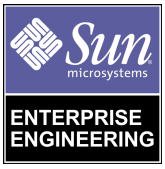
High Availability Best Practices

By Enrique Vargas - Enterprise Engineering

Sun BluePrints™ OnLine - December 2000



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.: 806-6858-10
Revision 01, December 2000

High Availability Best Practices

Introduction

Virtually every aspect of our lives are touched in some way by the computer revolution. We have come to rely on our computer systems for everyday life—email, pager services, pumping gas, and a myriad of other functions.

Because our trading systems represent the backbone of our financial stability, we have increasingly come to demand 100% availability. Any disruption to service is measured, not only in dollars and cents, but perhaps more importantly, in reputation.

Sun Microsystems, Inc. places system availability at the top of its business priority and has created two organizations that are responsible for high availability in Sun™ computer systems:

- The RAS (Reliability, Availability, and Serviceability) Engineering group concentrates on statistical analysis of system components to extract availability metrics. The RAS Engineering group develops tools that can assist hardware designers in achieving higher levels of availability in new and existing hardware designs.
- The SunUP™ organization is part of the SunLabs group, and is responsible for working with customers and third-party allies to develop products and services that enhance availability in Sun computer systems. The Sun BluePrints™ program works in alliance with the SunUP program to produce best practice documentation.

Availability has long been a critical component for online systems because business processes can quickly come to a halt when a computer system is down. For example, in the world of E-commerce, availability is critical due to a client's demand for instant access to a site—if a site is unavailable, for whatever reason, the competitors site is only a mouse click away.

Availability is directly affected by computer hardware and software, however, people and processes also have a significant impact.

This article introduces best practices that assist in minimizing the impact of people and processes in the datacenter—which helps to achieve higher availability goals.

Systems Management Principles

Systems should be managed by *highly trained* administrators who can appropriately configure the operating system and applications to fulfill a service level agreement (SLA). Because technology evolves rapidly there should be a training plan in place to keep system administrators' knowledge up-to-date.

System resources should be actively monitored by using platform monitors, (for example, Sun™ Management Center) and application monitors (for example, BMC Patrol/Best1 and Tivoli) to analyze resource usage trends which will help ensure existing resources are sufficient. Analysis of resource usage helps enable system administrators to predict when resources will be depleted—this enables upgrades to be scheduled before an application causes an outage.

Sun Microsystems, Inc. introduces the possibility of increasing system availability through the Sun™ Remote Services (SRS) service option which provides event monitoring and service management for critical production systems via a modem link. Remote monitoring of disk storage subsystems has long been adopted by high-end customers because they realize the value of replacing defective hardware and upgrading obsolete firmware before it becomes a critical issue in their datacenter.

The SRS option is sold as part of the standardized SunSpectrum Gold SM and SunSpectrum Platinum SM service bundles, and allows SunService SM engineers to monitor production systems on a 24 hour, 7 days a week basis to reduce the latency of a service action when a problem is detected. The SRS feature detects error conditions based on a set of predefined thresholds. When errors are detected, a corrective approach is taken in partnership with the customer.

Hardware Platform Stability

System hardware provides the base for the operating system and applications, and therefore must have a solid foundation. A common source of system problems can be related to loose mechanical connections that may result in intermittent problems

which are manifested through hard-to-track error messages at both operating system and application levels. If the hardware platform is subjected to mechanical vibrations, it could loosen components and change their electrical characteristics.

All connected components in a system must be correctly fitted and secured to help ensure maximum mechanical contact and help ensure the electrical characteristics remain constant. Memory and CPU components must be fully inserted and secured into their sockets. I/O cards must be fully inserted into their self-identifying bus (SBus) or protocol control information (PCI) connectors and fully secured. External cables need to be fully inserted and secured with the correct strain relief in place to ensure the cable weight does not place strain on internal components.

Consolidating Servers on a Common Rack

Consolidating multiple servers under a single rack introduces the possibility of an availability improvement through simplification of system management. It is recommended that each server within the same rack has independent I/O cabling, and an independent power source. These measures can help remove single points of failure (SPOF) and help prevent accidental outages when the servers that share the rack are serviced.

System Component Identification

When system components need to be replaced, availability is greatly improved if repairs take place in a swift and effective manner. Accurate system documentation and component identification is an investment which provides system administrators with increased control over existing resources. System documentation allows system administrators to plan future system changes without having to inventory existing resources repeatedly.

Each system platform comprises a large combination of components which may also be connected to network and storage devices. To prevent a technician from removing a wrong part, uniquely label each component including cables to assist the identification process. Identification can help ensure that only defective components are replaced—the replacement of wrong components can have expensive availability repercussions.

The Solaris™ Operating Environment (Solaris OE) creates controller instances for each device under its control (for example, `hme0`, `qfe1`, `c3`, etc.).

Note – It is a best practice to have controller instances identified on their physical controller ports.

The Solaris OE creates soft link entries in the `/dev/dsk` and `/dev/rdsk` directories for hard disk devices. Hard disk devices are also registered as SCSI device (sd) and serial SCSI device (ssd) instances (see the `sd`(7D) and `ssd`(7D) man pages for details). The Solaris OE displays disk errors using the controller and sd/ssd instances.

Note – It is a best practice to have disk devices labeled with the controller and sd/ssd instances at the replacement point to assist identification.

The Sun StorEdge™ A3500, A3500FC, and A1000 storage array platforms are hardware Redundant Array of Independent Disks (RAID) devices—a single Solaris OE disk instance may span several physical disks in the same storage array. The `rm6`(1M) utility can be used to identify any physical disk devices needing replacement.

The Solaris OE identifies tape devices through soft link entries in the `/dev/rmt` directory (see the `st`(7D) man page for details).

Note – It is a best practice to have tape devices labeled with the `rmt` instances at the replacement point to assist identification.

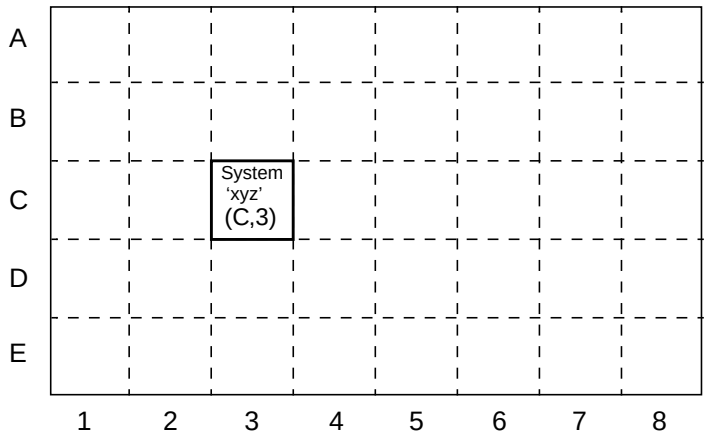


FIGURE 1 Datacenter Floor Plan Grid Diagram—System 'xyz' is Located at Grid Coordinate C3

Note – It is a best practice to have an interconnection diagram of the system platform showing all external equipment (for example, storage platforms, routers, hubs, and client machines). The diagram identifies each piece of equipment with unique name labels and its physical location in the datacenter. Physical locations are identified by creating a grid diagram of the datacenter—see FIGURE 1. This server location identification can have an additional entry to locate system or peripherals placed vertically within a rack (for example, C3F maps to a peripheral located on the sixth layer of a rack placed at the C3 coordinate). Devices on the same vertical layer *must not* have duplicate port IDs.

To minimize service time windows, label cables to reflect their destination location. As an example, one of the private interconnect patch cables will have a label of *qfe4C3A* on one end to reflect the `qfe4` connector located on the first vertical level of a rack placed at the C3 coordinate. The other end of the cable would have the *hme0C3B* label to reflect the `hme0` connector located on the second vertical level of a rack placed at the C3 coordinate. In order to minimize the service time when a cable needs to be replaced, a cable label needs to reflect the other end's location. One of the ends of the cable described above would have the *qfe4C3A- hme0C3B* label and the other end would have the *hme0C3B-qfe4C3A* label.

AC / DC Power

An alternating current (AC) power supply is a basic requirement for a system platform and is a critical component affecting system availability. The AC power supply must conform to the voltage settings established for a specific hardware platform. For example, if the direct current (DC) power supply subsystem fails, the whole system will fail. To help ensure voltages remain constant, they should be constantly monitored.

Mission-critical datacenter devices are commonly equipped with multiple power cords to access alternate AC sources. Systems with high availability requirements should have datacenter power that comes from different power stations to remove the potential of a SPOF.

Most power outages are brief in nature (brownouts), therefore, short term power needs can be fulfilled by using an uninterruptible power supply (UPS). The UPS battery should provide sufficient power to the interconnected system for a specified time period. Since a UPS converts a battery source to AC, it has the added advantage of smoothing any erratic output from a public utility power source (power conditioning).

Systems with higher availability requirements can use a diesel generator to supply power for extended power failures. Since a diesel generator requires time after starting to be fully functional, a UPS is still required to deliver power immediately following a public utility failure.

Do not position power cords where they are vulnerable to damage. It is also critical that power cords and circuit breakers are secured so that they are not accidentally tripped. Availability can sometimes be enhanced with simple actions such as applying duct tape over exposed circuit breakers to avoid accidental tripping resulting in a power outage. Keep in mind that when applying duct tape over exposed circuit breakers, there needs to be enough space provided to allow free tripping of the breaker during a current overload.

System Cooling

Datacenter equipment should be operated within a specific temperature range. When cooling in the datacenter is not available the only option is to shut all equipment down since high temperatures can trigger irreversible damage to

electronic components. Newer systems and devices generate more heat because they operate at faster clock speeds—plan to have additional air conditioning units to handle any additional heat output.

Systems with higher availability requirements should have standby or portable air conditioning units on hand to avoid an entire datacenter shutdown for times when the cooling infrastructure is unavailable.

Network Infrastructure

Network infrastructure comprising of switches, hubs, routers, and bridges, are considered external to the system and are commonly overlooked as being a key component of system availability. The network infrastructure enables communications between peer and client systems at the local and wide area network (WAN) levels, therefore, network infrastructure scaleability and availability directly affect a system's availability—even in a clustered environment.

Note – It is a best practice to use the Network Adapter Failover (NAFO) software provided with the Sun™ Cluster version 2.2 (SC 2.2) infrastructure to improve availability of the internal network controllers. This is accomplished by switching all network I/O to an alternate controller if a failure is detected. Most network product vendors are becoming aware of the high availability needs of network infrastructure and are providing redundancy schemes at switch and router levels to improve availability and scaleability.

Client connections establish the possibility of an availability improvement through an alternate WAN link which makes use of independent telephone switching stations to avoid single points of failure.

Customers with higher availability requirements should subscribe to the Internet using two different service providers, or at least use separate access points with the same provider to remove another potential single point of failure.

Security

In the past, security was mostly required by government and financial sectors, however, the E-commerce revolution is exposing more businesses to the potential of unauthorized access to confidential data, and attacks by hackers and viruses.

Malicious hackers attempt to modify business data for the purpose of satisfying their own egos while sophisticated hackers will steal business data in an attempt to profit financially. Overall, security breaches may result in partial outages or they may result in the total demise of an established business.

Denial of Service (DoS) attacks involve malicious, repetitive requests into an application with the intention of depleting application and system resources. DoS attacks are less severe from a security standpoint since they do not interfere with data integrity or privacy. However, since a regular end user is not able to access an expected service during a DoS attack, there is an immediate effect on end-to-end application availability.

Customers requiring Internet access should secure their external connections using:

- A secure application architecture with built-in functionality to resolve DoS attacks
- A secure, hardened operating environment with minimized system builds
- Up-to-date security patches
- Chokepoints and proxies where appropriate —Chokepoints are filtering routers, flow control devices, and application or packet filtering firewalls

For customers requiring higher levels of security, Sun Microsystems, Inc. offers the Trusted Solaris™ product which supplies security features and assurances supported by the National Computer Security Center (NCSC), and the Defense Intelligence Agency (DIA). An organization can implement the Trusted Solaris product with their own security levels options or they can enable full security so that users can perform administrative roles on their own workspaces.

Systems Installation and Configuration Documentation

All software applications start off with a system platform that can only be accessed through the OpenBoot™ Prom (OBP) prompt. After the system platform is correctly connected to the tape, disk, network, and printer devices, the application installation is accomplished through a series of sequential steps. The Sun Professional Services group provides Runbook services that produce detailed documentation and includes suggested procedures to assist in configuring and managing a system platform.

A simplified system installation sequence involves:

- Operating system installation
- Recommended Solaris OE patch installation
- Operating system tunable parameter configuration (`/etc/system`)
- Network, tape, printers, and disk configuration
- Volume manager installation and configuration

- Volume manager patch recommended patches installation
- System user configuration
- Software application installation and configuration
- Software application recommended patches installation

If all steps involved in the application installation are correctly documented, use this documentation to regenerate an entire system from scratch if disaster strikes. Additionally, system documentation can be used to educate other system administrators, or to provide material for an implementation review.

Note – It is important to have a documented backup and disaster recovery plan for all critical systems. Hardcopy documentation should be provided as a backup to online documentation.

Some system administrators use the `script`(1) utility to capture the input and output generated during a software installation session, however, the `script`(1) approach generates excessive text which can bury important content and it lacks comments explaining why things must be done a particular way. A `script`(1) output can be edited to create an HTML document with appropriate comments. User responses can be contrasted using an alternate font or text style.

Sun Microsystems, Inc. provides the JumpStart™ product as an option to install system software and application components from an OBP prompt without the risk of errors introduced by human interaction.

Note – It is a best practice to have JumpStart automatically regenerate the entire system software infrastructure to enable recovery from any local disaster.

Change Control Practices

Once the production system platform is stable and all applications are fully operational, it is crucial that any proposed system changes go through a peer review to identify the impact of such changes and to provide a strategy for implementation.

Note – It is critical for a system to be rebooted immediately after implementing any changes—it would be difficult to associate a failed reboot at some future time with modifications made today.

Datacenters with higher availability requirements should have a production system mirrored by a development system (the mirror could be a scaled down version of the production system) to implement and evaluate any changes prior to adoption. Care must be taken to isolate test operations on the development system from the real production environment. As an example of a poor test operation process, a customer mirroring an SAP manufacturing environment inadvertently generated a *test* production order which resulted in the building of an $1.5 million piece of equipment.

Note – It is imperative that a production system be backed up to tape before implementing any changes. This will enable the replication of the original system if the proposed modifications prove detrimental.

Maintenance and Patch Strategy

Sun Microsystems, Inc. introduces software patch releases on a continuous basis. Your local SunService provider can suggest the appropriate email alias subscription to notify you of all released patches and their contents. The provided patch information should be reviewed to identify if a patch is relevant for a particular production system. For example, some patches might involve I/O interfaces not used by the system, or involve a locality that does not apply.

It is recommended that all relevant patches are collected and stored, but only applied on a three or six-month schedule so as to keep systems current and minimize any outage impact. Only critical patches that affect the health of the production system should be applied immediately.

It is a best practice to always install all required or recommended patches for:

- Solaris Operating Environment Kernel Update
- Sun Enterprise™ 10000 System Service Processor (SSP)
- Sun Cluster Software
- Volume Manager (Solstice DiskSuite™/Veritas VxVM/RM6)
- Disk Controller/Disk Storage Array/Disk Drive Firmware

Datacenters with higher availability requirements may benefit from having patches applied to a development system to help analyze their full impact, and assist in a production rollout strategy.

Component Spares

Having spare components available in the datacenter introduces the possibility of an availability improvement by reducing the repair time window. Sun Microsystems, Inc. emphasizes the importance of engineering interchangeable components for use with different system platforms. This strategy allows a reduced spare component inventory.

System components have a higher probability of failure in their early and late life periods. Similar to the strategy used by rental car companies, customers with higher availability requirements can take a proactive approach of recycling components that are reaching their *wear-out* stage of life to avoid the exponential frequency of component failures (see section, “Failure Rate,” in the Sun BluePrints Online article, *High Availability Fundamentals*, November 2000 Edition (`http://www.sun.com/blueprints/1100/HAFund.pdf`)).

Note – Customers must contact a Service representative to arrive at a spare component strategy that best matches datacenter availability needs.

New Release Upgrade Process

Software products evolve over time—with new and enhanced features becoming available with each version release. Similar to the patch evaluation process, new software product releases should be evaluated by a review process to determine the business impact.

Additionally, systems with higher availability requirements may benefit from having new software releases applied to a development system to help analyze its benefits and plan the production rollout strategy.

Support Agreement and Associated Response Time

As discussed in section, “Serviceability Fundamentals,” in the Sun BluePrints Online article *High Availability Fundamentals*, November 2000 (`http://www.sun.com/blueprints/1100/HAFund.pdf`), logistic time can have a major impact on a system outage. As a general rule, the impact of logistic time on a system outage can be summarized by the following considerations:

- Logistic time significantly impacts the outage time caused by a SPOF
- Logistic time minimally impacts the outage time caused by additional system failures occurring within the same downtime period
- Logistic time has no impact on outage time caused by system interruptions triggered by Automatic System Recovery (ASR)
- Logistic time has no impact on outage time caused by maintenance related events

A service contract is a key element of system availability because it determines the maximum possible time it takes the service organization to assume ownership of a problem. It is important that customers understand the service options available in their geographic area to ensure that the appropriate service option is selected to meet their business availability requirements.

For mission-critical systems, it is important to assign a local person to be the central point of information gathering (in both directions) to avoid having too many people getting the status of a service request which can end up wasting the service engineers’ time.

System availability may be enhanced by being aware of the problem escalation process which ensures new bugs and problems are appropriately handled by the service organization. Whenever a problem is discovered with the functionality of a product, a request for enhancement (RFE) document should be filed with the service organization to initiate a review process.

Backup and Restore Testing

Tape backup of critical systems in a datacenter is routine. However, tape restores are not a common practice and it is often discovered after the fact that an inadequate backup ends up in an incomplete restore process.

Note – It is a best practice to routinely schedule tape restore fire drills to familiarize system administrators with the process, and to validate the adequacy of any backup procedure. Tape restore fire drills help evaluate the time involved to bring mission-critical systems back online—the business impact can be analyzed and reviewed for improvement.

Because the tape media makes physical contact with the read / write heads on tape drives, the tape becomes unreliable after a certain number of write cycles. It is recommended that the tape media manufacturer be contacted to obtain the best practice for retiring unreliable tapes from the backup pool.

Systems with higher availability requirements may benefit from triple mirroring disk data to make one of the mirrors available to an external system for backups. If the mirrored disk data is implemented with Veritas VxVM, then Dirty Region Logging (DRL) must be enabled to minimize a third mirror synchronization time after re-attachment.

Cluster Recovery Procedures

The SC 2.2 infrastructure introduces a failover server to execute a mission-critical application if an application failure is detected. Even though the SC 2.2 application failover is an automated process, it is a best practice to engage in fire drills by scheduling a manual failover to force system administrators to become familiar with the process. These fire drills can help enable a system administrator to make informed decisions if a real failover occurs.

Campus Cluster Recovery Procedures

Disasters come in many forms—for example, floods, lightning, earthquakes, and terrorism, any of which could disrupt a production site. The SC 2.2 infrastructure enables a campus cluster solution—a failover node can be located up to 10 Km (6 miles) away from a production node, which is an additional layer of protection in recovering from a production site disaster.

A campus cluster is more critical than a regular cluster. Any campus cluster site should engage in fire drills by scheduling routine failovers to force system administrators to become familiar with the failover process and to be able to make

informed decisions if a real failover occurs. If the campus cluster is a part of an established disaster recovery plan then the complete disaster recovery plan must be tested.

Summary

Customers increasingly require data and applications to be available around the clock to satisfy global operation demands, and they view availability as the single most important metric of their overall system performance. With the wider acceptance of E-commerce, and increasing dependency on computer systems—availability has become a requirement that applies not only to mission-critical applications, but to almost the whole IT infrastructure.

Availability is affected by computer hardware and software and also by people and processes. Availability can be optimized through a systematic approach, and by using best practices in the datacenter to minimize the impact of human error.

Author's Bio: Enrique Vargas

Enrique Vargas brings a wealth of large systems experience to Sun Microsystems and specializes in high-end UNIX® offerings, including the Sun Enterprise 10000 server. Enrique Vargas came to Sun Microsystems from Amdahl where he also focused on the high-end Solaris systems.