DPRO-95503
Jane Wright, Ann Katan

Product Report
3 October 2002

Sun Cluster Software

Summary

Sun Cluster 3.0 5/02 is the latest release of clustering software that is at the heart of the SunPlex system, which creates a highly available and scalable computing environment.

Table of Contents

List Of Tables

Gartner

Entire contents © 2002 Gartner, Inc. All rights reserved. Reproduction of this publication in any form without prior written permission is forbidden. The information contained herein has been obtained from sources believed to be reliable. Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Gartner shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve its intended results. The opinions expressed herein are subject to change without notice.

Corporate Headquarters

Sun Microsystems, Inc.

901 San Antonio Road

Palo Alto, CA 94303-4900, U.S.A.

Tel: +1 800 786 7638

Fax: +1 512 244 9222

Internet: www.sun.com

Overview

As information technology assumes an increasingly important role in major corporations, the requirements placed on the IT infrastructure have become more demanding. These requirements may include:

- Availability of critical applications on a 7×24 basis.
- Continuous access to key data.
- Confidentiality of electronically stored information, such as customer credit card numbers.
- Not-to-exceed response time Service-Level Agreements.
- Infrastructures that can rapidly adapt to changing usage patterns.

These requirements are often becoming the norm rather than the exception. Exacerbating these stringent service requirements, IT infrastructures have become increasingly complex. The increasing number of applications and the growth in the number of users (originally only internal to the corporation, but now increasingly include partners and customers) are also major contributors to IT structural complexity. Sun refers to this build-up in complexity as the “net effect,” and Sun touts its SunPlex system, based on Sun Cluster 3.0 5/02, as the solution for the “net effect.”

Sun Cluster software increases application availability by providing fast recovery for events (for example, application failure or network adapter failure) that can cause an application to become unavailable. Applications can be recovered locally by restarting the application or failing over a network adapter. For more serious errors, Sun Cluster 3.0 5/02 will restart all or part of a node’s workload on another node in the cluster.

Sun Cluster 3.0 5/02 can be used to simultaneously address requirements for high application availability, infrastructure scalability and continuous access to data while simultaneously reducing cost of ownership by drastically lowering administrative costs. Sun Cluster 3.0 5/02 provides the typical features of application failover and parallel database access that have long characterized Unix clustering products. More important, Sun provides three services in a way that allow the cluster to appear to most users and system administrators as a single system (single-system image), thus simplifying application development and deployment as well as day-to-day system administration. The three core services of Sun Cluster are:

- **Global File Service**, which allows any server in the cluster to access data that resides on any disk within the cluster. Using the global file service, all data in the cluster can be accessed as if it were local data (file system calls). The cluster can be configured for high availability so that there is always more than one copy of the data and more than one path to it. When configured for high availability, no single failure within the cluster will prevent an application from accessing data. A failure will simply cause an alternative path to be taken.

- **Global Network Service**, which allows external network access through a single IP address to a service that may reside on any node or on multiple nodes in the cluster. Alternatively, multiple IP addresses can be used for a single service, or multiple services can use a single IP address. These services are used to provide “horizontal” scalability. To increase application capacity, customers may simply run the application on more servers and load balance incoming requests among them.
- **Global Devices** are detected during system/cluster boot and are assigned global names that are known to all the cluster nodes.

Application Availability and Application Scalability

Sun Cluster 3.0 5/02 provides for application scalability, where multiple copies of the same application execute simultaneously on multiple nodes, as well as application availability, where only one instance of an application executes on the cluster.

Application scalability improves user response times for applications that do not require the state of the user transaction to be maintained (for example, Web server) while also providing a measure of availability, since the users can be “failed over” to another instance of the application using the global network services.

Where the “state” of a user transaction must be maintained following an application failure, Sun Cluster 3.0 5/02 provides application failover services. These services, which use Resource Group Manager technology, allow an instance of the application to execute on a different node with user connections re-directed using the global network services and file access restored using the global file service.

Cluster Services and Cluster Administration

The services that the SunPlex system delivers include:

- Failover services for business-critical applications and services.
- System-level monitoring and automatic reconfiguration of the network to ensure 100 percent network availability.
- System-level monitoring and automatic reconfiguration of the data paths to disk storage to ensure 100 percent data accessibility.
- Horizontal scalability for applications through easy addition of cluster nodes.

Global Devices

There are two classes of disks in a Sun Cluster:

- Local disks that are directly connected to a single node and hold the Solaris operating system for that node (each node in a Sun Cluster is required to have a local disk that holds a copy of the Solaris Operating Environment).
- Multihost disks that are connected to more than one node and typically hold application and client data.

When the cluster is first started, each node is probed to determine what storage devices (disk, tape and CDROM) are attached to the node and each mass storage device in the cluster is assigned a unique global device identifier for the cluster. Currently disk devices are the only multiported mass storage devices in a SunPlex environment. Devices that are multiported can automatically be made highly available (that is, the system automatically finds and uses an alternate path to the device if the node that is the device's master becomes unavailable).

Ease-of-Use Features

SunPlex Manager is included as part of SunCluster 3.0 5/02 software. It is a browser-based tool that helps simplify the configuration and installation of the cluster. Customers can use this tool to automatically detect the private interconnects in the cluster. In addition, the SunPlex Manager provides secure remote management of the cluster.

Sun offers scripts or agents for both high-availability applications and scalable services. Sun provides Sun Cluster Agents for HA-NFS (file systems) and HA-DNS (networking service) as part of the base Sun Cluster software. Other Sun Cluster Agents are available for a fee and are licensed on the entire cluster. These cluster agents provide procedures that still require customization (although much less development than creating an initial script) for many commonly used software packages. A new feature in Sun Cluster 3.0 5/02 is security hardening of all the supported agents.

The SunPlex Agent Builder works in either GUI or command-line modes and generates the source files and scripts that are necessary to build an agent for an off-the-shelf application. For applications meeting a well-defined set of prerequisites, the SunPlex Agent Builder is able to create either a highly available or a scalable application. The source files can be generated in either C or ksh. Developers can further modify the files generated by the SunPlex Agent Builder.

Disaster Recovery

Sun Cluster 3.0 5/02 supports two-node Campus Clustering with up to 10 km distances between the nodes. This enhanced clustering option is useful for building disaster-tolerant systems. Campus Clustering supports either a "two room" configuration (where the quorum disk is in one "room") or a "three room" configuration (with the quorum disk in a separate location). The "three room" configuration enhances the ability of the cluster to stabilize after a failure automatically, since a quorum can usually be re-established.

Analysis

Sun Cluster 3.0 5/02 makes the entire cluster system appear as a single computing resource, able to globally address its devices and files, although resources and files are distributed across multiple nodes. The cluster and its resources can be managed as if they were a single system, rather than a collection of servers on a private LAN.

A Sun Cluster can have up to eight nodes (where a node is a server in the cluster or a domain within a server). With Sun Cluster 3.0 5/02, Sun has integrated its clustering software into both the Solaris 8 and 9 Operating Environment, providing faster error detection and response but requiring that all the servers in the cluster run Solaris 8 or 9. Current Netra, Sun Enterprise and Sun Fire servers can be nodes in a SunPlex environment. Since a partition can also function as a cluster node on an Enterprise 10000, SunFire 12K or SunFire 15K, these enterprise-class cluster nodes can be reconfigured without downtime due to their dynamic reconfiguration capabilities being integrated with Sun Cluster 3.0 5/02.

The global devices and global file service found in the Sun Cluster make it easy to use cluster-wide resources. The downside to the transparent access to a file across the cluster is that there is additional latency (since data takes longer to reach its destination through the cluster interconnect than through a local attachment) and potential interconnect bandwidth saturation, both of which can negatively affect performance.

Sun Cluster Software

Sun Cluster 3.0 5/02, as is typical with most Unix clustering solutions, provides a proprietary environment with little portability. The only cluster interconnects that Sun supports are 100Mbit Ethernet, Gigabit Ethernet, prior generation networking technology and SCI (only on PCI adapters), a remote shared memory subsystem that is used with applications such as Oracle 9i RAC to achieve optimal performance. Again, this situation is common to most Unix clustering solutions. Given the proprietary nature of Unix cluster solutions, it would be desirable to implement the other cluster interconnects with leading-edge performance and better latency characteristics.

Pricing

Table 1: Pricing: Sun Cluster 3.0

	List Price per node (US$)
Sun Cluster 3.0	2,000-100,000 (depending on server)

GSA Pricing

Yes.

Competitors

Sun describes its clustering solution as being the solution for multiple problems that face IT managers designing and operating n-tier infrastructures. Sun Cluster 3.0 5/02 provides:

- A traditional failover clustering solution for business-critical applications and/or the back-end database tier, for availability.
- A highly available and easily scalable deployment platform for the application and Web-serving tiers, for scalability.
- A deployment environment that is easy to manage and that can easily adapt to change, for manageability.

All of the major Unix vendors provide products that address these capabilities; however, not all competitive vendors have chosen to provide these capabilities in one product under the umbrella of their clustering solution.

The products to choose and the vendor to use will depend on the problems to be solved. In terms of a traditional clustering solution, there are basic features and functions that each solution should be measured against. These basic features and functions include:

- Number and selection of available agents or scripts to allow ISV applications to take advantage of the cluster’s availability and/or scalability attributes.
- Number of nodes in the cluster.
- Whether or not the cluster presents itself as a single system image.
- Speed of the cluster interconnect (the faster, the better—especially for scenarios such as parallel database clusters or clusters offering concurrent file access).
- Online cluster reconfiguration capabilities.
- Types of load balancing supported by the clustering solution (the more extensive, the better).

Sun Cluster Software

The table "*Comparison: Unix Server Clustering Solutions*" compares Sun Cluster 3.0 5/02 against its major competitors based on these comparison factors. The clustering solutions that compete against Sun Cluster 3.0 5/02 are HP's MC/ServiceGuard and related products for HP 9000 servers, HP's (Compaq's) TruCluster Server for HP AlphaServer systems, IBM's Unix Clusters and HACMP for IBM pSeries servers and VERITAS Cluster Server, which is supported on a variety of servers.

Table 2: Comparison: Unix Server Clustering Solutions

Server Configuration	**Sun Cluster 3.0 5/02 Update**	**HACMP 4.4.1**	**HP MC/ServiceGuard a.11.14**	**HP (Compaq) TruCluster Server 5.1a**	**VERITAS Cluster Server 2.0**
Max. No. of Nodes Supported for Failover	8	32	16	8	32 *(7)*
Supports Failover Between Partitions	Yes	Yes	Yes *(4)*	Yes	No
Node Interconnect Technology (cluster interconnect; private network)	Ethernet (100 MB/s) Gigabit Ethernet, SCI	Gigabit Ethernet (125 MB/s), ATM, FDDI, Fibre Channel, Token Ring, SP Switch	Gigabit Ethernet, FDDI, Token Ring	PCI-based Memory Channel (800-1600 MB/s), Gigabit Ethernet	Ethernet, Gigabit Ethernet
Max. No. of Nodes Supported for Shared Database Access	8	8	16 (Oracle 9i RAC) 8 (OPS)	8	0 *(6)*
Operating System Supported	Solaris 8 Solaris 9	AIX 4.3.3 AIX 5L	HP-UX 11/11i	Tru64 Unix V5 and later	AIX 4.3.3, AIX 5L 5.1, HP-UX 11/11i, Solaris 8
Storage Configuration					
Server/Storage Interconnects	Fibre Channel, SCSI	Fibre Channel, SCSI, SSA	Fibre Channel, SCSI	Fibre Channel, SCSI	Fibre Channel, SCSI
Storage Vendors Supported	Sun (EMC supports SC 3.0)	IBM	HP XP & VA, EMC Symmetrix	HP (Compaq)	Many
Failover to Nodes With Mirrored Data	Yes	Yes (HAGEO)	Yes	Yes	Yes
Concurrent Disk Access Supported (no lock manager)	Yes	Yes	Yes	Yes	Yes

Copyright 2002

Table 2: Comparison: Unix Server Clustering Solutions

Server Configuration	**Sun Cluster 3.0 5/02 Update**	**HACMP 4.4.1**	**HP MC/ServiceGuard a.11.14**	**HP (Compaq) TruCluster Server 5.1a**	**VERITAS Cluster Server 2.0**
Concurrent Disk Access Supported With Lock Manager	Yes	Yes *(1)*	Yes *(4)*	Yes	No *(7)*
Single System Image					
Cluster-Wide File System (same file naming from all nodes)	Yes	Yes *(2)*	No	Yes	No
Cluster Alias/Cluster-Wide Network (connections to individual nodes not required)	Yes	No	Yes *(4)*	Yes	No
Cluster-Wide Ownership of I/O Devices (device naming the same from all nodes)	Yes	No	No	Yes	No
Cluster Management					
Server Administration Performed Separately on Each Node	Yes	Yes	No	No	Yes
Server Administrative Commands Can Be Replicated to All Cluster Nodes for Execution	Yes	Yes *(2)*	Yes	Yes *(5)*	Yes
Cluster Cloning Supported	No	Yes	No	Yes	No
Single Event Manager	Yes	Yes	Yes	Yes	Yes
Single-Error Log	Yes	Yes	No	Yes	Yes

Table 2: Comparison: Unix Server Clustering Solutions

Server Configuration	**Sun Cluster 3.0 5/02 Update**	**HACMP 4.4.1**	**HP MC/ServiceGuard a.11.14**	**HP (Compaq) TruCluster Server 5.1a**	**VERITAS Cluster Server 2.0**
Users Are Authorized Once for All Cluster Nodes	Yes	Yes	Yes *(4)*	Yes	Yes
Consolidated Management Station	Yes	Yes	Yes	Yes	Yes
Web-Based Management	Yes	Yes	No	Yes	Yes
Change Management					
Add/Remove Nodes Without Bringing Down the Cluster	Yes	Yes	Yes	Yes	Yes
Configuration Changes Applied Once for the Cluster	No	Yes *(2)*	No	Yes	No
Perform Rolling Upgrades of Operating System Across the Cluster (cluster remains available during process)	Yes	Yes	Yes	Yes	Yes
Perform Rolling Upgrades of Cluster Software Across the Cluster	Yes	Yes	Yes	Yes	Yes
Perform Rolling Upgrades of Applications Across the Cluster (applications remain available to users during process)	Yes	Yes	Yes	Yes	Yes

Sun Cluster Software

Table 2: Comparison: Unix Server Clustering Solutions

Server Configuration	**Sun Cluster 3.0 5/02 Update**	**HACMP 4.4.1**	**HP MC/ServiceGuard a.11.14**	**HP (Compaq) TruCluster Server 5.1a**	**VERITAS Cluster Server 2.0**
Application Failover					
Application Restart on Failover	Yes	Yes	Yes	Yes	Yes
Application Restart From Checkpoint	No	No	No	No	No
Application Can Be Locked to Specific Nodes (for performance)	Yes	Yes	Yes	Yes	Yes
Applications Automatically Load Balanced Across Cluster	Yes *(3)*	No	No	Yes	Yes
(1) Requires extra cost CRM option.					
(2) Requires Cluster 1600 and PSSP.					
(3) Prioritized Service Management provides policy-based service-level management integrated with the clustering software.					
(4) Requires add-on product.					
(5) With a shared root many commands are executed once for the entire cluster.					
(6) A two-node Oracle 8i OPS cluster is supported, but only on Solaris.					
(7) Eight nodes supported with Veritas High Availability Foundation Suite; two nodes supported with Oracle9i RAC.					

The table "*High-Availability Agents (aka Scripts) for Third-Party Applications*" lists the agent software, or scripts, that each vendor offers as a standard product with its clustering solution. These agents provide templates for application failover within the cluster. In some cases, the customer may choose to pay the vendor to further customize the standard agent to better fit their environment. When the required application agent is not already available from the vendor, the customer may hire the vendor's consulting division to develop the agent on a custom basis.

Table 3: High-Availability Agents (aka Scripts) for Third-Party Applications

	Sun Cluster 3 5/02	**IBM HACMP**	**HP MC/ServiceGuard**	**HP (Compaq) TruCluster Server**	**VERITAS Cluster Server**
Price	$4,000 per cluster *(1)*	No charge	$995 per cluster	No charge	Price varies
Available Agents					
Oracle 8	Yes *(2)*	Yes	Yes	Yes	Yes
Oracle 9i	Yes *(2)*	No	Yes	Yes	Yes (two-node only)

Sun Cluster Software

Table 3: High-Availability Agents (aka Scripts) for Third-Party Applications

	Sun Cluster 3 5/02	**IBM HACMP**	**HP MC/ServiceGuard**	**HP (Compaq) TruCluster Server**	**VERITAS Cluster Server**
Oracle OPS	Yes *(2)*	Yes	Yes	Yes	Yes
Oracle 9i RAC	Yes *(2)*	No	Yes	Yes	No
Informix	No (yes, this is supported by the ISV)	No	Yes	Yes	Yes
Sybase	Yes *(2)*	No	Yes	Yes	Yes
IBM DB2	No (yes, this is supported by the ISV)	Yes	No	No	Yes
Apache Server	Yes *(2)*	No	No	Yes	No
Netscape FastTrack/Enterprise	No	No	Yes	Yes	No
Netscape Directory Server	Yes	No	Yes	Yes	No
Netscape Collaboration Server	No	No	Yes	Yes	No
Sun ONE Messaging Server	Yes *(2)*	No	No	Yes	No
Sun ONE Web Server	Yes *(2)*	No	No	No	No
Sub ONE Directory Server	Yes *(2)*	No	No	No	No
BEA/Tuxedo	No	Yes	No	Yes	No
Baan IV	No	Yes	No	Yes	No
SAP R/3	Yes *(2)*	Yes	Yes	Yes	No
PeopleSoft	No	Yes	No	No	No
Lotus Domino	No	Yes	No	No	No
Oracle Financials	No	Yes	No	No	No
Unicenter TNG	No	No	No	Yes	No
NetBackup	Yes *(2)*	No	No	No	Yes
NFS	Yes *(2)*	Yes	Yes	Yes	Yes
DNS	Yes *(2)*	Yes	No	No	Yes

(1) NFS and DNS agents included at no additional charge.

(2) Security-hardened version available.

Strengths

Single System Image

Sun Cluster Software

Sun Cluster 3.0 5/02 provides significant advancements toward achieving a single system image. The cluster-wide file system and cluster alias make managing a Sun Cluster 3.0 5/02 cluster much easier to manage than clusters based on previous Sun Cluster versions, or some competing products that do not present a single system image.

Continuous Availability of Core File and Network Services

For any application running on any server in the cluster, there are at least two independent paths to disk data (multihost disks) and the public network. If one path fails, the clustering software will transparently reroute the traffic from the failed path. Any application running in a cluster will become more available as it is not affected by a failure of a single disk adapter or network card.

Fast and Accurate Failure Identification

Reducing the amount of time it takes to determine that there was a failure improves recovery time. The integration of Sun Cluster 3 5/02 with Solaris 8 and Solaris 9 Operating Environment provides improvements in the amount of time required for failure notification and analysis.

Familiar Solaris Environment

With the integration of Sun Cluster into Solaris, familiar Solaris commands execute just as if only a single system were being administered. Any resource in the cluster can be managed from anywhere on the network where the Sun Management framework is available. Sun Cluster 3 5/02 allows the management of the nodes in the cluster as if they were a single system.

Pre-Configured Clusters

Cluster configuration to ensure high availability (no single point of failure) can be complex. Sun offers three options of two-node high-availability cluster configurations based in the SunTone Cluster program. The latest pre-configured and pre-tested systems are the Cluster Platform 280/3 (dual 280R systems and mirrored StorEdge T3 storage arrays), the Clustered Database Platform 280/3 (dual 280R systems, mirrored StorEdge T3 arrays, with Oracle 9i and/or Oracle RAC) and the transaction processing-oriented Cluster Platform 15K/9960 (dual SunFire 15K, 24 CPUs and a shared StorEdge 9960).

Limitations

Cluster Size

When clusters are used primarily for application failover, cluster sizes of eight nodes are more than adequate; however, when a cluster is deployed for its horizontal scalability attributes, the number of nodes that can compose a cluster may become a significant factor. Additionally, in situations where a cluster is deployed to simplify administration by replicating management commands, the number of nodes supported in a cluster may also become a significant factor. Some of Sun's competitors can support 32 nodes in a cluster, compared to Sun's eight-node maximum.

Insight

With Sun Cluster 3.0 5/02, Sun has taken important steps to address both complexity in the data center as well as complexities in deploying high-availability infrastructures. Providing a cluster with a single system image simplifies application development and deployment by making data equally accessible from all nodes. From the perspective of day-to-day administration, making eight servers look like one has real appeal. With the 5/02 update release to Sun Cluster 3.0, Sun continues to integrate its management products and single system RAS features into a coherent framework that will eventually lead to automated, dynamic resource allocation based on user-defined policies. Customers must be aware that Sun Cluster 3.0 5/02 requires features found in Solaris 8 and Solaris 9, not in older versions of Solaris.

Background on Clustering

Clustering was first introduced on the proprietary operating systems of the 1980s and migrated to Unix in the early 1990s as a way to increase application availability. If the server on which the application was running went down, the application could be automatically switched to another server in the cluster.

A cluster consists of a number of servers linked together through a private network. Each server in the cluster, called a node, continuously checks the state (that is, health) of the other nodes. If one node becomes unavailable, its workload is automatically transferred to one or more other nodes in the cluster. To ensure that an application can execute on the node to which it has been moved, the resources (particularly the data) that it is dependent on must still be accessible.

When the disks are physically attached to more than one node, either each node has ownership of separate disks (shared nothing model) or all disks can be served to all nodes (shared everything model). In the shared nothing model, when a failure is detected in a server, the clustering software passes disk ownership to the other node that has physical access to it. The fact that there are common SCSI or fibre-channel connections makes this relatively straightforward.

If all the nodes in the cluster require access to the same database—as, for example, when using Oracle Parallel Server—a mechanism is required to synchronize access to the database to ensure that the data remains consistent. This is what a Distributed Lock Manager (DLM) does.

In addition to failover, Unix clustering solutions today usually support some degree of load balancing for better resource utilization and increased performance and capacity. This may involve balancing the incoming client connections or balancing the application requests.