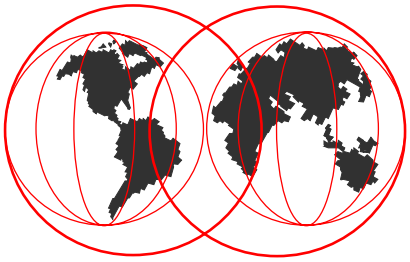
# S/390 Compared to The Sun E10000

Stephen Turner,
ITSO Poughkeepsie

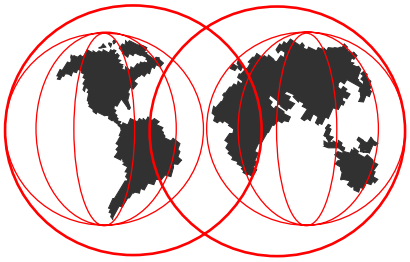




## Topics

- Introduction
- Sun E10000 Architecture Overview
- Solaris compared to OS/390
- Data Management
- S/390 and Sun clustering
- Availability
- Partitioning
- S/390 added value features

# Sun E10000 Architecture





## Sun E10000 Marketing Claims

- Industry-leading performance:
  - Business Intelligence (TPC-D)
  - OLTP (TPC-C)
  - ERP (SAP R/3, Baan)
- Mainframe-like partitioning - UNIX industry leadership
- High availability design
- Web Serving - platform of choice for ISPs
- "The network is the computer"
- Java, EJB, etc.

## Terminology Differences

| Term | S/390 | Solaris |
|---|---|---|
| **Dynamic reconfiguration** | Define, add, remove resources non-disruptively | Add, remove system boards from config (like 3084 SSI-PP) |
| **Workload balancing** | Workload manager with goals | Manually reconfigure or weighting |
| **Disk working set** | Actual disk used (assumes no DFHSM impact) | Disk required (approx. 2.5 x S/390 for RAID & buffer) |
| **Data Sharing** | No integrity exposures | Application responsibility |
| **Capacity** | The useable capacity for normal **operations** | The maximum capacity achieved in a benchmark |
| **High Availability** | Percentage of time application is available for productive use | Time to reboot system to UNIX prompt |
| **Hardware feeds/speeds** | In context of S/390 and LSPR | In context of cycle time or TPC |





## Terminology Differences

| Term | S/390 | Solaris |
|---|---|---|
| **Poor Availability** | Two outages a year | Two outages a week |
| **High Availability** | No interruption to service for a user at a workstation | Minimal time to restart |
| **Scalability** | Near linear increase in application throughput with increased processor capacity or # of processors | Increase in application throughput with increased resources in a system |
| **Connectivity** | Connection of I/O devices and other server systems within one center | Network connection to other remote systems with varying protocols |
| **Processor with multiple engines** | Multiprocessor | Symmetric Multiprocessor (SMP) |
| **Multiple Processors - loosely coupled** | Parallel Sysplex | Cluster |
| **Processor Memory** | Central Storage Memory | Memory |
| **External Disk Storage** | DASD | Disk |

## Sun E10000

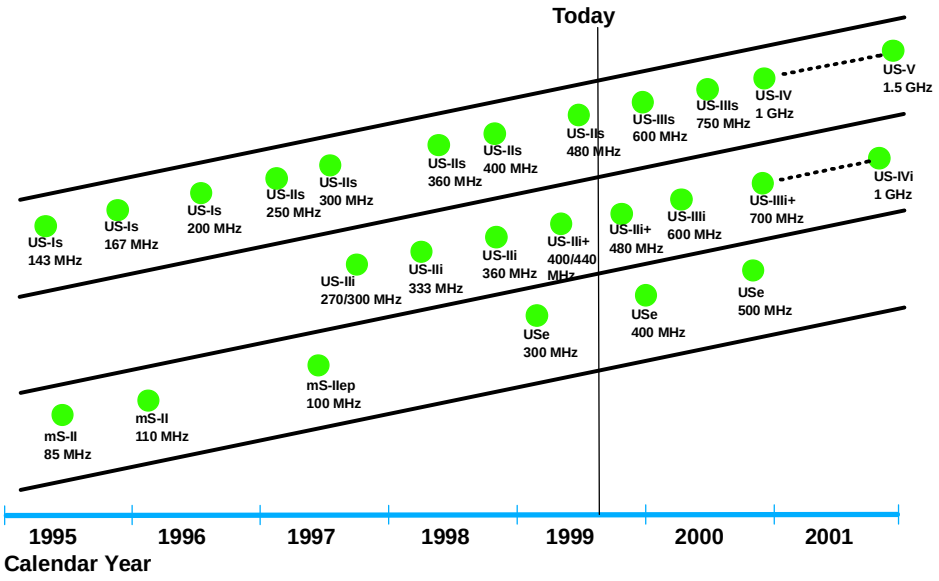




## E10000 Design

- UltraSparc II chip
  - designed by Sun, manufactured by Texas Instruments
- Three generations:
  - 250 MHz (January 1997)
  - 336 MHz (February 1998)
  - 400 MHz (February 1999)
  - 480 MHz expected late 1999

## Sun Sparc Chip Roadmap








## E10000 Packaging

- Single cabinet
  - 1 -16 boards (16 in x 20 in, 24 layers)
  - 0 - 4 CPs per board
  - 0 - 4 GB memory per board
  - 0 - 4 I/O ports per board
- Minimum System (4 boards):
  - 16 CPs
  - 512 MB memory
  - Smaller systems by special bid
- Maximum System (16 boards):
  - 64 CPs
  - 64 GB memory
  - 64 I/O ports

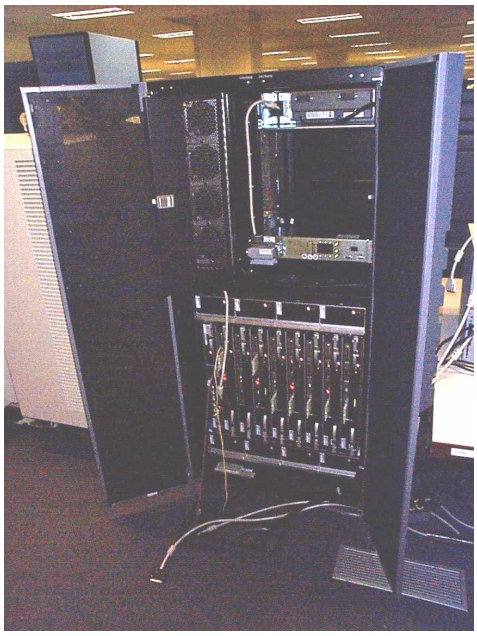
## E10000 Packaging (continued)

- Up to 3 racks for internal disks
- Up to eight N+1 power supplies
- 12 hot-pluggable fan trays (variable speed for noise control)
- Boards are located on either side of the Gigaplane
- Loss of a board means loss of all resources on it




## E10000 -Inside the Front Cover

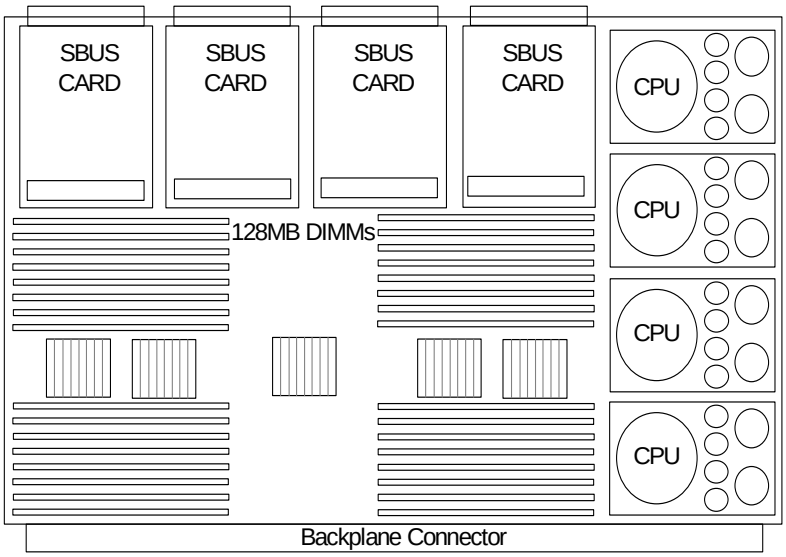
## E10000 Rear View



IBM Technical Support

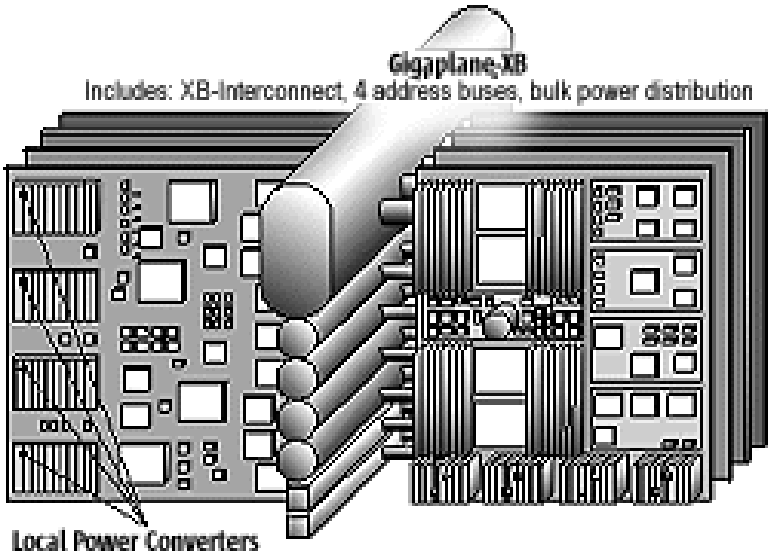


## E10000 Board Schematic




IBM Technical Support

## E10000 Gigaplane

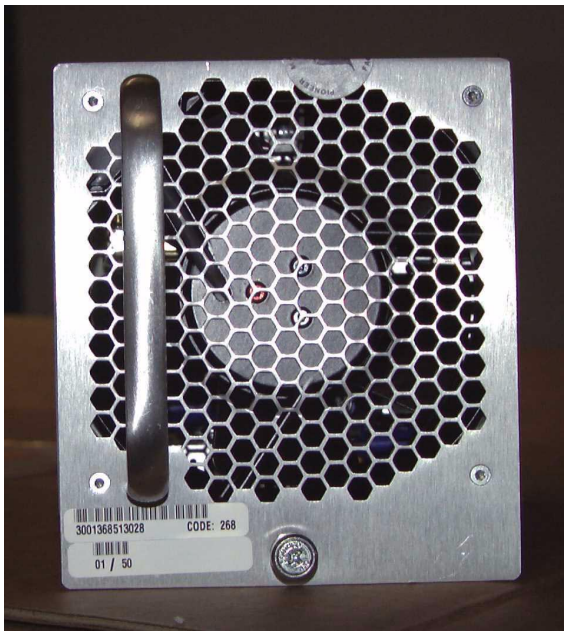







## E10000 - Back View

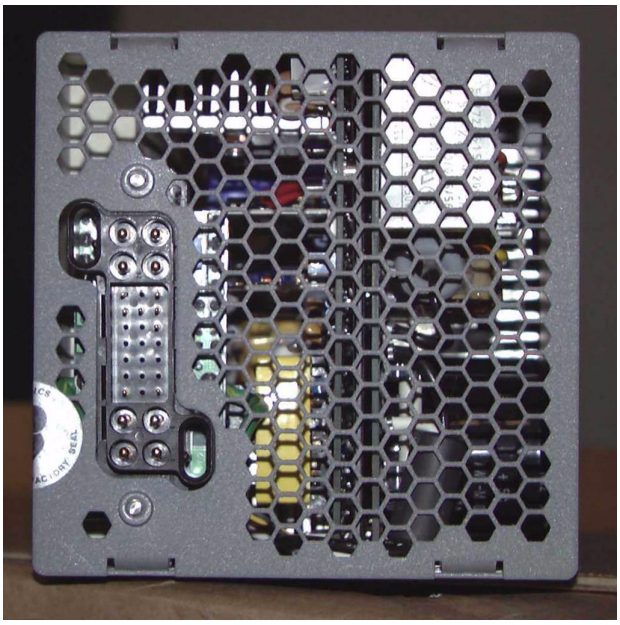
## E10000 Power



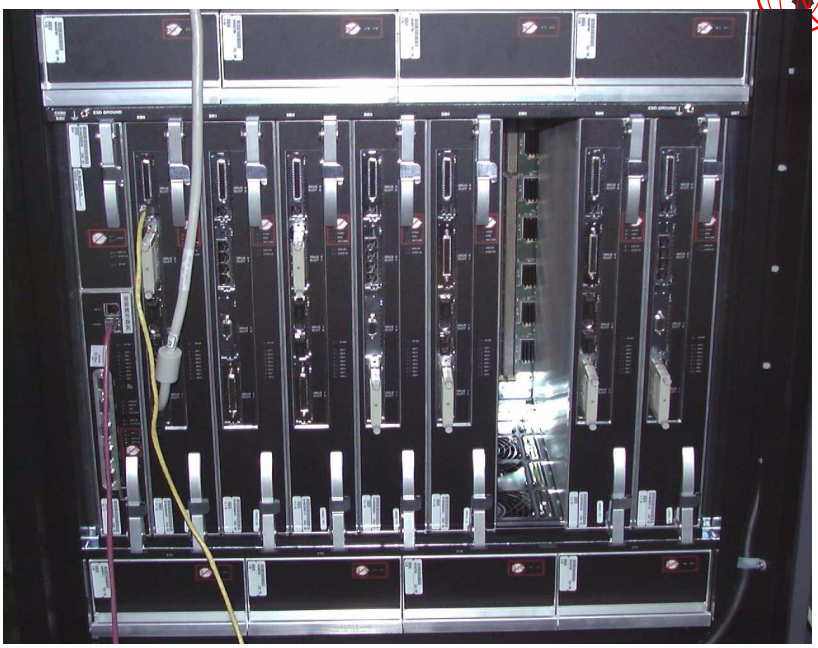





## E10000 Power (2)



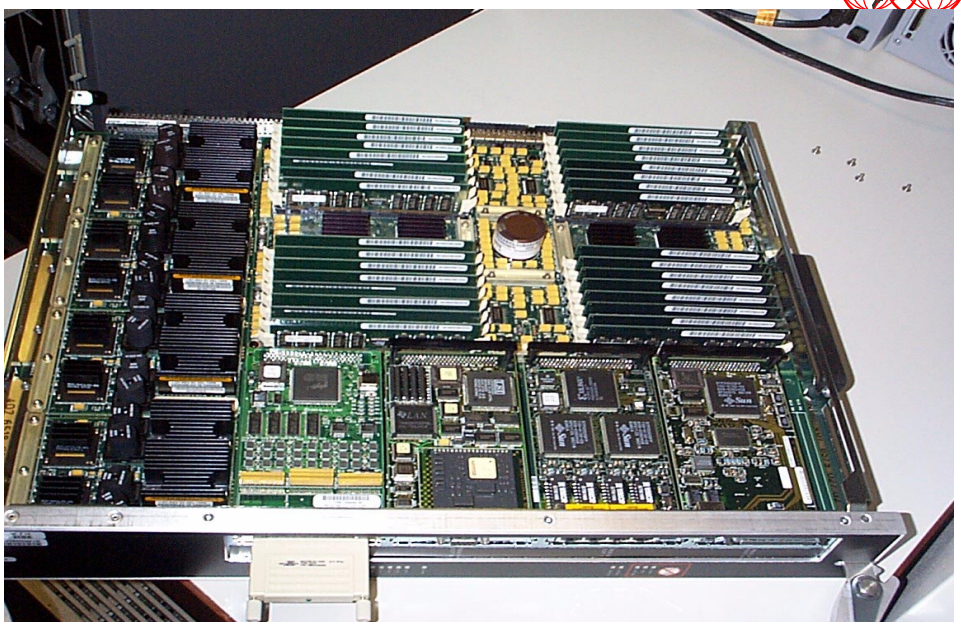
## E10000 Processor Boards



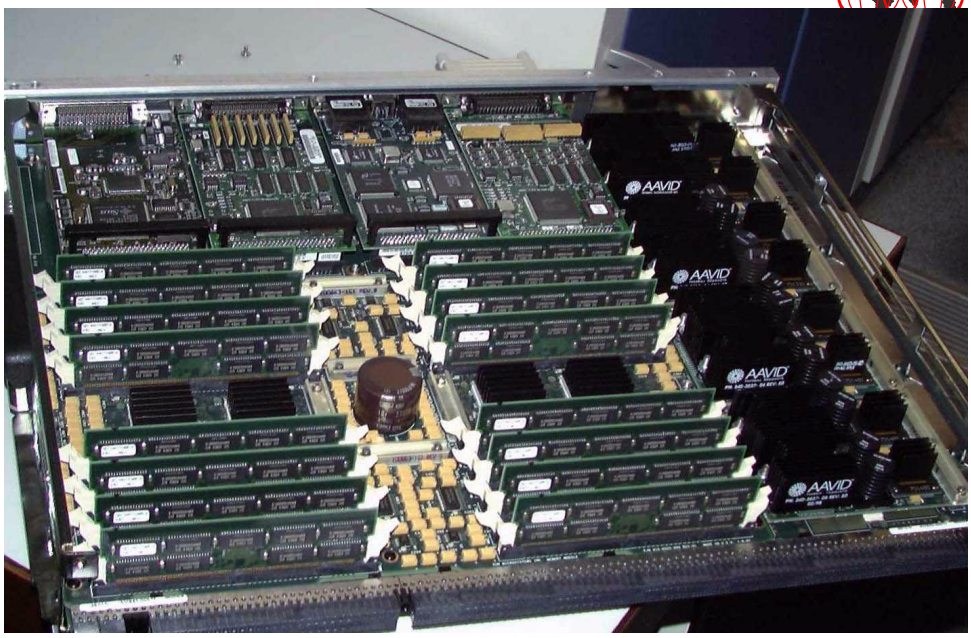




## E10000 System Board



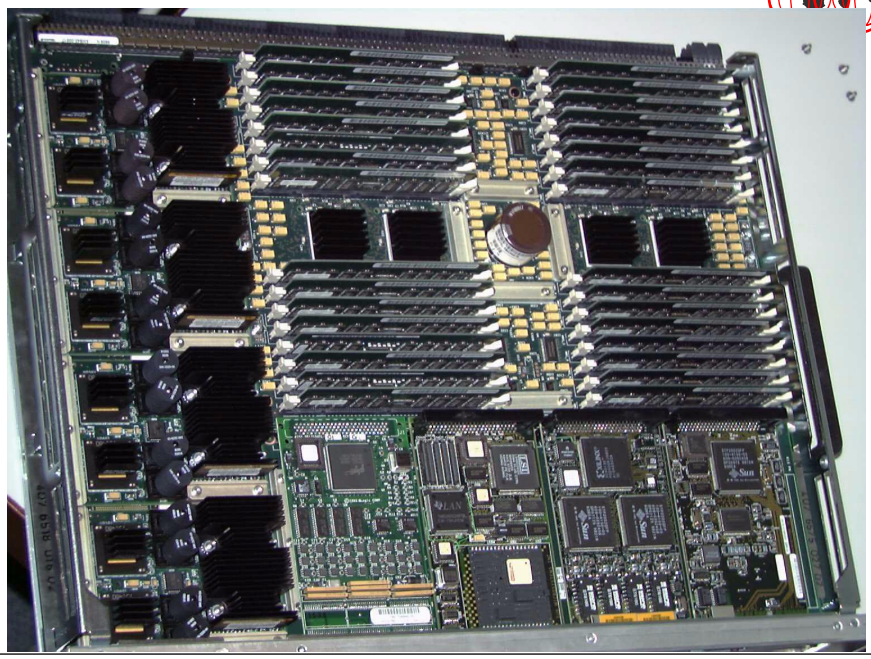
## E10000 System Board



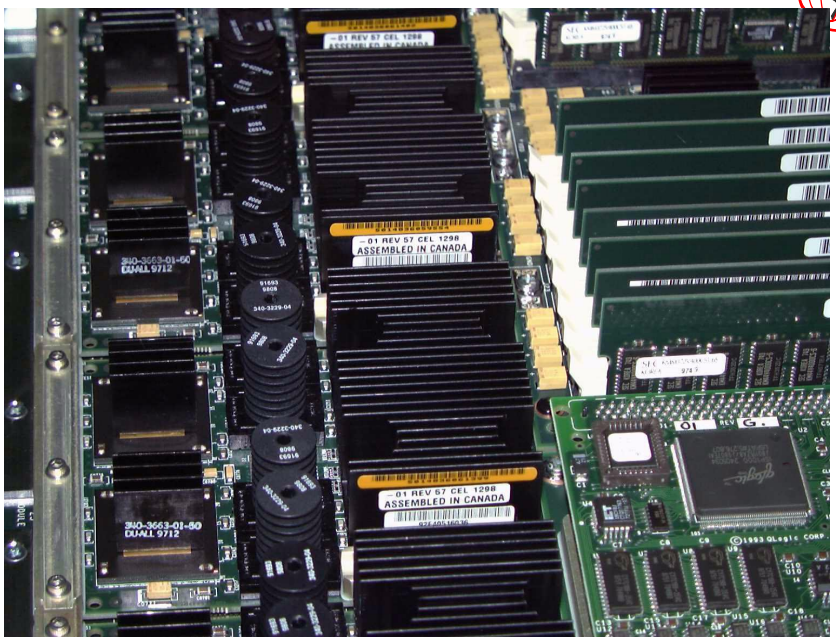





## E10000 System Board



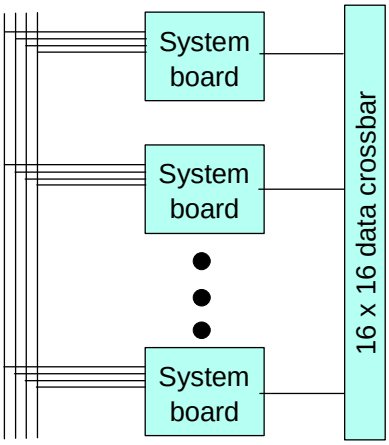
## E10000 System Board



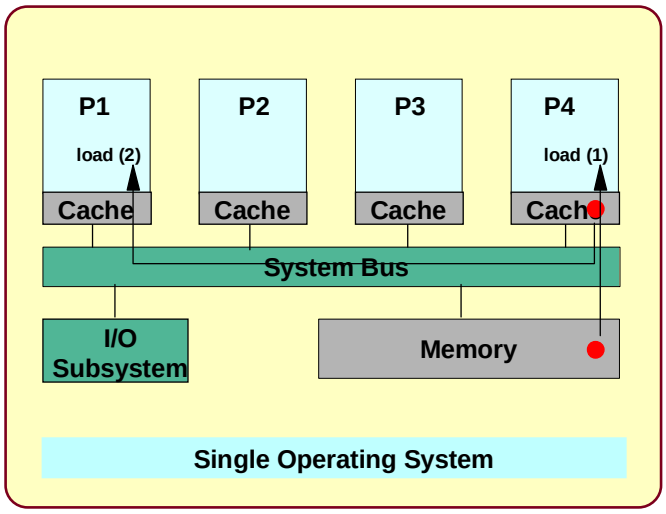





## E10000 Crossbar Switch



# Architectures - Traditional (Uniform) SMPs




Any processor can access data in the memory hierarchy with a single LOAD/STORE instruction.

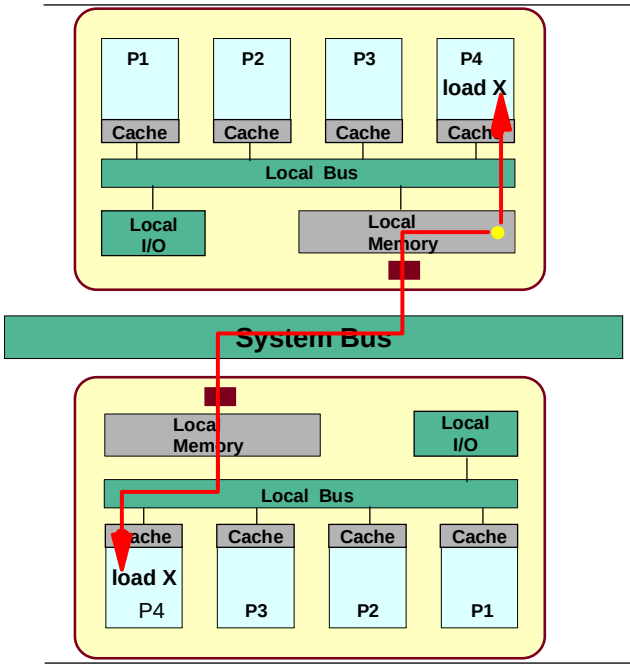
All processors can handle I/O interrupts from any device

Scalability depends on cache sizes, speed/type of interconnect and on workload characteristics





# Architectures - Non-Uniform SMPs




All processors still have access to all memory via LOAD/STORE

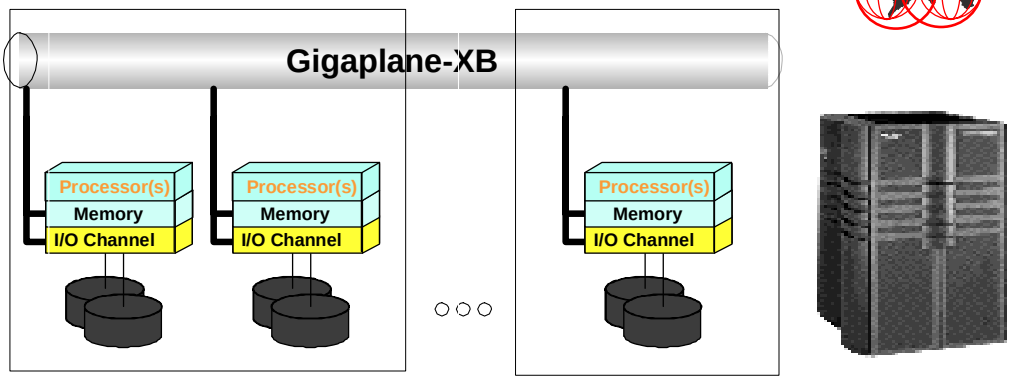
Access to local memory is faster than in traditional "big-bus" SMPs, but access to remote memory and I/O takes significantly longer

- Local adapter sits on memory bus, intercepts request and forwards to correct adapter
- Remote adapter returns current version of data

Scalability of large-scale systems depends critically on application's "locality of reference" to memory and I/O

Operating System calls to expose architecture to applications

## Architectures - Sun UE10000




- 64-way SMP using 4-way 'nodes', each with own I/O and memory
- All memory accesses via cross-bar switch
- Full NUMA possible future direction
- Single or multiple OS (in partitioned mode)
- Supports SMP programming model but generally benchmarked as a 'shared nothing' system





## Database Scalability - Sun E10000

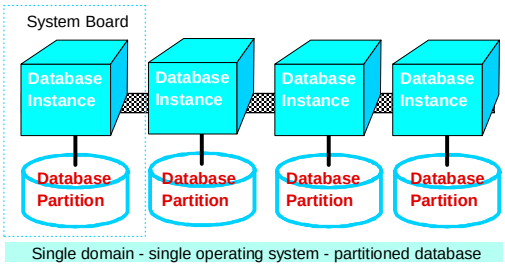
Early benchmarks used Informix (shared nothing) with multiple instances - each bound to a system board

Current benchmark uses Oracle with partitioned database

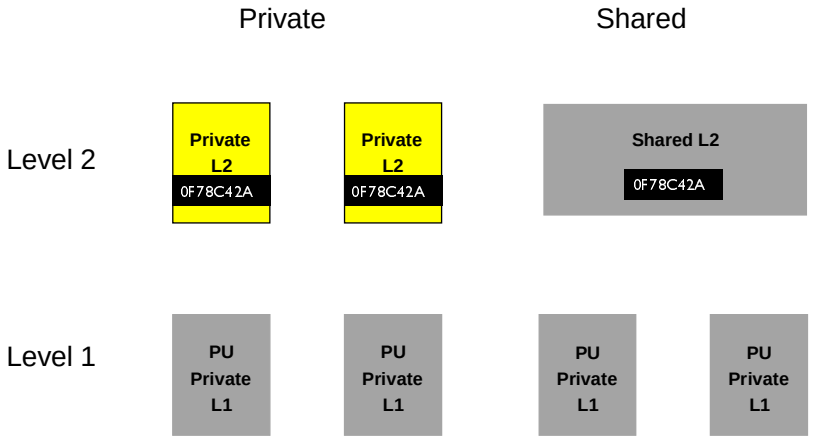
- Database admin is almost as complex as OPS

Database software problem brings down entire Oracle database

- Simpler administration vs lower availability



## Level 2 Cache Types








## E10000 Processor Specifications

- 16 KB L1 data cache
- 16KB L1 instruction cache
- 4MB L2 cache (private)
- 32 640bit integer registers
- 4 instructions/clock cycle (9 stage pipeline)
- 4 Integer execution units
- 4 floating point execution units (2 on 400 MHz models)
- 2 Graphics execution units)

## E10000 I/O

- Each board has an I/O module that supports:
  - Either a pair of Sbuses
  - or, a PCI bus
- Sbus and PCI can be mixed on a single board
- Sbus:
  - 25 MHz (for 83.3 MHz mother boards)
  - 30 MHz (for 100 MHz mother boards)
  - Supports 2 Sbus I/O cards or single PCI card
  - Maximum of 4 Sbus, or 2 PCI cards per board
  - Throughput of 100 MB/sec per bus (64-bit data path)
- PCI bus:
  - Either 32-bit or 64-bit data paths
  - 33 or 66 MHz
  - PCI module has 2x66 MHz buses, 1 PCI adapter per bus
  - Peak data rate 200 MB/sec (64-bit)
  - Sustained data rate 100 MB/sec





## E10000 I/O Controllers

- The E10000 supports:
  - SCSI
    - Disk, tape attachment
    - UltraSCSI, SCSI-2 - up to 40 MB/sec
  - Fiber
    - Disk, tape attachment
    - F/CAL standard
    - 64-bit - 100 MB/sec per GBIC (2 GBIC per card)
  - Ethernet
    - Networking between E1000, workstations etc.,
    - 32-bit 10 MBit/sec, 100 MBit/sec, or 1Gbit/sec
  - FDDI
    - 32-bit 100 MBits/sec
  - ATM
    - Interface to LAN/WAN
    - 64-bit, 155 MBit/sec or 622 MBit/sec
  - HIPPI
    - Point-point connections to "mainframe-style" peripherals
    - Dual interface (1 cable in, 1 cable out)
    - 32-bit, 80 MB/sec in each direction

## E10000 I/O Connectivity

- I/O devices connected via a single connection to an I/O card on an I/O board
- Second connection can be made (for most devices)
  - To an I/O card on a different board, or
  - To another I/O card on the same board
  - Improves availability when Sun Solaris Alternate Pathing function is configured
  - NO performance benefits
  - Second path ONLY used if the first path fails
- Some new I/O controllers (e.g. Sun A5100, A5200) support 2 paths concurrently due to the quantity of data behind them
  - Not dynamic like IBM DPR, APR etc.,
  - Can configure a backup path for each active path





## E10000 Interconnection

- On-board interconnection:
  - traffic from processors, SBus, and memory to off-board address and data ports
- Centerplane interconnection:
  - addresses and data between boards
- Memory accesses ALWAYS traverse the global interconnect, even if the memory location is physically on the same board
- Addresses must be sent off the board to accomplish global "snooping"
- Memory access time is uniform - independently of the board where the memory is located

## E10000 Centerplane (Gigaplane)

- All boards plug into the centerplane
- Bandwidth:
  - 10.7 GB/sec (83.3 MHz mother board)
  - 12.8 GB/sec (100 MHz mother board)
- **ALL** communication between processors and memory, or processors and I/O cards is always via the centerplane, even if both components are on the same board
- Uses ECC checking on the Gigaplane-XB
- Gigaplane is divided into 2 halves - one can continue to operate if the other fails





## E10000 RAS Features - Reliability

- Gigaplane:
  - ECC and parity checking for data and address buses
- Memory
  - DIMMs organized so that each DRAM chip contributes only one bit to each byte
  - DRAM failure causes only self-correcting single-bit error
  - ECC code on each 64-bit word
  - Memory scrubbing - Solaris accesses and rewrites each memory word every 24 hours, logging and correcting errors
- I/O
  - - All local and global I/O paths are protected by ECC and parity checks. Single bit errors are transparently corrected
- A fully redundant system can recover from a system crash. Automatic system recovery enables the system to reboot following a failure
- Sun claim there is no single point of failure that will prevent a reboot.

## E10000 RAS Features - Availability

- An E10000 can be optionally configured to have 100% redundancy of reconfigurable components:
  - Control boards
  - Support boards
  - System boards
  - Bulk power subsystems
  - Bulk power supplies
  - Cooling fans
  - Peripheral controllers
  - System service processors
- Centerplane can operate in degraded mode if one half fails
- Dynamic System Domains
  - Isolated copies of Solaris
  - Each domain is isolated from failures in an other
  - Some recovery procedures may require the entire system to be rebooted!





## E10000 RAS Features - Serviceability

- Service/maintenance process provided by the System Service Processor (a Sun UNIX workstation)
- SSP connected by ethernet
- Can be administered remotely
- System tries to recover without a service interruption if there is a failure of:
  - UltraSparc module, DIMM, SBus board, memory module, I/O module, system board, control board, centerplane support board, power supply, fan
  - Failed component can be hot-swapped out of the system and replaced

## E10000 Directions

- Serengeti
- Faster Chips
- Board-to-board communication via the backplane
- Use of Storage Area Networks



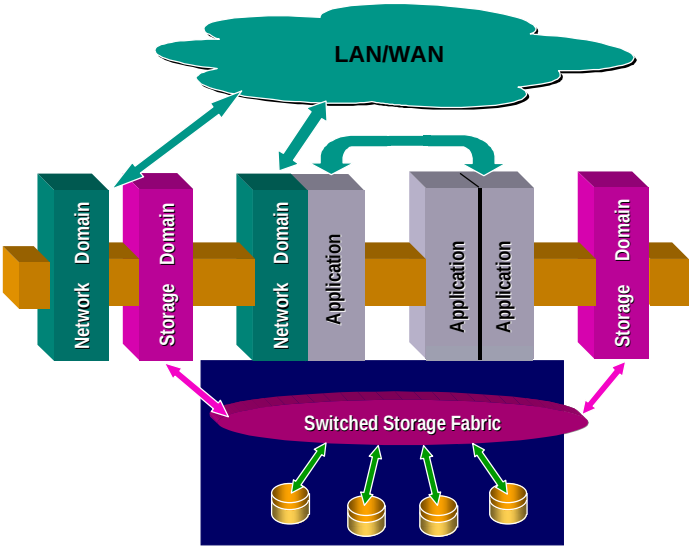


## Sun E10000 Serengeti

- Expected availability within next 18 months
- Entry price over $1 million US
- ccNUMA (cache-coherent Non-Uniform Memory Access) architecture.
- UltraSPARC3 microprocessors, up to 600 MHz.
- Early versions of this architecture expected to have 24 to 78 UltraSPARCs
- Clustering could boost the number of machines (and CPUs) that can be employed on a particular task to as many as 150
- Solaris 7.0 will be required
- Among other features, Sun are expected to employ Fiberconnect technology

As reported in Forbes magazine, February 1999

## Sun Futures - SANs








# Solaris compared to OS/390

## Comparing Operating Systems

- Feature by feature comparison charts are not usually helpful
- Sun is positioning the E10000 as a mainframe data center alternative
- Data Center concerns:
  - Workload management
  - Batch processing
  - Capacity planning, tuning and accounting
  - Security and integrity
  - Application development
  - Bandwidth management for IP networks





## Workload Management

### Efficient Use of Resources

**S/390 history**

- **Hardware resources were expensive, so use efficiently**
- **Mixed workloads since 1970s**
- **Memory was expensive, so use efficiently**
- **Efficient clustering techniques**

**S/390 now**

- **Sophisticated resource and workload management algorithms**
- **Efficient use of hardware**
- **Excellent performance, with mixed varying workloads**
- **Very powerful I/O capability**

# Workload Management

## UNIX history

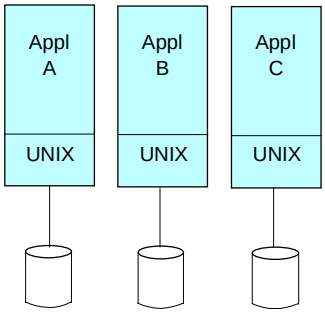
- Software free to universities and researchers
- Hardware resources were inexpensive, so don't use efficiently
- Single workload per server
- Low cost hardware and 'hackable software' meant low availability - therefore use fast restart
- Low-cost I/O with poor performance - use data-in-memory
- Clustering for availability

## UNIX now

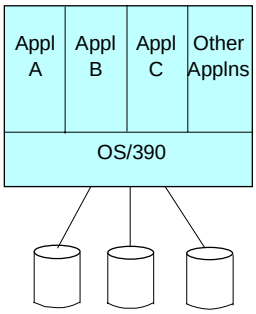
- Unsophisticated resource and workload management algorithms
- Inefficient use of hardware (low average utilizations)
- Poor stability and performance with mixed varying workloads
- Relatively poor I/O capability
- Need for large memories




# UNIX and S/390 Approaches




- Single application per server
- Separate/partitioned data bases
- Complex system management
- Complex to integrate applications




- Workload managed according to business priorities
- Multiple applications per server
- Data bases, shared with integrity
- Less complex system management
- Interoperability and integration between applications

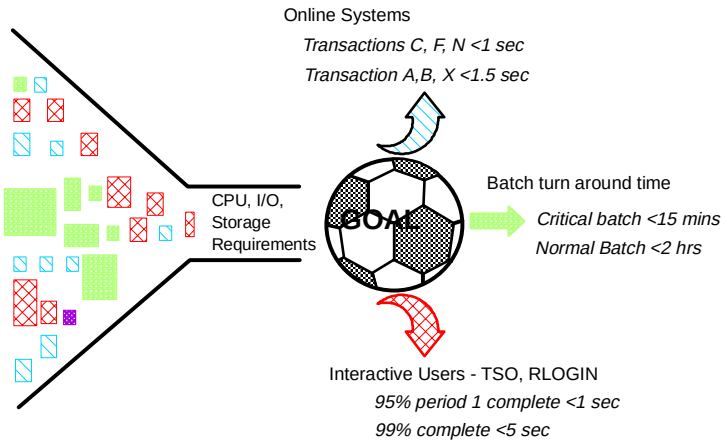
## Workload Management

- **Efficient workload management:**
  - **Multiple types of workload concurrently**
  - **All achieve their performance objectives**
  - **At processor utilizations of 90% +**
- **OS/390 workload management is far ahead of other platforms**
  - **Memory management**
  - **Data in memory**
  - **Workload management in an SMP**
  - **Workload management across a cluster**
  - **Operating system facilities**



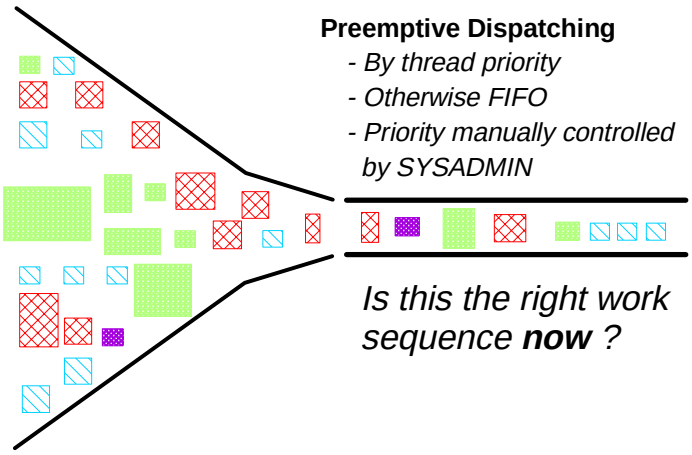

## OS/390 Workload Management

### OS/390 - Business Goal Oriented



## UNIX Workload Management

### UNIX - Limited Workload Management



**Preemptive Dispatching**

- *By thread priority*
- *Otherwise FIFO*
- *Priority manually controlled by SYSADMIN*

*Is this the right work sequence **now** ?*





## The Importance of Workload Management

- Need to access multiple applications from the web
- Need to operate large servers efficiently
  - Mixed workloads concurrently
  - Shared use of resources
- Trend towards clusters
  - Availability
  - Performance
  - Scalability

## OS/390 Workload Manager

- Workload management for both SMPs and sysplex
- Performance administration
- Performance management





## OS/390 Workload Management

- Different types of workloads have different Response Time requirements
- Installation needs to make best use of resources and maintain highest throughput possible
- OS/390 WLM enables definition of performance goals and business importance for each workload
- Workload management collects data on performance, delays, utilization, etc.,
  - Used for management of workload
  - Used for monitoring and reporting

## Performance Administration

- The process of defining and adjusting performance goals
- Introduces the job of Service Level Administrator
- Service Definitions form part of Service Level Agreements (SLAs) for:
  - CICS, IMS, DB2, TSO
  - UNIX System Services
  - JES
  - APPC
  - Component Broker
  - WebSphere
  - Etc.





## Performance Management

- Matching of resources to workload in order to meet performance goals
- WLM uses Service Definition information - internal monitoring feedback to check how well goals are being met
- Periodic adjustment and re-allocation of resources as workload level changes
- When congestion occurs, WLM makes tradeoff based on workload importance and how well goals are being met
- Address spaces can be dynamically started/stopped
- Period aging lowers priorities for lower importance workloads

## Solaris Scheduling

- Default is TS scheduler
  - Time shares
  - Approximately equal access for all processes
  - Limitations are applied to
    - processes without physical memory resources - not permitted to run
    - processes with pending I/O requests
- Realtime Scheduler (RT)
  - For users with sufficient priority
  - Processes get immediate access to the processor
  - Can co-exist with SRM, but SRM has NO control over RT processes
- Solaris Resource Manager
  - SHR scheduler class
  - Fair share scheduler
  - Specified by system administrator
  - NOT a job scheduler
  - Operates only on a single server





## Solaris Resource Manager (SRM)

- Stand-alone product or integrated into Solaris Enterprise Server (1H99)
- Based on Share II product from Software Pty Ltd, Australia

Sun state:

"*The Solaris Resource Manager software ensures resource availability for users, groups, and applications. It provides the ability to allocate and control major system resources such as CPU, virtual memory, and number of processes. It also implements administrative policies that govern which resources different users can access, and more specifically, what level of consumption of those resources each user is permitted. The Solaris Resource Manager product is a key enabler for server consolidation and increased resource utilization".*

## Claimed Benefits for Solaris Resource Manager

- Server Consolidation
  - Multiple concurrent applications
- Enterprise applications
  - *"Hosting multiple applications on a single system is simpler and more cost-effective than one application per system"*
- Web Site Hosting
  - Multiple (hundreds) of web servers on one machine versus large numbers of small servers
- Ensures Service Availability
- Resource Utilization and Allocation
  - Quote from Sun Executive





## Low UNIX Server Utilization

In an interview on the subject of Sun competing with mainframes, Mr. MacKay was asked,

"How are you improving high-end servers to make them more competitive with mainframes?"

Mr. MacKay replied:

*"Peak performance is an area we've been working on. One of the advantages the mainframe has over UNIX is that mainframes often run at 80% to 95% of capacity. UNIX servers usually run at 20% to 30% of the peak load."*

This is a disparity that MacKay went on to say that Sun is working on.

Mr. Steve MacKay, Chief Technical Officer of Sun Microsystems, Computer Systems Group, as quoted in the March 24, 1999, edition of Investor's Business Daily.

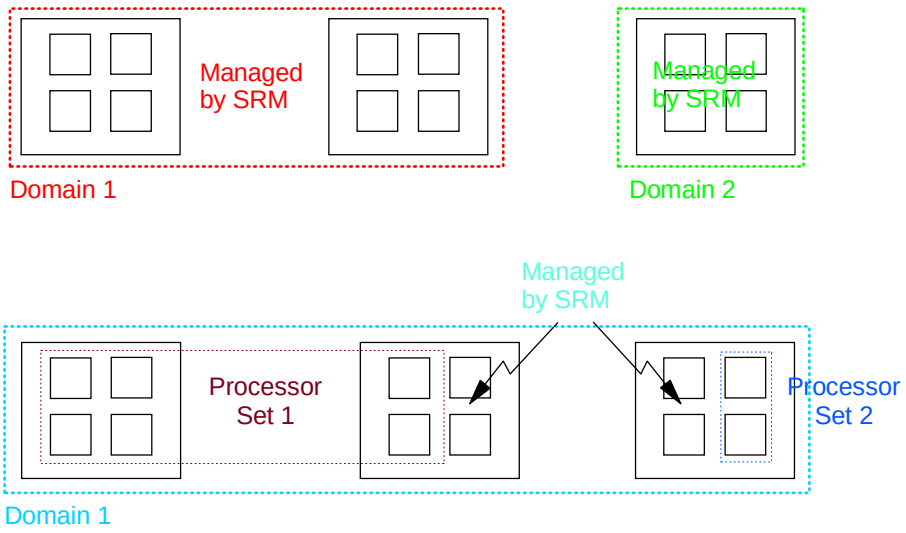
## Claimed Benefits for Solaris Resource Manager (2)

- Hierarchical Control
- Policy-based administration
  - Configure limits for fixed resources
- Resource Use Tracking
  - Detailed usage data
  - Can be input to user-defined tools for resource accounting, or capacity planning
- Complementary Controls
  - Augments Dynamic System Domains (partitioned on a board basis)
    - Solaris Resource Manager gives finer tuning within a domain
  - Processor Sets in Solaris
    - Binds applications to specific processors
    - Solaris Resource Manager can allocate resources on those CPUs which are not part of a processor set





## Sun Workload Management Options



## Solaris Resource Manager
## Resource Categories

- Categories:
  - Users
  - Groups
  - Applications
- Resources Controlled:
  - CPU
  - Virtual Memory
  - Number of processes
  - Number of logins
  - Connect time
- CPU Allocation based on:
  - Shares
  - Decay usage parameters




## Solaris Resource Manager
## Resource Control

- Controls:
  - CPU usage
  - Virtual memory
  - Number of processes
  - Number of logins
  - Terminal connect time
- Keeps track of each user's usage of each resource
- Users may be assigned a hard limit on resource usage (except for CPU)
  - Fails request for resource when limit is reached
- Progressively decays usage
  - Only most recent usage is significant
  - Half-life set by administrator
    - Long half-life favours even usage e.g. batch
    - Short half-life favours interactive users

## Dynamic Reconfiguration

- Dynamic reconfiguration allows addition/deletion of system boards
- Solaris resource Manager keeps track of available processor resource for scheduling purposes




## Sun versus OS/390 Workload Management

| Feature | Solaris Resource Manager | OS/390 Workload Manager |
|---|---|---|
| Manage work across partitions or domains | NO | YES |
| Manage work in a sysplex or cluster | NO | YES |
| Define service in SLA terms | NO | YES |
| Variable resource requirements | NO | YES |
| Job scheduling | NO | YES |

## OS/390 Batch Processing

- Long history of running batch efficiently
- Concurrent batch/online/BI/Web serving/etc.
- Job scheduling (JES)
- Job networking, automated systems
- Conditional execution and error situations
- Large complex daily/weekly/monthly suites of jobs handled easily
- Workload Manager ensures fair use of resources
- BatchPipes, SmartBatch





## UNIX Batch Processing

- Batch equals Background processing
- Limited abilities for automated scheduling (e.g. *cron* command)
- No inter-job correlation (no job networks)
- Pipes
- Solaris Resource Manager only gives share of processor, not prioritized by business goals

## Capacity Management and Planning

- RMF
- SMF
- Accounting
- OS/390 strength
  - lots of IBM and 3rd party tools and methodology
- Limited facilities in UNIX
  - "Buy a bigger server"

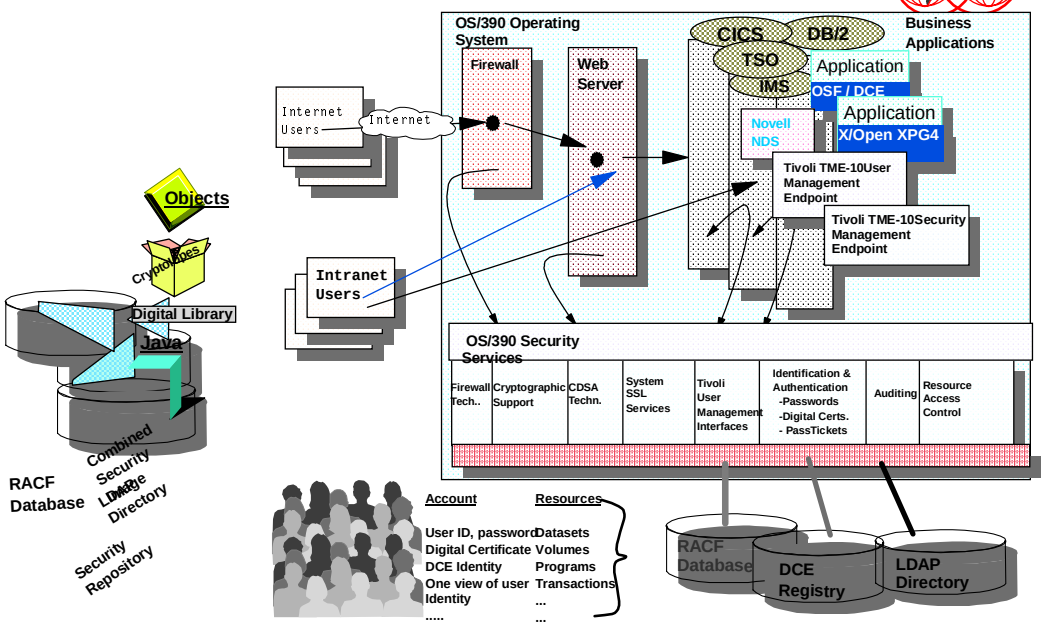




## Security and Integrity

- Successful security requires:
  - Strong hardware isolation functions
  - System integrity to ensure that misbehaving or malicious application/users cannot affect other user or the system
  - System level security to control and monitor the actions of users and applications
  - Network level security to protect from outside hackers
  - Integration of system security functions with system provided applications, vendor applications and user applications

# Security and Integrity








# S/390 Hardware Isolation

- Basic:
  - Provides mandatory separation of users/applications from each other and from the system:
    - Storage protect keys
    - Multiple address spaces
    - Program execution state
  - Users/applications can only affect each other, or the system, if appropriately authorized
- PR/SM:
  - Multiple operating system environments totally separated from each other
  - Can share external resources
  - E4 Security Certification (UK IT Security Evaluation and Certification Scheme)

## OS/390 Security Server

- S/390 Integrity Guarantee since 1973
- Includes:
  - Resource access control (RACF)
  - DCE Security Server
  - OS/390 Firewall Technology
  - eNetwork Communications Server (also provides firewall functions)
  - LDAP server - FTP proxy support etc.




## Security in Solaris

- Three levels:
  - Base Solaris
  - Trusted Solaris
  - Sun Enterprise Authentication Mechanism (part of Solaris Enterprise Server)

## Base Security in Solaris

- Typical UNIX security:
  - Access to files controlled by userid and group
    - Checked against file attributes in directory
    - Coarse level of control
  - Access Control lists:
    - Control access to files and directories on a per-user basis in addition to group filters
    - Similar to RACF dataset profiles and access rules
  - Password stored in a flat file (/etc/password)
    - Needs to be encrypted
    - Even with encryption respresents and exposure





## Trusted Solaris

- Trusted Solaris adds:
  - Labelling Information
    - Sensitivity and information labels assigned to files, devices, windows, hosts, networks and other system objects
    - Administrator indicates level of trust or job responsibility of anyone accessing the system by assigning a clearance level
      - Upper bound - highest level of clearance
      - Minimum sensitivity - lower bound
    - For example: Public, Internal users only, or Need to know engineering
  - Mandatory (MAC) and Discretionary Access Control (DAC)
    - see page 74
  - Privileges/Authorizations
    - Principle of "Least Privilege"
    - *root* user divided into 87 distinct privilege that can be tightly controlled

## Trusted Solaris Administrative Roles

- No powerful single superuser for administering the system
- Divides administrative tasks among a number of administrators
- Administrator activities are audited
- Three default roles
  - Security administrator
  - System administrator
  - Operator
- Fourth optional role
  - Software installation programming that requires UID of *root* to execute





## Sun Enterprise Authentication Mechanism

- Component of Solaris Enterprise Server
- Centralized authentication and management
  - Single repository for authentication information (Key Distribution Center - KDC)
  - Contains user, server and password information
  - Anyone attempting access must be checked against the KDC
- Strong encryption support
  - Information exchanged between the user and the KDC is encrypted
- Ease of use
  - Java-based administrator tool for easy access and configuration

## Application Development

- COBOL:
  - 3rd party vendor only (Merant)
- Database support:
  - IBM DB2 UDB
  - Oracle Database
  - Sybase Adaptive Server
  - Informix Dynamic Server
  - 32-bit versions only on Solaris 2
  - All vendors plan 64-bit support on Solaris 7
- Solaris on multiple platforms
  - Same Sparc processors from workstations to E10000
  - All run versions of Solaris





## Bandwidth Management for IP Networks

- Sun use their substantial installed base of ISPs as a marketing strength
- Growing importance of e-Business demands that vendors competing in this arena must provide a sound vehicle for operating and managing IP networks

## The Need for Bandwidth Management

- The Move to Intranet/Internet-based Business
  - Corporate networks are rapidly evolving from a classic client/server paradigm towards an intranet-based model, based on information-sharing and Web navigation
  - There is a mushrooming use of bandwidth.
  - Simply adding on new connections and fatter pipes is not cost-effective, and won't guarantee availability where it's most needed.
  - An intranet-based model implies the following factors:
    - Changing patterns of network use and unpredictable demands for bandwidth.
    - Demand for increased amounts of bandwidth. People may stay on the link for extended periods of time and download large amounts of data.
    - Demand for guaranteed quality of service in terms of bandwidth and minimum delay. Emerging Internet applications are both bandwidth intensive and time sensitive, often requiring support for voice, video, and multimedia applications across the network infrastructure.
    - Lack of control by MIS staff. Workgroups and departments generally create their Web sites without MIS approval, generating increased traffic without necessarily having the infrastructure to handle it.
    - A change in user attitude. Users expect instant access to information without delays or restrictions, especially if that information is critical to their work.





## The Need for Bandwidth Management

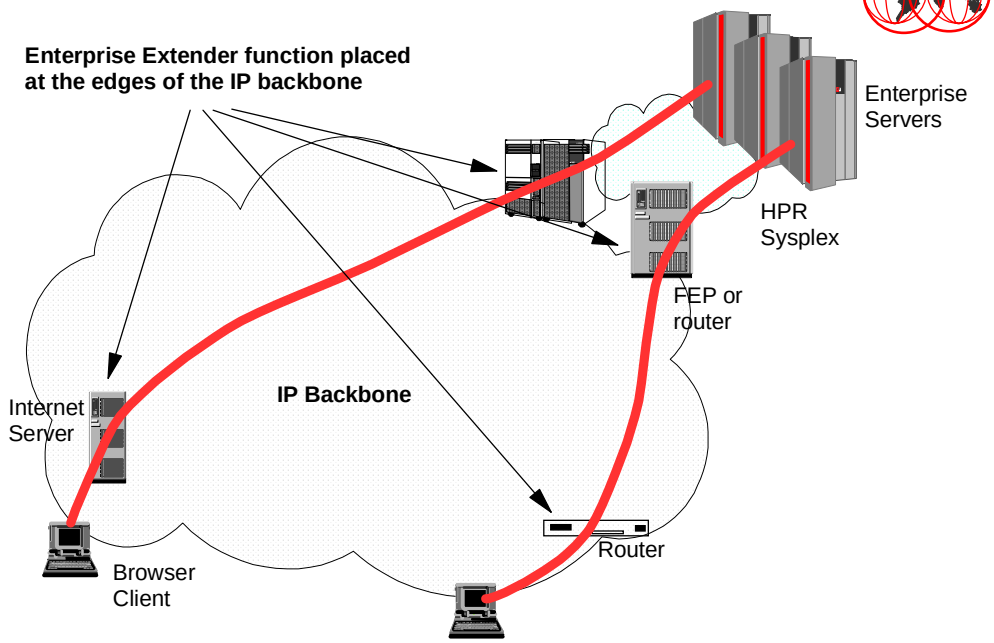
- Need for guaranteed bandwidth
  - Current networking technology has two major limitations:
    - The bandwidth available on a link at any given moment cannot be predicted.
    - It is difficult to control which applications or users get a share of the available bandwidth
- Service Level Agreements
- Service-driven networks
- Granularity

## IBM Enterprise Extender

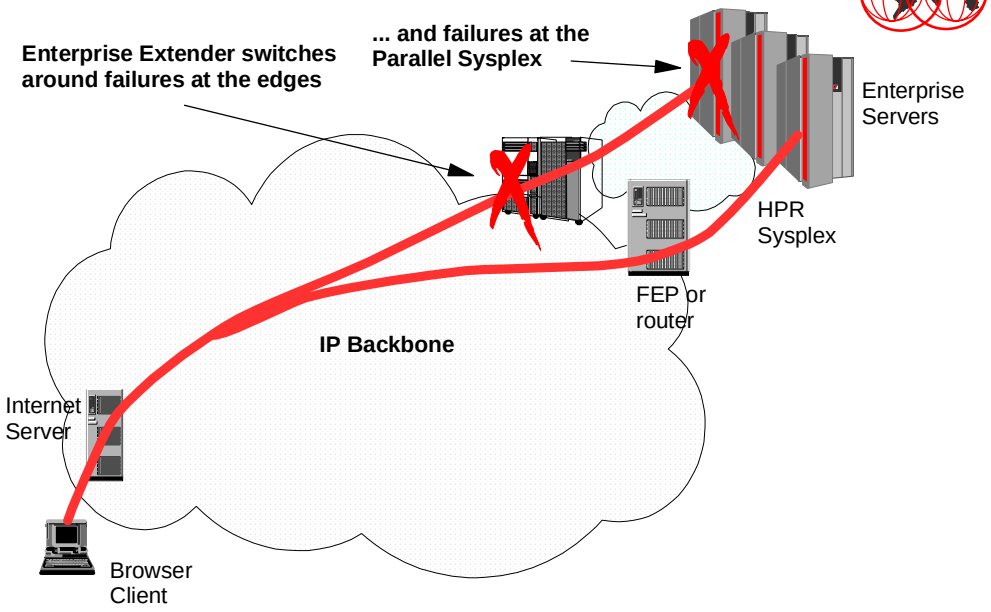
- Simple set of extensions to exisitng, open High Performance Routing (HPR) technology
- Provides:
  - Integration of networks
    - IP
    - SNA
  - Failure protection
  - Scalability and cost-efffectiveness
  - Traffic control




## Enterprise Extender



## Enterprise Extender Failure Handling



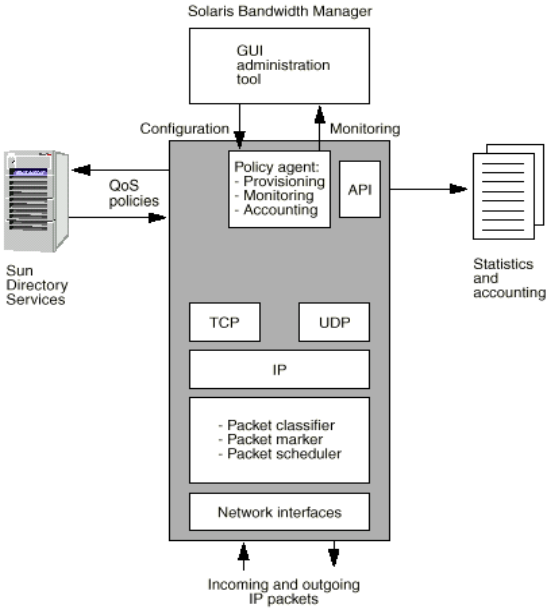





## Solaris Bandwidth Manager

- Component of Solaris Enterprise Server
  - Control of bandwidth for particular applications
  - Control of bandwidth for users
  - Control of department
  - Can distribute bandwidth evenly
  - Can prioritize traffic (prevention of resource hogging)

## Solaris Bandwidth Manager








## Solaris Bandwidth Manager Features

- Reduced congestion (increased efficiency)
- Quality of service
- Transparently homogeneous
- Ease of use
- Customized policies
- Flexible Deployment
- Accounting and reporting
- Extendible
- Integrated with Directory Services

# OS/390 versus Solaris

| OS/390 Parallel Sysplex features | OS/390 Benefits | Solaris |
|---|---|---|
| MIXED WORKLOAD MANAGEMENT within a single OS/390 system enables high processor utilization | Lower Costs | Single Workload Domains |
| DATA SHARING between OS/390 systems with full data integrity | All data always available to all applications | Shared-nothing discrete systems |
| DYNAMIC WORKLOAD BALANCING between OS/390 systems enables high processor utilization | Lower Costs | Wasted Capacity |
| LINEAR SCALABILITY through parallel Sysplex clustering | Nothing to stop Unlimited Growth | Technical inhibitors prevent growth |
| S/390 provides the ultimate in I/O PARALLELISM - e.g.. 4 path DPR, EMIF, DB2 data striping | I/O Bandwidth required for Enterprise Applications | SCSI single path protocols |
| FAULT TOLERANCE maintains service delivery when HW/SW faults occur | No Unplanned Outages | Crashes & Reboots quickly |





# OS/390 versus Solaris

| OS/390 Parallel Sysplex features | OS/390 Benefits | Solaris |
|---|---|---|
| CHANGE TOLERANCE maintains service delivery whilst HW/SW changes are being made | No Planned Outages | Many changes require outages |
| Incremental DB2 BACKUP/REORGANIZATION in parallel with online processing | No disruption to user service | Housekeeping requires planned outage |
| CICS STORAGE PROTECTION isolates systems code and transactions from inadvertent overwrite | No outages caused by new application code | New code can crash a domain |
| DISASTER/RECOVERY support via GDPS (Geographically Dispersed Parallel Sysplex) | No data loss Full recovery within 60 minutes | Not Available |
| HIGHEST SECURITY built into S/390 HW/SW provides bulletproof end to end security | Hackers can't break into S/390 | Many security exposures |

# OS/390 vs UNIX - Usable Capacity

| *Scalability* | *S/390 and OS/390* | *UNIX alternatives* |
|---|---|---|
| **Processor Utilization** | average at 75+% - peak at 95+%<br>Maximize use of resource<br>Lower cost | average 20-30% or even lower, peak at 60-70%<br>High utilizations can break UNIX systems.<br>White space = wasted resource |
| **Benchmarketing** | Thousands of real world mixed workload mission-critical high performing customers, and growing UNIX Services reference base | Vendors are excellent at building highly tuned benchmarks that cannot be replicated in any production customer. Look closely at customer ERP references. |
| **Workload Management Dynamic Workload Balancing** | Automatic via Resource Management Services and OS/390 Workload Manager | Manual |
| **Application Isolation** | Inherent function of S/390 architecture, OS/390 address spaces, and S/390 middleware exploitation | With multiple applications in a single UNIX instance, one application crash can take down all applications. Typically isolate worktypes using separate machines. |
| **Data Sharing** | All applications can share all data, and LPARs can access all data via EMIF | Shared nothing means each UNIX instance is limited to its own partitioned data, or must access other UNIX instances via ethernet |
| **Interdomain/LPAR Communication** | LPARs communicate via system bus speeds | UNIX instances communicate via ethernet speeds |
| **I/O Parallelism** | EMIF, Dynamic Path Reconnect, DB2 Data Striping, ESCON, FICON Better performance | SCSI single path protocols,<br>No DPR, No EMIF,<br>Alternate path for recovery only |
| **LPAR/Domain Management** | Automatic, dynamic acquisition of resource | Manual reassignment of system boards, with corresponding movement of I/O and memory |
| **Linear Scalability** | Near linear scalability with Parallel Sysplex | Subject to multiprocessor effect as CPUs are added to a UNIX instance. |

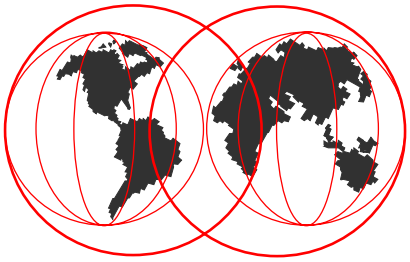




# OS/390 vs UNIX - Availability & Manageability

| Availability | S/390 and OS/390 | UNIX alternatives |
|---|---|---|
| **Fault Tolerance** | **Designed for no unplanned outages** | **Designed to reboot quickly after system crash** |
| **Clustering** | **Single S/390 system provides higher availability than any UNIX system. Parallel Sysplex is 99.999% available** | **No proven application exploitation of UNIX cluster, except for selected few that can support 4-node cluster** |
| **Mixed Workload Support** | **Architected for application isolation and dynamic workload balancing** | **Best they can do is isolate applications in separate machines** |
| **Single Points of Failure** | **No single points of failure** | **Several at system and domain level** |
| **Backup/Reorganization** | **Incremental DB2 B/R happens parallel to online processing** | **Housekeeping requires planned outages in most situations** |
| **CICS Storage Protection** | **Isolates system code and applications from errant new OLTP applications** | **New applications in a UNIX instance can overwrite other applications and crash the entire system.** |
| ***Manageability*** | | |
| **System Partitioning Flexibility** | **All LPARs share all resources, and can dynamically allocate them as workload demands fluctuate** | **Sun's domain architecture limited to manual allocation of entire system boards, and the memory and I/O they own.** |
| **SW Change Tolerance** | **SMP/E for receiving, applying, accepting fixes enables non-disruptive SW maintenance and backout.** | **No equivalent mechanism for SW maintenance causes planned outages, and can cause unplanned outages if fixes are bad** |
| **HW Change Tolerance** | **No single point of HW failure, hot pluggable components enable non-disruptive maintenance** | **Some components are hot pluggable, but other changes can take the entire system down for hours** |
| **Remote Operation/Management** | **Available through OS/390 and all major Middleware suites** | **Possible through an add on program with limited scope** |
| **Security** | **Acknowledged industry leader in all aspects of IT security driven by inate capability of OS/390 and RACF - "Can't hack S/390"** | **Questionable security - Any operation with access to the SSP can totally reconfigure the system.** |

# Data Management

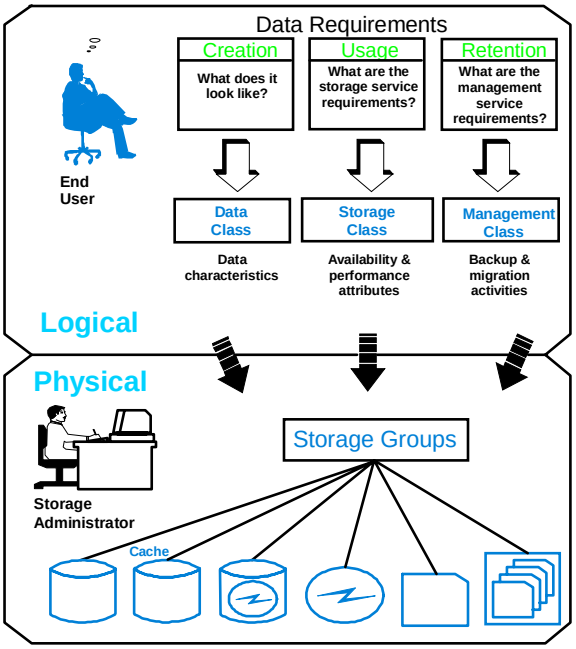




## S/390 Systems Managed Storage

- Developed in the early 1980s from IBM customer user group input
- Concept is for management of storage resources and data by system hardware and software
- First version in 1988, continually enhanced since then
- DFSMS for OS/390 provides:
  - Automated and construct-driven data management
  - Performance management
  - Program management
  - Backup and recovery
  - Space management
  - Tape management

## Systems Managed Storage




IBM Technical Support



## Systems Managed Storage Functions

- Policy management
- Space management
- Backup and recovery
- Disaster recovery
- Enterprise-wide data access

IBM Technical Support

## Sun Storage Management

- Basic Solaris storage management tools are similar to most other UNIX systems
- Solstice DiskSuite
  - Data and storage management for Solaris
  - Component of Solaris Enterprise Server (due 1H99)
- Veritas File System
- Veritas Volume Manager





## Solstice DiskSuite Functions

- Mirroring
- RAID-5
- Hot sparing
- UNIX File System Logging
  - Speeds up re-boot after a failure

## Claimed Benefits for Solstice DiskSuite

- Improved data reliability and integrity
  - Data mirroring,
  - RAID-5
  - Hot spare
  - UFS logging
  - Alternate paths
- Improved performance and system monitoring
  - Disk striping
- Increased file system capacity
  - Single file system across multiple physical disks
- Easy and flexible administration
- Centralized administration





## Other Sun Storage Management Products

- Sun Enterprise NetBackup
  - Backup/archive for multiple nodes
  - High performance database backup
- Sun Enterprise HSM
  - Optional module for Solaris backup clients
  - Archive/recall function similar to ADSM

## Veritas File System

- High performance, quick recovery file system
- Adds high availability, increased bandwidth and structural integrity to the UNIX File System (UFS)
- Veritas Claims:
  - High availability/manageability
    - Fast recovery after re-boot
    - On-line backup
    - On-line file system re-sizing
    - On-line defragmentation
  - Superior Performance
    - Allocate space in large contiguous extents, rather than small fixed sized blocks
    - Significant reduction in number of I/O operations





## Veritas Volume Manager

- Veritas Claims:
  - A disk management product that augments the UNIX disk partitioning model
  - Extends the operating system to improve reliability, performance and reliability
  - Includes RAID-0, RAID-1, RAID-0+1, and RAID-5
  - Online space allocation and configuration management, error handling, performance analysis and operator tracing

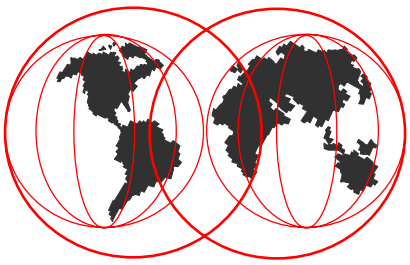
## Disk Partitioning

- UNIX allows administrators to partition disks
  - Different areas for different purposes
  - Normally have a fixed number of partitions
- Veritas Volume Manager
  - Two partitions:
    - Private - control data
    - Public - all remaining data
  - Allocates space in flexible partitions, or sub-disks, as small as one sector
  - Increased number of sub-disks available





# S/390 and E10000 Clustering



## Clustering Architectures

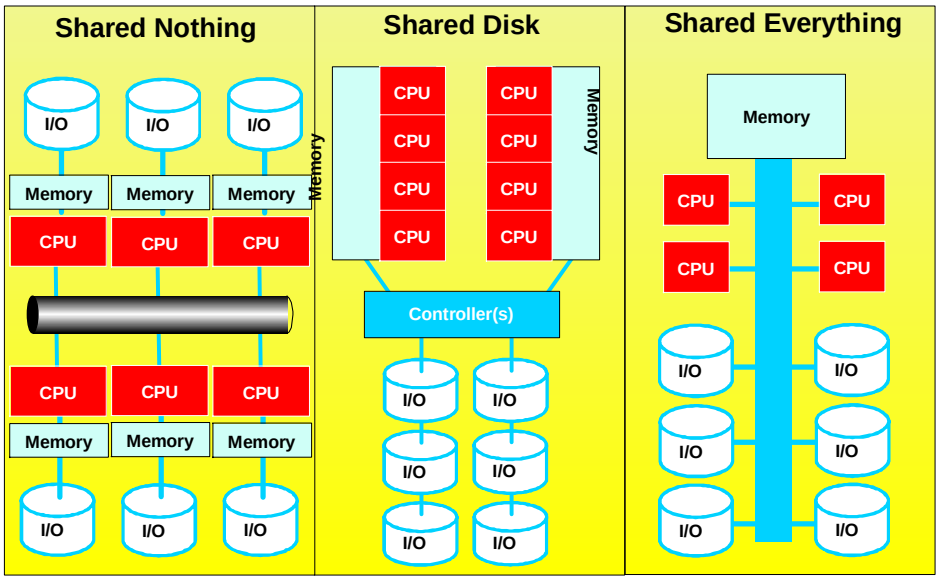
### Types of cluster

- High availability cluster
  - Fallback to designated systems
  - Improve application availability
- Parallel cluster
  - Scalability
  - Process workload in parallel
- Single system image cluster
  - All nodes appear as a single system to application and users
  - Workload balancing to distribute work
  - Single point for system administration
  - Users can still access application if one system fails





## Data Access in a Cluster



## Clustering for High Availability

- Clusters uniquely suited for high availability
  - Redundant processors, paths, I/O, network connections etc.
  - No single point of failure if configured properly
- Fault management software detects and manages failures
  - Parallel Sysplex
  - Oracle Parallel Server
  - Heart-beat monitoring (HACMP, Wolfpack etc.)





## Fault Tolerance versus High Availability

- High availability = % time system is available to the end-user
- Fault-tolerance = design techniques (error correction, majority voting, triple module redundancy etc.) used to hide failures from other parts of the system

## Causes of System Outage

- Environmental (facilities failure)
- Operations - system configuration, operations, administration
- System maintenance
- Hardware failures
- Software failures
- Process failures (administrative processes)




## Clustering for Scalability

- Not all workloads scale well with increasing numbers of engines in an SMP
- If the workload can be parallelized, then additional components (processors, memory, disk paths etc.) can lead to increased scalability

$$\text{Total execution time} = \frac{\text{parallel part}}{N} + \text{serial part}$$

- Good for engineering/scientific workloads
- Parallel Sysplex provides scalability for commercial workloads

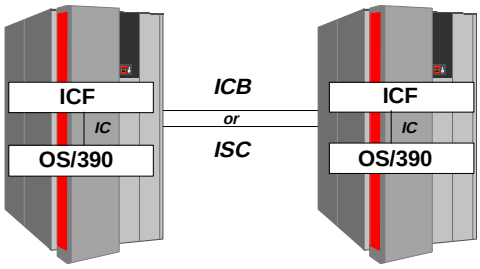
## Management of Cluster Resources

- Each node has its own resources:
  - Catalogs/directories
  - Console
  - Log files
  - Tape drives
- If resources like disk are shared, then need data integrity controls
- Management of each node's resources can:
  - Add to operational complexity
  - Decrease availability
  - Increase cost





## OS/390 Resource Sharing

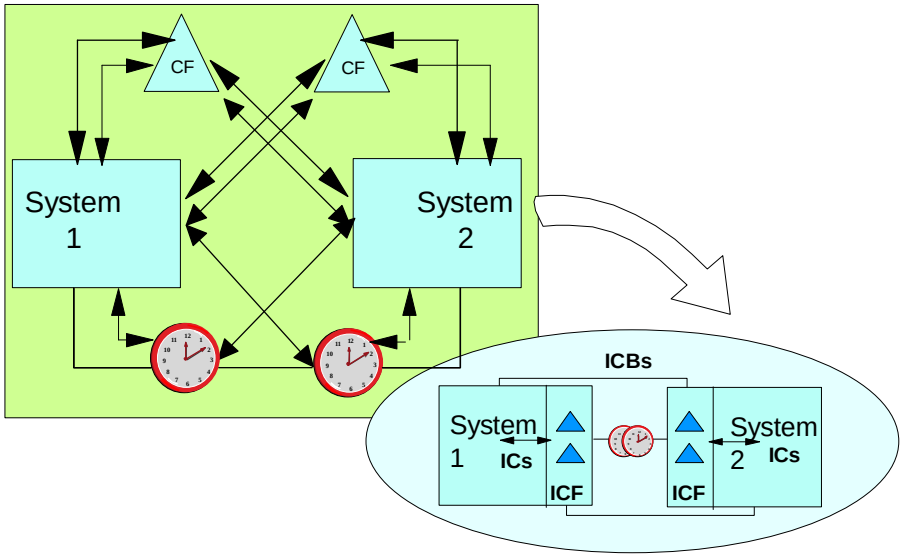



| | |
|---|---|
| **XCF Star** | **Multi-system signaling with reduced cost/management** |
| **GRS Star** | **Multi-system resource serialization for increased performance, recoverability and scalability** |
| **JES2 Chkpt** | **Multi-system chkpt for increased simplicity and reduced cost** |
| **RACF** | **Multi-system shared security profiles for simplicity/manageability** |
| **SharedTape** | **Multi-system tape sharing for reduced duplication/cost** |
| **Operlog*** | **Multi-system merged log for single system image/management** |
| **Log Rec*** | **Multi-system merged log for single system image/management** |
| **Shared Catalog** | **Multi-system shared Master Catalogs/User Catalogs for increased performance/simplicity and reduced cost** |

* Staging data sets could be used

# S/390 Hardware Technology Integration

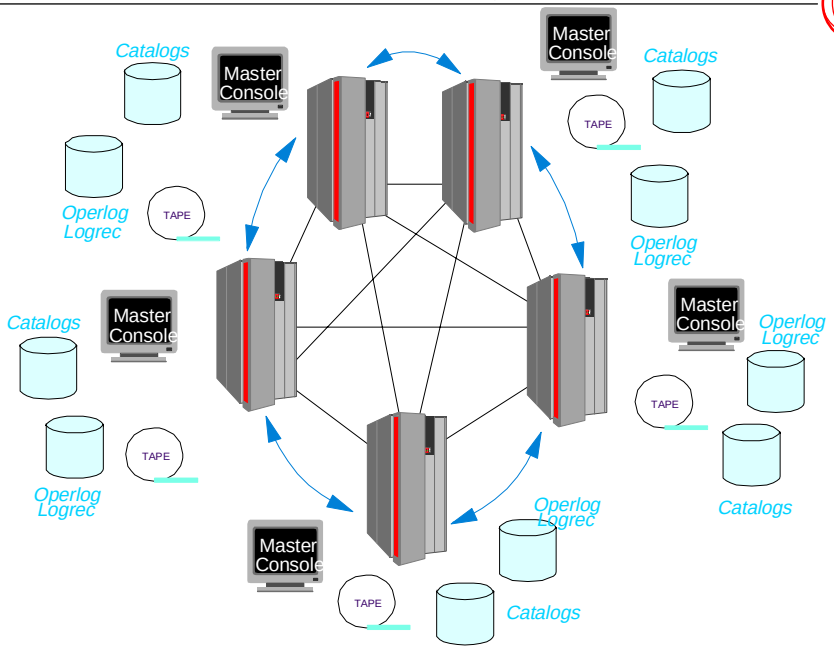
***G5/G6 Exclusives (ICBs/ICs) for an optimized Resource Sharing Environment***



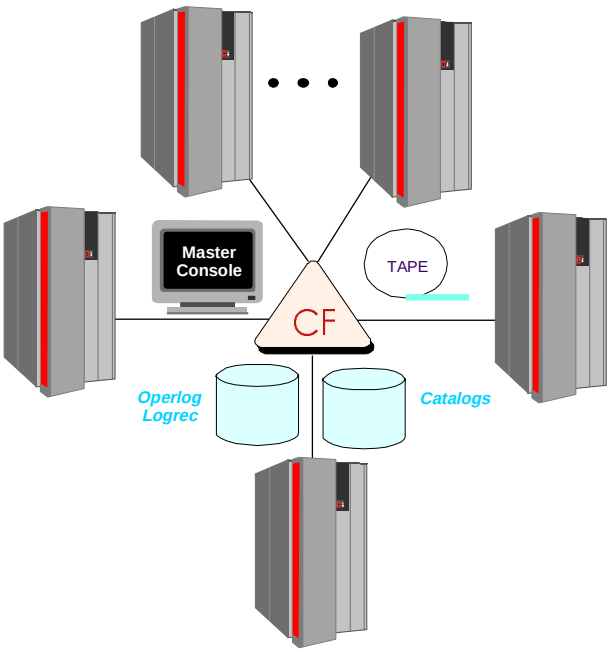





# S/390 Non-Resource Sharing



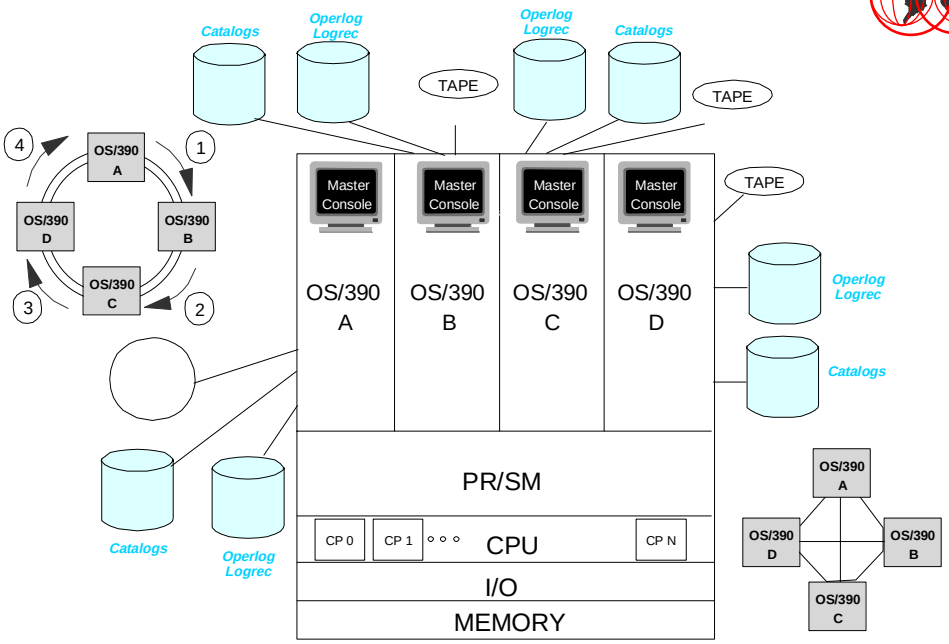
# S/390 Resource Sharing



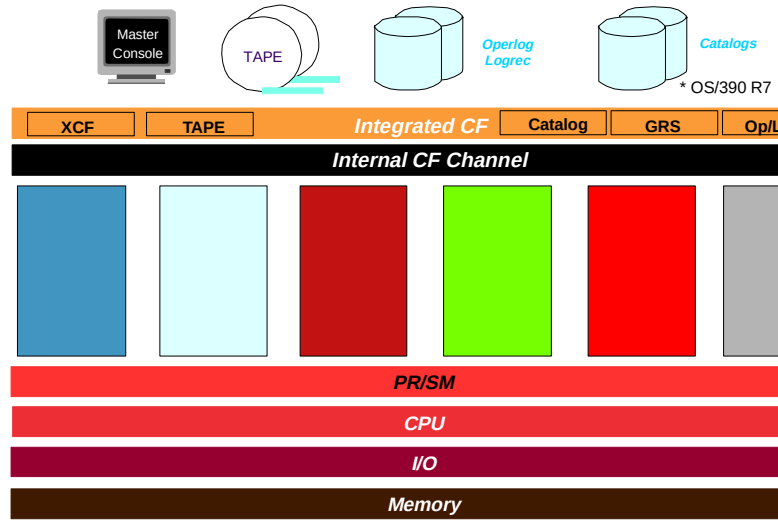





# OS/390 Images: Non-Resource Sharing



# OS/390 Images and Resource Sharing




**Single-Server Value**

**Improved Performance**
**Simplified Systems Mgmt**
**Shared Resources**
**Reduced Cost**



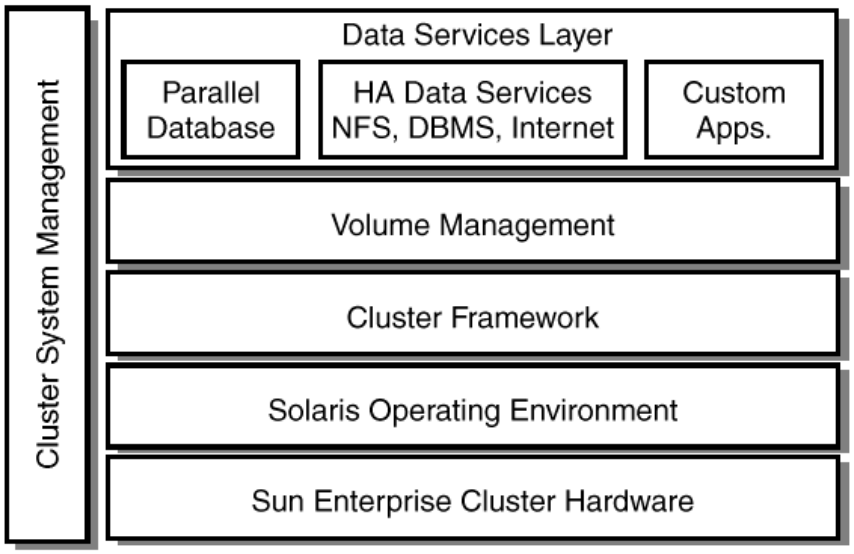


# Sun Clusters

Two Types:

- High Availability
- Parallel Database
  - Oracle Parallel Server
  - IBM DB2 UDB
  - Informix Adaptive Server

## Sun Enterprise Cluster Software Architecture



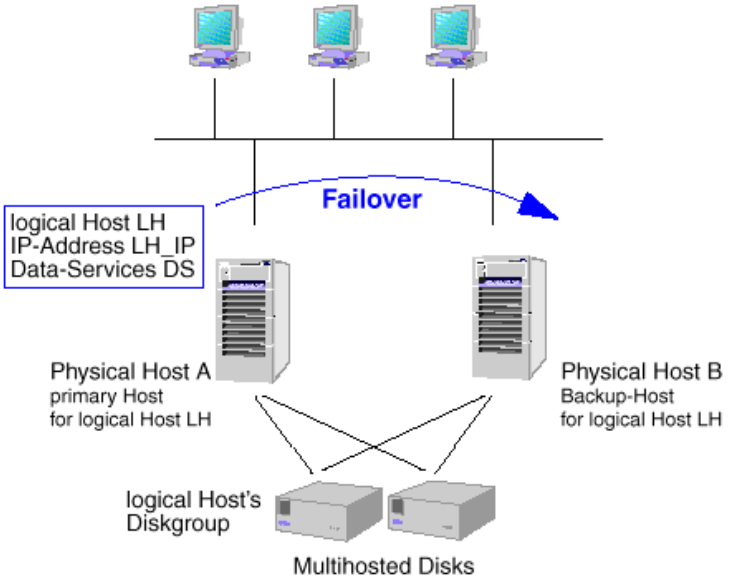





## Sun Enterprise Cluster Software Architecture

- Cluster Framework:
  - Core functionality
  - Fault monitoring, detection, recovery, failover and dynamic cluster reconfiguration
  - Able to recover from failures in CPU, memory, internal buses, disks, disk and network controllers, operating system, power and cooling, and cables
- Data Services Layer
  - Failover services for applications including NFS and databases
- High availability API
  - Allows integration of customer-written applications into the HA cluster
- Cluster System Management
  - GUI based tools
- Volume Manager
  - Used to configure storage as logical volumes
  - Enables definition of groups for easy addressing and failover

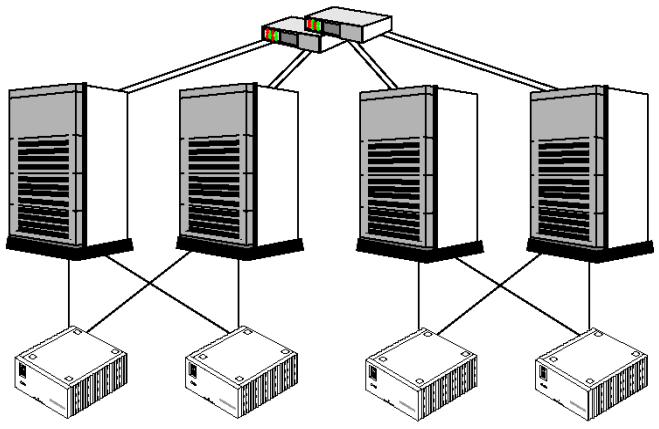
## Sun Cluster Failover





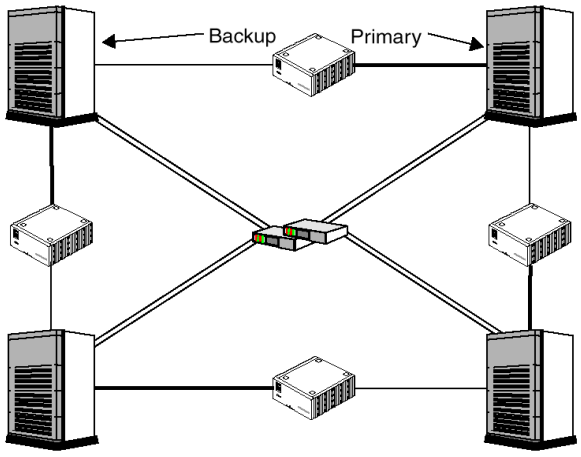



## Sun Cluster Topologies - Cluster Pairs



- Clustered in pairs
- No backup between pairs
- 4-node only for simpler administration

## Sun Cluster Topologies - Ring

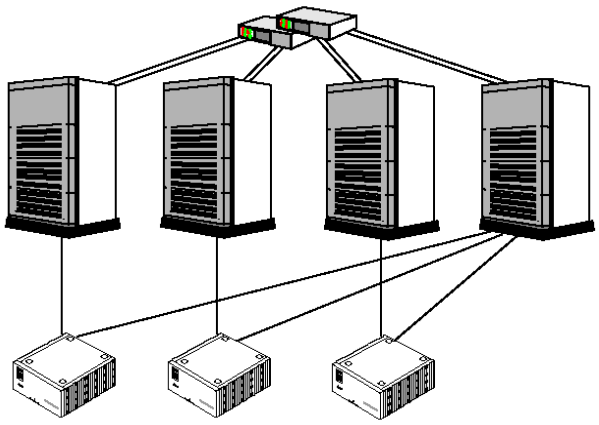



- 2, 3 or 4 nodes
- Failover only between nodes that are physically connected to the data
- Only one backup partner per node



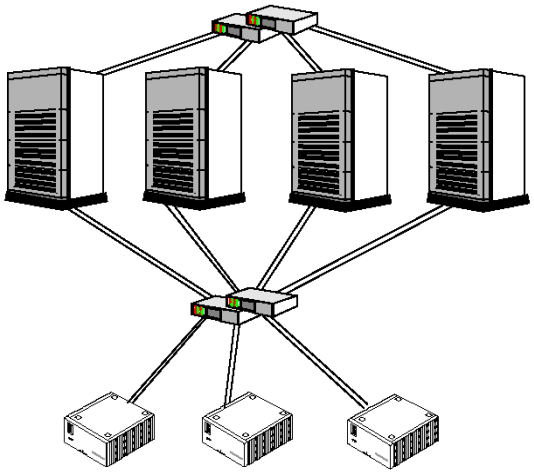


## Sun Cluster Topologies - N+1



- 2, 3 or 4 nodes
- Single server is backup for all nodes

## Sun Cluster Topologies - Scalable

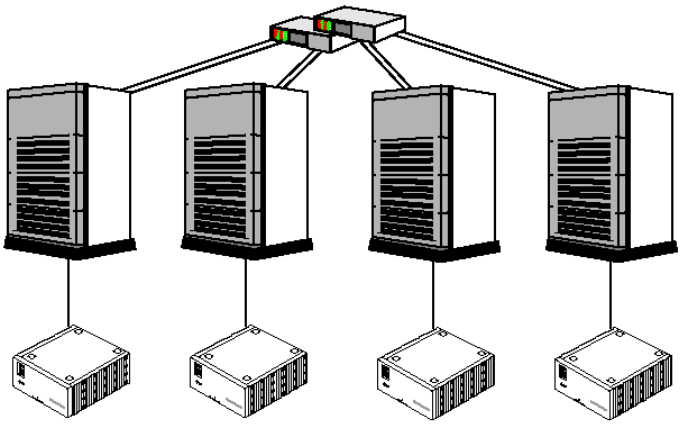


- 2, 3 or 4 nodes
- Requires Sun Network Array or RSM Array 2000 disks specially cabled to allow 4-node connection
- Most flexible and greatest redundancy of Sun clusters





## Sun Cluster Topologies - Shared Nothing



- 2, 3 or 4 nodes
- Relevant only for Informix XPS - shared nothing parallel database architecture

## Summary of Sun Cluster Topologies

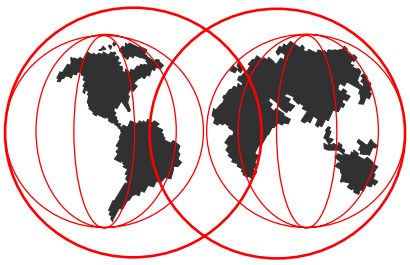
### Sun Summary of Cluster Topologies

| Topology | Highly available Data Services | Informix XPS | Oracle Parallel Server |
|---|---|---|---|
| Clustered pairs | Supported | Supported | Two-node OPS only |
| N+1 | Supported | Supported | Two-node OPS only |
| Ring | Supported | Supported | Two-node OPS only |
| Scalable | Supported | Supported | Recommended |
| Shared Nothing | Not supported | Recommended | Not supported |





# Availability



## S/390 Industry Leading Availability

### Gartner Group Survey

| Platform | Outages/Server/Year | Availability |
|---|---|---|
| **S/390 Parallel Sysplex** | **10 minutes** | **99.998%** |
| Tandem | 1.7 hours | 99.98% |
| AS/400 | 5.2 hours | 99.94% |
| **S/390 Standalone** | **8.9 hours** | **99.90%** |
| VAX servers (approx 50% clusters) | 18.9 hours | 99.78% |
| UNIX (< 10% clusters) | 23.6 hours | 99.73% |
| NT servers | 224.5 hours | 97.44% |

**Note: Availability is for un-scheduled outages ONLY**

*Platform Availability Data: Can You Spare a Minute?,* Gartner Group Research Report, 981029





## S/390 Hardware Availability Features

- ECC and parity checking
- Instruction retry
- Logging of errors
- Call-home to IBM
  - Auto-dispatch of IBM support
  - Down-loading of microcode patches
- Redundant components:
  - Memory arrays
  - Spare CPs
  - Spare SAP
- Application Preservation Facility
- Capacity Upgrade-on-demand
- Parallel Sysplex Coupling Facility

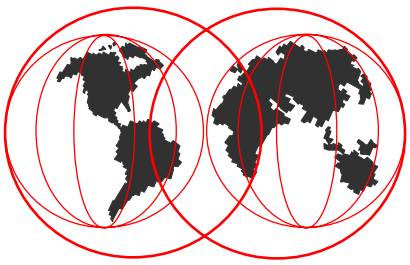
## OS/390 Software Availability Features

- Support for hardware instruction retry
- Logging of errors
  - Same log file as hardware errors
  - Down-loading of software fixes
- MVS Integrity Guarantee
- FRRs and ESTAEs
  - Used by system
  - Used by subsystems
- Address space protection mechanisms
  - Isolation of separate applications
- Use of storage protect keys
  - Base keys
  - CICS SSP and SGP
- System state and problem state
- Authorization mechanisms
- Parallel Sysplex





# Partitioning



## IBM Partitioning History

- Physical partitioning:
  - 1968 - S/360 MP
  - 1973 - 158MP, 168MP
  - 1976 - 3033MP
  - 1984 - 3084Q
  - 1985 - 3090-400
  - 1992 - 9021-900
  - 1993 - 9021-982
  - 1994 - 9021-9X2
- Logical Partitioning (software):
  - 1970 - CP67
  - 1972 - VM/370

  etc. to VM/ESA
- Logical Partitioning (hardware/microcode)

  1987 - PR/SM

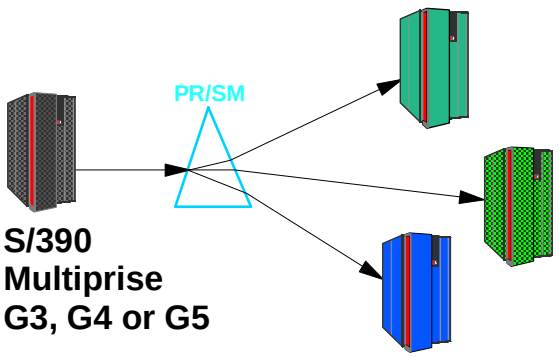




## S/390 Logical Partitioning

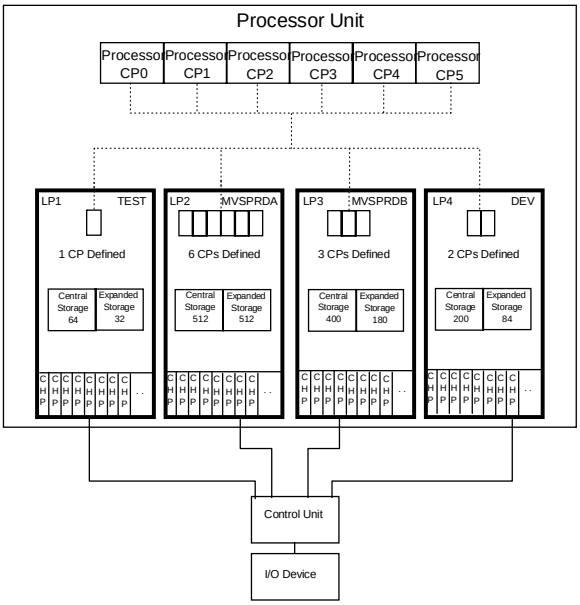
**PR/SM**

**Processor Resource/Systems Manager**




- Logical partitioning of physical processor
  - Processors, processor storage and channels can be partitioned
  - Up to 15 logical partitions
- Supports different architectures
  - S/370, 370-XA, ESA/370, ESA/390
- Maximizes response time and throughput
- Security using physical isolation
  - B1 security compliant
- ESCON Multiple Image Facility (EMIF)
  - Supports sharing of ESCON channels among PR/SM logical partitions *
- Concurrent LPAR LIC maintenance *
- Standard on S/390 family

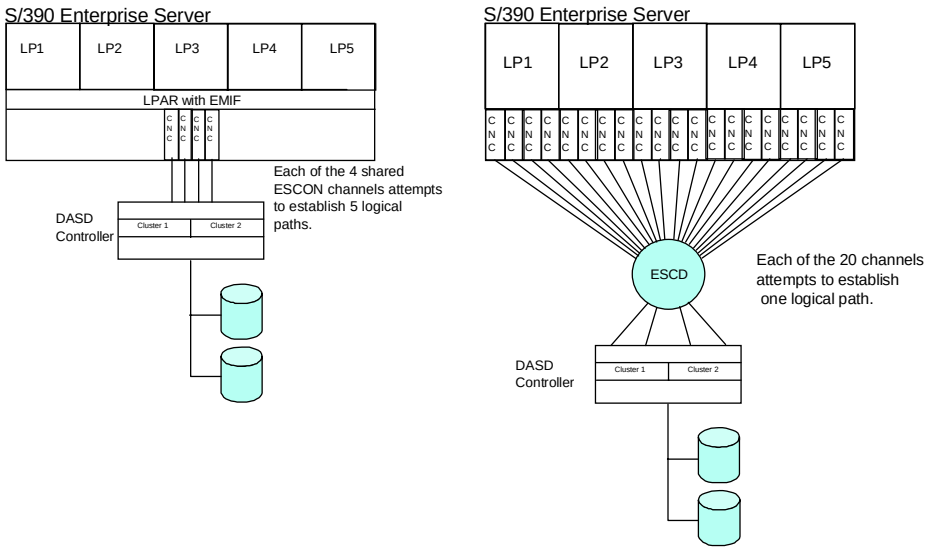
# S/390 PR/SM Logical Partitions








# PR/SM and Channels



## PR/SM Dynamic Change

- Partition weights and capping
- Number of CPS available to a partition
- I/O devices and paths
- Memory - Dynamic Storage Reconfiguration





## UNIX Vendor Partitioning

- Sun Dynamic System Domains:
  - Physical partitioning
- Compaq (DEC) Alpha ??
  - Software
- Sequent
  - Software

## Sun Dynamic System Domains

- Currently 8 physical domains
- Statement of Direction to increase to 16 domains
- Dynamic Reconfiguration Attach/Detach
  - Entire board
  - Can require careful planning to achieve reconfiguration

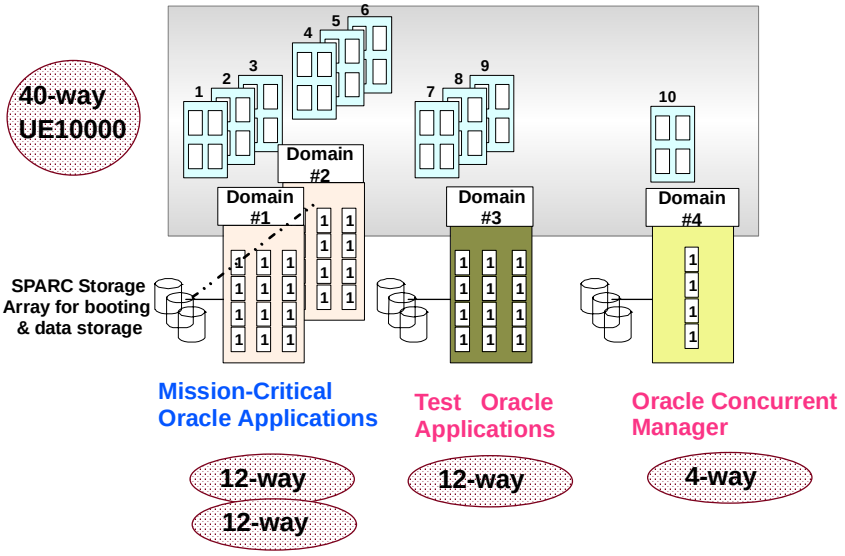




## Sun Dynamic System Domains

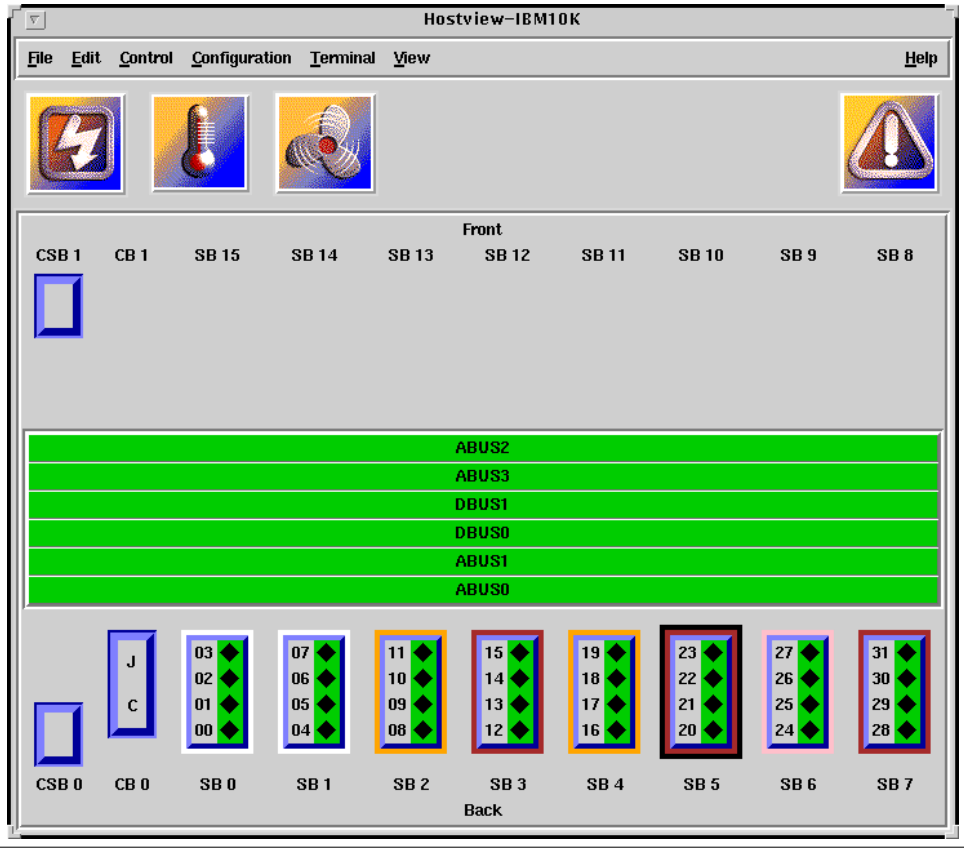
- Multiple environments on one server
- Single server for multiple departmental systems
- Server Consolidation
- Shared I/O or network functions via dynamic reconfiguration
- Software migration and upgrades
- Configuring for resource requirements
- Client-server applications on one server
- Internet Firewalls

# Sun UE10000 Partitioning



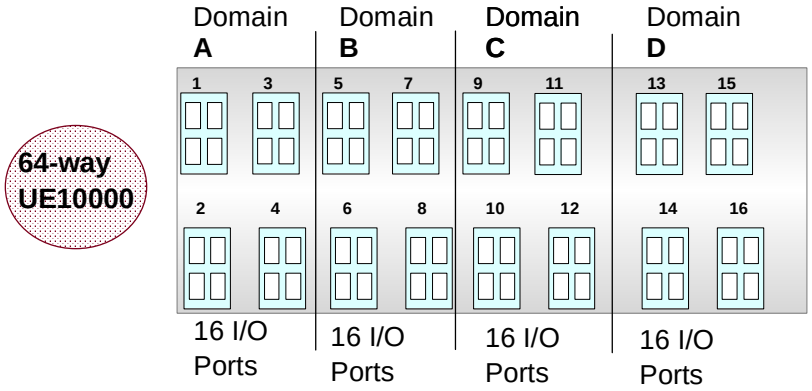

IBM Technical Support





## Sun UE10000 Multi-Domain Example (1)

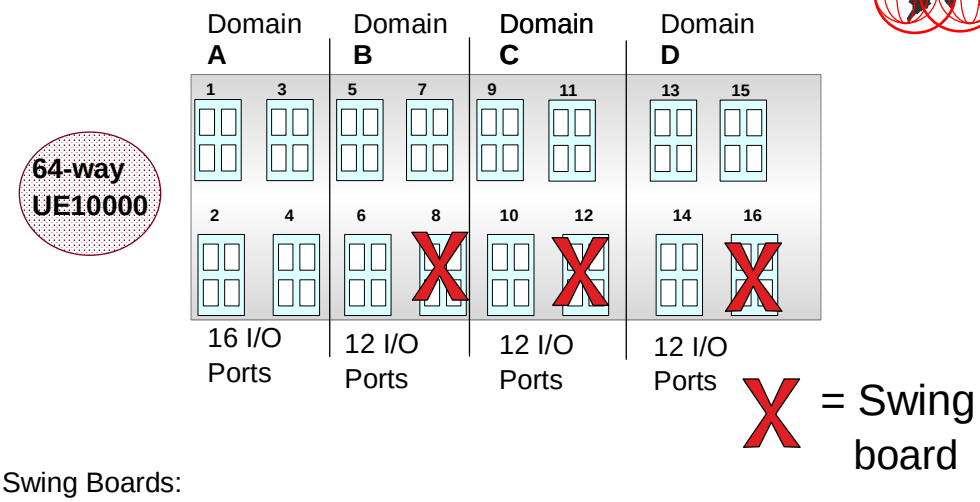
**4 domains**







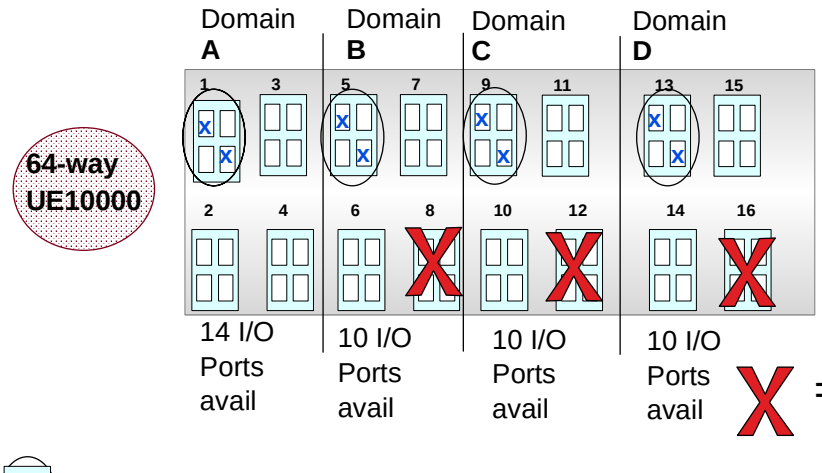

## Sun UE10000 Multi-Domain Example (2)




Swing Boards:

1) Not practical to have disk on swing board
2) Will use 3 boards for swing capacity
3) Will use I/O ports for Tape so that one domain at a time can be backed up

## Sun UE10000 Multi-Domain Example (3)




Anchor board

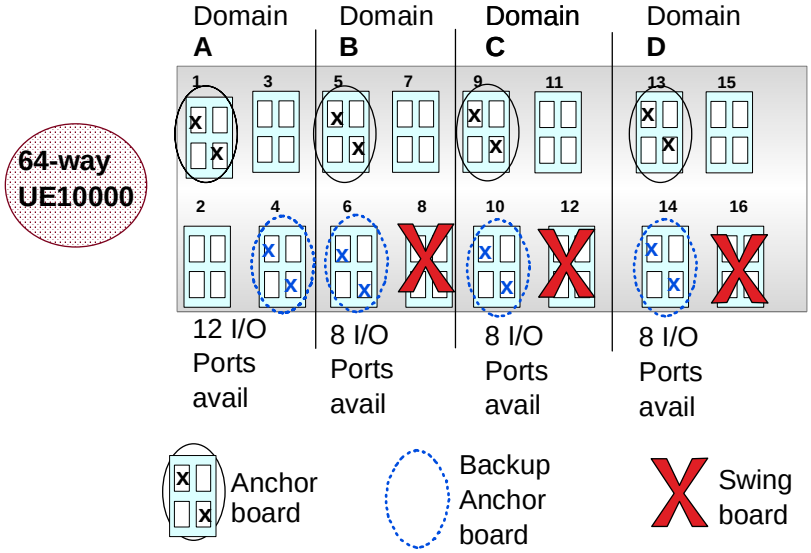
Port 0 = boot Disk

Port 1 = Quad Ethernet 1 port for SSP, other ports for private System Support

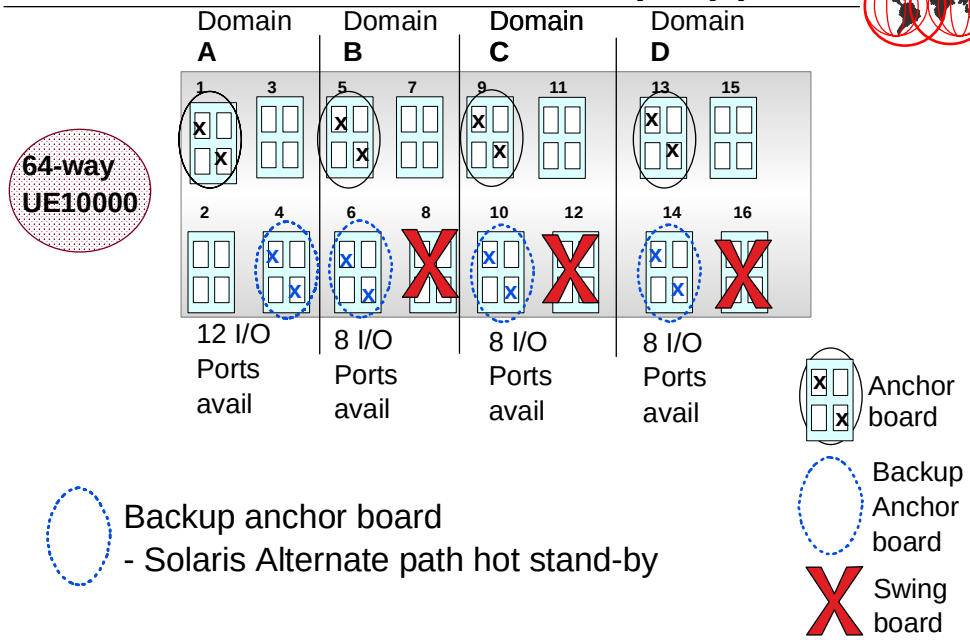




## Sun UE10000 Multi-Domain Example (4)



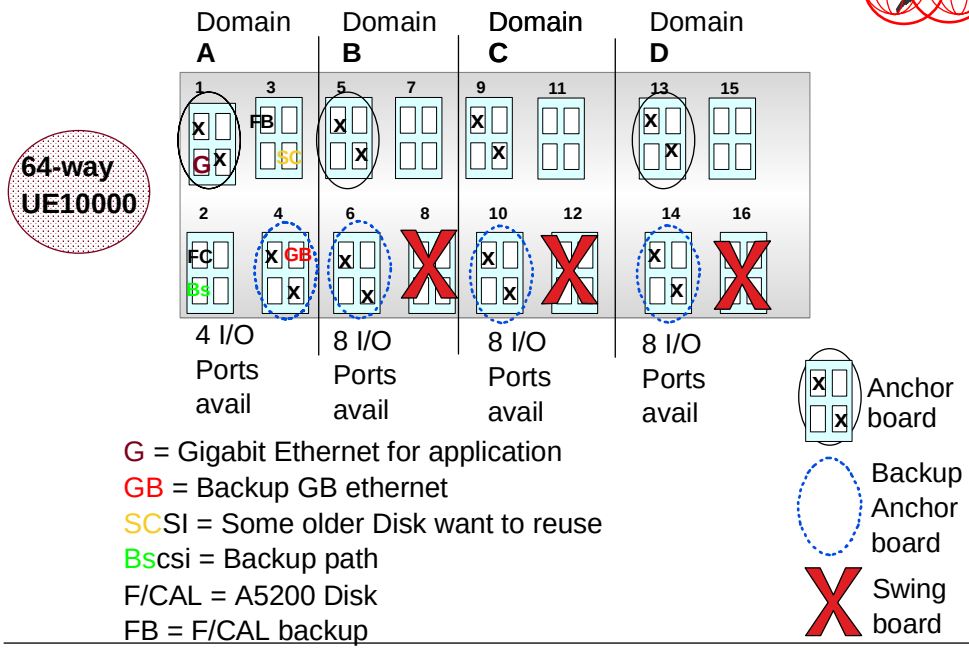
# Sun UE10000 Multi-Domain Example (5)





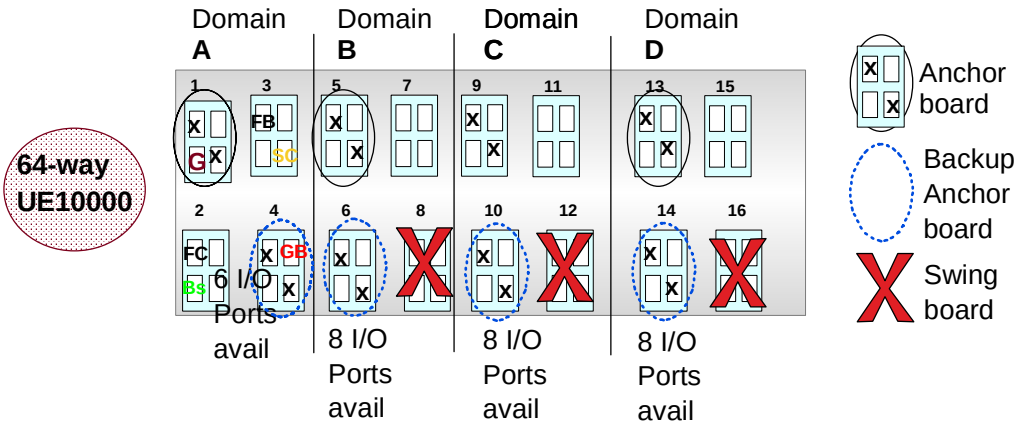



# Sun UE10000 Multi-Domain Example (6)




G = Gigabit Ethernet for application
GB = Backup GB ethernet
SCSI = Some older Disk want to reuse
Bscsi = Backup path
F/CAL = A5200 Disk
FB = F/CAL backup

# Sun UE10000 Multi-Domain Example (7)

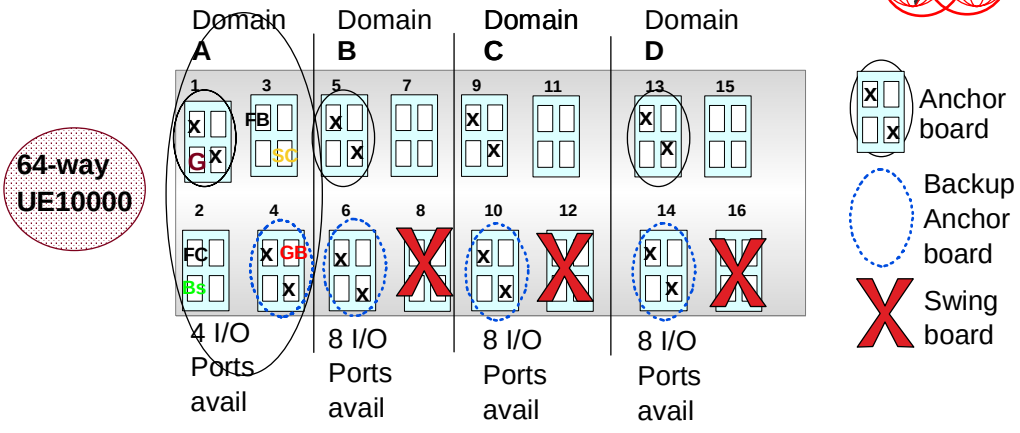



G = Gigabit Ethernet for application
GB = Backup GB ethernet
SCSI = Some older Disk want to reuse
Bscsi = Backup path
F/CAL = A5200 Disk
FB = F/CAL backup



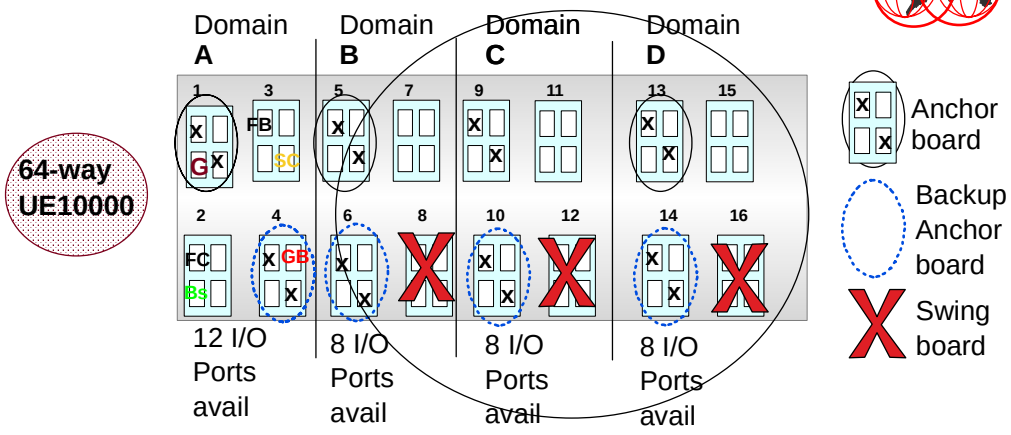


# Sun UE10000 Multi-Domain Example (8)




Domain A considerations
1) Configured for High Availability
2) Can survive any board failure
3) Board can be removed but will suffer risk of outage
4) Balanced I/O across Sbus

## Sun UE10000 Multi-Domain Example (9)



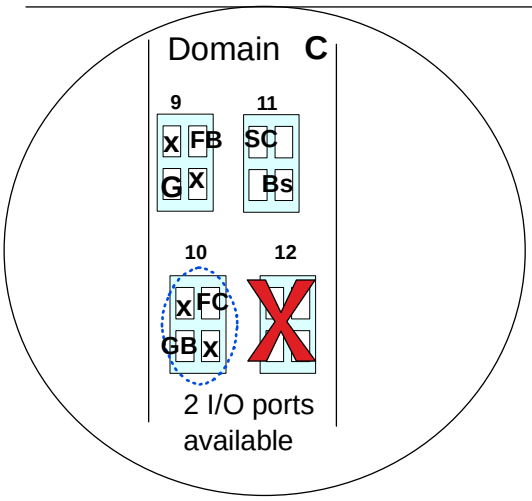

Domains B, C, D

1) Not so fast!
2) Tradeoffs must be made, either
   - Ability to either remove a board, or
   - Availability



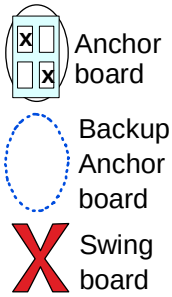


## Sun UE10000 Multi-Domain Example (10)




Want to add 6 devices

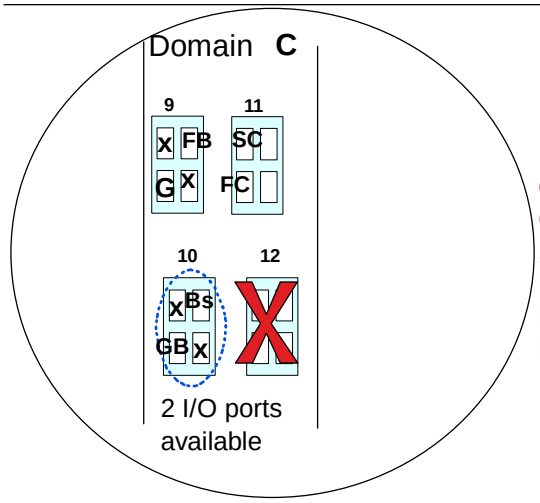
G = Gigabit Ethernet for application
GB = Backup GB ethernet
SCSI = Some older Disk want to reuse
Bscsi = Backup path
F/CAL = A5200 Disk
FB = F/CAL backup




Options

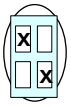
1) 9 & 10 backup each other
2) board failure 11 will crash system
3) Can swing board 9 or 10

## Sun UE10000 Multi-Domain Example (11)




Want to add 6 devices

G = Gigabit Ethernet for application
GB = Backup GB ethernet
SCSI = Some older Disk want to reuse
Bscsi = Backup path
F/CAL = A5200 Disk
FB = F/CAL backup




Anchor board

Backup Anchor board

Swing board

Options
1) Board failure won't crash system
2) No board can be removed





## Processor Sets

- Solaris Resource Manager permits an SMP to be effectively partitioned into Processor Sets
- All CPUs in Processor Set are dedicated to one application, thus permitting several different workloads on one server
- Default is a single system-wide Processor Set
- Basis for future support of NUMA architecture machine
  - Group processors with longer data access times

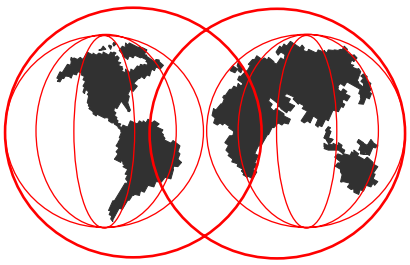
## S/390 PR/SM versus Sun UE10000 Partitioning

| | Sun UE10000 | S/390 PR/SM |
|---|---|---|
| Physical Partitioning | X | X |
| Logical Partitioning | (via processor sets) | X |
| Dynamic reconfiguration | X | X |
| Automatic reconfiguration | | X |
| Dedicated processors | X | X |
| Shared processors | | X |
| Memory increments independent of processor | | X |
| I/O sharing between partitions | (future) | X |
| Workload management | (via Solaris Resource manager and processor sets) | X |
| Weights and capping | | X |
| Single view management | | X |
| Different clocks for Y2K testing | X | X |





# S/390 Added Value Features



## S/390 Added Value Features

- Cryptography
- Hardware Data Compression
- Performance Assists
- Unique I/O functions





## Cryptographic Co-Processor

- Solaris Resource Manager permits an SMP to be effectively partitioned into Processor Sets
- All CPUs in Processor Set are dedicated to one application, thus permitting several different workloads on one server
- Default is a single system-wide Processor Set
- Basis for future support of NUMA architecture machine
  - Group processors with longer data access times

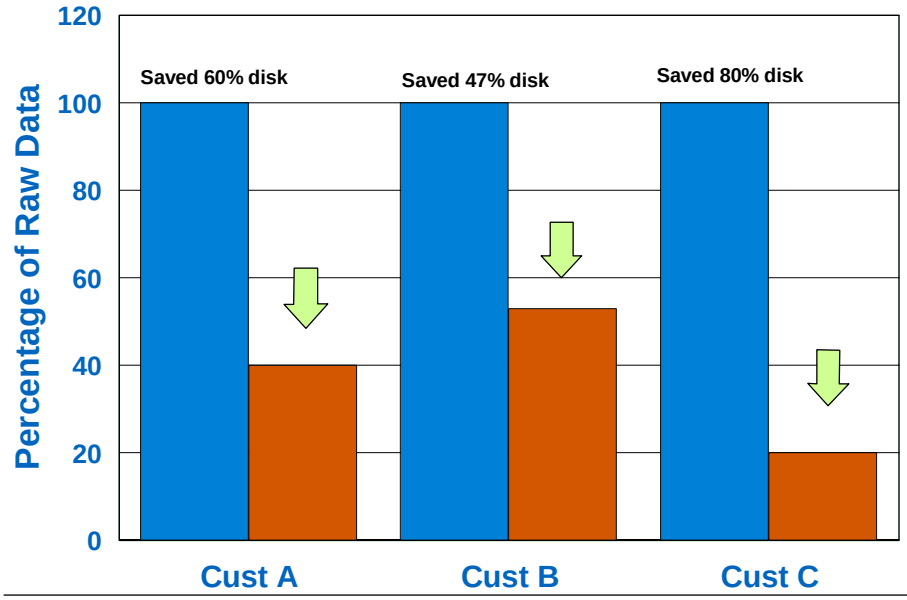
## Hardware Data Compression

- Data compression on disk can typically reduce space requirements by 50% or more
- Compression by software is VERY expensive in CPU time
- S/390 Hardware Data Compression:
  - Standard feature
  - Standard APIs for application and DBMS use
  - Overheads range from 3-5 time to 100-1000 times LESS than software compression
  - Benefits also in reducing channel load, and in increasing data-in-memory





## Typical Compression Savings



## Performance Assists

- ADMF
- Asynchronous PageOut Facility
- Checksum Facility
- Extended Sorting (DB2)
- Sorting instructions (DFSORT)
- Immediate and Relative Instruction family
- Logical String Assist
- Perform Locked Operation (PLO)
- Program-Call-Fast Facility
- Scalar Square Root
- Set Address Space Control Fast (SACF)





## Unique I/O Features

- Command Retry
- Concurrent Sense
- I/O device self-description
- RVA Disk and SnapShot
- Dynamic Path Reconnect
- Alternate Path Retry
- etc.,

# End of Presentation