SUN BLUEPRINTS APPROVED

Managing Systems and Resources in HPC Environments

By Omar Hassaine - CPREngineering-HPC
Sun BluePrints™ OnLine - February 2002

`http://www.sun.com/blueprints`

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303 USA
650 960-1300   fax 650 969-9131

Part No.: 816-4335-10
Revision 1.0, mm/dd/01
Edition: February 2002

DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

CETTE PUBLICATION EST FOURNIE "EN L'ETAT" ET AUCUNE GARANTIE, EXPRESSE OU IMPLICITE, N'EST ACCORDEE, Y COMPRIS DES GARANTIES CONCERNANT LA VALEUR MARCHANDE, L'APTITUDE DE LA PUBLICATION A REPONDRE A UNE UTILISATION PARTICULIERE, OU LE FAIT QU'ELLE NE SOIT PAS CONTREFAISANTE DE PRODUIT DE TIERS. CE DENI DE GARANTIE NE S'APPLIQUERAIT PAS, DANS LA MESURE OU IL SERAIT TENU JURIDIQUEMENT NUL ET NON AVENU.

Please
Recycle

Adobe PostScript

Managing Systems and Resources in HPC Environments

This article provides an overview of enterprise tools and features for managing systems and resources in compute-intensive environments typically found in high performance computing (HPC). This article provides recommendations to system administrators and users for taking advantage of these tools and features. The article covers the following:

- "Enterprise Server Environments" on page 2
- "Enterprise Server Tools" on page 2
- "Enterprise Server Features" on page 7

Introduction

The adoption of node clustering architecture in high performance computing (HPC) is welcomed by both customers and computer vendors such as Sun Microsystems, Inc. The compute intensive community is now able to build powerful systems using regular computers that typically were used in business-type environments. Customers from both business and technical computing environments can take advantage of each other's strengths to improve their computing capabilities. The business environment can take advantage of advances in algorithmic development that lead to higher computing performance. Technical and scientific environments can take advantage of legacy tools that matured in business environments.

Enterprise Server Environments

Although environments vary depending upon business needs, a typical HPC environment using Sun's products and technologies includes the following hardware and software products:

- UltraSPARC™ processor-based computer or cluster
- Sun HPC ClusterTools™ software that contains:
  - cluster runtime environment (CRE) for executing parallel programs
  - Parallel Debugger (Sun Prism™ software)
  - Parallel Filesystem (PFS) or other third-party fast file system
  - Scalable Scientific Subroutine Library (S3L)
- Forte™ application development environment that includes:
  - compilers
  - Sun Performance Library™ software
  - Sun Performance Workshop™ software
- network interconnect infrastructure for node clustering

Enterprise Server Tools

Enterprise server tools available on the Sun platform and supported by the Solaris™ Operating Environment (Solaris OE) provide easier system administration and better use of system resources. In this section, we introduce the Sun™ Management Center (Sun MC) software and the Solaris™ Resource Manager (Solaris RM) software tools, focusing on their use in HPC environments. This section contains the following:

- "Sun™ Management Center (Sun MC)" on page 3
- "Solaris™ Resource Manager (Solaris RM)" on page 5

Sun™ Management Center (Sun MC)

The Sun Management Center (Sun MC) software, formerly Sun Enterprise SyMON™, is an enterprise system management tool that supports the entire line of Sun servers and desktops. This tool provides a single point of management for the entire line of networked Sun systems.

Sun MC Features

The following list summarizes the most relevant features that apply to HPC environments:

- Manages thousands of Sun systems
- Provides a common GUI look and feel with the Java™ console
- Integrates with tools from leading third-party vendors to address heterogeneous environments
- Provides alarm management and predictive failure analysis
- Identifies faults before a system is affected, via comprehensive online hardware diagnostics testing
- Provides a powerful, easy-to-use interface for developing custom modules using the GUI module builder
- New filtering capabilities help pinpoint problems quickly, even in systems with thousands of objects or nodes
- Allows dynamic reconfiguration and domain management through secure management controls
- Supports new UltraSPARC III processor based systems

Sun MC Benefits and Recommendations

Typically, HPC systems require clustering of several nodes to achieve levels of performance required by scientific and engineering communities. The need for system monitoring and management increases with the number of nodes that form a compute cluster. This complexity makes a typical HPC site an excellent candidate for deploying tools such as Sun MC software.

As the system administrator, you would have a physical and logical view of the entire cluster or clusters on a web-based interface, accessible from a console desktop that is geographically located anywhere on the network.

We recommend that you integrate Sun MC software with a job management system (JMS). This combination allows you to monitor and control queues and jobs from within a Sun MC console. For example, a site system administrator can easily:

- check the status of jobs and resources
- receive alarms in case of overload situations or hard limit excesses
- suspend and resume queues as well as jobs

Another product that is useful to integrate with Sun MC software is the Load Sharing Facility (LSF) from Platform Computing Inc. This product supports the SNMP protocol and is easily integrated with Sun MC software.

Note – The Sun Grid Engine (SGE) job management system does not currently support SNMP. It does provide a partial integration with Sun MC software, using the Tcl scripting language interface at the agent level. This capability can make the administration of compute intensive tasks easier to perform, particularly if an Sun MC software is already deployed at the site.

The following figure shows a high-level diagram of how job management systems (JMS) integrate with Sun MC software.

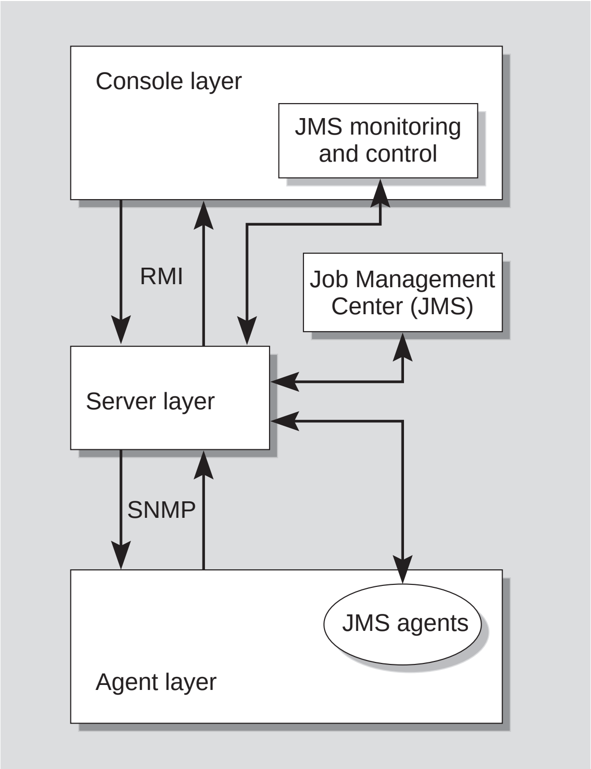

FIGURE 1 Integrating Job Management Systems With Sun MC Software

Solaris™ Resource Manager (Solaris RM)

The Solaris Resource Manager (Solaris RM) is a software product that is an extension to the Solaris OE. Sun MC software enhances resource availability for users, groups, and applications. It provides the ability to reserve and control major system resources such as CPU, virtual memory, and number of processes. Solaris RM software controls resource usage based solely on user ID.

Capabilities provided by Solaris RM software are regulated by a resource policy that is established according to a site's requirements. Users and applications receive a more consistent level of service on a single server, resulting in significant cost savings and greater administrative flexibility.

The Solaris RM software does not notify system administrators about usage limits; it provides resource usage reports and guarantees resources to key applications and users. It makes the performance of an application more predictable, and it ensures that system response times are not adversely affected by other tasks on a system.

Solaris RM Features

The following is a list of features that apply to HPC environments:

- Reserves and controls major system resources such as CPU, virtual memory, and number of processes.
- Provides users and applications with a more consistent level of service on a single compute server.
- Guarantees resources to key applications and users.
- Provides more predictability of application performance and ensures that system response times are not adversely affected by other tasks on a system.
- Provides resource usage reports.

Solaris RM Benefits and Recommendations

Most HPC sites use a job management system (JMS) product to monitor and schedule jobs submitted by their user community, according to resource usage and site configuration and policies.

Most of the functionality provided by the Solaris Resource Manager tool is already included in a job management system. For example, Solaris RM software supports hard limits on resources where a process fails if it exceeds resource limits. In comparison, most popular JMS products such as LSF and SGE support hard limits.

In some cases, combining Solaris RM software with other JMS products provides additional functionality and efficiency. An example where Solaris RM software can be used with a JMS for more efficient use of resources is assigning shares to workloads and users. Distributing shares allows for more fair share scheduling, and it prevents jobs from overusing more than their allotted shares. The following figure illustrates a simple example that includes a hierarchy with two layers. The first layer assigns CPU shares to three workloads as follows:

- compute intensive workload 60 shares
- application development workload 30 shares
- administration tasks workload 10 shares

The second hierarchical layer assigns CPU shares to users within a workload. The example in the figure illustrates Solaris RM software in an HPC environment with a single node. The same concept applies to a cluster of nodes when Solaris RM software is deployed at every node to ensure that applications or jobs running with a specific user ID only get allowed resources, according to a predetermined policy.

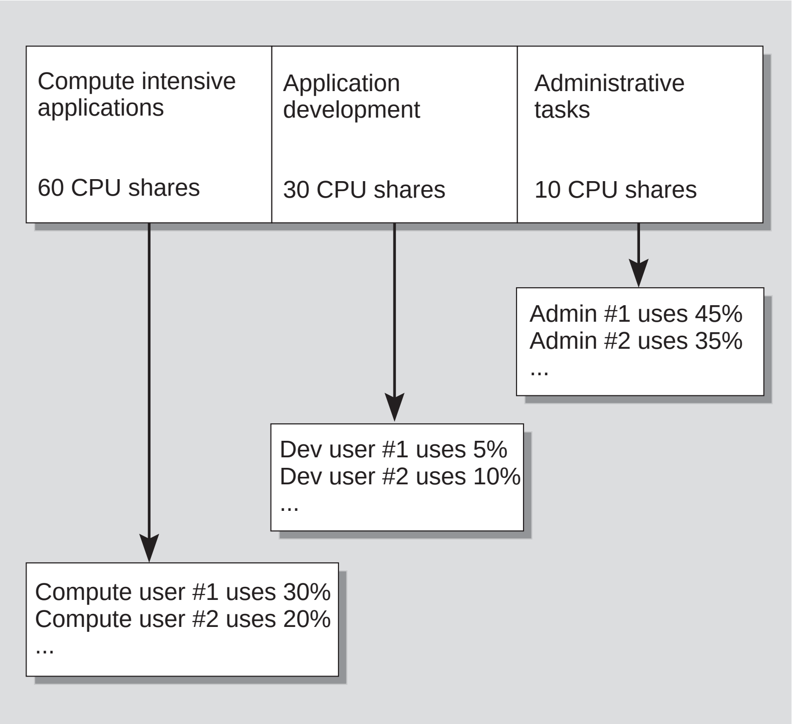

FIGURE 2 Solaris RM Software With a Single Node

Enterprise Server Features

Certain hardware and software features available on the Sun platform are beneficial in HPC environments. Using these features provides greater system flexibility and optimal use of hardware resources. This section describes the following:

Dynamic System Domains

This attractive feature allows a machine to be logically divided into several domains or machines where each runs its own copy of the Solaris OE. The effect is like having several machines combined into one hardware box. The latest Sun Fire™ mid range servers and the last two generations of high-end servers include Dynamic System Domains.

This feature provides a command-line and GUI interface for performing the following operations:

- creating domains
- removing domains
- showing status of domains

The following are some example uses of Dynamic System Domains:

1. Consolidating many servers in one small footprint
2. Creating a small domain for testing upgrades
3. Separating development domains and production domains
4. Partitioning domains to scale and improve performance of applications

The first three uses apply to both business and HPC environments. In particular, item three is a popular practice by customers who want two domains, where one is for developing code and the other is for submitting compute-intensive production code. The size of a domain varies with customer sites and is related to the workload required by development tasks and the size of projects at a site.

The fourth example improves system throughput and performance of applications by optimally managing hardware resources of a machine. Domain partitioning involves either a domain expansion or a domain splitting operation. The next two subsections describe both of these topics.

Domain Expansion

This operation expands the size of a domain by merging it with another domain or borrowing hardware resources from another domain. Expanding a domain can improve the performance of parallel applications that have great potential for scaling beyond the maximum number of processors in a domain.

Examples of applications that take advantage of domain expansion are parallel applications and other scalable compute-intensive code. Site administrators at large compute sites can merge whole or parts of idle domains into one larger domain, thereby making it available to candidate applications at appropriate times such as nights or weekends.

Note – Domain expansion incurs an overhead when a whole domain is merged, because it needs to be shut down and later restarted to its original state.

The following figure depicts a domain expansion scenario for scaling HPC applications.

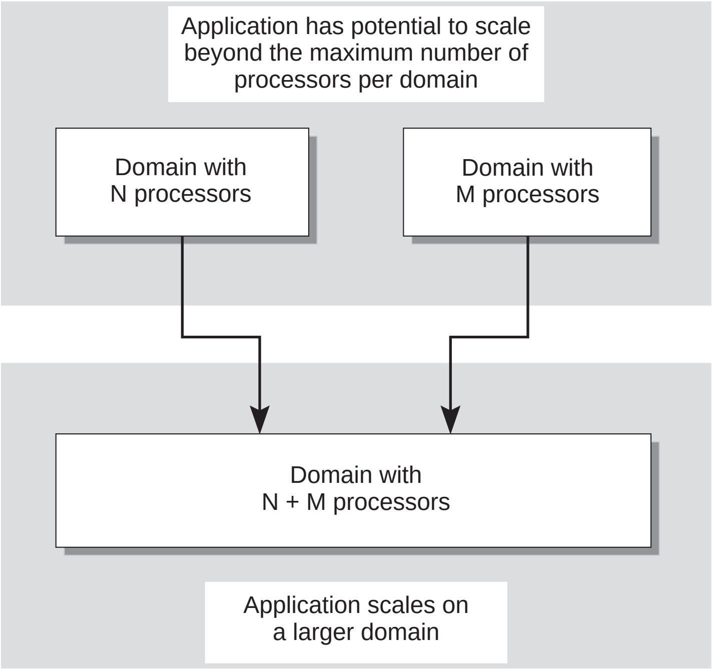
Application has potential to scale beyond the maximum number of processors per domain
Domain with N processors
Domain with M processors
Domain with N + M processors
Application scales on a larger domain

FIGURE 3 Merging Domains to Scale HPC Applications

Domain Splitting

This operation divides a large domain into two or more smaller domains. Splitting domains can improve overall performance of a system by configuring the most optimal size domains for running HPC applications.

Splitting a domain is beneficial when parallel applications would not scale beyond a certain number of processors (see the following figure). The example we illustrate is the Mesoscale Model weather program, also known as MM5. The MM5 is a public domain program developed by Pennsylvania State University and the National Center of Atmospheric Research (PSU / NCAR).

The MM5 program uses a variable number of cells that decide how many outer loops are used in the most time consuming part of the program. If only 25 cells are used, then the program needs 25 processors to optimally run because each cell (outer loop) is distributed to a separate processor.

When the program is run with a smaller number of cells, domain splitting may free up hardware resources by configuring the most optimal size domain, which includes only the needed number of processors. We see cases at customer sites where domain splitting is used to assign domains to groups at intervals to run their critical compute-intensive jobs.

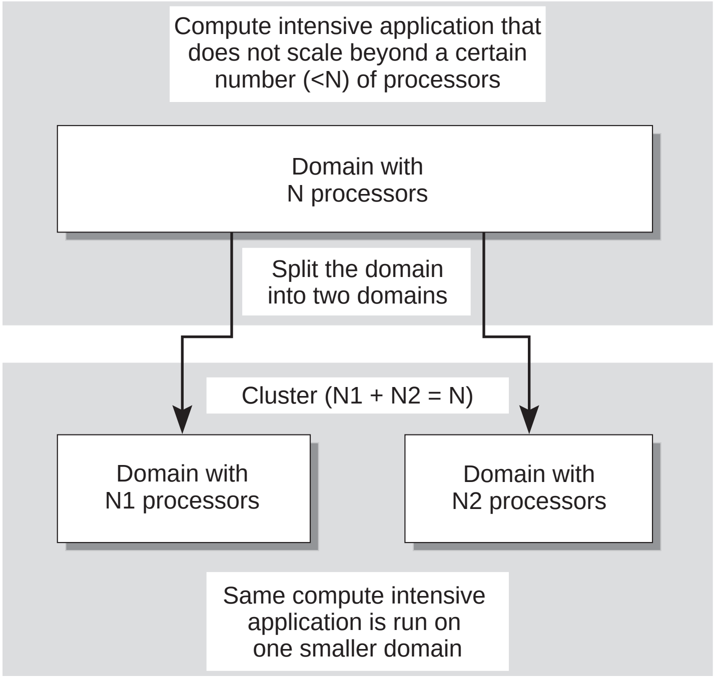
Compute intensive application that does not scale beyond a certain number (<N) of processors
Domain with N processors
Split the domain into two domains
Cluster (N1 + N2 = N)
Domain with N1 processors
Domain with N2 processors
Same compute intensive application is run on one smaller domain

FIGURE 4 Splitting Domains

Dynamic Reconfiguration

This feature allows a system administrator to add / delete system boards from either an idle domain or a domain running the Solaris OE. This feature was introduced with the Sun Enterprise™ 10000 server release and integrated with the Dynamic System Domains feature previously discussed.

Dynamic reconfiguration now supports automated dynamic configuration (ADR). This enhancement is attractive because the dynamic reconfiguration operation is performed automatically without requiring an operator's attendance.

ADR is helpful when workloads on domains vary at different times of the day, because it provides the capability to move system boards between domains to meet the requirements of load constraints.

ADR provides the following commands that you can execute either from a command line or a shell script:

- `addboard`: attach a board to a domain
- `deleteboard`: detach a board from a domain
- `moveboard`: detach a board from a domain and attach it to another domain
- `showusage`: display board and dynamic reconfiguration data

These commands are wrappers for lower level commands that perform manual dynamic configuration.

HPC sites that benefit most from dynamic reconfiguration are sites that have machines configured with a domain for application development and another domain for job execution (computing). These sites are excellent candidates because the development domain releases hardware resources to the compute domain, which desperately needs resources at night time or other non-peak usage times for scheduled batch jobs. The reverse operation happens in the morning or at peak usage times where the development domain needs to reclaim its original resources to serve the development community.

The following figure illustrates dynamic reconfiguration. This operation can be launched using scripts that monitor the load of domains and move resources according to dynamic needs of the system.

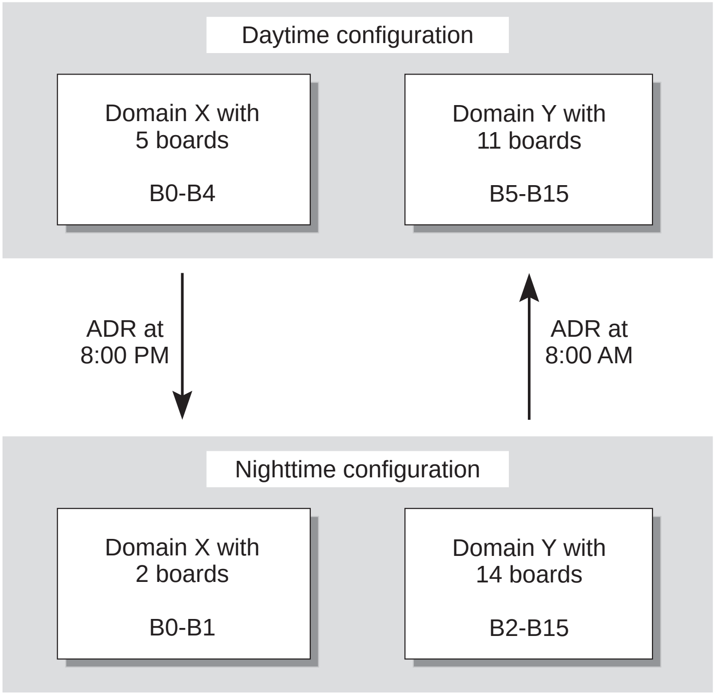

FIGURE 5 Dynamic Reconfiguration Example

Processor Sets

This feature allows a multiprocessor system to be divided into two or more logical groups of processors. Processor sets (also called processor partitions) provide a mechanism for scheduling processes to run exclusively on one processor set. This feature was introduced in Solaris OE, version 2.6. Processor sets serve the following uses:

- increase performance of a system by dividing a machine into processor sets
- assign and dedicate applications to a specific processor set

In this next section, we compare the Processor Sets feature with the Dynamic Systems Domains feature.

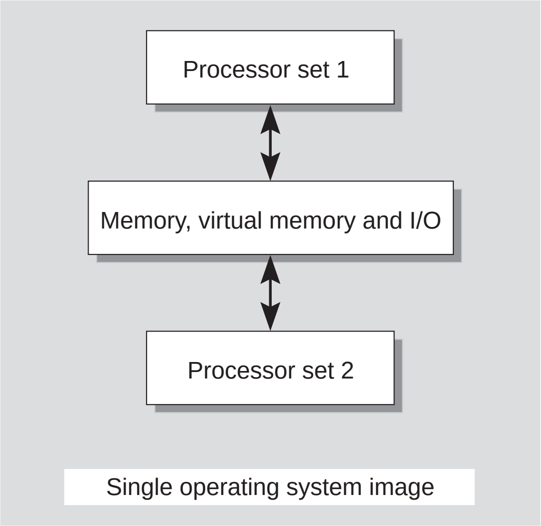

FIGURE 6 Processor Sets Architecture

Processor Sets Versus System Domains

The Processor Sets feature is somewhat similar to the Dynamic Systems Domains feature; a machine can be divided into groups of processors where applications can run exclusively. The Processor Sets feature is generally less robust than the Dynamic System Domains feature. The following information provides a brief comparison of the two features.

- Dynamic system domains divide a machine into two or more virtual systems. For example, a machine with two domains literally runs two copies of Solaris OE whereas Processor Sets operate within a single operating system instance. Both features permit isolation of one application from another, assuming that the two applications are operating in different processor sets.
- Memory, virtual memory, and I/O are shared by all processor sets. Dynamic system domains carve out their own memory and I/O that are used exclusively by the domain that owns them. All processor sets within a single Solaris OE instance use the same pool of memory, virtual memory, and I/O resources. If an application in one processor set stumbles on a bug and consumes all available memory, applications in other processor sets are affected.
- Applications that consume all available CPU affect only their own processor sets. The same also applies to dynamic system domains.

- From an administrative standpoint, it is a lot easier to create, destroy, and handle processor sets than to maintain dynamic system domains. To create a dynamic system domain, an administrator has to make sure that the required hardware is available, install the Solaris OE, and install additional software such as the Sun HPC ClusterTools software suite and the Sun Grid Engine software.
- The Dynamic System Domains feature allows an administrator to test a new version of software without affecting other domains. Unfortunately, the Processor Sets feature does not provide this capability. In contrast, it is possible to test a new version of an application in a separate processor set, however, there is always a risk that the application might affect hardware common to all processor sets.
- Both processor sets and dynamic system domains provide the system administrator the capability to set up a separate environment for HPC applications.

Extended Accounting

The Extended Accounting feature was introduced in the Solaris™ 8 Operating Environment, Update 1. This feature extends the Solaris OE environment's traditional system accounting with task and project ID concepts. The task and project IDs proved a way to tag a program so that it belongs to a job, which in turn can belong to a project.

In business environments, this feature is used by the third-party accounting tool PerfAcct 3.1, from Instrumental Inc. This product provides sophisticated accounting reports by gathering and processing data from machines within a network. Unfortunately, this accounting tool does not currently support parallel distributed environments.

A typical HPC site is characterized by multiple users who execute long-running programs, which compete for finite machine cycles that are strictly assigned to projects. HPC sites that lease computer time to external users need an advanced accounting infrastructure that provides an efficient charge-back accounting feature for a wide range of jobs. Also, they especially need the accounting feature for parallel jobs that span across nodes of computer clusters. To perform system accounting for typical HPC configurations, there is a dire need for a sophisticated accounting tool that provides this required functionality. The majority of HPC sites deploy a job management system product that regulates the use of resources by jobs and users. The underlying operating system provides the required hooks such as the project and task ID, which allow the job management system to provide more comprehensive job accounting.

The Sun Grid Engine software currently does not take advantage of extended accounting, and there is no accounting report tool that provides sophisticated reports and charge-back reports for SGE jobs in a Solaris OE.

Currently, the only product that satisfies the full compute-intensive job accounting requirements on a Sun platform is the LSF Analyzer tool, which reports accounting only for jobs that are submitted using the LSF job management system.

It is clear that the underlying API is available for future implementation of distributed job accounting on Sun platforms.

Summary

This article highlights the most important tools and features available on the Sun platform for system administration of HPC sites. It provides a brief description of the Sun Management Center software and the Solaris Resource Manager software tools. We provide recommendations on how system administrators can deploy these tools in high performance computing (HPC) environments to enhance system administration and to optimally use system resources. We briefly describe the following features: Dynamic Systems Domains, Dynamic Reconfiguration, Processor Sets partitioning, and Extended Accounting. Additionally, we compare the Processor Sets feature with the Dynamic System Domains feature. This article describes how these features can improve performance and throughput of HPC applications.

Acknowledgements

Thanks to everyone that offered advice, recommendations, and feedback during the writing of this paper.

Related Resources

- Sun HPC ClusterTools 4 documentation set: `http://docs.sun.com`.
- Forte Compilers and Performance library: `http://docs.sun.com`.
- Sun Management Center Whitepapers: `http://www.sun.com/sunmanagementcenter`.
- LSF SNMP Agent: `http://www.platform.com documentation`.
- Solaris Resource Manager: `http://docs.sun.com`.
- Mesoscale Model program home page: `http://www.mmm.ucar.edu/mm5`.
- Solaris Extended Accounting, Solaris Administration: `http://docs.sun.com`.
- Instrumental Inc.'s PerfAcct tool: `http://www.instrumental.com`.
- Platform Computing Inc. LSF Analyzer tool: `http://www.platform.com`.

Author's Bio: Omar Hassaine

Omar Hassaine is a senior HPC engineer with over twenty years experience in the computer industry. Omar worked on two consecutive high-end SPARC servers—including Sun Enterprise 10000 server—as a project leader in the system software group at Cray Research Business Systems Division and Sun Microsystems, respectively. Before Omar joined Cray, he was a compiler engineer in the area of source code optimizers for super-compilers at Kuck & Associates. Omar has authored and given several technical presentations at various Sun sponsored events. He helped develop and teach HPC administration and programming courses. Omar designed and developed a diagnosis tool that is being deployed at large HPC sites.