
An Oracle White Paper
June 2010

# Sun SPARC Enterprise Servers System and Resource Management

ORACLE

# Introduction

Though the importance of performance is indisputable, fast platforms alone are not enough to respond to the continually changing demands of today's high-pressure, globally competitive, cost-sensitive business and technical environments, or to respond to rising power consumption and other constraints in the datacenter. Sophisticated system, resource, and workload management tools are needed to harness the power of every system's performance, flexibility, and availability. Oracle's Sun SPARC Enterprise servers, with their complement of management tools, are designed specifically as general-purpose application, database, data warehousing, and consolidation servers. They are especially suited to address the needs of enterprise datacenters, with a goal of increasing performance and flexibility while consolidating systems to reduce datacenter costs and complexity.

Applications are no longer simply standalone, but all of these applications on separate systems create a level of complexity both from the perspective of administering and maintaining them, as well as from the perspective of successfully integrating application services. The Sun SPARC Enterprise servers help to manage consolidated applications by incorporating advanced manageability and innovative technologies designed to provide the agility to meet ever-changing demands for capacity and performance. Examples of these capabilities include fifth-generation Dynamic Domains supported at the sub-board level, Dynamic Reconfiguration, Oracle Solaris Containers, and full system redundancy.

The increased urgency and accelerated time frames of internal technical initiatives lead to server sprawl, or the deployment of multiple single-purpose servers to achieve needed availability or scalability. The result is underused resources and increased complexity. Consolidation collapses the functionality of multiple servers into a smaller number of systems, reducing costs and management complexity. The Sun SPARC Enterprise servers, together with Oracle Solaris 10, offer an extensive array of system, resource, and workload management capabilities to allow IT organizations to reduce costs, decrease complexity, and rapidly respond to changes in demand.

## Optimal Hardware for Consolidation

The Sun SPARC Enterprise servers are a uniquely robust class of servers equipped to manage the job of consolidation and resource management (see Figure 1). The Sun SPARC Enterprise M9000 server offers the highest availability, highest absolute performance, highest scalability, and the most sophisticated control of resources in Oracle's extensive server product line. It supports up to 64 dual-core SPARC64 VI 64-bit processors (up to 128 cores), or up to 64 quad-core SPARC64 VII 64-bit processors (up to 256 cores), or combinations of SPARC64 VI and SPARC64 VII processors. It also features 2 TB of memory; 128 hot-swappable PCI Express (PCIe) I/O slots; up to 64 internal disks; a new high-performance interconnect capable of up to 737 GB/sec; and support for external I/O, up to 24 domains, and thousands of software applications for the Oracle Solaris operating system (OS). Exhibiting 7.5 times greater system bandwidth than previous generations and scalability in every dimension, the Sun SPARC Enterprise M9000-64 server sets a new standard for performance and configuration flexibility in large symmetric multiprocessing platforms.

**Figure 1. Sun SPARC Enterprise server family.**

The SPARC64 VI processor contains two SPARC V9 cores running 2.15 GHz, 2.28 GHz, or 2.4 GHz; 6 MB on-chip shared L2 cache; two vertical threads per core; and 128 KB of I cache and 128 KB of D cache per core—and using the latest 90 nm process technology. It delivers up to 1.5 times the performance of 1.8 GHz UltraSPARC IV+ processors, with the largest gains exhibited in transactional workloads. The SPARC64 VII processor contains four SPARC V9 cores running either 2.4 GHz, 2.52 GHz, 2.53 GHz, 2.77 GHz, or 2.88 GHz, with up to 6 MB on-chip shared L2 cache, two simultaneous threads per core, 64 KB of I cache and 64 KB of D cache per core, and using 65 nm process technology. It delivers 1.4 to 1.7 times the performance of the SPARC64 VI processor.

The Sun SPARC Enterprise M9000-32 server offers high-end computing with a smaller system than the Sun SPARC Enterprise M9000-64 server, and the ability to upgrade to the Sun SPARC Enterprise M9000-64 server by adding another cabinet. The Sun SPARC Enterprise M9000-32 server offers half the maximum configuration of the Sun SPARC Enterprise M9000-64 server and, consequently, lower

acquisition and operating costs. Supporting up to 32 CPUs (up to 128 cores), 2 TB of memory, 64 hot-swappable PCIe I/O slots, 32 internal disks, external I/O, and up to 24 domains, the Sun SPARC Enterprise M9000-32 server is ideal for consolidation and mission-critical applications.

The Sun SPARC Enterprise M8000 server is the entry system for the high-end Sun SPARC Enterprise server family. Using the same building block components used by the Sun SPARC Enterprise M9000 servers, it is fully upgradeable to the Sun SPARC Enterprise M9000-32 server. The Sun SPARC Enterprise M8000 server supports up to 16 CPUs (64 cores), 1 TB of memory, 32 hot-swappable PCIe I/O slots, 16 internal disks, external I/O, and up to 16 domains.

The Sun SPARC Enterprise server family also includes two midrange, rackmount systems—the Sun SPARC Enterprise M4000 server and the Sun SPARC Enterprise M5000 server. These systems offer an economical price for the performance, enabling datacenter support for up to four domains and Dynamic Reconfiguration, and featuring greatly improved reliability, availability, and serviceability functionality. The Sun SPARC Enterprise M4000 server's six rack unit (6U) enclosure supports up to four dual-core SPARC64 VI processors (8 cores), or up to four quad-core SPARC64 VII processors (16 cores), 256 GB of memory, two internal disks, external I/O, and two domains. The 10U Sun SPARC Enterprise M5000 server supports up to eight dual-core SPARC64 VI processors (16 cores) or up to eight quad-core SPARC64 VII processors (32 cores), 512 GB of memory, four internal disks, external I/O, and four domains. Both systems feature a new higher-performance interconnect and industry standard PCIe I/O, providing more than four times the memory bandwidth and 5 to 10 times the I/O bandwidth of UltraSPARC IV+ systems. All Sun SPARC Enterprise M-Series servers can mix SPARC64 VI and SPARC64 VII processors in the same system, as well as in the same domain. Mixing processor speeds is supported, without clocking down the faster CPUs.

## Managing Consolidated Applications

In the past, datacenters focused on increasing predictability and discipline, while optimizing resources for maximum efficiency to process static workloads. Today, predictability has given way to extreme volatility, thus changing many of the core requirements for system and resource management. The advent of the global economy, e-commerce, Web-centric services, mobile devices, and dynamic workloads—along with the need to cut costs—is rapidly changing the computing environment.

These days, applications are generally components in an application service that is likely to be integrated into another application service. For example, an application service for a customer relationship management (CRM) system typically consists of a database, an application server, and Web server components, each deployed on its own server. The CRM system might then integrate with other systems in the company, such as order entry or marketing systems—adding to an already complex system. Reducing the complexity of the infrastructure by introducing a consistent, consolidated platform can free IT operations staff to focus on more-strategic projects, enable resources to be shared between applications for greater resource use, and provide a more-standardized environment to integrate applications services to improve business processes.

The Sun SPARC Enterprise servers offer the solution, delivering advanced reliability features, including instruction-level retry, protected static random-access memory and registers, extended error correction

code (ECC) memory and mirroring, end-to-end ECC protection, hot-swappable components, and hardware redundancy—all at open systems prices.

As consolidation platforms, the Sun SPARC Enterprise servers are easier to manage than similar systems from competitors, due to the following advantages:

- The Sun SPARC Enterprise M9000-32 and Sun SPARC Enterprise M9000-64 servers offer 24 hardware partitions or domains.
- The Sun SPARC Enterprise M4000, Sun SPARC Enterprise M5000, Sun SPARC Enterprise M8000, and Sun SPARC Enterprise M9000 servers support more than 8,000 Oracle Solaris Containers per domain.
- The Sun SPARC Enterprise M4000 to Sun SPARC Enterprise M9000 servers require only one Oracle Solaris instance for thousands of containers, while competitors require one OS instance per software partition. Fewer instances are easier to manage.
- Oracle Solaris Containers is free, while partitions from competitors must be purchased.

### Managing Consolidation

The problem of server sprawl not only requires greater capital investments, but it leads to significantly higher training, operational, and datacenter (floor space, heating, cooling) costs—as well as an inability to adapt to changing needs. Server consolidation can move the operation to a smaller number of systems while also increasing ROI by sharing existing resources across multiple applications. Server consolidation requires large servers that are capable of running more than one application simultaneously, an OS with proven scalability, the ability to change and grow configurations dynamically without impacting service, fine-grained control over system resources, and the ability to isolate applications from each other.

## Managing Sun SPARC Enterprise Servers

Oracle understands the difficult task that IT operators face today. Managers of large systems look for ways to automate, integrate, and quickly adapt to manage ever-increasing and changing workloads. That's why Oracle set out to make available a set of more-powerful management tools that simplify administration through streamlined procedures capable of enhancing existing skill sets. These tools include the following:

- **eXtended System Control Facility.** This is a GUI-based or command-line interface (CLI)–based set of system management applications.
- **Oracle Enterprise Manager Ops Center.** This a GUI-based single point of management for systems, storage, and operating systems.
- **Oracle Solaris Management Console.** Perform administrative tasks involving users, projects, and jobs with this tool.
- **Configuration and Service Tracker.** This tool monitors hardware configuration changes.

- **Sun Net Connect Services.** These services enable self-monitoring, configuration and patch collection, and reporting.
- **Oracle Solaris features.** These features enable upgrading of the OS and management of the system and application services.

## eXtended System Control Facility

The Sun SPARC Enterprise servers provide system management capabilities through the eXtended System Control Facility (XSCF) firmware, which is preinstalled on the service processor boards. XSCF firmware consists of system management applications and two user interfaces: XSCF Web, which is a browser-based GUI, and XSCF Shell, which is a terminal-based CLI. XSCF Web uses the secure version of HTTP and the Secure Sockets Layer (SSL) / Transport Layer Security protocols for connection to the server, which is connected to a network. It also uses these protocols for Web-based support of server status display, server operation control, and configuration information display.

XSCF firmware is a single centralized point for managing hardware configuration, controlling the hardware monitor and cooling system (fan units), monitoring domain status, powering on/off peripheral units, and monitoring errors. XSCF centrally controls and monitors the server. XSCF includes a partitioning function to configure and control domains. It has a function to monitor the server through an Ethernet connection to enable remote control. It also reports failure information to the system administrator.

XSCF provides the following functions:

- **Power control for the server system and domains.** XSCF has power-on and power-off control of the server and temperature control by fan operation. The IT operator can press the power switch button on the operator panel to turn the whole system on or off, or to turn on and off the supply of power to the whole system or individual domains.
- **Initial system configuration.** XSCF configures the initial hardware settings of the XSCF unit and initializes hardware to start the OS. It also controls the initial system configuration information.
- **Internal cabinet configuration, recognition, and domain configuration control.** XSCF displays the system configuration status, and it creates and changes domain configuration definitions. It also provides domain start and stop functions.
- **Dynamic reconfiguration.** XSCF supports dynamic system board configuration change operations and dynamic reconfiguration of a domain while the system is operating.
- **Console redirection.** XSCF provides a function that displays the Oracle Solaris console of each domain from XSCF through the LAN or serial port of the XSCF unit. With a secure shell (SSH) or telnet connection to XSCF, the IT operator can use the OS console function.
- **Component configuration recognition and temperature/voltage monitoring.** XSCF monitors component information such as the configuration status and the serial numbers of the components in the server. If an abnormality is detected in the component configuration, it is displayed and

reported. XSCF periodically monitors and displays the temperature inside the server, the ambient temperature, component temperatures, voltage levels, and fan speeds.

- **Firmware update.** The Web browser and commands can be used to download new firmware, such as XSCF firmware or OpenBoot programmable read-only memory firmware, without stopping the domain. It can also be used to update firmware without stopping other domains.
- **Monitoring server status and fault management.** XSCF displays the status and, if necessary, degrades the faulty parts, degrades the faulty domains, or resets the system to prevent another problem from occurring.
- **Hardware fault information collection.** XSCF collects hardware fault information quickly and saves it on the XSCF itself. The XSCF hardware failure log makes it possible to identify the location of a failure. The log also provides assistance in anticipating failures on the server and immediately reports precise information about failures.
- **Support of hot-swapping components.** XSCF supports maintenance work with XSCF Shell during hot-swapping.
- **Monitoring and notification during operation.** Using the network function of the cabinet, XSCF accesses the server to provide the following services:
  - Monitoring the server even when the OS is inactive.
  - Enabling remote operation of the server.
  - Reporting error messages by e-mail to specified addresses.
  - Trapping notification with the Simple Network Management Protocol (SNMP) agent functions. XSCF supports the SNMPv2 and SNMPv3 releases of management information base II (MIB-II) and the SNMPv1 release of MIB-I.
- **Security.** XSCF provides an encryption function using SSH or SSL. Any operation error or unauthorized attempt to access XSCF functionality is recorded in a log. Access details, such as which users logged in and the operations they executed, can also be recorded in an audit trail. This information can be used to troubleshoot system errors.
- **XSCF user account control.** XSCF controls the user accounts (system administrator, domain administrator, operator, and field engineer) for XSCF operations.
- **Capacity on Demand (COD) management.** XSCF firmware provides setup and management of COD boards and COD permits.
- **I/O box management.** Displays I/O box information, configures the I/O box, and can power on and off specific I/O boards or power supply units.

XSCF firmware has two networks for internal communication: the Domain to Service Processor Communications Protocol (DSCP) and XSCF. The DSCP network provides an internal communication link between the service processor and the Oracle Solaris domains. The DSCP service provides a secure TCP/IP–based and point-to-point protocol–based communication link between the

service processor and each domain. Without this link, the service processor cannot communicate with the domains.

The XSCF network provides an internal communication link between the two service processors in a high-end Sun SPARC Enterprise M8000 server or a Sun SPARC Enterprise M9000 server. In a high-end server with two service processors, one service processor is configured as active, and the other is configured as standby. This redundancy of two service processors enables them to exchange system management information and, in case of failover, to change roles. All configuration information on the active service processor is available to the standby service processor.

The SPARC Enterprise M-Series servers can be managed by Oracle Enterprise Manager Ops Center, or by third-party management tools. XSCF can communicate with the third-party tools using either built-in SNMP or Oracle Enterprise Manager Ops Center software agents. This allows administrators to simply add a Sun SPARC Enterprise M-Series server to their management tool of choice.

## Oracle Enterprise Manager Ops Center

Oracle Enterprise Manager Ops Center allows IT administrators to actively manage and monitor infrastructure resources from virtually anywhere on the network. Oracle Enterprise Manager Ops Center simplifies the management of Oracle Solaris, Linux, and Windows using an advanced knowledgebase while enabling automated lifecycle processes. It also provides full lifecycle management of virtual guests, including resource management and mobility orchestration. Oracle Enterprise Manager Ops Center helps customers streamline operations and reduce downtime.

### Asset Management and Discovery

Oracle Enterprise Manager Ops Center automatically draws out the relationship between servers and their associated service processors to hypervisors and operating system instances. Assets can be grouped manually based on location or business function or a smart groups feature can automatically sort the topology. Assets can be registered through inventory software services from Oracle, which provides details about the product's lifecycle so the user can take appropriate actions.

### Simplified Provisioning Process

After discovering and identifying systems and their components, Oracle Enterprise Manager Ops Center can automatically filter through the operating system images and available firmware and present only those that are appropriate to the target system. Oracle Enterprise Manager Ops Center automatically creates and maintains the underlining technologies used during OS and firmware provisioning so administrators can focus on more important tasks. By having Oracle Enterprise Manager Ops Center deploy only the relevant firmware and operating system across larger and diverse asset groups, guesswork is eliminated. Oracle Enterprise Manager Ops Center also bridges the gap between the embedded Oracle VM Server for SPARC and its controlling domain by automatically verifying the appropriate firmware is installed. This process simplifies virtual machine creation later on and is completely transparent to the user.

### Reduced Administrative Costs

By automating most of the deployment process for physical and virtual systems, user involvement is minimized. Oracle Enterprise Manager Ops Center offers a facility to create and store the approved operating system or firmware profile required by business services. Jumpstart, JET modules, kickstart, and yast customizations can be stored under named profiles to be deployed more easily later by personnel less familiar with the underlying technology. This allows the business to more efficiently leverage IT skill sets across functional units.

### Rapid Deployment

Oracle Enterprise Manager Ops Center deploys a complete stack on a bare-metal server to make the system production ready within a short time. It can create a snapshot of a system catalog and restore an operating system to a previous state. It can compare inventories of multiple systems and make target systems match source inventories. This process can be applied to a single system, multiple systems, or multiple datacenters. Since Oracle Enterprise Manager Ops Center has out of the box automation and in-depth knowledge of Oracle systems, operational staff can spend more time focusing on driving greater business value.

### Fault and Event Management

Hardware status and operating system performance is tracked to check the overall health of the system. When a predefined threshold or hardware condition is reached, a notification is sent to the user interface and an e-mail is auto generated.

### Comprehensive Reports

Oracle Enterprise Manager Ops Center's rich UI and functionality presents information based on user definitions. Monitored information can be presented at a per-system/per-virtual resource level or can be aggregated across a group of servers or virtual pools. Historical information for parameters such as CPU, memory, network I/O, and WATT consumption can be monitored and stored for future reference. Moreover, any and all gathered data can be exported for further analysis or to create custom reports.

### Automated Patching Using a Unique Knowledgebase

Oracle's Knowledge Services is a hosted metadata knowledgebase of Oracle Solaris, Oracle Unbreakable Linux, Red Hat, and SuSE operating systems. This knowledgebase is a very powerful capability unique to Oracle. It is served down to customers through a web service or in a disconnected mode. Leveraging the knowledgebase metadata improves patch accuracy and reduces downtime. It maintains advanced patch, rpm, and package dependency information that has been discovered through unique methods exclusively owned by Oracle. Oracle Enterprise Manager Ops Center uses this knowledgebase to download only the required patches the first time (not all new patches)—saving

both network bandwidth and compute resources. It applies those patches and performs appropriate actions (single/multiuser mode, reboot option) as required (See Figure 2).

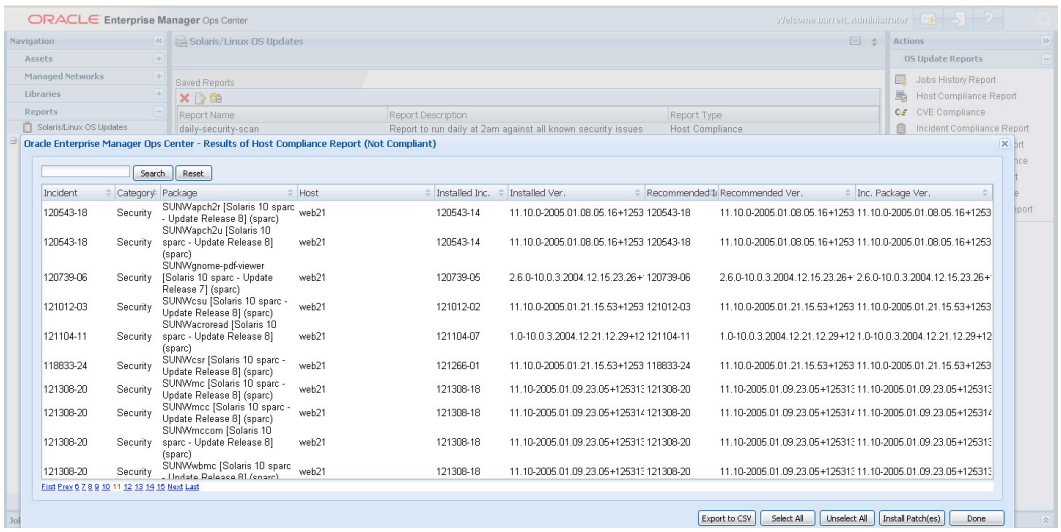


**Figure 2. Comprehensive reports cover patch requirements and gaps with other systems, enabling system updates and compliance.**

### Reduce Downtime

Oracle Enterprise Manager Ops Center helps system administrators meet their maintenance windows in three ways. Leveraging its unique knowledgebase, the product first examines the installed software to see if any broken dependencies exist. Next it searches against vendor bugs, Common Vulnerability databases, or customer profiles to discover if updates are needed. With every action, it automatically takes snapshots of the inventory on the box in case rollbacks or time comparisons are needed. It automatically resolves required patch trees and groups them correctly during patch installation. Lastly, it will cache patch payloads on the agents and simulate installation to insure the operating system commands and directories are healthy enough to install additional software. Now the platform can be reliably patched with a higher level of confidence that nothing will go wrong. Oracle Enterprise Manager Ops Center will also automatically discover Oracle Solaris Live Upgrade alternate boot environments and display them for selection during patching allowing for zero downtime patching.

### Compliance

Multiple compliance reports are possible with Oracle Enterprise Manager Ops Center.

- Compare all the servers against a business project's requirements.
- Compare against an older vendor provided baseline or always test against the latest information from the vendor.
- Test against a government approved common vulnerability database.
- Continuously schedule reports to help discover the server sprawl across the datacenter.

- Compare servers to one another or compare previous snapshots of the same server.
- Report who installed or uninstalled what, when, and where via Oracle Enterprise Manager Ops Center.

### Manages Oracle Virtualization Technologies

Oracle Enterprise Manager Ops Center manages the lifecycle of Oracle Solaris Containers and Oracle VM Server for SPARC. Their resources are monitored continuously to provide up-to-date information on usage. Based on the dynamic needs of the applications, new Oracle Solaris Containers and Oracle VM Server for SPARC virtual guests can be created, deleted, cloned, or reconfigured.

### Centralizes Management of Resources

As the central management console for all relevant infrastructures, Oracle Enterprise Manager Ops Center tracks hardware, virtualization components, and operating systems. It provides the appropriate components to keep physical and virtualization assets up to date. Oracle Enterprise Manager Ops Center ensures that the system using Oracle VM Server for SPARC has the appropriate firmware.

### Lifecycle Management—Simple Deployment and Maintenance

With Oracle Enterprise Manager Ops Center, you can install and manage all relevant components in a virtualized stack.

- Asset discovery. Oracle Enterprise Manager Ops Center can discover all assets, such as hardware, firmware, virtual systems, and operating systems. You can view them in a usable format.
- Provisioning. With Oracle Enterprise Manager Ops Center, you can provision firmware and operating systems on bare-metal and virtual systems.
- Patching. With its unique knowledgebase, Oracle Enterprise Manager Ops Center keeps all components (physical as well as virtual) in the stack up to date.
- Monitoring. Oracle Enterprise Manager Ops Center monitors physical and virtual systems to provide end-to-end monitoring of the complete stack. It monitors individual as well as aggregate resources to get a complete view of the system.

### Eco-friendly

Oracle Enterprise Manager Ops Center monitors the power of all servers and aggregates consumption patterns. Based on the outcome, administrators can balance the resources by shutting servers down, migrating workloads, or leveraging power capping capabilities in the servers.

### Reduces Resource Management Complexity

By providing access to all assets, such as hardware, virtual systems, and operating systems (through one interface), Oracle Enterprise Manager Ops Center makes managing these resources simple. Assets can

be logically grouped, automated through smart groups, tagged for custom grouping, and filtered with multiple options.

### Investment Protection

Oracle Solaris 8 and Oracle Solaris 9 servers can easily migrate to Oracle Solaris 10. This allows Oracle Solaris 8 and Oracle Solaris 9 implementations to leverage the latest capabilities in Oracle Solaris 10 and newer servers. Virtual resource management is a seamless extension of physical system management. System management software users can easily adjust to virtual resource management through Oracle Enterprise Manager Ops Center's rich user interface (See Figure 3).

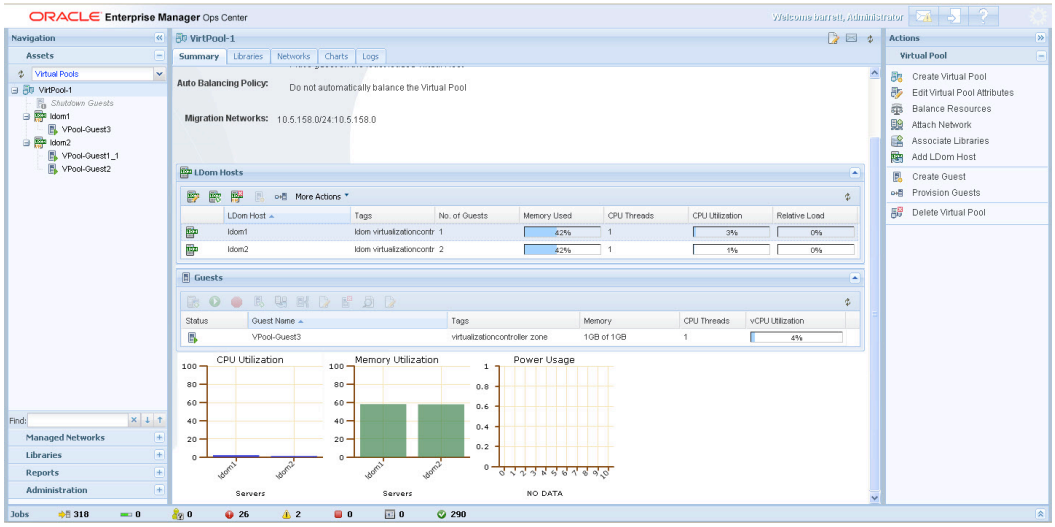


**Figure 3/ Oracle Enterprise Manager Ops Center delivers comprehensive lifecycle management of physical and virtual resources in an easy-to-use interface.**

### Scale with Resources

Virtualization makes it easy to deploy resources on demand. A proliferation of resources requires a robust management platform that can not only manage the compute elements in one location, but can also scale with different geographic locations. With its three-tier architecture, Oracle Enterprise Manager Ops Center scales its performance and usability from a single datacenter to distributed datacenters. With proxies deployed closer to the managed systems, performance more-easily meets required service levels. Oracle Enterprise Manager Ops Center uses the latest Web technologies and fits into existing datacenter designs reducing the need for configuration changes.

### Efficient Resource Use

Oracle Enterprise Manager Ops Center pools virtual resources to cater to similar applications. Appropriate policies can be applied to virtual resources that are generated on demand and placed in pools of physical machines that meet the application requirements. Once in the pool, resource policies continue to watch over the workload and auto balance the environment automatically by migrating Oracle VM Server for SPARC guests within the pool to assure the most optimal use of physical resources.

#### Availability

Virtual resources are offered depending on the application needs. Virtualization mobility features such as cold migration for Oracle Solaris Containers and warm migration for Oracle VM Server for SPARC can be used to move virtual resources to other systems to improve utilization with minimal downtime.

#### Security

Oracle Solaris Container and Oracle VM Server for SPARC technologies ensure isolation between different containers and domains. In addition, role-based mechanisms are implemented to allow multiple users to manage systems with different access controls. Oracle Enterprise Manager Ops Center can patch physical as well as virtual systems that have security vulnerabilities.

### Oracle Solaris Management Console

The Oracle Solaris Management Console makes it easy for IT operators to configure and manage systems running Oracle Solaris. Based on Java technology, the Oracle Solaris Management Console is a container for GUI-based management tools that are stored in collections called toolboxes. The console includes a default toolbox with tools to manage users and projects, and with jobs for managing file system mounts, disks, and serial ports. The console helps IT operators to easily view activity on all managed servers and to modify applications and services running on them. The Oracle Solaris Management Console also delivers powerful capabilities to make the process of adding users, hosts, or applications as simple as pointing and clicking from any client on the network.

The Oracle Solaris Management Console allows users and IT operators to register other Oracle Solaris Management Console servers and applications on the network. It dynamically configures a hierarchical or flat tree view of registered hosts and services when the console is accessed, making it easier to manage each server.

The Oracle Solaris Management Console helps improve productivity for IT departments, IT systems, and network administrators by providing several important capabilities:

- **Support for all experience levels.** Inexperienced IT operators can complete tasks by using the GUI, which includes dialog boxes, wizards, and context help. Experienced IT operators find that the console provides a convenient, secure alternative to using vi, the UNIX Visual Editor.
- **Administration tools.** These tools can be integrated and run from one location.

- **Centralized management.** All servers on a network can be managed with centralized management.
- **Single login.** Use single login to access applications launched by the Oracle Solaris Management Console.
- **Instant access.** Gain access instantly to existing administration tools.
- **Secure communication.** Ensure secure communications with support for HTTPS and SSL.

The Oracle Solaris Management Console also includes a set of graphically based wizards to streamline and simplify these complex or frequently performed administration tasks:

- Monitor and manage system information such as date, time, and time zone.
- Monitor and manage system processes.
- Monitor system performance.
- Manage users, rights, roles, groups, and mailing lists.
- Create and manage projects.
- Manage patches.
- Create and manage scheduled cron jobs.
- Mount and share file systems.
- Create and manage disk partitions, volumes, hot spare pools, state database replicas, and disk sets.
- Set up terminals and modems.
- Create and monitor computer and network information.
- Shut down and boot the system.

## Configuration and Service Tracker

The **Configuration and Service Tracker** (CST) software application is a Web-enabled application designed to help achieve higher levels of system availability and manageability. It uses a lightweight agent, running on each monitored system, to send updates whenever a system configuration changes. With CST, IT operators are able to monitor hardware changes to a system or a domain, including dynamic reconfiguration actions and service repairs, and to track configurations down to the field replaceable unit level.

To help IT operators better manage their systems, CST probes the monitored systems and collects data in the following categories:

- System information, including system data such as system model, host ID, host name, Internet Protocol address, OS installed, number of CPUs, total memory, and number of disks
- Installed hardware, including system boards, controllers, and devices (excluding external network and communication devices)

- Cluster configuration
- Software packages, including Oracle and third-party packages (those that are installed using the pkgadd facility)
- Software patches
- Pertinent information about Oracle Solaris
- Volume manager configuration

By continuously gathering this data, CST provides a means to track and manage systems for compliance with corporate configuration standards including OS upgrades and patch migration across the install base.

## Sun Net Connect Services

**Sun Net Connect** is a free, Web-based, self-monitoring, configuration/patch collection, and reporting service for SPARC systems that helps drive system uptime and reduce system management costs. Sun Net Connect offers self-monitoring of the CPU, memory, swap disk, and root disk, as well as complete system-level configuration and patch reporting. All alerts and reports are viewed through the Sun Net Connect Web portal. The following services are available:

- **Performance self-monitoring.** This service allows the IT operator to actively manage system resources through a Web portal system dashboard.
- **Hardware failure self-monitoring.** This service provides hardware failure alarms to the Web portal and notification through pager, e-mail, or both.
- **Event alarming and notification.** This service provides system performance alarms to the Web portal and notification through pager, e-mail, or both.
- **Web view reporting.**
  - **Trend reporting.** This service displays system performance trends over time.
  - **Configuration and patch reporting.** This service displays system configuration and patch information.
  - **Availability reporting.** This service provides a high system availability percentage to assist in determining system management and maintenance effectiveness.

### Sun Update Connection

This Web application is hosted at Oracle and allows IT operators to manage updates remotely on one or more Oracle Solaris systems. Sun Update Connection services allow IT operators to remotely monitor and manage all update activities for each of their registered systems. These services are available through a Web application that runs at Oracle. This tool can be used to create jobs to run on systems as they check in to the service. A job either installs an update or uninstalls an update. The Web application can also be used to view the update status of systems and jobs.

### Update Manager

The Update Manager GUI and the smpatch CLI allows IT operators to manage updates (patches) locally on Oracle Solaris systems. Sun Update Connection, System Edition has the same functionality as the Sun Patch Manager tools, with the addition of some new features and enhancements.

Update Manager is one part of Sun Update Connection, System Edition 1.0, which allows IT operators to locally manage updates on systems. Update Manager offers a GUI for updating systems with updates. It can be used to analyze a system, apply selected updates, remove updates, and configure the update management environment. Systems that are managed with Sun Update Connection services can still be managed locally by using Update Manager.

Update Manager incorporates PatchPro (automated patch management technology) functionality. PatchPro performs update analyses on systems, and then downloads and applies the resulting updates. PatchPro uses signed updates, which improves the security of Oracle Solaris updates by ensuring that they have not been modified. Update Manager's update management processes include

- Analyzing the system to obtain a list of appropriate updates
- Downloading the appropriate updates to the system
- Applying the appropriate updates to the system
- Configuring the update management environment on the system
- Tuning the update management environment on the system
- Removing updates from the system
- Using Sun Update Connection services to remotely manage the system

## Oracle Solaris: Support for Efficient Management

With millions of registered licenses worldwide and more than 600 innovative features, such as DTrace, predictive self healing, Containers, and Oracle Solaris ZFS, the open source Oracle Solaris 10 is the most advanced OS available. These technologies supply the foundation for building, deploying, and managing efficient, secure, and reliable enterprise-class service-oriented architectures for today's demanding business processes.

In addition to providing a world-class platform for deploying applications, Oracle Solaris includes sophisticated services to improve manageability and resource use. Many of the most difficult problems occur as a result of changes to the server OS configuration. Oracle Solaris offers unique tools, such as Oracle Solaris JumpStart, Oracle Solaris Live Upgrade, and the flash archive feature, to help automate tasks, thus improving consistency. Oracle Solaris also features ZFS to effectively manage file systems.

### Oracle Solaris JumpStart

Oracle Solaris JumpStart automates the process of installing or upgrading multiple systems based on custom profiles and optional preinstallation and postinstallation scripts. A **profile** is a text file that defines how to install Oracle Solaris on a system. A profile and a rules file must be created for each

group of systems to be installed. A **rules file** is a text file that contains a rule for each group of systems or single systems. Each rule distinguishes a group of systems, based on one or more system attributes. Each rule also links each group to a profile.

When installing, Oracle Solaris JumpStart searches for the first rule with defined system attributes that match the system on which Oracle Solaris JumpStart is attempting to install Oracle Solaris. If a match occurs, Oracle Solaris JumpStart uses the profile that is specified in the rule to install Oracle Solaris.

### Oracle Solaris Live Upgrade

Oracle Solaris Live Upgrade provides a method of upgrading a system while the system continues to operate. While the current boot environment is running, the IT operator can duplicate the boot environment, and then upgrade or patch the duplicate. Alternatively, rather than upgrading, a flash archive can be installed on a boot environment. The original system configuration remains fully functional and unaffected by the upgrade or installation of an archive. With Oracle Solaris Live Upgrade, an upgrade or new patches can be tested without affecting the current OS. When ready, the IT operator can activate the new boot environment by rebooting the system. If a failure occurs, the system can be quickly reverted to the original boot environment with a simple reboot. Thus, the normal downtime of the test and evaluation process can be eliminated. This process is illustrated in Figure 4.

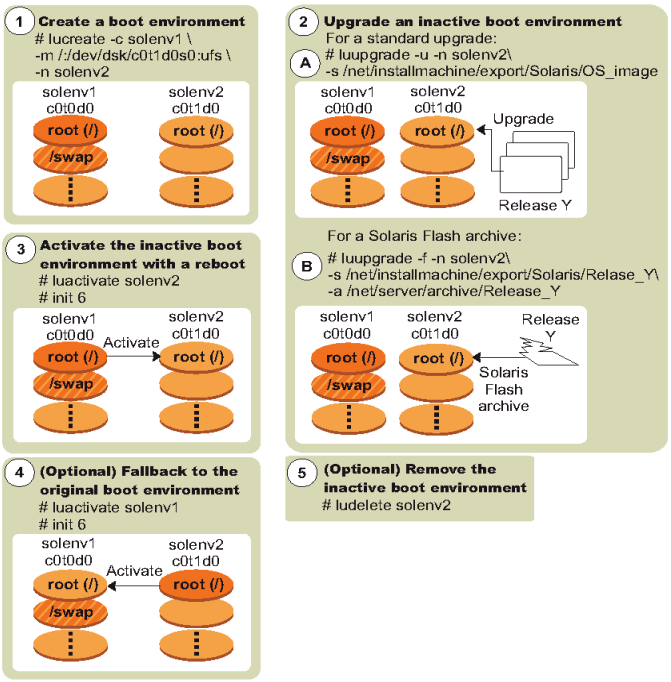


**Figure 4. Oracle Solaris Live Upgrade process.**

With Oracle Solaris Live Upgrade, a boot environment can be duplicated without affecting the currently running system. IT operators can then

- Upgrade a system
- Change the current boot environment's disk configuration to different file system types, sizes, and layouts on the new boot environment
- Maintain numerous boot environments with different images; for example, one boot environment can contain current patches, and another boot environment can contain an update release

### Flash Archive

The flash archive feature in Oracle Solaris allows IT operators to use a single reference installation of Oracle Solaris on a system. This reference installation is called a *flash archive*. The flash archive can then be installed on multiple systems, which are called *clone systems*. The installation is an initial installation that overwrites all files on the clone system. Flash archives save installation time by installing all Oracle Solaris packages at one time. Other programs install each individual Oracle Solaris package and update the package map for each package.

A flash installation can update a clone system with minor changes. If a clone system needs to be updated, a differential archive can be created that contains only the differences between two images: the original master image and an updated master image. When a clone system is updated with a differential archive, only the files that are specified in the differential archive are changed. The installation is restricted to clone systems that contain software that is consistent with the original master image. The custom Oracle Solaris JumpStart installation method can be used to install a differential archive on a clone system—or Oracle Solaris Live Upgrade can be used to install a differential archive on a duplicate boot environment.

### Sun Patch Manager

Sun Patch Manager is the standard tool for applying and managing patches on Oracle Solaris systems. It offers many useful features, including the following:

- **PatchPro analysis engine.** This tool incorporates PatchPro functionality to automate the patch management process. This process includes performing patch analysis on systems, then downloading and applying the resulting patches.
- **Local-mode CLI.** This CLI allows IT operators to use smpatch to apply patches while the system is in single-user mode.
- **Patch list operations.** This feature allows IT operators to generate, save, edit, order, and resolve patch lists. These lists can be used to perform patch operations, such as downloading and applying patches.

Sun Update Connection has the same functionality as the Sun Patch Manager tools, with the addition of some new features and enhancements.

### Service Management Facility

Included in Oracle Solaris 10, the service management facility provides an infrastructure that augments the traditional UNIX startup scripts and configuration files that IT operators must modify to start system and application services. The fundamental unit of administration in the service management facility framework is the service instance. Each service management facility service has the potential to have multiple versions of it configured. Multiple instances of the same version can run on a single Oracle Solaris system. An **instance** is a specific configuration of a service. For example, a Web server is a service, and a specific Web server daemon that is configured to listen on port 80 is an instance. The service management facility simplifies the management of system and application services by delivering new and improved ways to control services. It provides the ability to define a service that is dependent on other services so that it does not run unless the other services it requires are running.

The service management facility provides the following functions:

- Automatically restarts failed services in dependency order, whether they fail as the result of IT operator error, software bug, or an uncorrectable hardware error
- Makes it easy to backup, restore, and undo changes to services by taking automatic snapshots of service configurations
- Makes it easy to debug and ask questions about services by providing an explanation of why a service is not running
- Allows for services to be enabled and disabled
- Enhances the ability of IT operators to securely delegate tasks to nonroot users, including the ability to modify properties and enable, disable, or restart services on the system
- Boots faster on large systems by starting services in parallel according to the dependencies of the services (the opposite process occurs during shutdown)

### Oracle Solaris ZFS

Oracle Solaris ZFS offers a dramatic advance in data management with an innovative approach to data integrity, tremendous performance improvements, and an integration of file system and volume management capabilities. The centerpiece of this new architecture is the concept of the *virtual storage pool*, which decouples the file system from physical storage in the same way that virtual memory abstracts the address space from physical memory—enabling more-efficient use of the storage devices (see Figure 5).

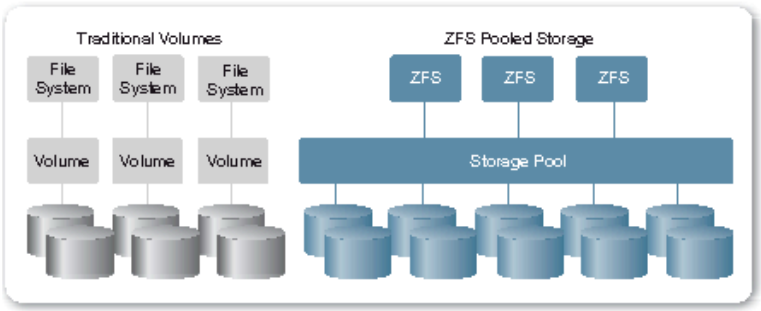

**Figure 5. The Oracle Solaris ZFS architecture offers the virtual storage pool, which decouples the file system from physical storage.**

In the ZFS, space is shared dynamically between multiple file systems from a single storage pool and is parceled out from the pool as file systems request it. As a result, physical storage can be added to or removed from storage pools dynamically, without interrupting services. This provides new levels of flexibility, availability, and performance. In terms of scalability, the ZFS is a 128-bit file system, so its theoretical limits are virtually unlimited—2128 bytes of storage and 264 for everything else, such as the number of file systems, snapshots, directory entries, or devices.

Other benefits and features of the ZFS include the following:

- **Maintains data integrity.** The ZFS combines proven and cutting-edge technologies like copy-on-write and end-to-end 64-bit checksum, providing extreme reliability to ensure that the data on the disk is self-consistent at all times. In addition, the ZFS hot spares feature allows users to identify disks that could be used to replace a failed or faulty device in one or more storage pools.
- **Improves performance.** The ZFS optimizes and simplifies the code paths from the application to the hardware, producing sustained throughput at near-platter speeds. New block allocation algorithms accelerate write operations and consolidate what would traditionally be many small random writes into a single more-efficient sequential operation. Additionally, the ZFS implements intelligent prefetch, performing read ahead (in either direction) for sequential data streaming and adapting its read behavior on the fly for more-complex access patterns. The ZFS also eliminates bottlenecks and increases the speed of both reads and writes by striping data across all the available storage devices—balancing I/O and maximizing throughput.
- **Reduces costs.** Unlike traditional file systems that require a separate volume manager, the ZFS architecture integrates volume management functions.
- **Simplifies management.** The ZFS integrates devices, storage, and file system structures into a single structure, simplifying file system management and providing a reliable and flexible solution that can help reduce cost, complexity, and risk.
- **Provides low-overhead RAID.** The ZFS provides software RAID through RAID-Z. **RAID-Z** is a virtual device that stores data and parity on multiple disks, similar to RAID-5. RAID-Z uses variable-width RAID stripes so that all writes are full-stripe writes. Replicated RAID-Z configuration can have either single parity or double parity, which means that one or two device failures can be sustained, respectively, without any data loss.

Oracle Solaris 10 integrates a fault manager diagnostic engine for the ZFS that is capable of diagnosing and reporting pool failures and device failures. Checksum, I/O, device, and pool errors associated with pool or device failures are also reported to help quickly identify and resolve failures.

## Simple Network Management Protocol Service

An SNMP agent can be configured and enabled on the service processor. The service processor SNMP agent monitors the state of the system hardware and domains and exports the following information to an SNMP manager:

- System information such as chassis ID, platform type, total number of CPUs, and total memory
- Configuration of the hardware
- Dynamic reconfiguration information, including which units are assigned to which domains
- Domain status
- Power status
- Environmental status

The service processor SNMP agent can supply system information and fault event information using public MIBs. SNMP managers, such as third-party manager applications, use any service processor network interface with the SNMP agent port to communicate with the agent. The SNMP agent supports concurrent access from multiple users through SNMP managers.

SNMP can be configured using SNMPv1, SNMPv2, or SNMPv3. XSCF supports two MIBs for SNMP:

- **XSCF extension MIB.** This is used to get information on the status and configuration of the platform. If there is a fault, it sends a trap with the basic fault information.
- **Fault management MIB.** This is used only when there is a fault. It sends the fault trap, but it includes all of the same detailed information as the fused multiply-add MIB in an Oracle Solaris domain. The information has the information needed by the service technician when placing a service call. It is also useful if the domain crashed due to a part failure.

There are two methods of FMA reporting on XSCF: through SNMP and through the DSCP to the affected domain. To have XSCF report all platform faults through SNMP using fused multiply-add descriptors, enable SNMP on XSCF.

## Managing Resources and Domains

Delivering computing services over the internet brings with it an emphasis on finding ways to provide needed levels of service such as throughput, availability, and response time—with an economical option. Many consumers and internal customers now demand guarantees in the form of **service-level agreements (SLAs)**: contracts defining the services to be provided, along with metrics for determining if they have been adequately delivered.

One way to meet the terms of an SLA is to overprovision resources. Although this approach provides predictable service levels, it is expensive and wasteful. Another approach is to provide each customer or application with a dedicated system. This might prevent other applications or users from impacting performance, but it creates a proliferation of systems that require more administration, and it can be costly, inflexible, and inefficient to manage and maintain.

A better tactic for efficient resource management is to consolidate tasks and applications onto larger, more-powerful servers with tools to help manage service levels and resource use. Oracle offers powerful features that help control resource use on the Sun SPARC Enterprise servers by partitioning

the system's resources into isolated domains and providing a process to dynamically reconfigure those resources to meet changing demands.

## Dynamic Domains

Two key technologies for enabling consolidation are Dynamic Domains and Dynamic Reconfiguration (DR), which is the enabling technology behind Dynamic Domains. With DR, Sun SPARC Enterprise servers help keep applications up and running, while more-cost-effectively managing IT resources.

A **domain** is an independent system resource that runs its own copy of Oracle Solaris. Domains divide a system's total resources into separate units that are not affected by each other's operations. Domains can be used for different types of processing; for example, one domain can be used to test new applications, whereas another domain can be used for production purposes.

Each domain uses a separate boot disk with its own instance of Oracle Solaris, and each uses I/O interfaces to network and disk resources. CPU/memory boards and I/O boards can be added separately and removed from running domains by using DR, provided they are not also assigned to other domains. Domains run applications in strict isolation from applications running in other domains. Security between domains is maintained through role-based access control, assigning unique privileges per domain and restricting the platform and root administrators from domain control and data access.

Domains, in conjunction with modular systems like the Sun SPARC Enterprise servers, can help decrease costs and reduce overhead when employed to consolidate applications or when supporting multiple components—such as the Web, application, and database components of an application service—of a service on a single platform (see Figure 6). Each domain is completely protected from hardware or software faults originating in other domains.

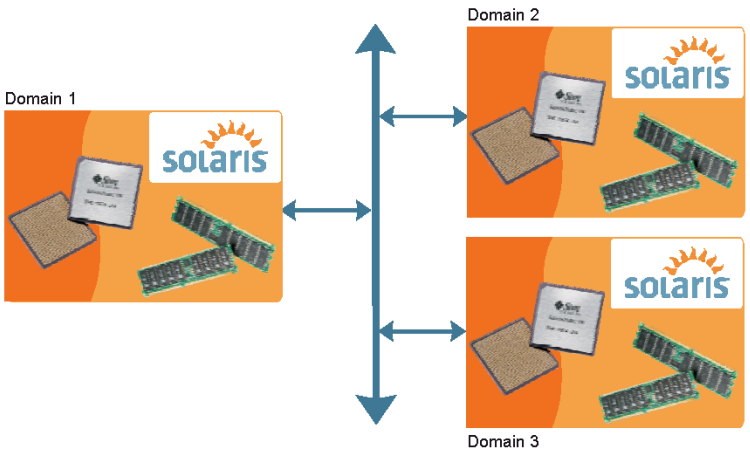


**Figure 6. Oracle Solaris enables the consolidation of applications into domains.**

### How Domains Work

A physical system board (PSB) consists of up to four CPUs, 32 memory modules, and I/O. The Sun SPARC Enterprise servers have a unique partitioning feature that can divide one PSB into one logical board or into four logical boards. The number of PSBs in the Sun SPARC Enterprise servers varies from 1 to 16, depending on the model. In the high-end systems, one PSB consists of four CPUs, 32 dual inline memory modules (DIMMs), and I/O. The I/O varies with the server and can include PCIe slots, PCI-X slots, and built-in I/O.

In the midrange systems, the CPU module consists of two CPU chips. The Sun SPARC Enterprise M4000 and Sun SPARC Enterprise M5000 servers support up to two and four CPU modules, respectively. Each memory board in the systems contains eight DIMMs, with the Sun SPARC Enterprise M4000 server supporting up to four memory boards and the Sun SPARC Enterprise M5000 server supporting up to eight memory boards. Each I/O unit contains five PCI slots. The Sun SPARC Enterprise M4000 server supports one I/O unit and the Sun SPARC Enterprise M5000 server supports up to two I/O units.

To use a PSB in the system, the hardware resources on the board must be logically divided and reconfigured as eXtended System Boards (XSBs), which support two types: Uni-XSB and Quad-XSB. These XSBs can be combined freely to create domains.

A **Uni-XSB** is a PSB that is not logically divided and is configured into one XSB. It contains all of the resources on the board—including four CPUs, 32 DIMMs, and I/O—and is suitable for domains requiring a large quantity of resources. A Uni-XSB provides physical domaining, where the domain boundaries are at the board level. The Uni-XSB provides the best fault isolation because it can only be assigned to one domain. Therefore, if the board fails, it only affects one domain. A Uni-XSB for midrange systems is illustrated in Figure 7.

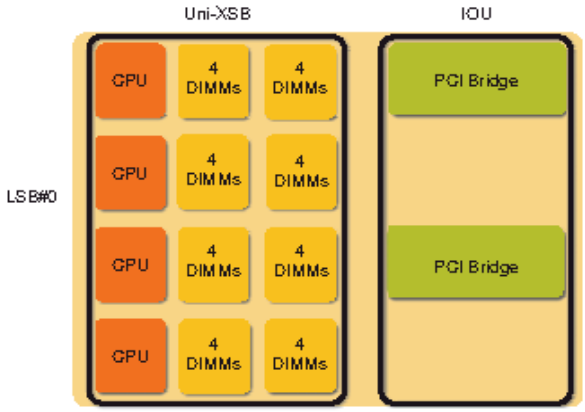


**Figure 7. A Uni-XSB on a midrange system.**

A Uni-XSB for high-end systems is illustrated in Figure 8. Uni-XSBs can also be configured for memory mirror mode. In this mode, the PSB has two memory units—one mirroring the other. Saving the same data in a separate memory unit improves data security.

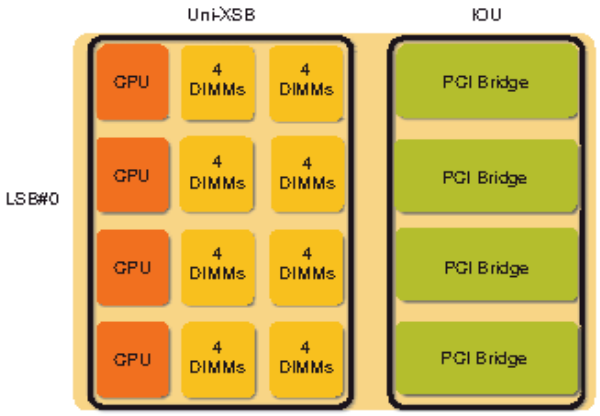


**Figure 8. A Uni-XSB on a high-end system.**

A **Quad-XSB** is a PSB that is logically divided and configured into four XSBs. Each of the four XSBs contains one-quarter of the total board resources—for example, on high-end systems, each XSB in a Quad-XSB contains one CPU, eight DIMMs, and two PCIe cards (one-quarter of the I/O unit), as illustrated in Figure 9.

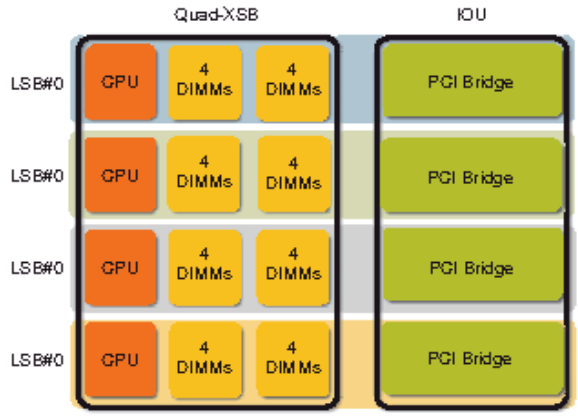


**Figure 9. A Quad-XSB on a high-end system.**

On midrange servers, shown in Figure 10, only two XSBs have I/O. Quad-XSBs enable subboard domain granularity and, therefore, better resource use. However, if a board fails, so do the domains to which it is assigned. Memory mirror mode can be enabled for Quad-XSBs in the midrange systems only.

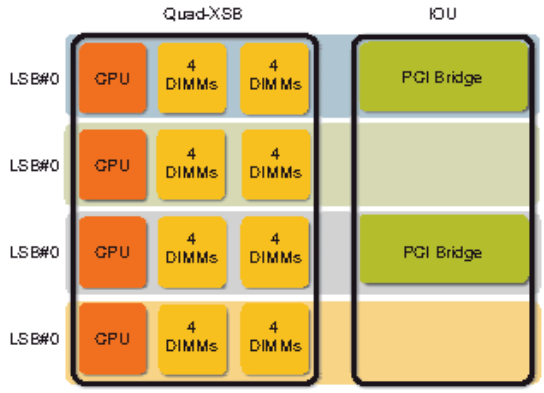


**Figure 10. A Quad-XSB on a midrange system.**

A domain consists of one or more XSBs. Each domain runs its own copy of Oracle Solaris and must have a minimum of one CPU, eight DIMMs, and I/O. The number of domains allowed depends on the server model. The default is 1 domain and the maximum number of domains is 24. The maximum number of XSBs in a domain is 16. Domains can be set up to include both Uni-XSBs and Quad-XSBs.

A domain component list identifies the potential resources for a domain. A single XSB can potentially be used by multiple domains. However, a single XSB can be assigned to only one specific domain. The domain configuration software maps each XSB number to a logical system board (LSB) number.

Figure 11 illustrates a configuration of four domains using one Uni-XSB and two Quad-XSBs. Note that each domain requires I/O, but not every LSB requires it.

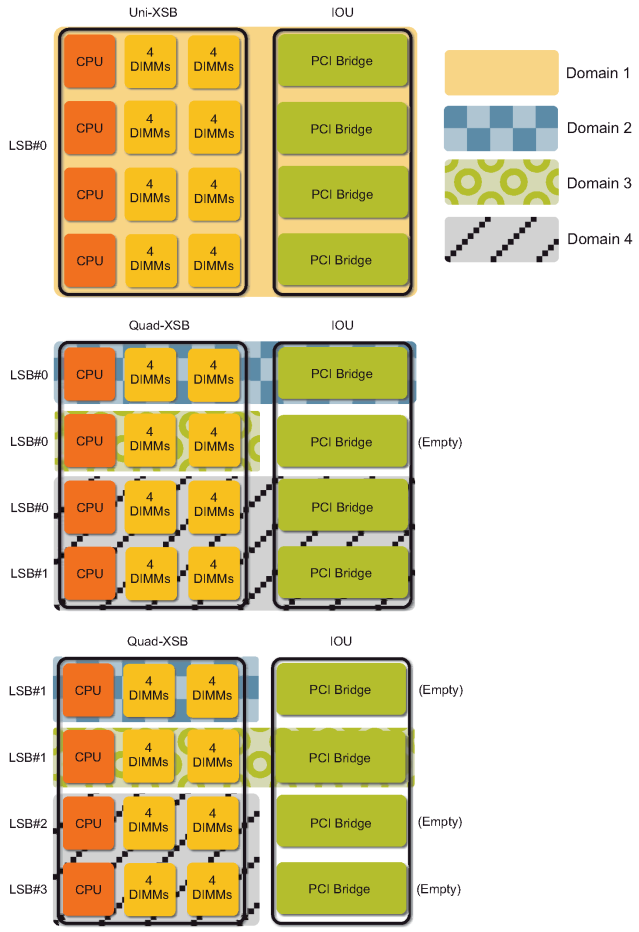


**Figure 11. Domain configuration example using a Uni-XSB and Quad-XSBs.**

Oracle Solaris is installed on a per-domain basis. The OS image is installed on internal disks. On midrange systems, the disks are available only for the first (top) I/O device and the third (third from the top) I/O device. The second and fourth I/O devices do not have the capability to support internal hard disks.

### Fault Isolation and Error Management

Domains are protected against software or hardware failures in other domains. Failures in hardware shared between domains cause failures only in the domains that share the hardware. When a domain encounters a fatal error, a domain stop operation occurs that cleanly and quickly shuts down only the domain with the error. The domain stop operates by shutting down the paths in and out of the system address controller and the system data interface ASICs. The shutdown is intended to prevent further corruption of data, and to facilitate debugging by not allowing the failure to be masked by continued operation.

When certain hardware errors occur in a Sun SPARC Enterprise server system, the system controller performs specific diagnosis and domain recovery steps. The following automatic diagnosis engines identify and diagnose hardware errors that affect the availability of the system and its domains:

- **XSCF diagnosis engine.** This engine diagnoses hardware errors associated with domain stops.
- **Oracle Solaris diagnosis engine.** This engine identifies nonfatal domain hardware errors and reports them to the system controller.
- **Power-on self-test diagnosis engine.** This engine identifies any hardware test failures that occur when the power-on self-test (POST) is run.

In most situations, hardware failures that cause a domain crash are detected and eliminated from the domain configuration either by the POST or by the OpenBoot PROM during the subsequent automatic recovery boot of the domain. However, there can be situations where failures are intermittent or the boot-time tests are inadequate to detect failures that cause repeated domain failures and reboots. In those situations, XSCF uses configurations or configuration policies supplied by the domain administrator to eliminate hardware from the domain configuration in an attempt to get a stable domain environment running.

### Using Dynamic Domains

Domains have been used with great success in the mainframe world for years. With the Sun SPARC Enterprise servers, even-greater numbers of applications can benefit from the advantages of domains. Key advantages are listed here:

- **Consolidation.** One Sun SPARC Enterprise server can replace multiple smaller servers. Consolidated servers are easier to administer, more robust, and offer the flexibility to shift resources freely from one application to another. Increased flexibility is especially important as applications grow or when demand reaches peak levels, requiring additional resources to be rapidly deployed.
- **Development, production, and test environments.** In production environments, many sites require separate development and test systems. Isolating these systems with domains helps to enable development work to continue without impacting production runs. With the Sun SPARC Enterprise servers, development and test functions can safely coexist on the same platform.
- **Software migration.** Dynamic Domains can be used to help migrate systems or application software and their users to updated versions. New, or perhaps more-experienced, users can employ the latest versions in one isolated domain, while others waiting to be trained can continue to use older versions in another domain. This approach applies equally well to Oracle Solaris, database applications, new administrative environments, and other applications.
- **Special I/O or network functions.** A system domain can be established to deal with specific I/O devices or functions isolated within its domain, or it can be further managed using Oracle Solaris Resource Manager within the domain. For example, a high-end tape device can be attached to a dedicated system domain, which can be added to other system domains when there is a need to make use of the device.

- **Departmental systems.** Multiple projects or departments can share a single Sun SPARC Enterprise server, increasing economies of scale and easing cost justification and accounting requirements.
- **Configuring for special resource requirements or limitations.** Projects that have resource requirements that might starve other applications can be isolated to their own system domain. For applications that lack scalability, multiple instances of the application can be run in separate system domains, or in containers within one domain.
- **Hardware repairs and rolling upgrades.** Because each domain runs its own instance of Oracle Solaris and has its own peripherals and network connections, domains can be reconfigured without interrupting the operation of other domains. Domains can be used to remove and reinstall boards for repair or upgrade, to test new applications, and to perform OS updates.

## Dynamic Reconfiguration

DR and automated DR (ADR) allow resources to be dynamically reallocated, or balanced, between domains by enabling a physical or logical restructuring of the hardware components of the Sun SPARC Enterprise servers while the system is running and the applications remain available. This high degree of resource flexibility allows the domain or platform administrator to reconfigure the system easily in order to provision the resources to meet changing workload demands. Domain configurations can be optimized for workloads that are either compute intensive, I/O intensive, or both. DR can also be used to remove and replace failed or upgraded hardware components while the system is online.

DR functions of the Sun SPARC Enterprise servers are performed on XSB units. DR operations are performed and managed through XSCF. The XSCF security management restricts DR operations to administrators who have proper access privileges. Three types of system board components can be added or deleted by DR: CPU, memory, and I/O devices.

DR allows the domain or platform administrator to

- Display the DCL and domain status
- Display the status of system or I/O boards and components to help prepare for DR operations
- Test live boards
- Register system or I/O boards to the DCLs of domains
- Delete (electrically isolate) system or I/O boards from a domain in preparation for moving to another domain or for removal from the system while the domain remains running
- Add system or I/O boards to a domain to add resources or to replace a removed board while the domain remains running
- Configure or deconfigure CPU or memory modules on system boards to control power and capacity of a domain or to isolate faulty components
- Enable or disable PCI cards or related components and slots
- Reserve a system board to a domain

For example, the IT operator can use DR to delete a faulty system board, and then use the system's hot-plug feature to physically remove it. After plugging in the repaired board or a replacement, DR can be used to add the board into the domain. System or I/O boards can also be associated with multiple domains for load balancing or to provide extra capabilities for specific tasks. However, resources can only be assigned to one domain at a time.

In addition, with DR and multipathing solutions such as IP network multipathing (see "IP Network Multipathing"), Oracle Solaris Fibre Channel, and storage multipathing (see "Solaris Fibre Channel and Storage Multipathing"), IT operators can manage networks or storage subsystems that have redundant access paths with automatic failover, load balancing, and DR capabilities.

### Basic Dynamic Reconfiguration Functions

All system boards that are targets of DR operations must be registered in the target domain's DCL through XSCF. The basic functions of DR are as follows:

- **Add.** DR can be used to add a system board to a domain without stopping Oracle Solaris, provided the board is installed in the system and not assigned to another domain. A system board is added in three stages: assign, connect, and configure. In the add operation, the selected system board is assigned to the target domain so that it is connected to the domain. Then, the system board is configured to the Oracle Solaris instance of the domain. At this point, the system board is added to the domain, and its CPU and memory resources can be used by that domain.
- **Delete.** DR can be used to delete a system board from a domain without stopping the instance of Oracle Solaris running in that domain. A system board is deleted in three stages: unconfigure, disconnect, and unassign. In the delete operation, the selected system board is unconfigured from its domain by Oracle Solaris. Then, the board is disconnected to unassign it from the domain. At this point, the system board is deleted from the domain.
- **Move.** DR can be used to reassign a system board from one domain to another without stopping the instance of Oracle Solaris running in either domain. The move function changes the configurations of both domains without physically removing and remounting the system board. The move operation for a system board is a serial combination of the delete and add operations; in other words, the selected system board is deleted from its domain, and then added to the target domain.
- **Replace.** DR can be used to remove a system board from a domain, and either add it back later or replace it with another system board, without stopping the instance of Oracle Solaris running on that domain—provided both boards satisfy DR requirements (such as not making up an entire domain and no processes are running on the CPU). In the replace operation, the selected system board is deleted from the OS of the domain. Then, the system board is removed when it is ready to be released from its domain. After field parts replacement or some other such task, the system board is reinstalled and added. DR cannot be used to replace a system board in a midrange system, because replacing a system board replaces a motherboard unit. To replace a system board in a midrange system, turn off the power of all domains, and then perform hardware replacement.

In the example shown in Figure 12, System Board #2 is deleted from Domain A and added to Domain B. In this way, the physical configuration of the hardware (mounting locations) is not changed, but the logical configuration is changed for management of the system boards.

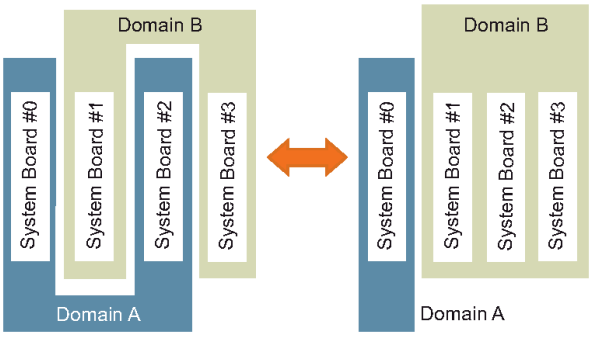


**Figure 12. DR changes logical configuration for system board management, but physical configuration is unchanged.**

### Using Dynamic Reconfiguration to Change a CPU Configuration

Upon adding a CPU, it is automatically recognized by Oracle Solaris and becomes available for use. To delete a CPU, it must meet the following conditions:

- No running process is bound to the CPU to be deleted.
- The CPU to be deleted does not belong to any processor set.
- If the resource pool facility is in use by the domain, the CPU to be deleted must not belong to a resource pool. (See "Oracle Solaris Resource Manager" for more information on resource pools and processor sets.)

A Sun SPARC Enterprise server domain runs in one of the following CPU operational modes:

- **SPARC64 VI–compatible mode.** All processors in the domain, which can be SPARC64 VI processors, SPARC64 VII processors, or any combination of them, behave like and are treated by the OS as SPARC64 VI processors. The new capabilities of SPARC64 VII processors are not available in this mode.
- **SPARC64 VII–enhanced mode.** All boards in the domain must contain only SPARC64 VII processors. In this mode, the server uses the new features of these processors.

DR operations work normally on domains running in SPARC64 VI–compatible mode. DR can be used to add, delete, or move boards with either or both processor types, and all boards are treated as if they are SPARC64 VI processors. DR also operates normally on domains running in SPARC64 VII–enhanced mode, with one exception: DR cannot be used to add or move into the domain a system board that contains any SPARC64 VI processors. To add a SPARC64 VI processor, the domain must be powered off, changed to SPARC64 VI–compatible mode, and then rebooted.

### Using Dynamic Reconfiguration to Change Memory Configurations

DR functions classify system boards by memory usage and by two types: kernel memory board and user memory board. A **kernel memory board** is a system board on which kernel memory (that is, memory internally used by Oracle Solaris and containing an OpenBoot PROM program) is loaded. Kernel memory is allocated in the memory on a single system board as much as possible. If all memory on the system board is not allocated to kernel memory and more kernel memory must be added, the memory on another system board is also used.

DR operations can be performed on kernel memory boards. When a kernel memory board is deleted, the system is suspended and kernel memory on the system board to be deleted is copied into memory on another system board. The copy destination board

- Cannot have any kernel memory
- Must have the same or more memory
- Must have the same memory configuration as the system board to be deleted

**Kernel cage memory** is a function used to minimize the number of system boards to which kernel memory is allocated. Kernel cage memory is enabled by default in Oracle Solaris 10. If the kernel cage is disabled, the system might run more efficiently, but kernel memory is spread among all boards and DR operations do not work on memory if the kernel cage is disabled.

A **user memory board** is a system board on which no kernel memory is loaded. Before deleting user memory, the system attempts to swap out the physical pages to the swap area of disks. Sufficient swap space must be available for this operation to succeed.

Some user pages are locked into memory and cannot be swapped out. These pages receive special treatment by DR. **Intimate Shared Memory (ISM) pages** are special user pages that are shared by all processes. ISM pages are permanently locked and cannot be swapped out as memory pages. ISM is usually used by database software to achieve better performance. Locked pages cannot be swapped out, but the system automatically moves these pages to the memory on another system board. Deleting user memory fails if there is not sufficient free memory size on the remaining system boards to hold the relocated pages.

### Using Dynamic Reconfiguration to Change I/O Configurations

In the domain where DR is performed, all device drivers must support the addition of devices by DR. When DR adds an I/O device, it is reconfigured automatically. An I/O device can be deleted when the device is not in use in the domain where the DR operation is to be performed and the device drivers in the domain support DR. In addition, all PCI cards and I/O device interfaces on a system board must support DR. If not, DR operations cannot be executed on that system board. In this case, the power supply to the domain must be turned off before performing maintenance and installation.

In most cases, the device to be deleted is in use. For example, the root file system or any other file systems required for operation cannot be unmounted. To solve this problem, the system can be

configured using redundant configuration software to make the access path to each I/O device redundant. One way to accomplish this for disk drives is to employ disk mirroring software.

PCI slots support hot-plug. Before a PCI card can be removed, it must be unconfigured and disconnected. XSCF controls DR events, but because hot-plug is controlled entirely within Oracle Solaris, XSCF is not aware of hot-plug events—including I/O box hot-plug events (see "I/O Box"). The service manager feature in Oracle Solaris includes a new daemon, oplhpd, that listens for I/O DR events and sends messages to XSCF. XSCF uses this information to keep track of faulty I/O cards and when they are replaced.

#### Replacing Quad-eXtended System Boards

If a domain is configured by only the XSBs in the PSB to be replaced, the DR operation for replacement is disabled, and the domain must be stopped for replacement. In the example in Figure 13, Domain #1 has a configuration that requires it to be stopped before the XSB can be replaced.

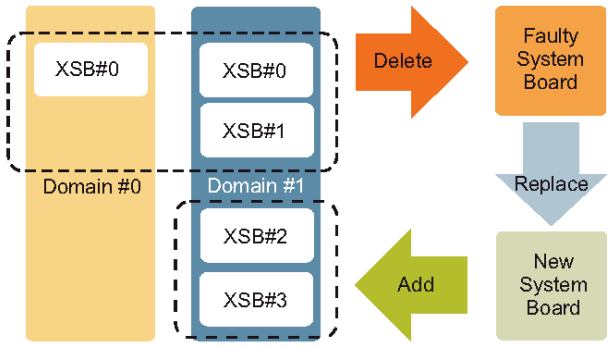


**Figure 13. Domain #1 must be stopped before the PSB can be replaced.**

### System Board Pooling

The system board pooling function assigns a specific system board a status indicating that the board does not belong to any domain. This function can be effectively used to move a system board among multiple domains, as needed. For example, a system board can be added from the system board pool to a domain when CPU or memory experiences a high load. When the added system board becomes unnecessary, it can be returned to the system board pool. A system board that is pooled can be assigned to a domain only when it is registered in the DCL for that domain.

### Reserving Domain Configuration Changes

A domain configuration change is reserved when a system board cannot be added, deleted, or moved immediately for operational reasons. The reserved add, delete, or move of the system board is executed when the power of the target domain is on or off, or when the domain is rebooted. If a system board used as a floating board is pooled in the system board pool, a domain configuration change can be reserved to assign the system board to the intended domain in advance, preventing the system board from being acquired by another domain.

### eXtended System Control Facility Settings

DR functions provide some options to avoid the complexities of reconfiguration and memory allocation with Oracle Solaris, and to make DR operations smoother. These options can be set up in XSCF.

#### Configuration Policy

DR operations automatically diagnose hardware to add or move a system board safely. Degradation of components occurs when the components are set according to the configuration of this option, and a hardware error is detected. This option specifies the range of degradation. Moreover, this option can be used for initial diagnosis by domain startup in addition to DR operations. The unit of degradation can be a component that is located where a hardware error (CPU and memory) is detected, where the component is mounted to the XSB, or at a domain.

#### Floating Board

The floating board option controls kernel memory allocation. When a system board on which kernel memory is loaded is deleted, the OS is temporarily suspended. The suspended status affects job processes and might disable DR operations. To avoid this problem, the floating board option can be used to set the priority of kernel loading into the memory of each system board, which increases the likelihood of successful DR operations. To move a system board among multiple domains, this option can be enabled.

The value of this option is TRUE (to enable the floating board setting) or FALSE (to disable the floating board setting). The default is FALSE. A system board with TRUE set for this option is called a **floating board**. A system board with FALSE set for this option is called a **nonfloating board**.

Kernel memory is allocated to the nonfloating boards in a domain by priority in ascending order of LSB number. When only floating boards are set in the domain, one of them is selected and used as a kernel memory board. In that case, the status of the board is changed from floating board to nonfloating board.

#### Omit-Memory

When the omit-memory option is enabled, the memory on a system board cannot be used in the domain. Even when a system board actually has memory, this option makes the memory on the system board unavailable through a DR operation to add or move the system board. This option can be used when the target domain needs only the CPU (and not the memory) of the system board to be added.

If a domain has a high load on memory, an attempt to delete a system board from the domain might fail. This failure results if a timeout occurs in memory deletion processing (saving of the memory of the system board to be disconnected onto a disk by paging), when many memory pages are locked because of high load. The omit-memory option can be enabled to prevent this situation. However, enabling the omit-memory option reduces available memory in the domain and could lower system performance.

### Automatic Dynamic Reconfiguration

ADR enables an application to execute DR operations without requiring user interaction. This ability is provided by an enhanced DR framework that includes the Reconfiguration Coordination Manager (RCM) and the system event facility. RCM executes preparatory tasks before a DR operation, error recovery during a DR operation, and cleanup after a DR operation. The ADR framework enables applications to automatically give up resources prior to unconfiguring them, and to capture new resources as they are configured into the domain.

### Global Automatic Dynamic Reconfiguration

Remote DR and local ADR functions are building blocks for a feature called *global ADR*. Global ADR introduces a framework that can be used to automatically redistribute the system board resources on a Sun SPARC Enterprise server system. This redistribution can be based upon factors such as production schedule, domain resource uses, and domain functional priorities. Global ADR accepts input describing resource use policies, and then uses those policies to automatically marshal the Sun SPARC Enterprise server system resources to produce the most effective use.

### Reconfiguration Coordination Manager

RCM is designed to provide a public API for DR removal events, helping programs such as databases and system management tools to take predetermined actions when hardware configuration or OS events occur. For example, applications can be notified when CPUs, memory, or interface cards are removed, so that actions can be taken to optimize performance based on the new status of the domain. (Note: On the Sun SPARC Enterprise servers, RCM scripts cannot be used for DR add operations.)

RCM is a daemon process that coordinates DR operations on resources that are present in the domain. The RCM daemon uses generic APIs to coordinate DR operations between DR initiators and RCM clients. RCM consumers consist of DR initiators (which *request* DR operations) and DR clients (which *react to* DR requests). Normally, the DR initiator is the configuration administration command cfgadm(1M). However, it can also be a GUI such as Oracle Enterprise Manager Ops Center. DR clients can be any of the following:

- Software layers that export high-level resources comprising one or more hardware devices (for example, multipathing applications)
- Applications that monitor DR operations (for example, Oracle Enterprise Manager Ops Center)
- Entities on remote systems, such as the system controller on a server

In Oracle Solaris 10, RCM is supported in Oracle Solaris Volume Manager.. RCM support gives the volume manager the ability to respond appropriately to DR requests. This helps ensure that removal of devices under the volume manager's control is blocked with an appropriate warning. The block remains in effect until the devices are no longer in use. This warning prevents IT operators from accidentally removing active volumes from a DR-configured system.

### System Events Framework

DR uses the Oracle Solaris system events framework to notify other software entities of the occurrence of changes that result from a DR operation. DR accomplishes this by sending DR events to the system event daemon, syseventd, which in turn sends the events to the subscribers of DR events.

## Capacity on Demand

Capacity on Demand (COD) is an innovative procurement model enabled by Dynamic Domains. With COD, fully configured systems are shipped with only a portion of their resources enabled—in accordance with current needs. Additional processors and memory are installed but initially disabled with a hardware enablement mechanism. Under certain conditions, COD boards can be used before actually purchasing a hardware activation option.

When a system encounters a resource constraint, additional capacity can be quickly enabled by purchasing a hardware activation option. Processors and memory can then be added to existing or new domains on the system. This approach

- Helps avoid the potentially costly possibility of overburdening critical systems when workload increases
- Helps reduce system outages
- Reduces upgrades that take valuable time to execute

COD allows IT operators to configure new or existing systems with additional processor capacity at lower acquisition costs. Additional resources can be activated when the business demands them—without disrupting operations—by purchasing COD hardware activation options. COD is designed for organizations that need to scale quickly with extra capacity to meet future and continuous increases in demand. COD permits are conveniently tracked and managed through XSCF or through Oracle Enterprise Manager Ops Center.

### Capacity on Demand Boards

A **COD board** is a system board that is configured at the factory for COD capability. COD boards come in the same configurations as standard system boards. The number of CPUs per COD board depends on the Sun SPARC Enterprise server model. A Sun SPARC Enterprise server can have any combination of COD and system boards. It can even be configured entirely with COD boards. However, COD and non-COD CPU modules cannot be mixed on the same board.

COD boards can be configured into a domain after the permits are updated. Once a COD board permit is purchased, the board can be configured into domains in the same way as a system board (including Quad-XSB) and can fully support DR operations.

### Capacity on Demand Permits

A COD permit is assigned to a specific server, one permit per CPU. A maximum of 50 permits can be installed on a Sun SPARC Enterprise server. A COD permit has no expiration date. All of the permits

assigned to a server are handled as a floating pool of permits for all of the COD processors installed on that server. For example, if a server has two COD boards with four processors each, but only six of those processors are going to be used, only six permits are required. Those six permits can be used by all eight processors, but only by six at a time.

COD uses DR to add and remove COD resources so that a reboot is not required. When a COD board is removed from a domain through a reconfiguration operation, when a domain containing a COD board is shut down normally, or when the service processor detects a fault and unconfigures a board from the domain, the COD permits for those boards are released and added to the pool of available permits. All permits remain allocated to their resources during a service processor reboot or failover.

The software allocates COD permits automatically on a first-come, first-served basis. However, permits can be reserved to domains to make sure a specific number of COD permits are allocated to a particular domain. After power on, reserved permits are first allocated to their domains, and then remaining permits are allocated on a first-served basis to the remaining resources. When a domain is powered off, the reverse happens: first, the unreserved permits are released to the pool, and then the reserved permits are released.

#### Headroom Management

**Headroom** is the capability to use up to four COD processors per server before actually purchasing a hardware activation option. With headroom, a COD board can be activated as a hot spare to replace a failed system board—or when permit purchase is imminent but the resources are needed immediately. Using headroom to activate a COD resource also prompts a contractual obligation to purchase the hardware activation options.

By default, headroom is disabled on COD resources. Headroom can be enabled, reduced, or disabled at any time. While in use, warning messages appear on the console every four hours. Once a hot-spared COD board is deactivated or a permit is purchased and the keys entered, the warnings stop.

### I/O Box

The Sun External I/O Expansion Unit (I/O box) for Sun SPARC Enterprise servers provides additional slots for PCI cards. A single I/O boat within the I/O box provides six additional slots. A two I/O boat configuration provides 12 slots. The I/O box includes two power supplies and fans.

The I/O box supports two types of I/O boats: PCI-X and PCIe. PCI cards are not interchangeable between the two types of boats. The PCI-X I/O boat accepts PCI-X cards and some older types of PCI cards. The PCIe I/O boat accepts PCIe cards up to x8 lanes wide. PCIe x16 cards do not fit in this boat. PCI card slots are hot-pluggable. I/O boxes are added to the system by inserting a link card into a PCIe slot in an I/O unit and using a cable to connect the link card to the I/O box.

XSCF monitors I/O boxes for voltage, current, and temperature and can power down the I/O box if parameters are exceeded. Hot-swapping of PCI cards requires a combination of commands through

the Oracle Solaris domain DR commands and XSCF, which detects and enables PCI cards and I/O box units.

## Fine-Grained Resource Management

Traditionally, dedicated servers are configured to match the peak resource needs of a single critical application. Oracle Solaris Containers helps to enable application consolidation on Sun SPARC Enterprise servers and more-efficient system use through granular resource management within the server domains.

Whereas Dynamic Domains enable consolidation of several servers into one Sun SPARC Enterprise server, Containers allows consolidation of several applications into one domain. For example, a single Sun SPARC Enterprise server, partitioned into domains and each running an instance of Oracle Solaris, can support application, file, and print services for heterogeneous clients; messaging/mail; and Web services, application services, and mission-critical databases for entire application services. Because the Sun SPARC Enterprise M9000-64 server can scale to 64 processors with up to 256 cores, one server can easily be shared by a number of applications or application services. And, by combining application workloads with different usage profiles, resources can be better optimized for optimal application performance.

In other server consolidation efforts, development, prototype, and production environments might be combined on a single large server, rather than on three separate servers. Still other consolidation projects combine multiple database instances and application servers within a single system, sharing the same OS instance and providing cost savings in administrative tasks such as data management and archive.

Containers can be configured to favor certain users in mixed workload environments. For example, in large brokerage firms, traders intermittently require fast access to execute a query or perform a calculation, whereas other system users have more-consistent workloads. Using Containers, traders can be granted a proportionately larger number of shares of resources to give them the system resources they require.

### Oracle Solaris Containers

A primary objective of Oracle Solaris 10 is to deliver tools to help IT departments do more with less by consolidating applications onto fewer servers. The Containers functionality in Oracle Solaris 10 enables multiple, software-isolated applications to run on a single server or domain, allowing IT to easily consolidate servers. IT operators can also gain tight control over allocation of system and network resources, significantly improving resource use. Combined with the predictive self healing feature, user and process rights management in Oracle Solaris, and DTrace, the capabilities of Oracle Solaris 10 can help consolidate applications without compromising the service levels, privacy, or security of individual applications or users. Containers allows organizations to

- Build customized, isolated containers—each with its own IP address, file system, users, and assigned resources—to safely and easily consolidate systems

- Guarantee sufficient CPU and memory resource allocation to applications while retaining the ability to use idle resources as needed
- Reserve and allocate a specific CPU or group of CPUs for the exclusive use of the container
- Automatically recover from potentially catastrophic system problems by leveraging the combined functionality of the predictive self healing feature and Containers

An Oracle Solaris Container is a virtualized OS environment created within a single instance of Oracle Solaris. Applications within containers are isolated, preventing processes in one container from monitoring or affecting processes running in another container. Even a superuser process from one container cannot view or affect activity in other containers. A container also provides an abstract layer that separates applications from the physical attributes of the system on which they are deployed. Examples of these attributes include physical device paths.

Containers enable more-efficient use of the system. Dynamic resource reallocation allows unused resources to be shifted to other containers as needed. Fault and security isolation means that poorly behaved applications do not require a dedicated and underused system. With containers, these applications can be consolidated with other applications. Containers also allow the IT operator to delegate some administrative functions while maintaining overall system security.

Oracle Solaris Containers is designed to provide fine-grained control over the resources that applications use, enabling multiple applications to operate on a single server while maintaining specified quality of service levels. Fixed resources such as processors and memory can be partitioned into pools on multiprocessor systems, with different pools shared by different projects (a specified collection of processes) and isolated application environments. Dynamic resource sharing enables different projects to be assigned different ratios of system resources. The Oracle Solaris IP quality-of-service (IPQoS) feature can be employed to manage network bandwidth used by multiple, competing network applications and is covered in more detail in the section "Managing Other Resources, Monitoring, and Accounting." When resources such as CPUs and memory are dynamically allocated, resource capping controls can be used to set limits on the amount of resources used by a project. With all of these resource management capabilities, organizations can consolidate many applications onto one server, helping to reduce operational and administrative costs while increasing availability (see Figure 14).

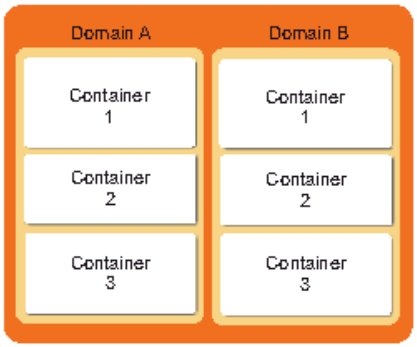


**Figure 14. Consolidate many applications on one server with Containers by placing multiple containers in domains.**

Resource management is provided by Oracle Solaris Resource Manager. Every service is represented by a project, which provides a networkwide administrative identifier for related work. All the processes that run in a container have the same project identifier, also known as the project ID. The Oracle Solaris kernel tracks resource usage through the project ID (see Figure 15). Historical data can be gathered by using extended accounting.

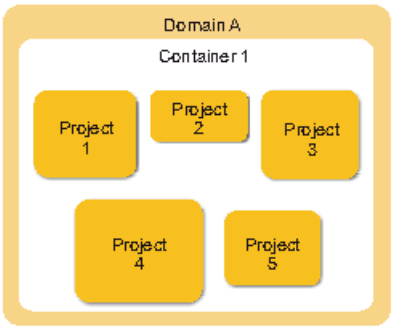


**Figure 15. With Containers, processes running in a single container are tracked for resource usage by the same project ID.**

## Oracle Solaris Resource Manager

Modern computing environments have to provide a flexible response to the varying workloads that are generated by different, consolidated applications on a system. A **workload** is an aggregation of all processes of an application or group of applications. Oracle Solaris provides a facility called *projects* to name workloads once they are identified. For example, one project would be named for a sales database and another project for a marketing database. If resource management features are not used, Oracle Solaris responds to workload demands by adapting to new application requests dynamically. This default response generally means that all activity on the system is given equal access to resources.

Oracle Solaris Resource Manager enables the system to treat workloads individually by

- Restricting access to a specific resource
- Offering resources to workloads on a preferential basis
- Isolating workloads from each another
- Denying resources or preferring one application over another for a larger set of allocations than otherwise permitted
- Preventing an application from consuming resources indiscriminately
- Changing an application's priority based on external events
- Balancing resource guarantees to a set of applications against the goal of maximizing system usage

These capabilities allow Containers to deliver predictable service levels. Effective resource management is enabled in Oracle Solaris by offering control, notification, and monitoring mechanisms. Many of

these capabilities are provided through enhancements to existing mechanisms, such as the proc(4) file system; processor sets; scheduling classes; and new mechanisms, such as dynamic resource pools.

### Dynamic Resource Pools

Resource pools allow the IT operator to separate workloads so that they do not consume overlapping resources. They provide a persistent configuration mechanism for processor sets and, optionally, scheduling classes, as illustrated in Figure 16. Resource pools provide a mechanism for dynamically adjusting each pool's resource allocation in response to system events and application load changes. Dynamic resource pools simplify and reduce the number of decisions required from the IT operator. Pools are automatically adjusted to preserve system performance goals. The software periodically examines the load on the system and determines whether intervention is required to enable the system to maintain optimal performance.

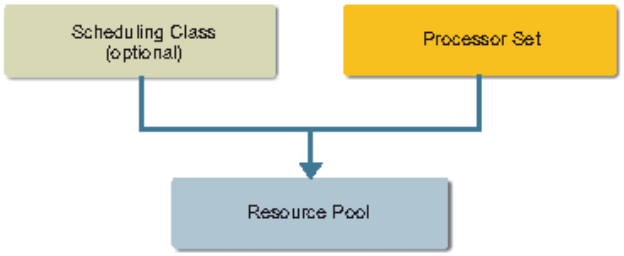


**Figure 16. Resource pools provide a persistent configuration mechanism for processor sets and, optionally, scheduling classes.**

On a system that has zones enabled, a nonglobal zone can be associated with one resource pool, although the pool does not need to be exclusively assigned to a particular zone. Moreover, individual processes in nonglobal zones cannot be bound to a different pool by using the poolbind command from the global zone. If a scheduling class is set for a pool and a nonglobal zone is associated with that pool, the zone uses that scheduling class by default.

### Resource Management Control

Oracle Solaris provides three types of control mechanisms to control resource usage:

- **Constraint.** Constraint is a resource-sharing mechanism that sets bounds on the amount of specific resources a workload can consume. It can also be used to control ill-behaved applications, such as applications with memory leaks, which can otherwise compromise performance or availability through unregulated resource requests.
- **Scheduling.** Scheduling is a resource-sharing mechanism that refers to making a sequence of resource allocation decisions at specific intervals, based on a predictable algorithm. An application that does not need its current allocation leaves the resource available for another application's use. Scheduling-based resource management enables full use of an undercommitted configuration, while providing controlled allocations in a critically committed or overcommitted scenario. The algorithm

determines the level of control; for example, it might guarantee that all applications have some access to the resource. The fair share scheduler is an example of a scheduling mechanism that manages application access to CPU resources in a controlled manner.

- **Partitioning.** Partitioning is a more-rigid mechanism used to bind a workload to a subset of the system's available resources. This binding guarantees that a known amount of resources is always available to the workload. Resource pools as a partitioning mechanism limit workloads to a specific subset of the resources of the system. Partitioning can be used to avoid systemwide overcommitment. However, in avoiding this overcommitment, the ability to achieve high usage can be reduced because resources bound to one pool are not available for use by a workload in another pool when the workload bound to them is idle—unless a policy for dynamic resource pools is employed. A good candidate for this type of control mechanism might be transaction processing systems that must be guaranteed a certain amount of resources at all times.

### Managing CPU Resources with Resource Pools

The ability to partition a server using processor sets has been available since Oracle Solaris 2.6. Every system contains at least one processor set: the system or default processor set that consists of all of the processors in the system. Additional processor sets can be dynamically created and removed on a running system, provided that at least one CPU remains for the system processor set.

Resource pools allow IT operators to create a processor set by specifying the number of processors required, rather than CPU physical IDs. The definition of a processor set is, therefore, not tied to any particular type of hardware. It is also possible to specify a minimum and maximum number of processors for a pool. Multiple configurations can be defined to adapt to changing resource requirements, such as different daily, nightly, or seasonal workloads. Resource pools can have different scheduling classes. Scheduling classes work per resource pool. The two most common are the fair share scheduler and the time share scheduler.

### Fair Share Scheduler

The fair share scheduler (FSS) allocates CPU resources by using CPU shares. The FSS helps ensure that CPU resources are distributed among active zones or projects based on the number of shares each zone or project is allocated. Therefore, more-important workloads should be allocated more CPU shares. A CPU share defines the portion of the CPU resources available to a project in a resource pool. It is important to note that CPU shares are not the same as CPU percentages. Shares define the relative importance of projects with respect to other projects. If Project A is twice as important as Project B, then Project A should be assigned twice as many shares as Project B. The actual number of shares assigned is largely irrelevant—2 shares for Project A versus 1 share for Project B yields the same results as 18 shares for Project A versus 9 shares for Project B. Project A is entitled to twice the amount of CPU as Project B in both cases. The importance of Project A relative to Project B can be increased by assigning more shares to Project A, while keeping the same number of shares for Project B.

The FSS calculates the proportion of CPU allocated to a project by dividing the shares for the project by the total number of active projects. An **active project** is a project with at least one process using

the CPU. Shares for idle projects—that is, projects with no active processes—are not used in the calculations. An important point is that the FSS only limits CPU usage if there is competition for the CPU. A project that is the only active project on the system can use 100 percent of the CPU, regardless of the number of shares it holds. CPU cycles are never wasted—if a project does not use all of the CPU it is entitled to because it has no work to perform, the remaining CPU resources are distributed between other active processes.

### Fair Share Scheduler and Processor Sets

The FSS can be used in conjunction with processor sets to provide more-fine-grained control over allocation of CPU resources among projects that run on each processor set than would be available with processor sets alone. When processor sets are present, the FSS treats every processor set as a separate partition. CPU entitlement for a project is based on CPU usage in that processor set only. The CPU allocations of projects running in one processor set are not affected by the CPU shares or activity of projects running in another processor set, because the projects are not competing for the same resources. Projects only compete with each other if they are running within the same processor set, as illustrated in Figure 17.

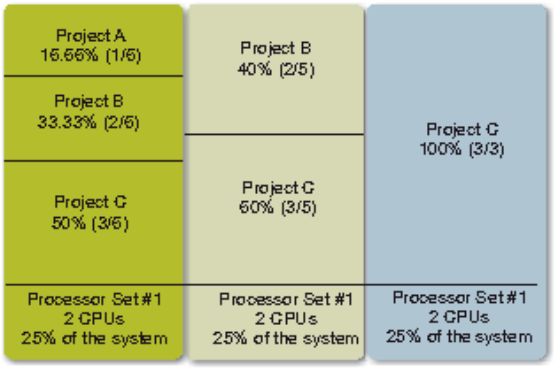


**Figure 17. The FSS ensures projects only compete with each other if they are running within the same processor set.**

### Resource Pools and Dynamic Reconfiguration Operations

DR enables the hardware to be reconfigured while the system is running. A DR operation can increase, reduce, or have no effect on a given type of resource. Because DR can affect available resource amounts, the pools facility must be included in these operations. When a DR operation is initiated, the pools framework acts to validate the configuration. If the DR operation can proceed without causing the current pools configuration to become invalid, then the private configuration file is updated. An invalid configuration is one that cannot be supported by the available resources.

If the DR operation causes the pools configuration to be invalid, then the operation fails and the IT operator is notified by a message to the message log. The configuration can be forced to complete by using the DR force option. The pools configuration is then modified to comply with the new resource configuration.

#### Using DTrace with Oracle Solaris Containers

DTrace is zone aware, making it especially useful for troubleshooting problems between applications running in Containers. For example, on OSs other than Oracle Solaris 10, it is very difficult to find and rectify performance issues between different systems—such as a Web server, an application server, and a database server all running on separate systems. With Oracle Solaris 10 and Containers, all three of these applications can run in different (or the same) containers on a single system. Using DTrace, it is now possible and easy to pinpoint performance issues or transient problems occurring between applications—issues that are extremely difficult, time consuming, and expensive to uncover in other OSs. In addition, DTrace can be used to ascertain the level of resources an application uses, helping to fine-tune resource allocation on a zone or project level.

## Managing Other Resources, Monitoring, and Accounting

Resource management helps to ensure the applications that are consolidated and running on a single system (as well as single application servers) meet required response times and service levels. It can also be used to increase resource use. By categorizing and prioritizing usage, reserve capacity can be used during off-peak periods, often eliminating the need for additional processing power. Oracle Solaris 10 includes a variety of features for managing, monitoring, and accounting for usage of memory, network bandwidth, and storage resources between multiple applications running on the same system. It also includes a facility to assign rights to processes.

### User and Process Rights Management in Oracle Solaris

User and process rights management, introduced in Oracle Solaris 10, gives IT operators the ability to limit and selectively enable applications to gain access to just enough system resources to perform their functions. This capability dramatically reduces the possibility of attack from a poorly written application by eliminating inappropriate access to the system. Even if hackers gain access to an application server, they are unable to increase operating privileges, thus limiting the opportunity to inject malicious code or to otherwise damage data. In Containers, user and process rights management helps to ensure that applications—even those run with privileges—are constrained to access resources only in their own Containers.

### Resource-Capping Daemon

A **resource cap** is an upper boundary placed on the consumption of a resource such as physical memory. Per-project physical memory caps and CPU caps are supported. The resource-capping daemon and its associated utilities provide mechanisms for physical memory and CPU resource cap enforcement and administration.

The resource cap daemon repeatedly samples the resource use of projects that have physical memory or CPU caps. When the system's physical memory usage exceeds the threshold for cap enforcement and other conditions are met, the daemon takes action to reduce the resource consumption of projects with memory caps to levels at or below the caps.

Physical memory control that uses the resource capping daemon is an optional feature. The resource-capping rcapd daemon regulates the consumption of physical memory by processes that run in projects with defined resource caps. Associated utilities provide mechanisms for administering the daemon and reporting related statistics.

CPU caps provide absolute fine-grained limits on the amount of CPU resources that can be consumed by a project or a zone. CPU caps are provided as a zonecfg resource, and as project and zonewide resource controls.

### IP Quality-of-Service

With the release of Oracle Solaris 9, new network resource management technology was introduced, superseding Oracle Solaris Bandwidth Manager, which was available for previous releases of Oracle Solaris. This technology is usually referred to as simply IP quality-of-service, or IPQoS, as discussed earlier. IPQoS is ideal for controlling, monitoring, and accounting for network resources. IPQoS forms the foundation of managing the network resource aspects of Containers with features that can help make network performance more efficient. It is implemented at the IP level of the TCP/IP stack and is configured for the global zone, unless nonglobal zone traffic is routed outside of the system. IPQoS can be a central means to offer SLAs, allowing IT to provide

- Guaranteed bandwidth for mission-critical business applications
- Reduced traffic congestion and increased network efficiency
- Controlled user and application access to network resources
- Detailed network use statistics and accounting data for billing purposes
- Differentiated classes of network service to users

IPQoS helps to enable workload-centric datacenter management, which is essential for consolidating applications onto a single server. It helps ensure that a group or application does not consume more than its allotted bandwidth. Users can be charged for the exact amount of network resources they consume, and resources can be dynamically assigned to the workloads that require them—when they need them.

### IP Network Multipathing

IP network multipathing, or IPMP, is a feature in Oracle Solaris that enables IP failover and IP link aggregation. It helps manage network workloads and failures on the Sun servers in the following ways:

- **Outbound load spreading.** IPMP spreads outbound network packets across multiple network adapters, without affecting the ordering of packets, to achieve higher throughput. Load spreading occurs only when the network traffic is flowing to multiple destinations using multiple connections.
- **Failure detection.** IPMP offers the ability to detect a network adapter failure and automatically switch (fail over) its network access to an alternate network adapter.

- **Repair detection.** IPMP enables the ability to detect repair or replacement of a previously failed network and to automatically switch back (fail back) network access from an alternate network adapter.
- **Dynamic reconfiguration.** On systems that support DR, IPMP can be used to transparently fail over network access, providing uninterrupted network access to the system.

## Managing Storage Resources in Oracle Solaris Containers

Storage is managed at the global zone level. Each zone is configured with a portion of the file system hierarchy that is located under the zone root. Because each zone is configured to its subtree of the file system hierarchy, a workload running in a particular zone cannot access the on-disk data of another workload running in a different zone.

Oracle Solaris Containers enables sharing of the file system data, especially read-only data such as executables and libraries. Parts of the file system can be shared between zones in the system by using the read-only loopback file system, which allows a directory and its contents to be inserted into another part of the file system. The loopback file system is improved in Oracle Solaris 10 to support read-only mounts, preventing nonglobal zones from writing in the shared directory. This not only substantially reduces the amount of disk space used by each container, but also reduces the time to install zones and apply patches and enables greater sharing of text pages in the virtual memory system.

In addition, multiple applications in one zone or multiple zones can access the same data by implementing Sun QFS (multireader/multiwriter capability).

### Using Oracle Solaris ZFS in Oracle Solaris Containers

Oracle Solaris ZFS data sets (file systems, snapshots, volumes, or clones) can be added to a zone either as a generic file system or as a delegated data set. Adding a file system enables the nonglobal zone to share space with the global zone, although the zone administrator cannot control properties or create new file systems in the underlying file system hierarchy. This is identical to adding any other type of file system to a zone, and should be used when the primary purpose is solely to share common space.

ZFS also enables data sets to be delegated to a nonglobal zone, giving complete control over the data set and all of its children to the zone administrator. The data set is then visible and mounted in the nonglobal zone and no longer visible in the global zone. The zone administrator can create and destroy file systems within that data set, and modify properties of the data set. The zone administrator cannot affect data sets that have not been added to the zone or exceed any top-level quotas set on the data set assigned to the zone.

## Oracle Solaris Fibre Channel and Storage Multipathing

Oracle Solaris Fibre Channel (FC) and Storage Multipathing software is integrated into Oracle Solaris 10. In Oracle Solaris 10, fabric-connected devices are configured and made available to the system automatically during install and boot time. Oracle Solaris FC and Storage Multipathing software provides the following key features:

- **Dynamic storage discovery.** This feature automatically recognizes devices and any modifications made to device configurations. This makes devices available to the system without requiring a reboot or a manual change to information in configuration files.
- **Persistent device naming.** This feature enables devices to maintain their device naming through reboots, reconfiguration, or both. The only exception to this is the tape device, found in /dev/rmt. Tape devices are not changed unless they are removed and later regenerated.
- **Fabric booting.** Oracle-supported host bus adapters can boot from fabric devices as well as from nonfabric devices.
- **Path management.** This feature dynamically manages the paths to any storage devices the software supports. Adding or removing paths to a device is done automatically when a path is brought online or removed from a service. This enables systems configured with Solaris FC and Storage Multipathing software to begin with a single path to a device and add more host controllers, increasing bandwidth and availability—without changing device names or modifying applications. For Sun storage, there are no configuration files to manage or databases to keep current.
- **Single device instances.** Unlike other multipathing solutions, Oracle Solaris FC and Storage Multipathing software is fully integrated with Oracle Solaris 10, enabling Oracle Solaris FC and Storage Multipathing software to display multipath devices as single device instances instead of as one device, or device link, per path. This reduces the cost of managing complex storage architectures, because it enables utilities such as format(1M)—or higher-level applications such as Oracle Solaris Volume Manager—to access one representation of a storage device instead of a separate device for each path.
- **Failover support.** Implementing higher levels of reliability and availability requires redundant host connectivity to storage devices. Oracle Solaris FC and Storage Multipathing software manages the failure of storage paths while maintaining host I/O connectivity through available secondary paths.
- **I/O load balancing.** In addition to providing simple failover support, Oracle Solaris FC and Storage Multipathing software can use any active path to a storage device to send and receive I/O. With I/O routed through multiple host connections, bandwidth can be increased by adding host controllers. Oracle Solaris FC and Storage Multipathing software uses a round-robin load-balancing algorithm, by which individual I/O requests are routed to active host controllers in a series, one after the other.
- **Dynamic reconfiguration.** Oracle Solaris FC and Storage Multipathing software supports the Oracle Solaris 10 Dynamic Reconfiguration feature.

## Monitoring and Accounting

When running many applications on one system, it is important to constantly monitor the system—and the resource pools, zones, and projects within the system—in order to efficiently manage the resources of the system to best meet the needs of the applications and users on that system.

### Oracle Solaris Process Accounting and Statistics

**System accounting software** in Oracle Solaris is a set of programs that can collect and record data about user connect time, CPU time charged to processes, and disk usage. With the collected data, reports can be generated to help IT departments charge fees for system usage. The system accounting programs can be used to

- Monitor system usage
- Locate and correct performance problems
- Maintain system security

The system accounting software provides C language programs and shell scripts that organize data into summary files and report. Daily accounting helps IT departments perform four types of auditing:

- **Connect accounting.** This type of auditing helps determine the length of time a user is logged in, and keeps track of such data as the number of reboots on the system and creation and termination of user processes.
- **Process accounting.** This type of auditing keeps track of data about each process that runs on the system—including user and group IDs of users, beginning and elapsed times, CPU time, amount of memory used, and the commands run by the process.
- **Disk accounting.** This type of auditing gathers and formats data about the files each user has on disk, including the number of blocks that are used by the user's files.
- **Fee calculation.** This type of auditing enables a chargefee utility to store charges for special services that are provided to a user. An example of such a special service is file restoration.

### Extended Accounting

The extended accounting subsystem acctadm(1M) within Oracle Solaris provides a flexible way to record system and network resource consumption on a task or process basis, or on the basis of selectors provided by the IPQoS flowacct module. The extended accounting subsystem collects and reports information for the entire system (including nonglobal zones) when run in the global zone. It labels usage records with the project for which the work is performed, and the global administrator can also determine resource consumption on a per-zone basis.

The wracct(1M) process writes extended accounting records for active processes and tasks. The files that are produced can be used for planning, charging back, and billing. With extended accounting data available, it is possible to develop or purchase software for resource chargeback, workload monitoring, or capacity planning. There is also a Perl interface to libexacct that enables the development of customized reporting and extraction scripts.

### Monitoring Resource Pools

The poolstat utility is used to monitor resource usage when pools are enabled on the system. This utility iteratively examines all of the active pools on a system and reports statistics based on the selected output mode. The poolstat statistics help to determine which resource partitions are heavily used.

These statistics can be analyzed to make decisions about resource reallocation when the system is under resource pressure.

### Monitoring Network Usage

The IPQoS flowacct module can be used to collect information about traffic flows on networks; for example, source and destination addresses, the number of packets in a flow, and similar data can be collected. The process of accumulating and recording information about flows is called *flow accounting*. The results of flow accounting on traffic of a particular class are recorded in a table of flow records. Each flow record consists of a series of attributes. These attributes contain data about traffic flows of a particular class over an interval of time.

Flow accounting is particularly useful for billing clients as defined in their SLAs. Flow accounting can also be used to obtain flow statistics for critical applications to observe their behavior.

### Monitoring Capped Memory

IT operators can use rcapstat to monitor the resource usage of projects that are configured with memory caps. The rcapstat report includes information such as the project ID of the capped project, the project name, virtual memory size of all processes in the project, the total resident set size of the project's processes, and the total resident set size cap for the project.

### Monitoring Resource Controls

Often, the resources that a process consumes are unknown. To obtain more information, global resource control actions are available with the `rctladm(1M)` command, which can be employed to establish a syslog action on a resource control such as shared memory. Then, if any entity managed by that resource control encounters a threshold value, a system message is logged at the configured logging level.

For example, it might be necessary to determine whether a Web server application is allocated sufficient CPU resources for its typical workload. The sar data could be analyzed for idle CPU time and load average, or the extended accounting data could be examined to determine the number of simultaneous processes that are running for the Web server process. However, an easier approach is to place the Web server in a task, and then set a global action to notify the IT operator whenever a task exceeds a scheduled number of light-weight processes appropriate for that application.

## Conclusion

In today's exceedingly competitive environment, where profit margins continue to shrink, every IT department operates under the mandate to reduce complexity, reduce costs, increase return on investment, provide a more-consistent environment to support compliance initiatives, and adapt quickly to changes in demand and business processes. By consolidating applications onto Sun SPARC Enterprise servers, IT departments can do all of this, and more.

Oracle's Sun SPARC Enterprise servers are the most powerful and innovative enterprise-class systems available today. With the ability to partition the system into subboard level domains, isolate applications into containers, and manage resources with fine-grained and dynamic control, the systems are ideally suited for consolidating applications and optimally using resources.

Borrowing from the mainframe world, the systems also include GUI-based tools for administering, monitoring, and managing the hardware, OS, storage, and applications. These tools streamline and automate many tasks, thus decreasing complexity and IT operations costs while providing a more-consistent environment.

Oracle is committed to developing practices and products that help protect the environment

Sun SPARC Enterprise Servers System and Resource Management
June 2010

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com



**SOFTWARE. HARDWARE. COMPLETE.**