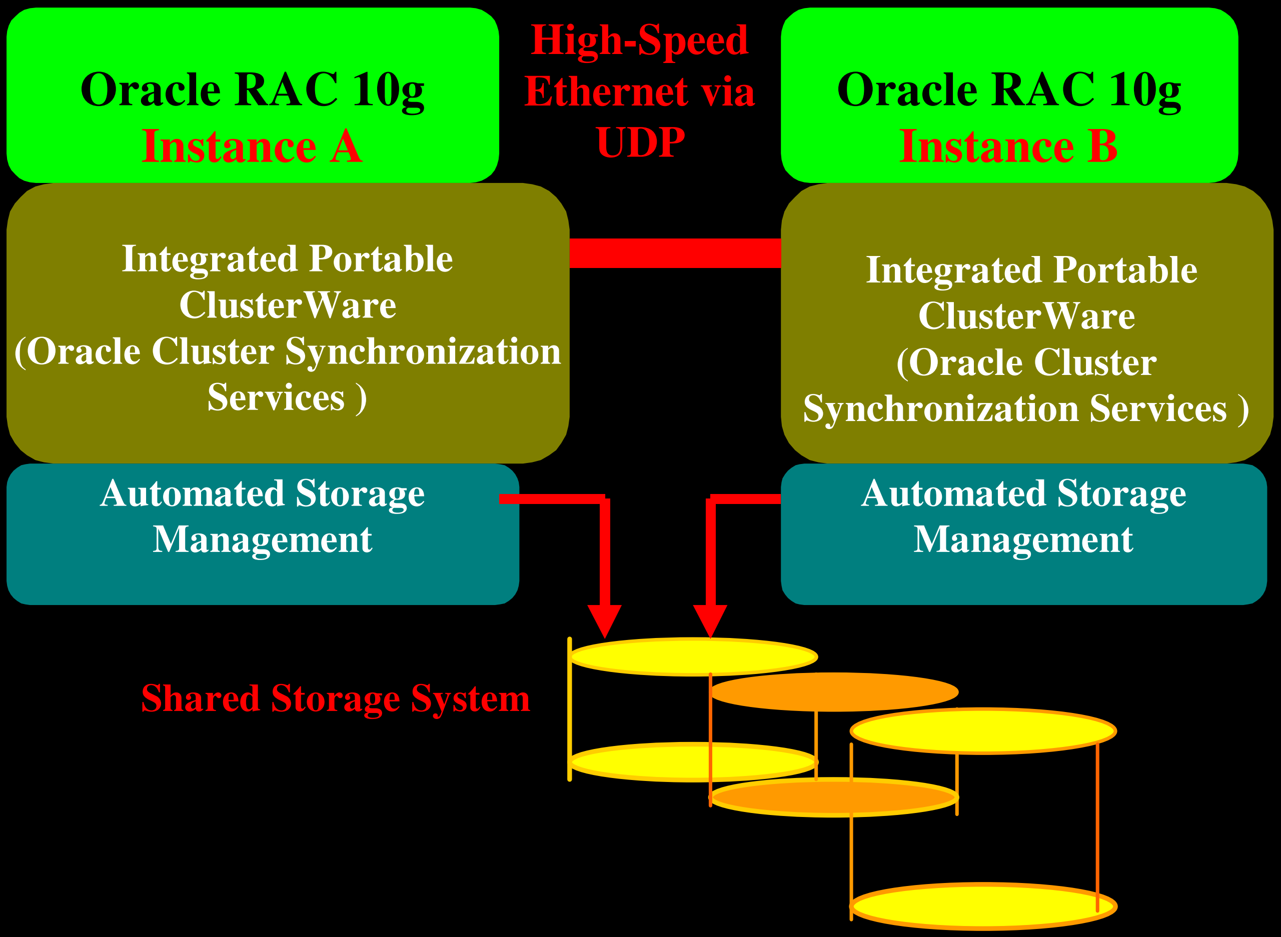
- Cache Fusion Technology
  - provides an expanded database cache for queries and updates
  - reduced disk I/O synchronization which overall speeds up database operations

- Oracle Integrated Portable Clusterware
  - Cluster Ready Services, Cluster Synchronization Services
  - Cluster Ready Services with enhanced coordination of the Global Cache Service (GCS) leads to advantages like resource affinity, data integrity, application transparency and fault tolerance
  - Cluster Synchronization Services allow Node Monitoring and Heartbeat Monitoring for split-brain configurations across the network
  - Supports vendor clusterware

Cache Fusion: requesting a changed block for modification



Oracle RAC 10g: Shared Disk Architecture



Oracle RAC 10g
Instance A

High-Speed Ethernet via UDP

Oracle RAC 10g
Instance B

Integrated Portable ClusterWare (Oracle Cluster Synchronization Services)

Integrated Portable ClusterWare (Oracle Cluster Synchronization Services)

Automated Storage Management

Automated Storage Management

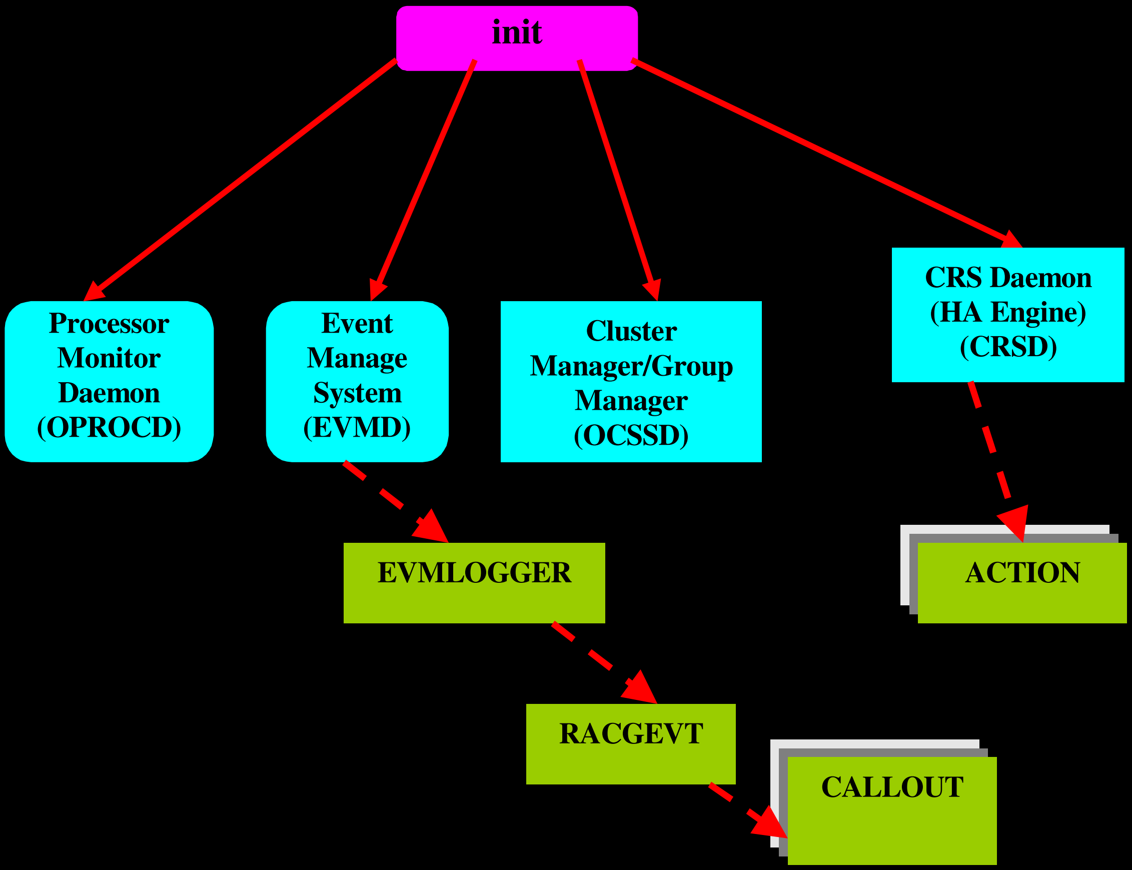
Shared Storage System

Oracle RAC 10g: Cluster Ready Services

The Integrated Portable Cluster Ware
- a unique solution to all needs of
Cluster Manageability and High Availability

HP WORLD 2004
Solutions and Technology Conference & Expo

Oracle Cluster Ready Services: The Run Time View


init
Processor Monitor Daemon (OPROCD)
Event Manage System (EVMD)
Cluster Manager/Group Manager (OCSSD)
CRS Daemon (HA Engine) (CRSD)
EVMLOGGER
ACTION
RACGEVT
CALLOUT

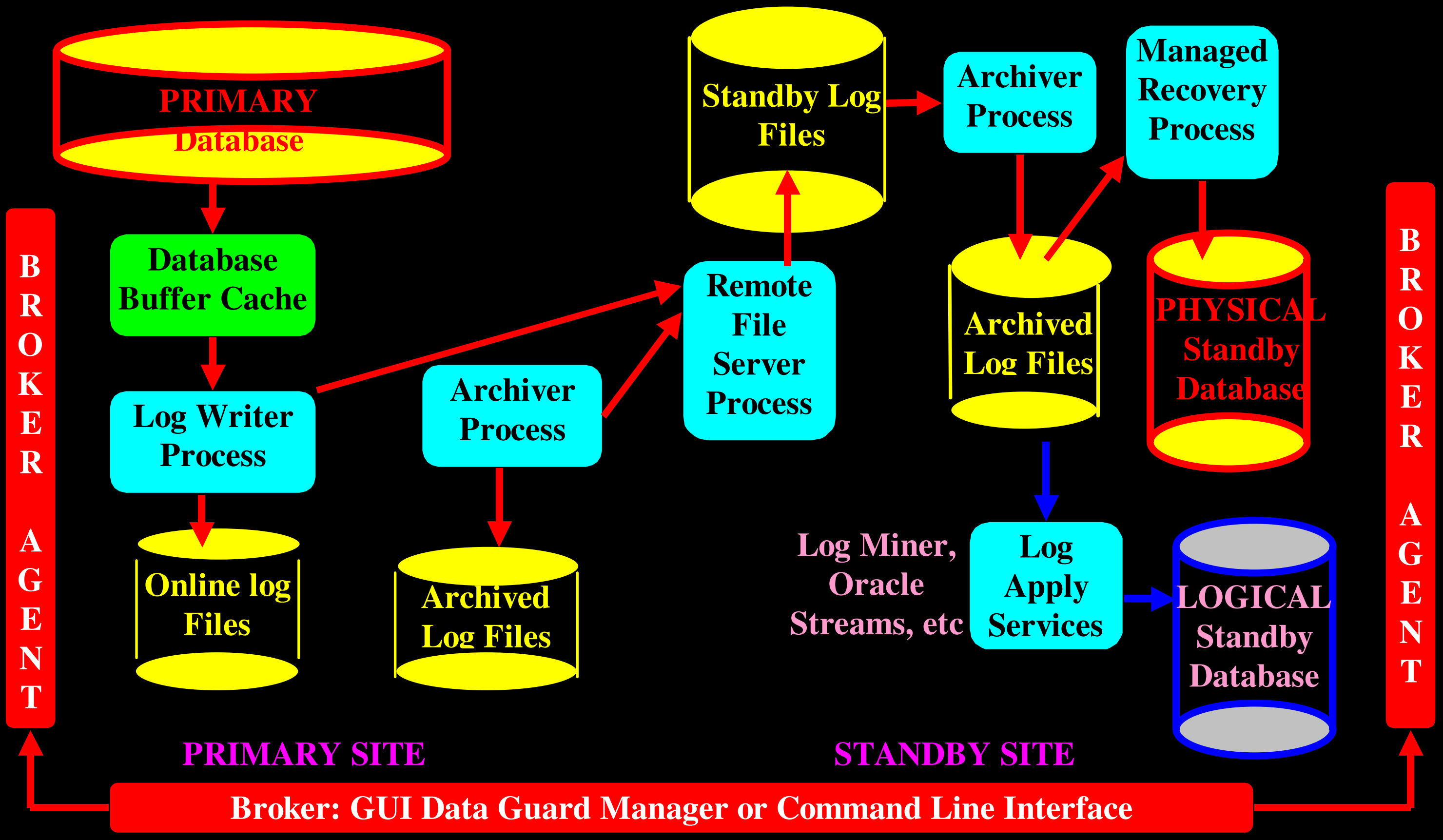
Oracle RAC 10g: High Availability

- Oracle RAC 10g provides high availability
  - Node and instance failover in seconds
  - Integrated and intelligent connection and service failover across various instances
  - Planned node, instance, and service switchover and switchback
  - Rolling patch upgrades
  - Multiple active instance availability and scalability across multiple nodes
  - Comprehensive manageability integrating database and cluster features

Oracle 10g Data Guard – a solution for Disaster recovery, Data Protection and High Availability

Now extended support for RAC :
RAC Primary databases and RAC Standbys
- a more robust enterprise solution

Oracle 10g Data Guard - supports RAC



HP WORLD 2004 Solutions and Technology Conference & Expo

Challenges of Flexibility for Growing Enterprises

Virtualization and Dynamic Provisioning – The “Adaptive” Concept – an enterprise solution

- “Virtualization” - breaking the hard coded association of resources to applications
- “Provisioning” - dynamically making resources available to applications when they need them
- Provisioning more resources at peak, and then re-provisioning them elsewhere when the peak is past
- Adoption of Blades:
  - addresses the needs of large scale computing centers to reduce space requirements for application servers and lower costs
  - Can be added as required, often as "hot pluggable" units of computing as they share a common high speed bus
- Re-architecting the IT infrastructure of the enterprise is an easy migration – not a start over – incremental steps only
- Clustering is the best solution - Server blades are interconnected, so they form a cluster. Ease of capitalization on this new standard hardware enable pooling of compute resources

Oracle RAC 10g provides Flexibility

- Oracle RAC 10g lets one run everything on industry standard, modular components – the Server Blades - no engineering for SMP, a lot cheaper, have the newest CPUs - so they're faster
- Oracle RAC 10g - designed to dynamically reallocate resources as needed. Responds to changing business conditions with no downtime
- Ease of alignment with the business - staying aligned as the business changes
- Higher availability of applications by eliminating server and storage as single points of failure - centralized user provisioning instead of fragmented security
- Solution of incremental 'scale up' cost – Oracle RAC 10g on Server Blades - More information, more processing, higher availability, scalability and lower cost. No trade-off

Flexibility of Storage Management – a challenge for Growing Enterprises

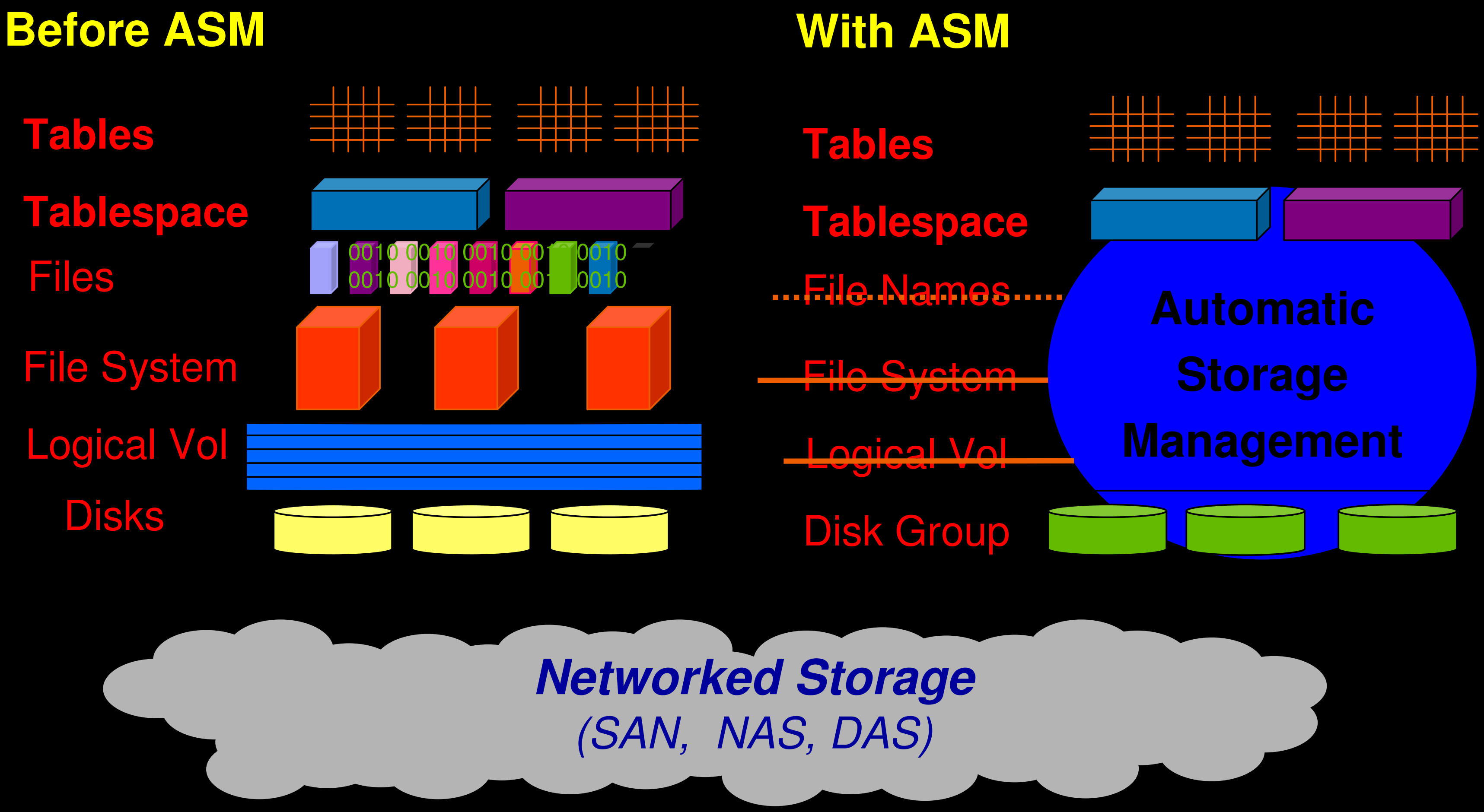
The Storage Management Challenge

- Demand for database uptime along with the increasing rate of database size is a continuous challenge
- Disk density and database size keep doubling leading to challenges in storage management
- Provisioning storage for a single database instance can be complex
  - Estimate of disk space, mapping out data files to avoid hot spots, creating logical volumes, create file systems, load database, etc
  - Moving data files around to reduce contention, add more disks, rebalance again
- Manual I/O performance and tuning tasks by DBAs

Oracle Database 10g allows Storage Virtualization

- Oracle Database 10g provides Automatic Storage Management (ASM)
- ASM virtualizes storage into a set of disk groups and provides redundancy options to enable a high level of protection
- ASM facilitates non-intrusive storage configuration changes with automatic rebalancing
- It spreads database files across all available storage to optimize performance and resource utilization
- It is a capability that saves DBA’s time
  - By automating manual storage management
  - By increasing their ability to manage larger databases and more of them with increased efficiency

Storage Flexibility with ASM



Before ASM

Tables

Tablespace

Files

File System

Logical Vol

Disks

With ASM

Tables

Tablespace

File Names

~~File System~~

~~Logical Vol~~

Disk Group

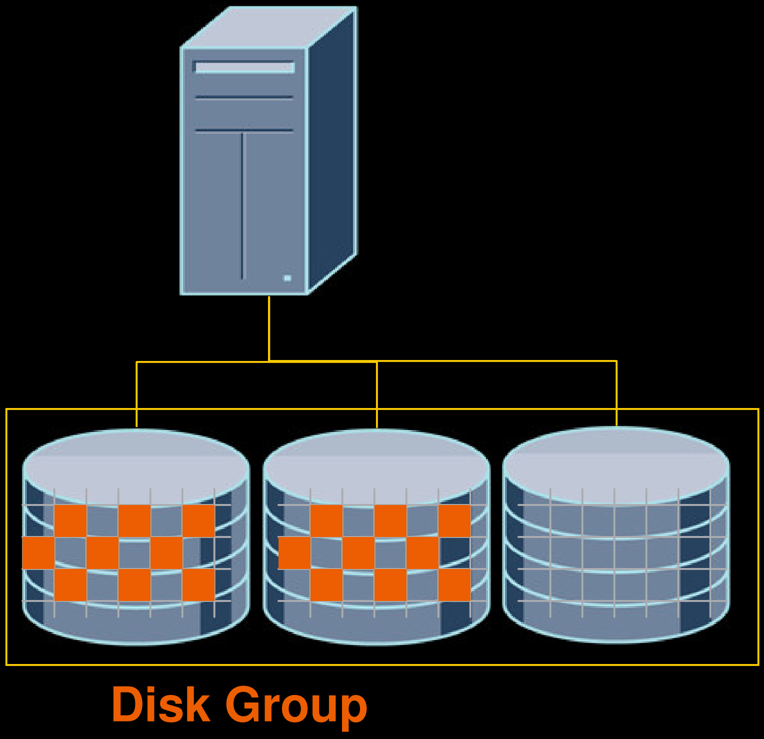
Automatic Storage Management

Networked Storage

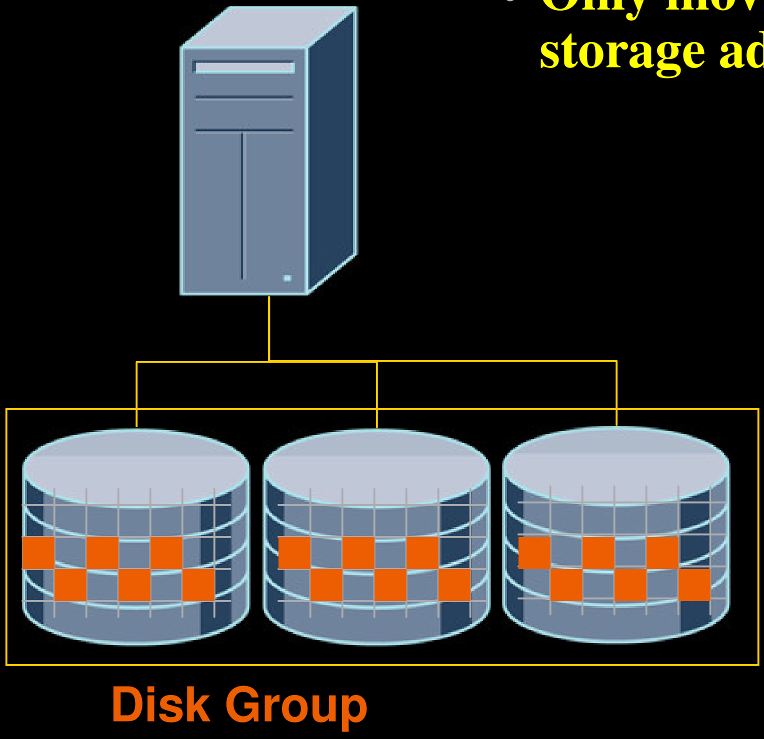
(SAN, NAS, DAS)

Dynamic Rebalancing with ASM

- **Automatic online rebalance whenever storage configuration changes**

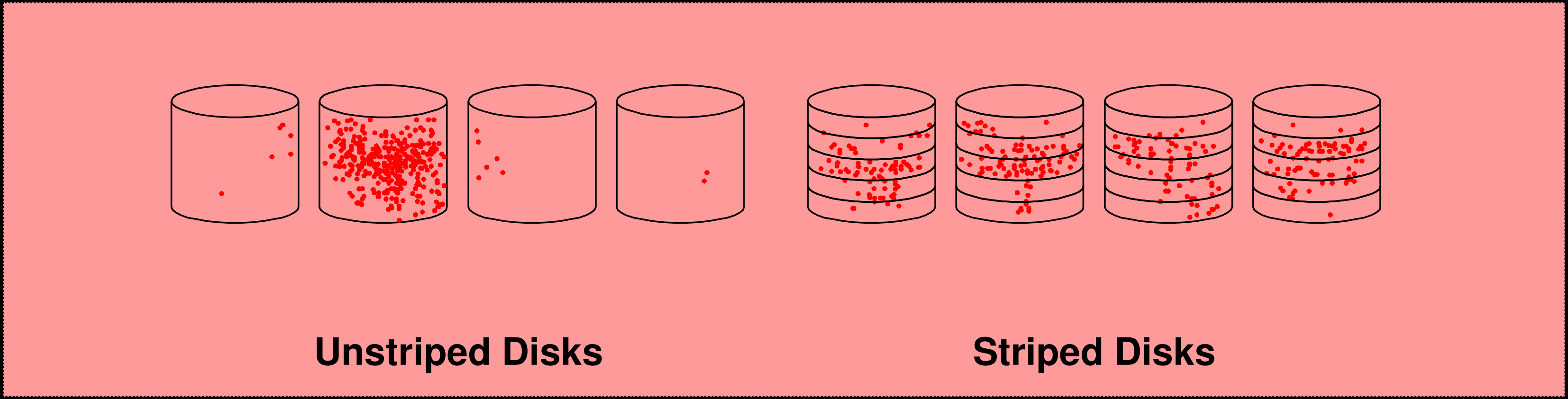


Dynamic Rebalancing with ASM



- Automatic online rebalance whenever storage configuration changes
- Only move data proportional to storage added

ASM Striping for even Data Distribution

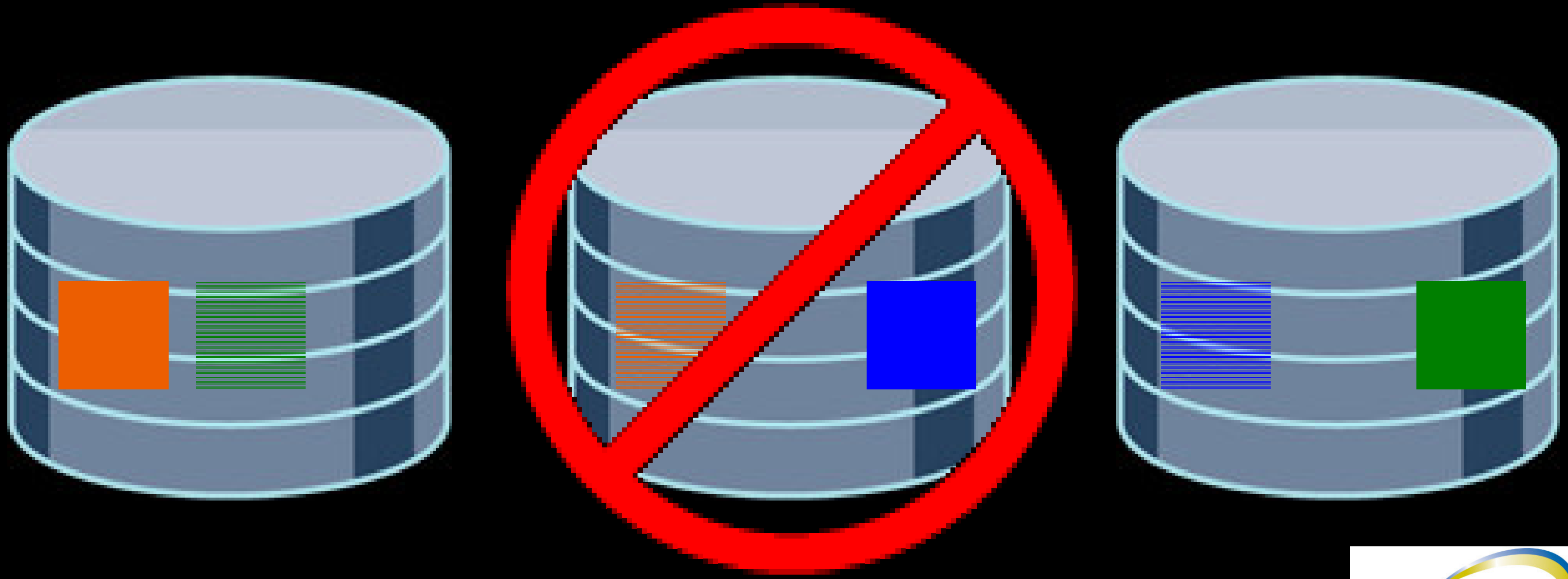


Automatic Storage Management

- Allows dynamic online Storage reconfiguration
- Efficient relocation of data during rebalance
- Eliminate manual I/O tuning in all storage configurations

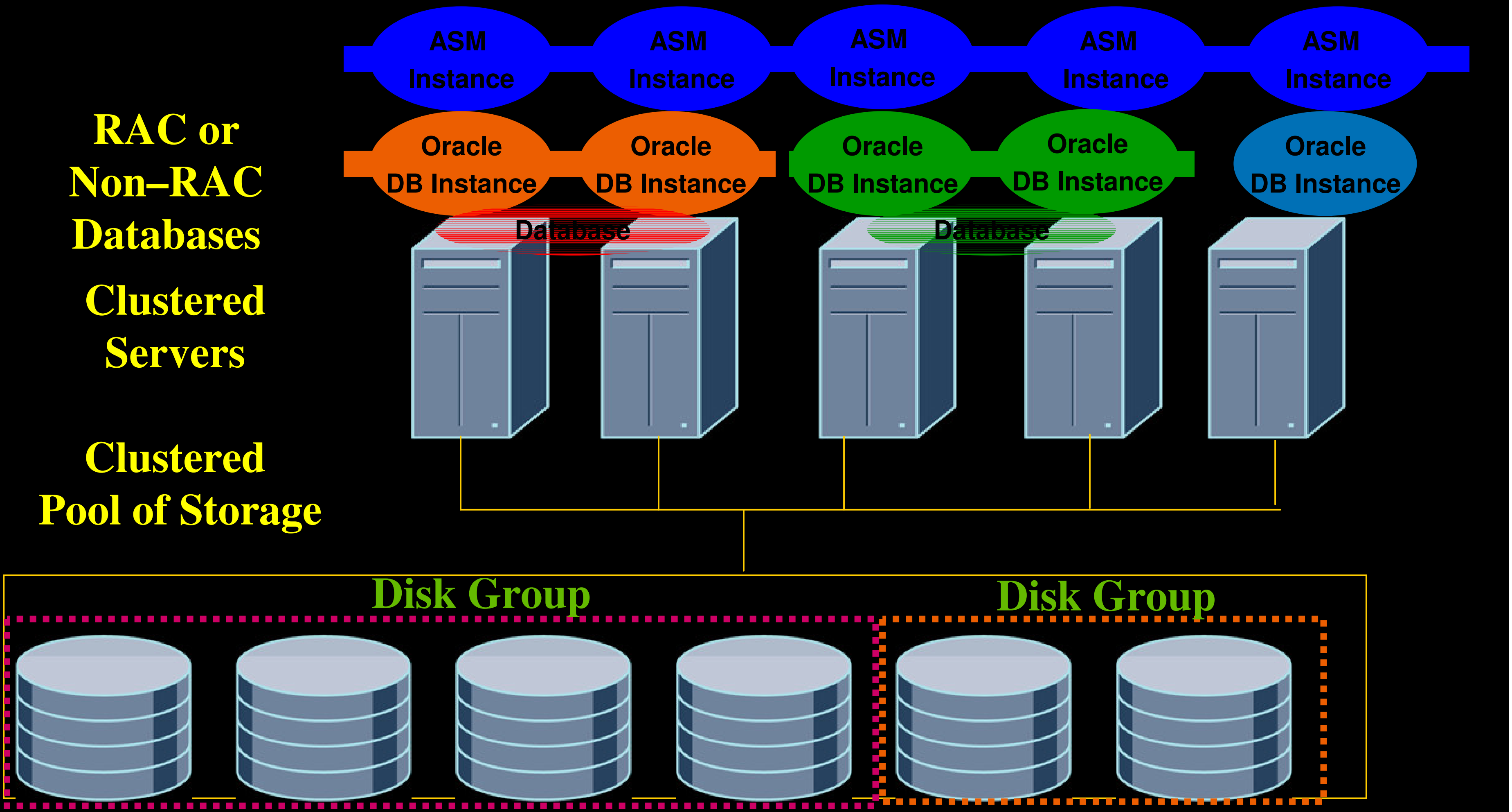
ASM Mirroring allows fault tolerance

- **Mirror at extent level**
- **Mix primary & mirror extents on each disk**



HP WORLD 2004
Solutions and Technology Conference & Expo

ASM Enables Consolidated Clustered Storage



HP WORLD 2004
Solutions and Technology Conference & Expo

Today’s Enterprise Grid

The Grid Solution for Enterprises

IT Challenges:		Grid Solution:
• Separate servers	→	• Shared servers
• High h/w & s/w costs	→	• Low cost components
• Difficult to reconfigure	→	• Allocate as needed
• Difficult to manage	→	• Unified management

Oracle Database 10g provides a Complete, integrated grid infrastructure for enterprises

Oracle Database 10g: Enterprise Grid Control

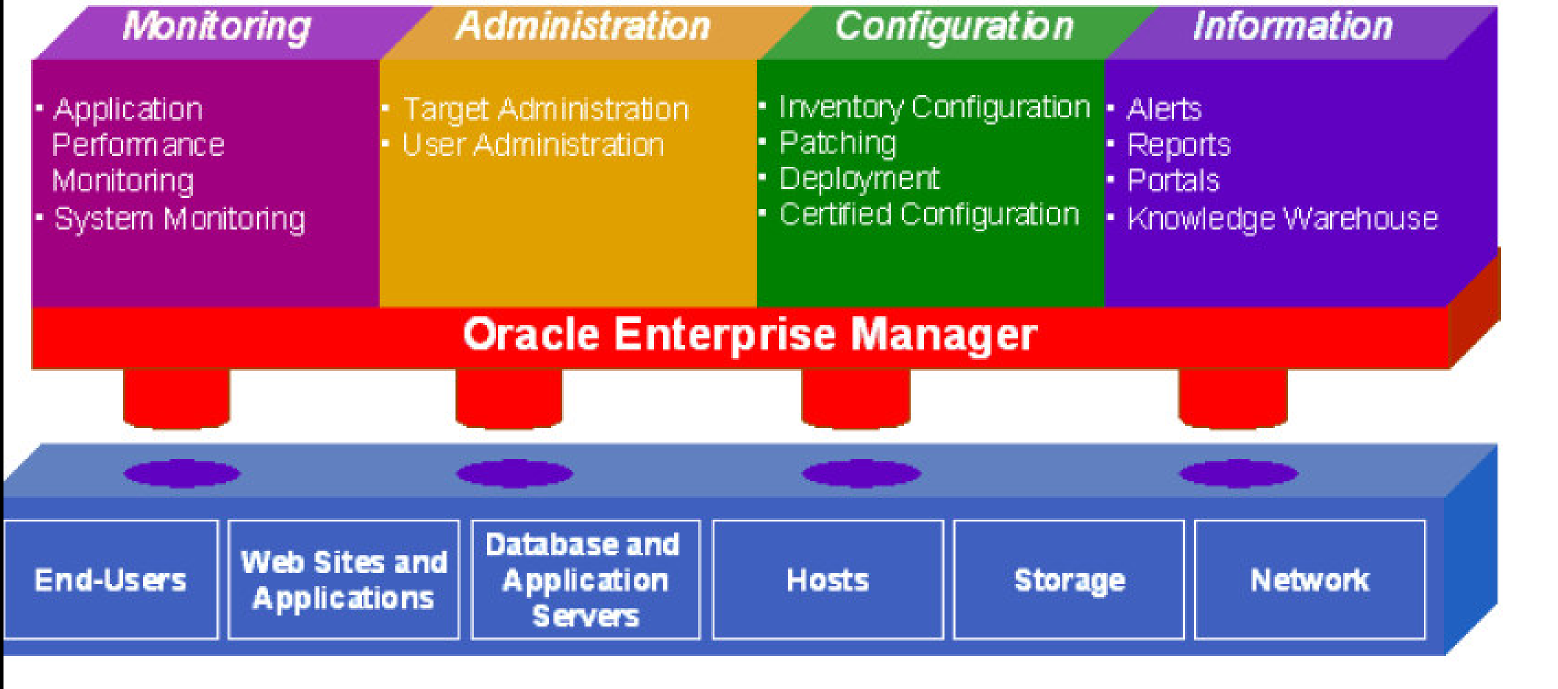
- Enterprise Grid Control Software automates management of Oracle Database software platform

- It manages by exception. Highlights what needs attention. Alerts when things are going wrong. It is of less effort providing better results

- It allows administrators to set up policy based provisioning to prioritize and manage resources in a grid

- Oracle 10g resonates with the grid
  - It is the ability of Oracle 10g to automatically meet service level demands in terms of performance and availability
  - It automatically adds capacity on demand. One can automatically rebalance load across that capacity. One can automatically scale out across these standard components.

Oracle Database 10g Grid Control



Monitoring

- Application Performance Monitoring
- System Monitoring

Administration

- Target Administration
- User Administration

Configuration

- Inventory Configuration
- Patching
- Deployment
- Certified Configuration

Information

- Alerts
- Reports
- Portals
- Knowledge Warehouse

Oracle Enterprise Manager

End-Users | Web Sites and Applications | Database and Application Servers | Hosts | Storage | Network

Oracle Database 10g is ready for today’s enterprise

- Oracle Database 10g integrated with RAC meets all the challenges of today’s enterprises.
- Users determine the policies and priorities. Oracle Database 10g does the heavy lifting. Oracle Database 10g surfs blades.
- Oracle RAC 10g Database makes enterprises adaptive, proactive and agile.

HP/Oracle Benchmarks on IA-64 Linux

- **OLTP Benchmark (TPC-C)**
  - Cluster Result: 16-node cluster of 4-way Itanium 2 (Madison: 1.5GHz, 6MB L3) HP Integrity rx5670 running Red Hat Enterprise Linux AS 3.0 and Oracle Database 10g with Real Application Clusters
    - Performance metric – 1,184,893.38 tpmC; price/performance - $5.52/tpmC
  - Single Instance Result: On a 4-way Itanium 2 (Madison: 1.5GHz, 6MB L3) HP Integrity rx5670 running Red Hat Enterprise Linux AS 3.0 and Oracle Database 10g
    - Performance metric – 536,783 tpmC; price/performance - $3.94/tpmC

Work in Progress…..

- Architect robust enterprise solutions for the Linux Grid
- Investigate new features in the Linux kernel for Oracle 10g Release2
  - Clustering solutions with Oracle 10gR2 using Linux kernel features
- Investigate & integrate new features of HP Integrity Server Family suitable for Linux
  - Numa-awareness & cell-local memory, dynamic expandability, high performance features
- Make efficient use of new Intel compiler optimizations for Oracle RDBMS on Linux architecture
  - Optimizer flags, new prefetching techniques, enhanced dynamic profiling, etc
- Investigate performance of new Oracle features with new Linux kernel (RHEL4 , etc)
  - Infiniband, etc

Acknowledgments

- Hewlett Packard: Development Alliance Lab
- Hewlett Packard: Performance Benchmarking Group
- Oracle: Product Line Engineering
- Oracle: Product Management and Alliance
- Oracle: Release documentation Group
- Oracle: Server Technologies Division
- Oracle: Platform Technologies Division

“If I had eight hours to cut down a tree, I would have spent six hours in sharpening my axe”

- Abraham Lincoln

HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:

interex

shared knowledge • shared power

encompass

AN HP USER GROUP