SUN SPARC® ENTERPRISE SERVERS SYSTEM AND RESOURCE MANAGEMENT

White Paper
April 2007

Abstract

The next generation Sun SPARC® Enterprise servers are the most powerful systems offered by Sun Microsystems. 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. Sophisticated system, resource, and workload management tools are needed to harness the power of their performance, flexibility, and high availability. This paper discusses the system and resource management tools available for the Sun SPARC Enterprise servers.

Table of Contents

Chapter 1
Hardware Alone is Not Enough

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, as well as rising power consumption and other constraints in the data center. Sophisticated system, resource, and workload management tools are needed to harness the power of every system's performance, flexibility, and availability. The Sun SPARC® Enterprise servers (Figure 1), 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 data centers with a goal of increasing performance and flexibility, while consolidating systems to reduce data center costs and complexity.

The Sun SPARC Enterprise M9000-64 server offers the highest availability, highest absolute performance, highest scalability, and the most sophisticated control of resources in Sun Microsystem's extensive server product line. It supports up to 64 dual core SPARC64 VI 64-bit Chip Multithreading (CMT) processors (up to 128 cores), 2 TB of memory, 128 hot-swappable PCI-Express (PCIe) I/O slots, up to 64 internal disks, support for external I/O, a new high performance interconnect capable of up to 304.2 GB/sec., support for up to 24 domains, and thousands of software applications for the 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 (SMP) platforms.

Figure 1. Sun SPARC Enterprise server family

Sun's new SPARC64 VI CMT processor contains two SPARC V9 cores running 2.15, 2.28, or 2.4 GHz, 6 MB on-chip shared L2 cache, two vertical threads per core, 128 KB of I cache and 128K of D cache per core, 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 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 Enterprise M9000-32 server offers half the maximum configuration of the Sun Enterprise SPARC M9000-64 server, and consequently, lower acquisition and operating costs. Supporting up to 32 CPUs (up to 64 cores), 1 TB of memory, 64 hot-swappable PCIe I/O slots, 32 internal disks, support for external I/O, and support for 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 systems family. Using the same building block components of 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 (32 cores), 512 GB of memory, 32 hot-swappable PCIe I/O slots, 16 internal disks, support for external I/O, and up to 16 domains.

The Sun SPARC Enterprise server family also includes two midrange, rack-based systems — the Sun SPARC Enterprise M4000 and M5000 servers. These systems offer economical price/performance for the datacenter with support for up to four domains, Dynamic Reconfiguration, and featuring greatly improved RAS functionality. The Sun SPARC Enterprise M4000 server's 6 RU enclosure supports up to four dual core SPARC64 VI processors (eight cores), 128 GB of memory, two internal disks, support for external I/O, and two domains. The 10 RU Sun SPARC Enterprise M5000 server supports up to eight dual core SPARC64 VI processors (16 cores), 128 GB of memory, four internal disks, support for 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.

Managing Consolidated Applications

In the past, data centers focused on predictability, discipline, and optimizing resources for maximum efficiency in order 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.

Applications are no longer simply stand-alone applications, they are generally a component 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 consist of database, 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. All of these systems create a level of complexity both from the perspective of administering and maintaining the systems and applications, and from the perspective of successfully integrating application services. 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 utilization, and provide a more standardized environment to integrate applications services to improve business processes.

The Sun SPARC Enterprise servers fulfill these requirements 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, Solaris Containers, and full system redundancy. In addition, they deliver advanced reliability features including instruction-level re-try, protected SRAMS and registers, extended ECC memory and mirroring, end-to-end ECC protection, hot-swappable components, and hardware redundancy, at open systems prices.

As consolidation platforms, the Sun SPARC Enterprise servers are easier to manage than similar systems from competitors.

- The Sun SPARC Enterprise M9000-32 and M9000-64 servers offers 24 hardware partitions or domains.
- The Sun SPARC Enterprise M4000 to M9000 server supports over 8000 Solaris Containers per domain.
- The Sun SPARC Enterprise M4000 to M9000 servers require only one Solaris OS instance for thousands of containers, while competitors require one OS instance per software partition. Fewer instances are easier to manage.
- Solaris Containers are free, while partitions from competitors must be purchased.

Managing Consolidation

The increased urgency and accelerated time frames of internal technical initiatives leads to server sprawl, or the deployment of multiple single-purpose servers to achieve needed availability or scalability. The result is under utilized resources and increased complexity, leading to significantly higher training, operational, and data center (floor space, heating, cooling) costs, as well as an inability to adapt to changing needs.

Consolidation collapses the functionality of multiple servers into a smaller number of systems, reducing costs and management complexity. Server consolidation can also

increase return on investment (ROI) by sharing existing resources across multiple applications. Server consolidation requires large servers that are capable of running more than one application simultaneously, an operating system 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. The Sun SPARC Enterprise servers, together with the Solaris 10 Operating System, offer an extensive array of system, resource, and workload management capabilities to enable IT organizations to reduce costs, decrease complexity, and rapidly respond to changes in demand.

Chapter 2

Managing Sun SPARC Enterprise Servers

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

- **eXtended System Controller Facility (XSCF)** — a graphical user interface (GUI)-based or command line interface (CLI)-based set of system management applications.
- **Sun™ Management Center software** — a GUI-based single point of management for systems, storage, and operating systems.
- **Solaris Management Console** — used to perform administrative tasks involving users, projects, and jobs.
- **Configuration and Service Tracker** — monitors hardware configuration changes.
- **Sun℠ Net Connect Services** — self monitoring, configuration and patch collection, and reporting.
- **Solaris Operating System features** — for upgrading the OS and managing system and application services.

eXtended System Controller Facility (XSCF)

Sun SPARC Enterprise servers provide system management capabilities through XSCF firmware, which is pre-installed on the Service Processor boards. XSCF firmware consists of system management applications and two user interfaces: XSCF Web (a browser-based GUI) and XSCF Shell (a terminal-based CLI). The XSCF Web uses the secure version of the HyperText Transfer Protocol (https) and the Secure Sockets Layer/ Transport Level Security (SSL/TLS) protocols for connection to the server connected to a network and for Web-based support of server status display, server operation control, and configuration information display.

The 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. The XSCF centrally controls and monitors the server. The XSCF includes a partitioning function to configure and control domains. It also has a function to monitor the server through an Ethernet connection to enable remote control.

The 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 on or off the whole

system, or 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 OS console of the Solaris OS of each domain from XSCF via the LAN or serial port of the XSCF unit. With an SSH (Secure Shell) 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 (XSCF firmware or OpenBoot PROM firmware) without stopping the domain and 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 Management Information Bases-II (MIB-II) (SNMPv2 and SNMPv3) and MIB-I (SNMPv1).

- *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 licenses.
- *I/O Box management* — Displays I/O box information, configures the I/O box, and can power on/off specific I/O boards or power supply units.

XSCF firmware has two networks for internal communication: DSCP and XSCF. The Domain to Service Processor Communications Protocol (DSCP) network provides an internal communication link between the Service Processor and the Solaris domains. The DSCP service provides a secure Transmission Control Protocol/Internet Protocol (TCP/IP)- and Point-to-Point Protocol (PPP)-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 SPARC Enterprise M8000 or 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 allows 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.

Sun Management Center Software

To help increase service levels and decrease administrative costs in enterprise computing environments, Sun Management Center software offers a single point of management for Sun system and storage components, the Solaris Operating System, Linux, and applications. With Sun Management Center software, organizations can deliver monitoring and management capabilities that optimize performance, enhance application availability, and simplify management of the IT environment. Because IT operators are able to control the enterprise infrastructure from virtually anywhere on the network or on the Web, they can actively manage hundreds of Sun systems at once through a Java™ technology interface or Web browser. While providing the most comprehensive instrumentation and administrative knowledge for Sun environments, open interfaces enable information to be shared with other management platforms, thus lowering the total cost of system management.

Sun Management Center software supports full integration with Dynamic Domains, Dynamic Reconfiguration, Reconfiguration Coordination Manager, Solaris Containers, and IP Network Multipathing. It also tightly integrates with major enterprise management frameworks such as CA Unicenter and IBM Tivoli, enabling easy integration into heterogeneous IT environments.

Sun Management Center software for Sun SPARC Enterprise systems is an extensible monitoring and management tool that integrates standard SNMP-based management structures with intelligent and autonomous agent and management technology. Sun Management Center software is used as the GUI and SNMP manager and agent infrastructure for the Sun SPARC Enterprise system. The SNMP agent is part of the XSCP software, and can be loaded into any domain. The agent exports all information relevant for monitoring at the system, component, and domain levels to Sun Management Center software.

Sun Management Center software is accessed via a Java technology-based GUI that is intended to provide a common look and feel for all Sun Management Center applications, and offers the flexibility to monitor and manage the enterprise from any platform supporting Java technology. It allows proactive and automated management of complex and common administrative tasks, reducing the likelihood of costly errors and promoting high availability. Another important feature is the ability to perform predictive failure analysis, alerting operators to potential memory and disk failures on a statistical basis, aiding decision making, and increasing availability.

Sun Management Center software includes the following features.

- *System management* — monitors and manages the system at the hardware and operating system levels. Monitored hardware includes boards, tapes, power supplies, and disks.
- *Operating system management* — monitors and manages operating system parameters that include load, resource usage, disk space, and network statistics.
- *Application and business system management* — provides enabling technology to monitor business applications such as trading systems, accounting systems, inventory systems, and control systems.
- *Scalability* — provides an open, scalable, and flexible solution to configure and manage multiple management administrative domains. These domains consist of many systems and span across an enterprise. IT operators can configure the software in a centralized or distributed fashion so that the product supports multiple users.

This solution uses SNMP, the Java Remote Method Invocation (RMI), and http. These tools enable Sun Management Center software to provide integrated, comprehensive enterprise-wide management of Sun products and their subsystems, components, and peripheral devices.

Sun Management Center software employs autonomous SNMP-base agent technology, a technique where agents are not dependent on other software components. The agents collect and process data locally on the nodes that Sun Management Center software manages, and can act on data to send SNMP traps and run processes. The agents can also initiate alarms, send notification, or initiate specific actions through customizable rules and thresholds.

System Administration Tools

Sun Management Center software includes a number of system administration tools to help IT operators perform their jobs more efficiently:

- *Physical view* — displays photo-realistic images of hardware components and points to components with an associated event. This tools helps IT operators to quickly determine which components need to be replaced.
- *Logical view* — presents a tree hierarchy of the managed host or domain, including all hardware and operating system components. When an event is associated with a particular component, the logical viewer identifies its exact location within the hierarchy.
- *Alarm manager* — displays alarm information for managed objects and enables the IT operator to manually run a registered action after an alarm is triggered.
- *Log-file scanner* — enables IT operators to search and parse logs and registers for a particular status.
- *Domain details* — displays information about a domain's hardware. Only the boards and components in the specified domain are included.

Integration Features

Sun Management Center software publishes the SNMP MIBs for the modules. These MIBs can be used as the data definition by third-party SNMP-based management stations, such as CA Unicenter, IBM Tivoli, HP Openview, BMC, SMARTS InCharge Suite, etc.

Scalability and Customization Features

Sun Management Center software's common management platform is scalable from a single system to thousands of server and desktop systems with the flexibility to allow it to be easily configured to best fit the needs of the environment. Its extensible, autonomous agent architecture and rules-writing documentation enable IT operators to add functionality and management features, and easily create and customize rules.

The developer environment helps organizations to plan, design, develop, and integrate Sun Management Center modules as well as third-party applications, tools, and customized solutions. The development environment includes all of the tools needed to implement dynamically loaded modules, which are encapsulated sets of monitoring functions focusing on a particular aspect of system or application health and

performance. In particular, organizations can generate module configuration files, create and realize a data model, implement data acquisition, write data acquisition code, and define threshold rules with Tool Command Language (Tcl) procedures, shell scripts, or shared object libraries. Sun Management Center software also includes a GUI module builder that allows developers to quickly build new Sun Management Center modules or incorporate existing scripts into Sun Management Center software.

Security Features

Sun Management Center software includes enterprise-wide security measures, such as Java security classes, SNMPv2usec (SNMP version 2, user-based security model), authentication, data integrity, and access control lists for managing data and active management functions. Sun Management Center 3.6 software also supports SNMPv3, which enables Sun Management Center agents to communicate with third-party management applications securely. These security features help minimize the possibility of loss of services due to internal or external attack. In addition, community-based SNMP access to data in SNMP-based components can be disabled.

Ease of Use Features

Sun Management Center software includes many features designed to ease administration and improve productivity, including:

- *Dynamic agent modules* — allows asynchronous and dynamic insertion or removal of functionality.
- *Domain-aware agents* — allows independent monitoring of system domains and their associated resources.
- *Agent update* — automates the deployment of Sun Management Center agents and add-on packages and patches to managed systems. Agent Update provides great flexibility in distributing and fine-tuning agent software with minimal impact on other applications installed on managed systems. This allows Agent Update to operate on production systems with no downtime of mission-critical applications.
- *Automatic discovery* — automatically finds network resources for Sun Management Center software to monitor and manage. Discovery criteria can be based on several values including IP addresses, subnet addresses, host names, loaded modules, operating system, hardware type, or some combination of these values.
- *Manage jobs* — enables the IT operator to create persistent jobs for objects. This feature provides a way to manage a collection of agents (or groups) with the same ease as a single agent.
- *Topology import and export* — enables the topology database to be imported and exported from or to an ASCII file that uses eXtensible Markup Language (XML) markup. These features provide a convenient way to migrate management domains from one Sun Management Center server to another server or to backup information for a server.

Sun Management Center Web Console and Command-Line Interface

The Sun Management Center Web console is a Web-based management interface for the Sun Management Center platform. This host management system uses the secure HTTPS protocol to provide easy access to Sun Management Center management information. The Web console can be accessed across the firewall to monitor and manage Sun Management Center information from any location. It supports widely available browsers, such as Mozilla1.4 or higher, Netscape Navigator™ 4.5.1 or higher, and Internet Explorer 5.0 or higher. The Web console is an optional Sun Management Center component and depends on the Sun Management Center server for retrieving and manipulating managed objects. The Sun Management Center Web console provides many of the same host management functions as the Sun Management Center Java Console provides.

The Sun Management Center CLI is a character-driven console application for monitoring and managing the system. The CLI offers low overhead for use on simple data terminal equipment. Sun Management Center CLI provides a set of commands that are a subset of the functionality provided by the GUI console.

Solaris 10 Operating System Features

Sun Management Center software incorporates the following new features in the Solaris 10 OS:

- *Solaris Service Manager* — The Solaris 10 OS introduces the concept of a Service Management Facility (SMF) to simplify the management of system services. SMF creates a supported, unified model for services and service management on a Solaris system. SMF is an integral part of the Solaris Predictive Self Healing technology available in the Solaris 10 OS. It provides a mechanism for failed services to restart automatically in the dependency order, and allows the enabling and disabling of services with the changes persistent across patches and upgrades.
- *Fault Management Architecture* —The Solaris 10 OS provides a Fault Management Architecture (FMA) framework, which deals with historical OS issues of limited resilience to hardware faults, ad-hoc error reporting and handling, and dependence on human fault diagnosis. FMA provides a cohesive architecture for fault management, with pluggable diagnosis engines consuming the error event stream. The FMA framework also includes a mechanism to track dependencies between system components aimed at limiting fault impact. On Sun SPARC Enterprise servers, the FMA is also supported on the Service Processor, reporting to the OS through the DSCP network.
- *Solaris Dynamic Tracing (DTrace)* — An innovative feature introduced in the Solaris 10 OS, DTrace provides dynamic instrumentation and tracing for both applications and kernel activities. It enables developers to track down performance problems and locate the cause of aberrant behavior. This module launches DTrace scripts and manages the alarms generated by DTrace. DTrace enables knowledgeable IT

operators and developers to examine and rectify application bottlenecks to improve performance. DTrace can be especially useful when trying to pinpoint problems between components of an application service that are running on the same system.

Advanced System Monitoring

Advanced System Monitoring is a licensed value-added software. It provides additional modules that support more complete system monitoring capabilities. Advanced System Monitoring includes the following modules:

- *Directory Size Monitoring module* — provides the capability to isolate and monitor the size of any directory and its subdirectories on a system on which an agent is installed.
- *Fault Manager module* — handles hardware and software faults effectively. This module also displays a detailed fault report or a message article for the selected fault.
- *File Scanning module* — scans files on a system for user-specified patterns. Multiple instances of the File Scanning module can be loaded to scan multiple files.
- *Hardware Diagnostic Suite* — tests the system for hardware faults.
- *Health Monitor module* — monitors the health of the system. When alarm conditions occur, this module offers suggestions, if necessary, on how to improve the performance of the system.
- *Kernel Reader module* — monitors kernel statistics and all kernel information including CPU statistics, system load statistics, disk statistics, file system usage, and so on.
- *Process Monitoring module* — monitors statistics, such as percentage of CPU and virtual size, of processes that match user-defined criteria.
- *Service Management Facility module* — monitors and displays the services running on a system.

System Reliability Manager

System Reliability Manager is an add-on to the Sun Management Center software that includes additional capabilities to help proactively monitor and manage Sun servers and the Solaris Operating System.

System Reliability Manager consists of the following modules:

- *Patch Monitoring module* — uses the Solaris Patch Manager to identify and produce alarms for missing patches within the Solaris Operating System, helping to ensure optimal system performance through easier management and tracking of patches. It helps improve system reliability by notifying the IT operator of any un-installed, recommended, security, or general patches.
- *File Watch module* — monitors an enterprise-wide list of system or configuration files for record deletion, addition, or modification, and detects syntax errors, resulting in improved system reliability and availability. It proactively alerts IT operators about

missing system and configuration files, automatically monitors default system files, and allows IT operators to specify additional files to be monitored.

- *Script Repository and Launcher modules* — The Script Repository module provides a view of the scripts available to be run by the Script Launcher module. The Script Repository module also contains commonly used DTrace scripts for the Solaris 10 OS. The Script Launcher enables IT operators to manage shell and Java language programs that monitor Sun systems and applications. Scripts can be remotely launched and monitored, and their execution verified, from any Sun Management Center console. These modules help provide improved system reliability and security by reducing operator errors and offering simpler management procedures.
- *OS Crash Dump Analyzer module* — monitors and detects system crash dumps and alerts the IT operator. It presents information about corresponding system and `vmcore` files, helping to find root causes of failures and allowing IT operators to take corrective action to prevent future problems.
- *Installed Packages Audit module* — checks the integrity of the specified packages. It reduces the number of commands needed to collect audit information.

Performance Reporting Manager

Performance Reporting Manager is an add-on to Sun Management Center software that enables IT operators to create reports detailing the status of systems. Performance data is gathered by the history logging capability that is part of Sun Management Center software. Performance reports generated with Performance Reporting Manager use the data collected in this manner. The software includes 16 standard reports for which the report options are predefined. Custom reports can be created by selecting options from the provided report templates.

Performance Reporting Manager can automatically generate performance or configuration reports in a number of graphical formats on any data, including alarms, hardware configurations, performance, software packages and patches, and system uptime. Performance report data is especially useful for determining trends in performance.

Solaris Container Manager 3.6

Solaris Container Manager 3.6 is an add-on software to the Sun Management Center 3.6 release. It helps to consolidate servers to control the cost of large networks of servers and software. Solaris Container Manager enables the following tasks:

- Partition the resources of a system
- Allocate, control, and organize resources
- Isolate applications from each other
- Analyze resource consumption for specific applications
- Monitor resource usage and gather extended accounting data for CPU and memory usage

Solaris Container Manager enables the IT operator to create and manage containers, projects, resource pools, and zones, as well as organize existing resource management utilities that run on the Solaris OS. See “Solaris Containers” on page 40 for more information on containers, zones, and resource management.

If the Performance Reporting Manager add-on is installed with the Solaris Container Manager, the IT operator can create reports that provide historical resource utilization data per container, resource pool, zone, project, or system. CPU data, memory utilization data, and CPU extended accounting data are stored in the database by the Performance Reporting Manager data collection service. From the GUI, a graph report detailing resource usage can be requested, or the data can be exported to a text file in comma-separated value (CSV) format. The latter can be used in a billing and accounting application, for example.

Service Availability Manager

Service Availability Manager, a value-added package for Sun Management Center software, increases the availability of network services running locally or remotely on Sun systems. By continuously measuring and validating the delivered quality of service, this software simplifies the management of systems running the Solaris Operating System platform. To monitor services remotely, Service Availability Manager uses Synthetic Transaction modules. Synthetic (dummy) transactions simulate the use of the services. The synthetic transactions can be used to measure performance statistics such as domain name service (DNS) resolve time, the total time a transaction takes, or connect time. To monitor services on the local system, Service Availability Manager uses Service Element modules. A Service Element module sends service requests periodically according to specified configuration parameters. In this way, both service availability and response time can be determined.

Solaris Management Console

Solaris Management Console makes it easy for IT operators to configure and manage systems running the Solaris Operating System. Based on Java technology, the 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, projects, and jobs for managing file system mounts, disks, and serial ports. The console helps enable IT operators to easily view activity on all managed servers and modify applications and services running on them. The 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 Solaris Management Console enables users and IT operators to register other 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 Solaris Management Console helps improve productivity for IT departments and 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`.
- Administration tools that can be integrated and run from one location.
- Centralized management — all servers on a network can be managed.
- Single login into applications launched by the Solaris Management Console.
- Instant access to existing administration tools.
- Secure communication via support for https and SSL.

The 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

Configuration and Service Tracker 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 Configuration and Service Tracker, 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 (FRU) level.

In order to help IT operators better manage their systems, Configuration and Service Tracker probes the monitored systems and collects data in the following categories:

- System information — system model, host ID, host name, IP address, OS installed, number of CPUs, total memory, number of disks, etc.
- Installed hardware — system boards, controllers, devices (excluding external network and communication devices)
- Cluster configuration
- Software packages — Sun and third-party packages (those that are installed using the `pkgadd` facility)
- Software patches
- Pertinent information about the Solaris Operating System
- Volume manager configuration

By continuously gathering this data, Configuration and Service Tracker 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 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* — enables the IT operator to actively manage system resources through a Web-portal system dashboard
- *Hardware failure self-monitoring* — provides hardware failure alarms to the Web-portal and notification through pager and/or E-mail
- *Event alarming and notification* — provides system performance alarms to the Web-portal and notification through pager and/or E-mail
- *Web-view reporting*
  - Trend reporting — displays system performance trends over time
  - Configuration and patch reporting — displays system configuration and patch information
  - Availability reporting — provides high-level system availability percentage to assist in determining system management and maintenance effectiveness

Sun℠ Update Connection

This Web application is hosted at Sun and enables IT operators to manage updates remotely on one or more Solaris systems. The Sun Update Connection services enable 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

Sun. 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 un-installs an update. The Web application can also be used to view the update status of systems and jobs.

Sun™ Update Manager

The Sun Update Manager GUI and the `smpatch` command-line interface enables IT operators to manage updates (patches) locally on Solaris systems. The Sun Update Connection, System Edition software has the same functionality as the Sun Patch Manager tools, with the addition of some new features and enhancements.

The Sun Update Manager software is one part of the Sun Update Connection, System Edition 1.0 software that enables IT operators to locally manage updates on systems. Update Manager offers a graphical user interface 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 the Sun Update Connection services can still be managed locally by using Sun Update Manager.

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

Sun Update Manager enables an update management process that includes the following tasks:

- 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 the Sun Update Connection services to remotely manage the system

Solaris Operating System—Support for Manageability

With over 6 million registered licences worldwide and more than 600 innovative features such as DTrace, Predictive Self Healing, Solaris Containers, and Solaris ZFS (zettabyte file system), the open source Solaris 10 Operating System is the most advanced operating system available. These technologies supply the foundation for building, deploying, and managing efficient, secure, and reliable enterprise-class service-oriented architectures (SOAs) for today's demanding business processes.

In addition to providing a world-class platform for deploying applications, the Solaris Operating System includes sophisticated services to improve manageability and

resource utilization. Many of the most difficult problems occur as a result of changes to the server operating system configuration. The Solaris Operating System offers unique tools, such as JumpStart™, Solaris Live Upgrade, and Solaris Flash Archives to help automate tasks, thus improving consistency. The Solaris OS also features Solaris ZFS to more effectively manage file systems.

JumpStart

JumpStart automates the process of installing or upgrading multiple systems based on custom profiles, as well as optional pre-installation and post-installation scripts.
A profile is a text file that defines how to install the Solaris software on a system.
A profile and a rules file must be created for each group of systems to be installed. The 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 that are based on one or more system attributes. Each rule also links each group to a profile.

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

Solaris Live Upgrade

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, then upgrade or patch the duplicate. Or, rather than upgrading, a Solaris 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 Solaris Live Upgrade, an upgrade or new patches can be tested without affecting the current operating system. 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 2.

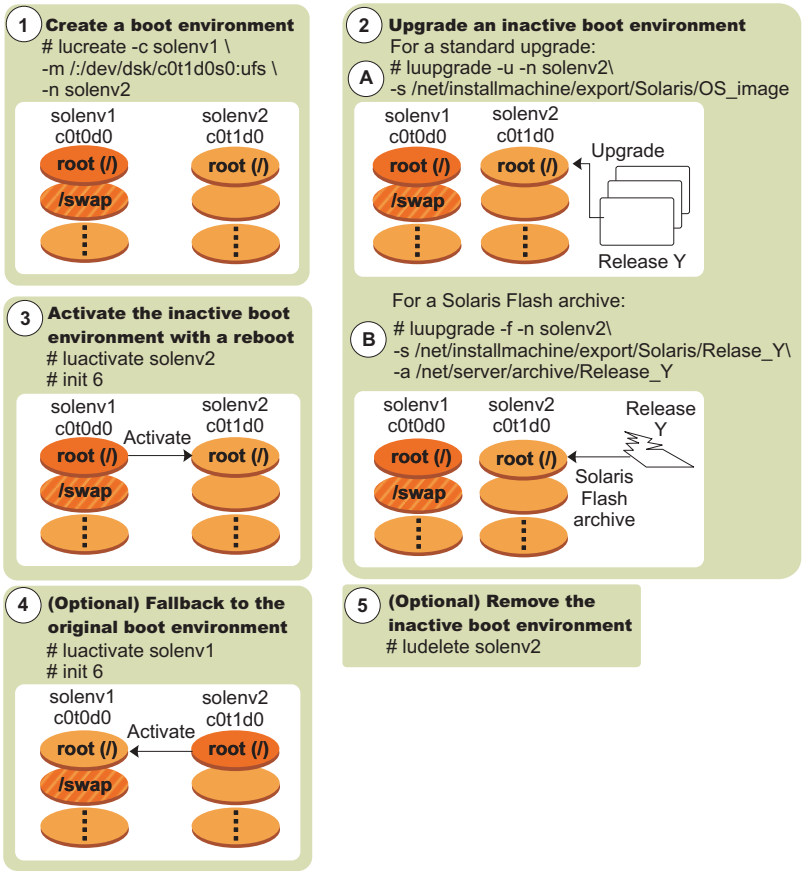

Figure 2. Solaris Live Upgrade process

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

- 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.

Solaris Flash Archives

The Solaris Flash installation feature enables IT operators to use a single reference installation of the Solaris software on a system. This reference installation is called a Solaris Flash archive. The 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. Solaris Flash archives save installation time by installing all Solaris

packages at one time. Other programs install each individual Solaris package and update the package map for each package.

A Solaris 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 JumpStart installation method can be used to install a differential archive on a clone system. Or, 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 Solaris systems.

- *PatchPro analysis engine* — 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 command-line interface* — enables IT operators to use `smpatch` to apply patches while the system is in single-user mode.
- *Patch list operations* — enables 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.

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

Service Management Facility

Included in the Solaris 10 OS, Service Management Facility provides an infrastructure that augments the traditional UNIX® start-up 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 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. 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.

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 un-correctable 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 non-root 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.

Solaris ZFS

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 de-couples the file system from physical storage in the same way that virtual memory abstracts the address space from physical memory, as illustrated in Figure 3, allowing for much more efficient use of the storage devices.

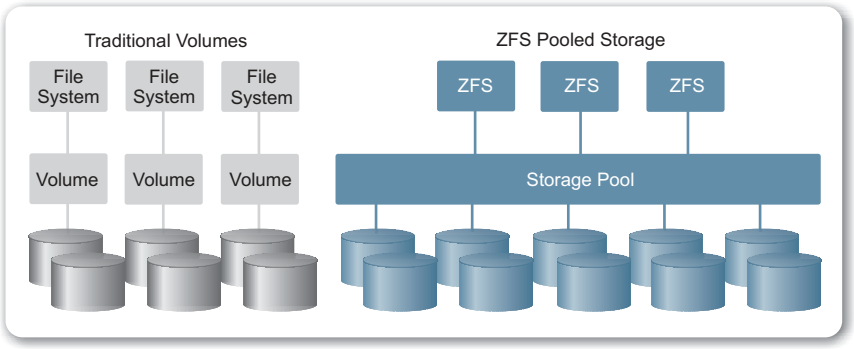

Figure 3. Solaris ZFS file system

In Solaris 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, Solaris ZFS is a 128- bit file system, so its theoretical limits are virtually unlimited — 2^{128} bytes of storage and 2^{64} for everything else, such as the number of file systems, snapshots, directory entries, devices, etc.

Other benefits and features of Solaris ZFS include:

- *Data integrity* — Solaris ZFS combines proven and cutting edge technologies like copy-on-write and end-to-end 64- bit checksumming, providing extreme reliability to ensure that the data on the disk is self-consistent at all times.
- *Improves performance* — Solaris 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, Solaris ZFS implements intelligent prefetch, performing read ahead (in either direction) for sequential data streaming, and can adapt its read behavior on-the-fly for more complex access patterns. Solaris 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 Solaris ZFS architecture introduces the integration of volume management functions.
- *Simplifies management* — Solaris 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.
- *Low overhead RAID* — Solaris 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.

The Solaris 10 OS integrates a Fault Manager diagnostic engine for Solaris 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.

Chapter 3
Managing Resources and Domains

The advent of delivering computing services over the Internet brings with it an emphasis for IT operations staff to find ways to economically provide needed levels of service such as throughput, availability, and response time. 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 a SLA is to over-provision resources. While 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 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 coupled with tools to help manage service levels and resource utilization. Sun offers powerful features to help control resource utilization 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 the Solaris OS. 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, while another domain can be used for production purposes.

Each domain uses a separate boot disk with its own instance of the Solaris Operating System, as well as I/O interfaces to network and disk resources. CPU/memory boards and I/O boards can be separately added and removed from running domains using DR, providing 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 via 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 of a service on a single platform (such as the Web, application, and database components of an application service), as illustrated in Figure 4. Each domain is completely protected from hardware or software faults originating in other domains.

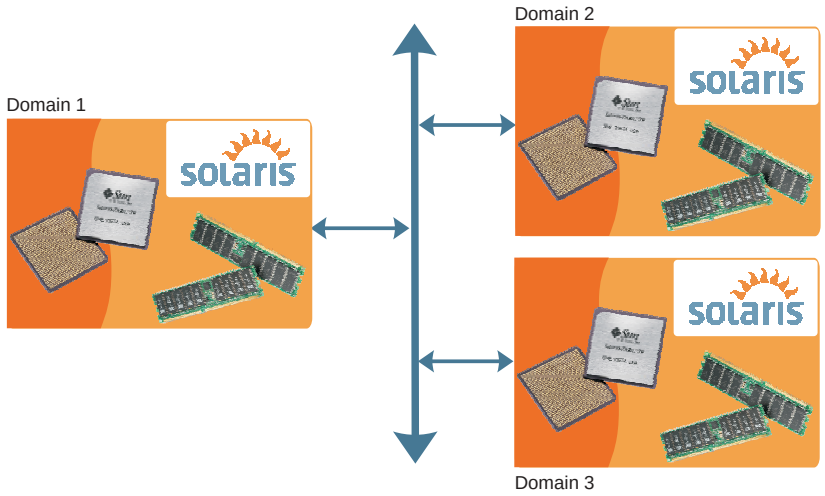

Figure 4. Consolidating applications into domains

PSB — A physical system board consisting of up to four CPUs, 32 memory modules, and I/O.

How Domains Work

Sun SPARC Enterprise servers have a unique partitioning feature that can divide one physical system board (PSB) into one logical board or four logical boards. The number of physical system boards in the Sun SPARC Enterprise servers vary from 1 to 16, depending on the model. In the high-end systems, one PSB consists of 4 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 M5000 servers support up to two and four CPU modules respectively. Each memory board in the systems contain 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 (IOU) contains five PCI slots. The Sun SPARC Enterprise M4000 server supports one IOU and the Sun SPARC Enterprise M5000 server supports up to two IOUs.

XSB — hardware resources on boards are logically divided and reconfigured as eXtended System Boards.

In order 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.

Uni-XSB — a PSB that is not divided and is configured into one XSB. It contains all of the resources on the board.

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 i.e., 4 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. 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. Uni-XSBs for the midrange and high-end systems are illustrated in Figure 5 and Figure 6. 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.

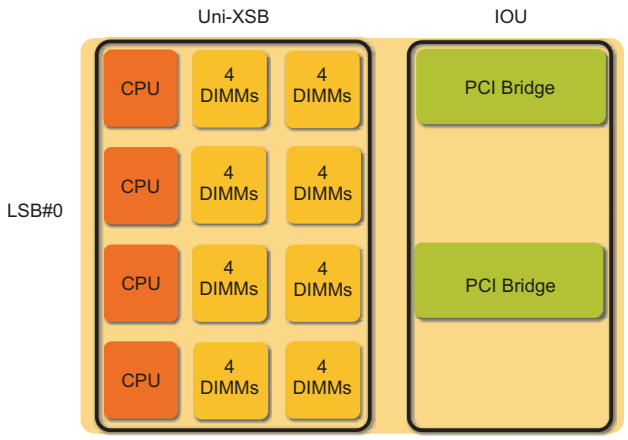

Figure 5. Uni-XSB on midrange system

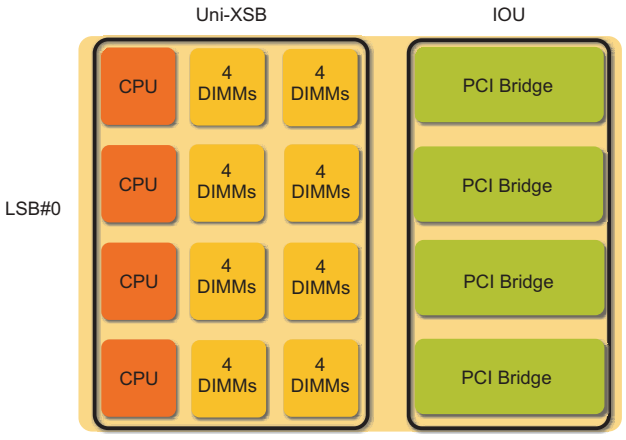

Figure 6. Uni-XSB on high-end system

Quad-XSB — a PSB that is logically divided and configured into four XSBs. Each XBS contains 1 CPU, 8 DIMMs, and I/O.

A Quad-XSB is a PSB that is logically divided and configured into four XSBs. Each of the four XSBs contain one-quarter of the total board resources, e.g., on high-end systems, 1 CPU, 8 DIMMs, and 2 PCIe cards (1/4 of the I/O unit or IOU), as illustrated in Figure 7.

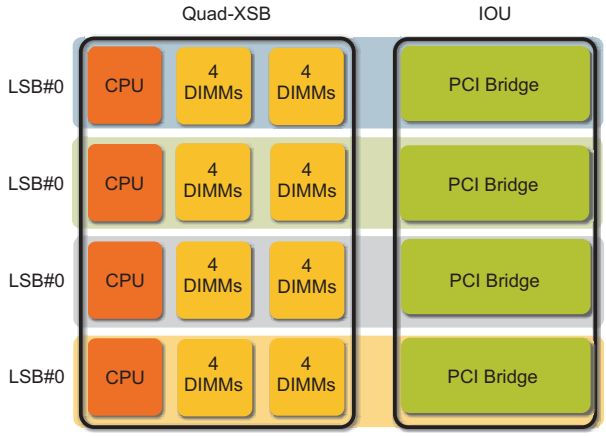

Figure 7. Quad-XSB on high-end systems

On midrange servers, shown in Figure 8, only two XSBs have I/O. Quad-XSBs enable sub-board domain granularity, and therefore, better resource utilization. However, if a board fails, so do the domains it is assigned to. Memory mirror mode can be enabled for Quad-XSBs in the midrange systems only.

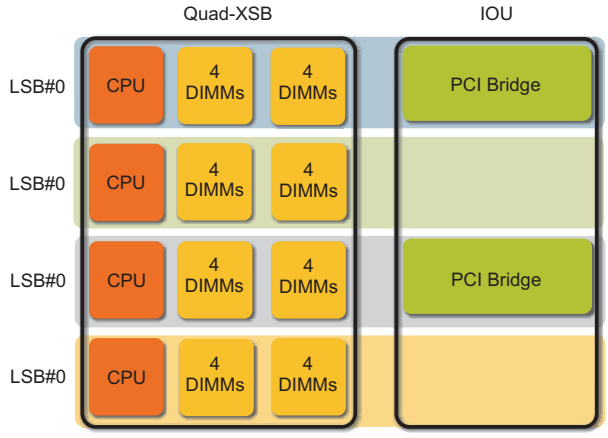

Figure 8. Quad-XSB on midrange systems

A domain consists of one or more XSBs. Each domain runs its own copy of the Solaris OS and must have a minimum of 1 CPU, 8 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- and Quad-XSBs.

A domain component list (DCL) identifies the *potential* resources for a domain. A single XSB can potentially belong to multiple domains. However, a single XSB can be *assigned*

LSB — logical system board mapping to XSB number for domain configuration purposes.

to only one specific domain. The domain configuration software maps each XSB number to a logical system board (LSB) number. Figure 9 illustrates a configuration of four domains utilizing one Uni-XSB and two Quad-XSBs. Note that each domain requires I/O, but that it is not a requirement for each LSB.

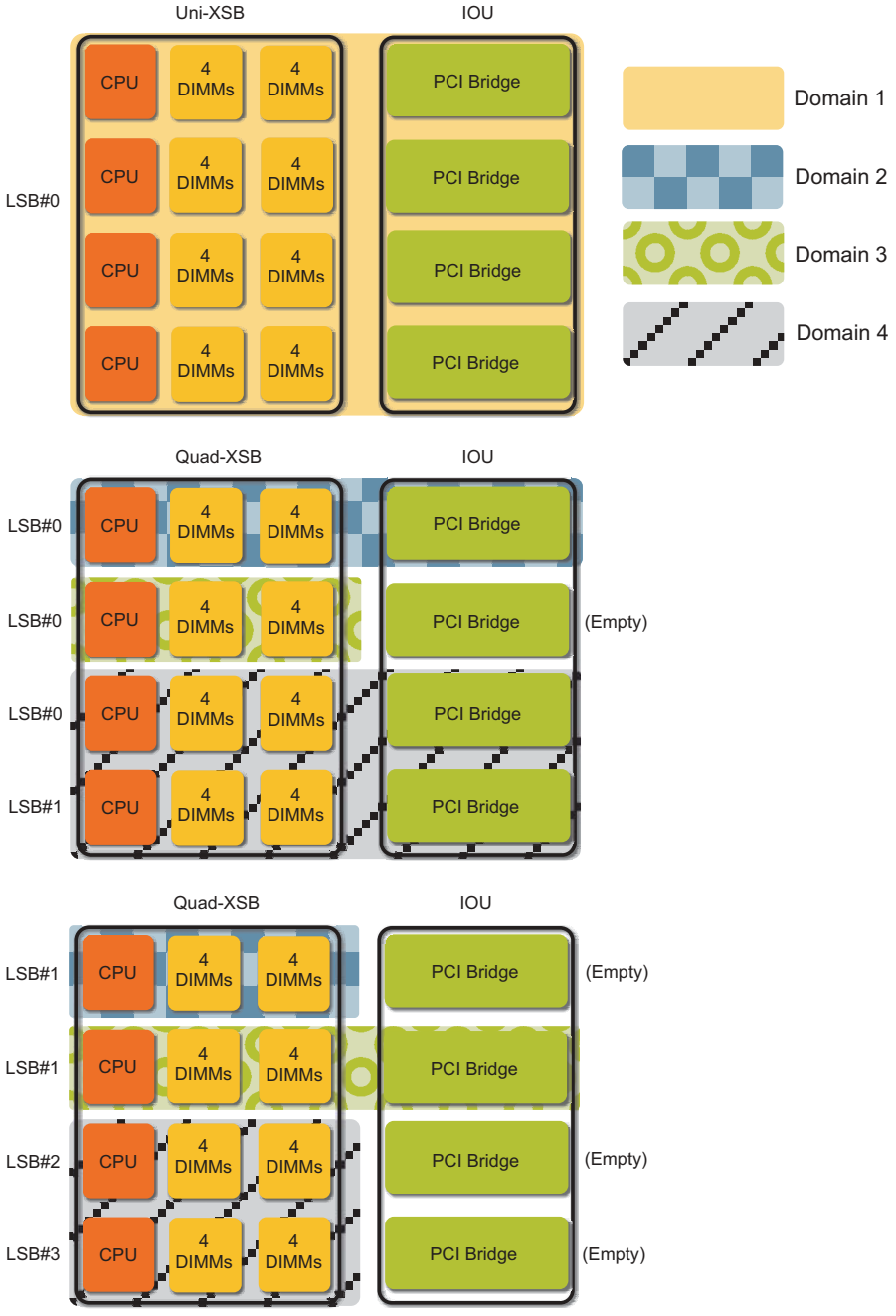

Figure 9. Domain configuration example using a Uni-XSB and Quad-XSBs

The Solaris OS 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 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 domainstop operation occurs that cleanly and quickly shuts down only the domain with the error. Domainstop operates by shutting down the paths in and out of the system address controller and the system data interface application-specific integrated circuits (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 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* — diagnoses hardware errors associated with domain stops
- *Solaris Operating System diagnosis engine* — identifies non-fatal domain hardware errors and reports them to the system controller
- *POST diagnosis engine* — identifies any hardware test failures that occur when the power-on self-test is run

In most situations, hardware failures that cause a domain crash are detected and eliminated from the domain configuration either by power-on self test (POST) or 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:

- *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 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 the Solaris Operating System, 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 further managed using Solaris Resource Manager software 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 the Solaris Operating System 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 operating system updates.

Dynamic Reconfiguration

Dynamic Reconfiguration (DR) and Automated Dynamic Reconfiguration (ADR) allow resources to be dynamically reallocated, or balanced, between domains by enabling a physical or logical restructuring of the hardware components of 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 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 that have the proper access privileges. Three types of system board components can be added or deleted by DR: CPU, memory, and I/O devices.

Dynamic Reconfiguration allows the domain or platform administrator to perform the following functions:

- Display the DCL and domain status
- Display the status and state of system or I/O boards and some components to help prepare for DR operations, including whether the board is COD or not
- 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 removal from the system while the domain remains running
- Add system or I/O boards to a domain to add resources or replace a removed board, while the domain remains running
- Configure or unconfigure CPU or memory modules on system boards to control power and capacity of a domain or isolate faulty components
- Enable or disable peripheral component interconnect (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, 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 multi-pathing solutions such as "IP Multipathing" on page 50 and the "Solaris Fibre Channel (FC) and Storage Multipathing Software" on page 51, IT operators can manage networks or storage subsystems that have redundant access paths with automatic failover, load balancing, and dynamic reconfiguring capabilities.

Basic DR 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: add, delete, move, and replace.

- *Add* — DR can be used to add a system board to a domain without stopping the Solaris OS provided that 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 Solaris OS 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 Solaris OS 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 the Solaris OS. 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 Solaris OS 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 Solaris OS running on that domain provided both boards satisfy DR requirements (such as not making up an entire domain, no processes are running on the CPU, etc.). 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 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 (MBU). 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 10, 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.

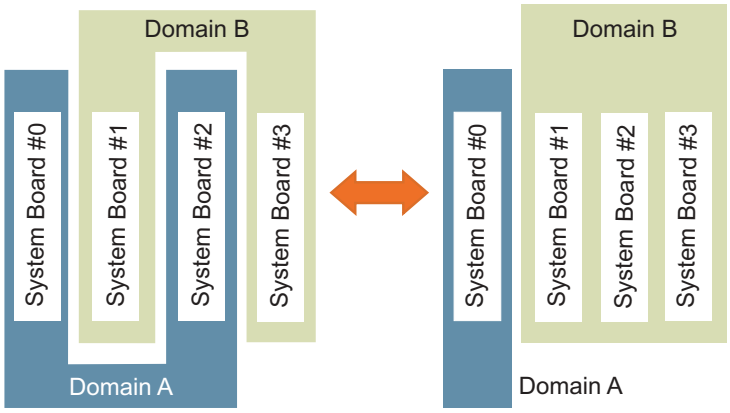

Figure 10. Reconfiguration example

Using DR to Change a CPU Configuration

Upon adding a CPU, it is automatically recognized by the Solaris OS 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 pools facility is in use by the domain, the CPU to be deleted must not belong to a resource pool (see “Dynamic Resource Pools” on page 44 for more information on resource pools and processor sets)

Using DR to Change Memory Configurations

The DR functions classify system boards by memory usage into two types: kernel memory board and user memory board. A kernel memory board is a system board on which kernel memory (i.e., memory internally used by the Solaris OS 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, and 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 the Solaris 10 OS. 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 DR 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 since hot-plug is controlled entirely within the Solaris OS, XSCF is not aware of hot-plug events, including I/O Box hot-plug events (see “I/O Box” on page 39). The Solaris Service Manager 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-XSB 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 11, domain #1 has a configuration that requires it to be stopped before the system board can be replaced.

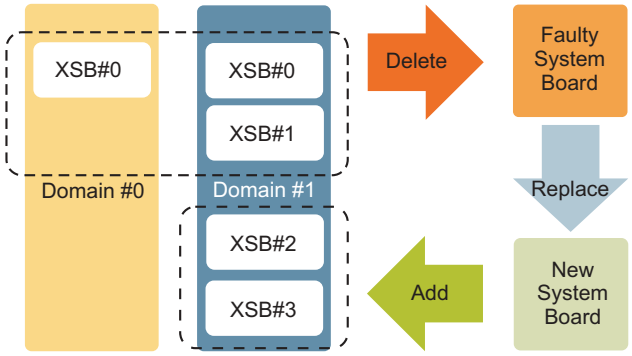

Figure 11. Domain 1 must be stopped before the PSB can be replaced

System Board Pooling

The system board pooling functions assigns a specific system board in a status where that 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 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 the domain is rebooted. If a system board used as a floating board (see“Floating Board” on page 35) 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.

XSCF Settings

The DR functions provide some options to avoid the complexities of reconfiguration and memory allocation with the Solaris OS, and 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

where a hardware error (CPU and memory) is detected, the system board (XSB) where the component is mounted, or 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 non-floating board.

Kernel memory is allocated to the non-floating 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 non-floating 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 might lower system performance.

Automatic Dynamic Reconfiguration

Automatic DR 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 and the system event facility.
The Reconfiguration Coordination Manager executes preparatory tasks before a DR operation, error recovery during a DR operation, and cleanup after a DR operation.
The automatic DR 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 automatic DR functions are building blocks for a feature called global automatic DR. Global automatic DR introduces a framework that can be used to automatically redistribute the system board resources on a Sun SPARC Enterprise system. This redistribution can be based upon factors such as production schedule, domain resource utilizations, domain functional priorities, and so on. Global automatic DR accepts input describing resource utilization policies and then uses those policies to automatically marshal the Sun SPARC Enterprise system resources to produce the most effective utilization.

Reconfiguration Coordination Manager

Reconfiguration Coordination Manager is designed to provide a public application program interfaces (API) for Dynamic Reconfiguration removal events, helping programs such as databases and system management tools to take predetermined actions when hardware configuration or operating system 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 Sun SPARC Enterprise servers, RCM scripts cannot be used for DR add operations.

The reconfiguration coordination manager is a daemon process that coordinates DR operations on resources that are present in the domain. The Reconfiguration Coordination Manager daemon uses generic APIs to coordinate DR operations between DR initiators and Reconfiguration Coordination Manager clients. The Reconfiguration Coordination Manager 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 Sun Management Center software. The DR clients can be:

- Software layers that export high-level resources comprised of one or more hardware devices (for example, multipathing applications)
- Applications that monitor DR operations (for example, Sun Management Center software)
- Entities on a remote systems, such as the system controller on a server

In the Solaris 10 OS, Reconfiguration Coordination Manager is supported in the Solaris Volume Manager. Reconfiguration Coordination Manager support gives Solaris Volume Manager the ability to respond appropriately to DR requests. This helps ensure that

removal of devices under Solaris Volume Manager 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 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 licensing mechanism. Under certain conditions, COD boards can be used before actually purchasing a license.

When a system encounters a resource constraint, additional capacity can be quickly enabled by purchasing a right-to-use (RTU) license. 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, and reduces upgrades that take valuable time to execute.

COD 2.0 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 Right to Use (RTU) licenses. COD 2.0 provides a pay-as-you-use model, and is designed for organizations that need to scale quickly with extra capacity to meet future and continuous increases in demand. COD licenses are conveniently tracked and managed through XSCF or Sun Management Center software.

COD 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 they are licensed. Once a COD board license is purchased, the board can be configured into domains in the same way as a system board (including Quad-XSB), and fully support DR operations.

COD Licenses

A COD license is assigned to a specific server, one license per CPU. A maximum of 50 license keys can be installed on a Sun SPARC Enterprise server. A COD license has no expiration date. All of the licenses assigned to a server are handled as a floating pool of licenses 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 licenses are required. Those six licenses can be used by all eight processors, but only six at a time.

COD uses DR to add and remove COD resources so 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 licenses for those boards are released and added to the pool of available licenses. All licenses remain allocated to their resources during a Service Processor reboot or failover.

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

Headroom Management

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

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

Temporary Capacity on Demand

Temporary COD (T-COD) also provides access to additional system capacity and the ability to activate resources when needed. In this case, however, temporary RTU

licenses are purchased and are valid for one month. The resources can be deactivated when they are no longer needed. Organizations that purchase T-COD RTU licenses for six consecutive months receive a permanent RTU license. Like COD 2.0, T-COD is designed to minimize disruptions to operations.

T-COD provides a pay-per-use model, making it ideal for organizations that sustain periods of peak workload demand followed by periods of lower workload demands. COD 2.0 and T-COD are designed to provide key business advantages such as improved availability, increased system utilization, and reduced hardware acquisition costs:

- Higher utilization of resources — COD 2.0 allows additional resources to be activated in single CPU increments, which helps increase system utilization by better matching system usage with asset acquisition. This enables IT organizations to instantly scale to meet peak demands, without having to pay for the entire amount until the extra capacity is needed. COD 2.0 also helps reduce additional third-party software charges that are applied to active CPUs. By utilizing only the specific number of CPU resources needed, it is possible to minimize hardware and software acquisition and maintenance costs.
- Reduced cost — COD 2.0 and T-COD enable organizations to purchase larger system configurations than are currently required at lower up-front costs, allowing them to be prepared for growth without having to pay in advance. There is no premium price for COD 2.0 or T-COD options over standard options, and there is no time requirement to purchase RTU licenses. For data center environments with unpredictable growth or sudden increases in peak usage, COD 2.0 and T-COD provide a cost-effective method for adding resources.

I/O Box

The Sun SPARC Enterprise External I/O Expansion Unit (I/O Box) provides additional slots for PCI cards. A single I/O boat within the I/O Box provides 6 additional slots. A two I/O boat configuration provides 12 slots. The I/O Box includes two power supplies and fans for redundancy.

The I/O Box supports two types of I/O boat, PCI-X and PCI Express (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 PCI Express 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 PCI Express 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 Solaris domain DR commands and XSCF, which detects and enables PCI cards and I/O BOX units.

Chapter 4

Fine-Grained Resource Management

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

While Dynamic Domains allow consolidation of several servers into one Sun SPARC Enterprise server, Solaris Containers allows consolidation of several applications into one domain. For example, a single Sun SPARC Enterprise server, partitioned into domains, each running an instance of the Solaris Operating System can support application, file, and print services for heterogeneous clients, messaging/mail, as well as Web services, application services, and mission-critical databases for entire application services. Since the Sun SPARC Enterprise M9000-64 server can scale to 64 processors with 128 cores, one server can easily be shared by a number of applications or application services. And, by combining application workloads with different utilization 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 operating system instance and providing cost savings in administrative tasks such as data management and archive.

Solaris Containers can be also 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, while other system users have more consistent workloads. Using Solaris Containers, traders can be granted a proportionately larger number of shares of resources to give them the system resources they require.

Solaris Containers

A primary objective of the Solaris 10 Operating System is to deliver tools to help IT departments to more with less by consolidating applications onto fewer servers. The Solaris Containers functionality in the Solaris 10 OS 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 utilization. Combined with Solaris Predictive Self Healing, Solaris Process Rights Management, and DTrace, the capabilities of the Solaris 10 OS can help consolidate applications without compromising the service

levels, privacy, or security of individual applications or users. Solaris Containers enable 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 Solaris Predictive Self Healing and Solaris Containers

A Solaris Container is a complete runtime environment for applications. Solaris 10 Resource Manager and Solaris Zones software partitioning technology are both parts of the container. These components address different qualities the container can deliver and work together to create a complete container. The zones portion of the container provides a virtual mapping from the application to the platform resources. Zones allow application components to be isolated from one another even though the zones share a single instance of the Solaris Operating System. The Solaris Resource Manager provides the capability to allocate the quantity of resources that a workload in a zone receives.

Containers establish boundaries for resources, such as CPUs, and can be expanded to adapt to the changing processing requirements of the application or applications running in the container. A container is a virtualized operating system environment created within a single instance of the Solaris Operating System. 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 can not 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 utilization of the system. Dynamic resource reallocation permits 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 under-utilized 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.

Solaris Containers are designed to provide fine-grained control over the resources that applications use, allowing multiple applications to operate on a single server while maintaining specified Quality-of-Service (QoS) levels. Fixed resources such as processors and memory can be partitioned into pools on multi-processor systems, with different pools shared by different projects (a specified collection of processes) and isolated application environments. Dynamic resource sharing allows different projects to be

assigned different ratios of system resources. Solaris IP Quality-of-Service (IPQoS) can be employed to manage network bandwidth used by multiple, competing network applications and is covered in more detail in “IPQoS” on page 50. 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, as illustrated in Figure 12, helping to reduce operational and administrative costs while increasing availability.

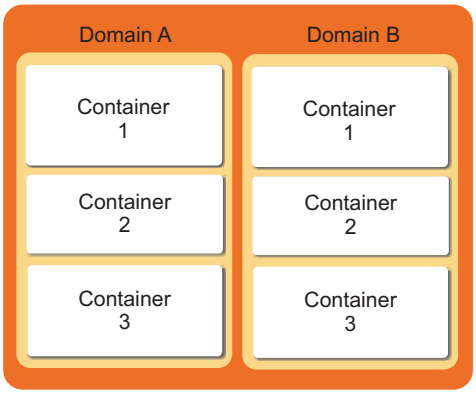

Figure 12. Example of containers in a domain

In the Solaris 9 and Solaris 10 releases, resource management is provided by the Solaris Resource Manager. Every service is represented by a project, which provides a network-wide 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 Solaris kernel tracks resource usage through the project ID. This relationship is depicted in Figure 13. Historical data can be gathered by using extended accounting.

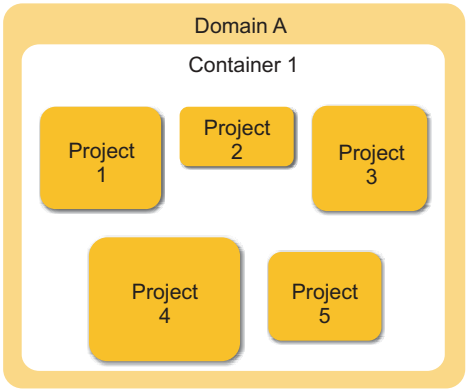

Figure 13. Example of projects in a container

Solaris Zones

The Solaris Zones software partitioning technology, a component of the Solaris Containers environment, is a software partitioning technology used to virtualize operating system services and provide an isolated and secure environment for running applications. A zone is a virtualized operating system environment created within a single instance of the Solaris Operating System. Zones basically provide the standard Solaris interfaces and application environment. Each zone can provide a customized set of services.

A zone can be thought of as a box. One or more applications can run in this box without affecting the rest of the system. This isolation prevents processes that are running in one zone from monitoring or interfering with processes that are running in other zones. Even a process with superuser credentials that is running inside a zone cannot view or affect activity in other zones.

The single instance of the Solaris Operating System running on the system or on a domain is the global zone. The global zone is both the default zone for the system or domain and the zone used for system- or domain-wide administrative control. One or more non-global zones can be created by an IT operator working in the global zone. Once created, these non-global zones can be administered by individual zone administrators. The privileges of a zone administrator are confined to a non-global zone.

Non-global zones provide isolation at almost any level of granularity required. A zone does not need a dedicated CPU, a physical device, or a portion of physical memory. These resources can either be multiplexed across several zones running within a single domain or system, or allocated on a per-zone basis using the resource management features available in the operating system. Zones can be booted and rebooted without affecting other zones on the system.

Each zone can provide a customized set of services. To enforce basic process isolation, a process can see or signal only those processes that exist in the same zone. Basic communication between zones is provided by giving each zone at least one logical network interface. Applications running in different zones on the same system can bind to the same network port by using the distinct IP addresses associated with each zone or by using the wildcard address. An application running in one zone cannot observe the network traffic of another zone. This isolation is maintained even though the respective streams of packets travel through the same physical interface.

Each zone that requires network connectivity is configured with one or more dedicated IP address. A process assigned to a zone can manipulate, monitor, and directly communicate with other processes that are assigned to the same zone. The process can not perform these functions with processes that are assigned to other zones in the system or with processes that are not assigned to a zone. Processes that are assigned

to different zones are only able to communicate through network APIs (application programming interfaces).

For more information on consolidating applications using Solaris Containers see: *http:/www.sun.com/software/solaris/reference_resources.jsp*.

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. The Solaris OS provides a facility called projects to name workloads once they are identified. For example, one project for a sales database and another project for a marketing database. If resource management features are not used, the Solaris Operating System 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.

Solaris Resource Manager enable 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
- Deny resources or prefer one application over another for a larger set of allocations than otherwise permitted
- Prevent an application from consuming resources indiscriminately
- Change an application's priority based on external events
- Balance resource guarantees to a set of applications against the goal of maximizing system utilization

These capabilities enable Solaris Containers to deliver predictable service levels. Effective resource management is enabled in the Solaris OS 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 enable 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 14. 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 the system performance goals. The software

periodically examines the load on the system and determines whether intervention is require to enable the system to maintain optimal performance.

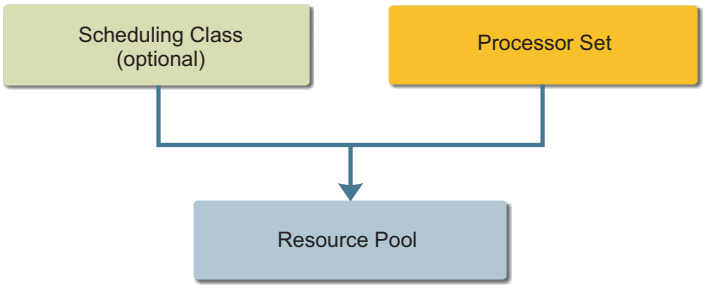

Figure 14. Resource pools

Resource Management Control

The Solaris Operating System provides three types of control mechanisms to control resource usage:

- *Constraint* — 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, that can otherwise compromise performance or availability through unregulated resource requests.
- *Scheduling* — 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 utilization of an under-committed 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* — 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 are always available to the workload. Resource pools are a partitioning mechanism that limit workloads to a specific subset of the resources of the system. Partitioning can be used to avoid system-wide over commitment. However, in avoiding this over commitment, the ability to achieve high utilizations 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 version 2.6 of the Solaris OS. 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, providing that at least one CPU remains for the system processor set.

Resource pools enable 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 Sharing Scheduler.

Fair Share Scheduler (FSS)

The fair share scheduler allocates CPU resources 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 the project A is twice as important as the 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 fair share scheduler 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, i.e., projects with no active processes, are not used in the calculations. An important thing to note is that the fair share scheduler 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.

FSS 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 15.

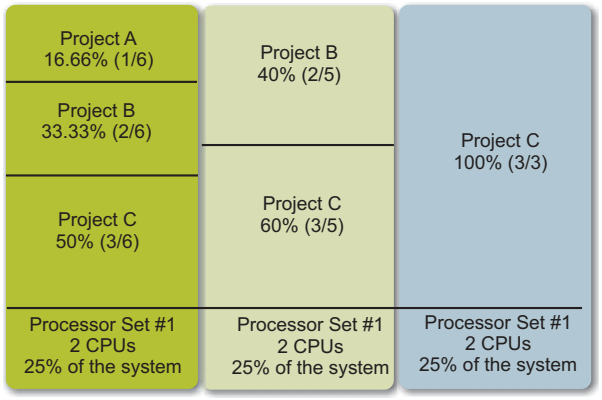

Figure 15. Allocating shares in processor sets

Resource Pools and DR Operations

Dynamic Reconfiguration allows 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 can not 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 using the DR force option. The pools configuration is then modified to comply with the new resource configuration. For information on the DR process and the force option, see the Dynamic Reconfiguration user guide for the Sun hardware that the pools are running on.

Using DTrace with Containers

DTrace is zone aware, making it especially useful for troubleshooting problems between applications running in Solaris Containers. For example, on operating systems other than the Solaris 10 OS, it is very difficult to find and rectify performance issues between different systems — such as a Web server, application server, and database server all running on separate systems. With the Solaris 10 OS and Solaris 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 pin-point performance issues or transient problems occurring between applications that are extremely difficult, time-consuming, and expensive to uncover in other operating systems. In addition, DTrace can be used to ascertain the level of resources an application uses, helping to more finely tune resource allocation on a zone or project level.

Chapter 5

Managing Other Resources, Monitoring, and Accounting

Resource management helps ensure that 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 utilization. By categorizing and prioritizing usage, reserve capacity can be used during off-peak periods, often eliminating the need for additional processing power. The Solaris 10 Operating System 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.

Solaris Process Rights Management Software

Solaris Process Rights Management software, introduced in the Solaris 10 Operating System, 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 otherwise damage data. In Solaris Containers, Solaris Process Rights Management software helps ensure that applications — even those run with privileges — are constrained to access resources only in their own Solaris Containers.

Resource Capping Daemon

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

The resource cap daemon repeatedly samples the resource utilization of projects that have physical memory caps. When the system's physical memory utilization 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 daemon `rcapd` regulates the consumption of physical memory by processes that run in projects that have defined resource caps. Associated utilities provide mechanisms for administering the daemon and reporting related statistics.

IPQoS

With the Solaris 9 OS 9/02 release, Sun introduced new network resource management technology, superseding the Solaris Bandwidth Manager software available for previous Solaris OS releases. This technology is usually referred to as IP quality of service (QoS), or simply IPQoS. IPQoS is ideal for controlling, monitoring, and accounting of network resources. IPQoS forms the foundation of managing the network resource aspects of Solaris Containers with features that can help make network performance more efficient. It is implemented at the IP level of the TCP/IP protocol stack and is configured for the global zone unless non-global zone traffic is routed outside of the system. IPQoS can be a central means to offer service-level agreements, enabling 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 data center 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. For more information on IPQoS, see *System Administration Guide: IP Services*.

IP Multipathing

IP Network Multipathing (IPMP) is a feature in the Solaris Operating System that enables IP fail-over and IP link aggregation. It helps manage network workloads and failures on the Sun servers in the following ways:

- *Outbound load spreading* — outbound network packets are spread 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* — the ability to detect a network adapter failure and automatically switch (fail over) its network access to an alternate network adapter.
- *Repair detection* — the ability to detect repair or replacement of a previously failed network and 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 Solaris Zones

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 can not access the on-disk data of another workload running in a different zone.

Solaris Zones allow 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 (`lofs`), which allows a directory and its contents to be inserted into another part of the file system. `lofs` is improved in the Solaris 10 OS to support read-only mounts, preventing non-global 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 allows for 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 StorageTek™ QFS software (multi-reader/writer capability).

Using Solaris ZFS in Solaris Containers

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 allows the non-global zone to share space with the global zone, though 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.

Solaris ZFS also allows data sets to be delegated to a non-global 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 non-global 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.

Solaris Fibre Channel (FC) and Storage Multipathing Software

The Solaris FC and Storage Multipathing software is integrated in the Solaris 10 OS. In Solaris 10 OS, fabric-connected devices are configured and made available to the system automatically during install and boot time. The Solaris FC and Storage Multipathing software provides the following key features:

- *Dynamic storage discovery* — 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* — devices maintain their device naming through reboots and/or reconfiguration. The only exception to this are tape devices, found in `/dev/rmt`, which are not changed unless they are removed and later regenerated.
- *Fabric booting* — Sun supported host bus adapters (HBAs) can boot from fabric devices as well as non-fabric devices.
- *Path management* — 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 allows systems configured with the 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, the Solaris FC and Storage Multipathing software is fully integrated with the Solaris 10 OS, allowing the 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, since it enables utilities, such as `format`(1M) or higher level applications such as the 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. The 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 fail over support, the Solaris FC and Storage Multipathing software can use any active paths 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. The 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* — supports the Solaris 10 OS Dynamic Reconfiguration feature.

Monitoring and Accounting

When running many applications on one system it is important to constantly monitor the system, 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.

Solaris Process Accounting and Statistics

System accounting software in the Solaris OS 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
- Located 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, process accounting, disk accounting, and fee calculations.

- *Connect accounting* — helps determine the length of time a user is logged in, number of reboots on the system, creation and termination of user process, etc.
- *Process accounting* — 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* — 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* — the chargefee utility stores charges for special services that are provided to a user. For example, a special service is file restoration.

Extended Accounting

The extended accounting subsystem `acctadm`(1M) within the Solaris OS 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 non-global 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, charge back, and billing purposes. With extended accounting data available, it is possible to develop or purchase software for resource charge back, 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 utilization 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 utilized. 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. For more information on flow accounting see, *Using Flow Accounting and Statistics Gathering (Tasks) in System Administrator Guide: IP Services.*

Monitoring Capped Memory

IT operators can use `rcapstat` to monitor the resource utilization of projects that are configured with memory caps. The `rcapstat` report includes the project ID of the capped project, the project name, virtual memory size of all processes in the project, total resident set size (RSS) of the project's processes, the RSS cap for the project, etc.

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 may 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 LWPs (light weight processes) appropriate for that application.

Chapter 6

Conclusion

The Sun SPARC Enterprise servers are the most powerful and innovative enterprise-class systems available from Sun today. With the ability to partition the system into sub-board 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 utilizing resources.

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

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 to be able to quickly adapt to changes in demand and business processes. By consolidating applications onto Sun SPARC Enterprise servers, IT departments can do all of this, and more.

References

Sun Microsystems posts product information in the form of data sheets, specifications, and white papers on its Web site page at: http://www.sun.com/.

And on http://www.docs.sun.com:

- *System Administration Guide: Advanced Administration, Part No. 817-0403*
- *System Administration Guide: Basic Administration, Part No. 817-1985*
- *Sun SPARC Enterprise M4000/M5000/M8000/M9000 Servers SCP Administration Guide for eXtended System Control Facility (XSCF), Part No. 819-3601*
- **Sun SPARC Enterprise M8000/M9000 Systems Overview, Part No.*
- *Sun SPARC Enterprise M4000/M5000 Servers Overview, Part No. 819-2204*
- **Sun SPARC Enterprise Dynamic Reconfiguration User's Guide, Part No.*
- *Sun SPARC Enterprise M4000/M5000/M8000/M9000 Servers Capacity on Demand (COD) Administration Guide, Part No. 819-7417*
- *Sun SPARC Enterprise External I/O Expansion Unit Installation and Service Manual, Part No. 819-1141*
- *System Administration Guide: Solaris Containers, Resource Management, and Solaris Zones, Part No. 819-2450*
- *System Administrator Guide: IP Services, Part No. 816-4554*
- *Solaris 10 6/06 Installation Guide: Solaris Live Upgrade and Upgrade Planning, Part No. 819-5777*
- *Sun Management Center 3.6.1 User's Guide, Part No. 819-5417*

- *Sun Management Center 3.6.1 Performance Reporting Manager User's Guide, Part No. 819-5423*
- *Sun Management Center 3.6 System Reliability Manager User's Guide, Part No. 819-4100*
- *Sun Management Center 3.5 Service Availability Manager User's Guide, Part No. 816-7416*
- *Installing and Administering Solaris Container Manager 3.6.1, Part No. 819-5420*
- *Solaris 10 6/06 Installation Guide: Solaris Flash Archives (Creation and Installation), Part No. 819-5779*
- *Solaris ZFS Administration Guide, Part No. 819-5461*
- *Solaris Fibre Channel and Storage Multipathing Administration Guide, Part No. 819-0139*

Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA **Phone** 1-650-960-1300 or 1-800-555-9SUN (9786) **Web** sun.com

Printed in USA 04/07