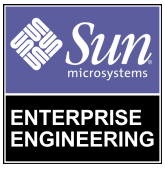
High Availability: Configuring Boot, Root and Swap

By Jeannie Johnstone Kobert - Enterprise Engineering

Sun BluePrints™ OnLine - June 1999



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.: 806-4395-10
Revision 01, June 1999

High Availability: Configuring Boot, Root, and Swap

Highly available solutions are in demand today because more businesses depend on the availability of their applications for business operations and employee productivity. As users come to expect their computers to work more reliably, every component of a system must be taken into consideration, including the boot/root/swap device. The range of solutions that have been used in the past are becoming unacceptable in many computing environments. For example, although you can create a simple backup of the root filesystem, if the root disk fails you will need to bring down the system to perform the restore operation.

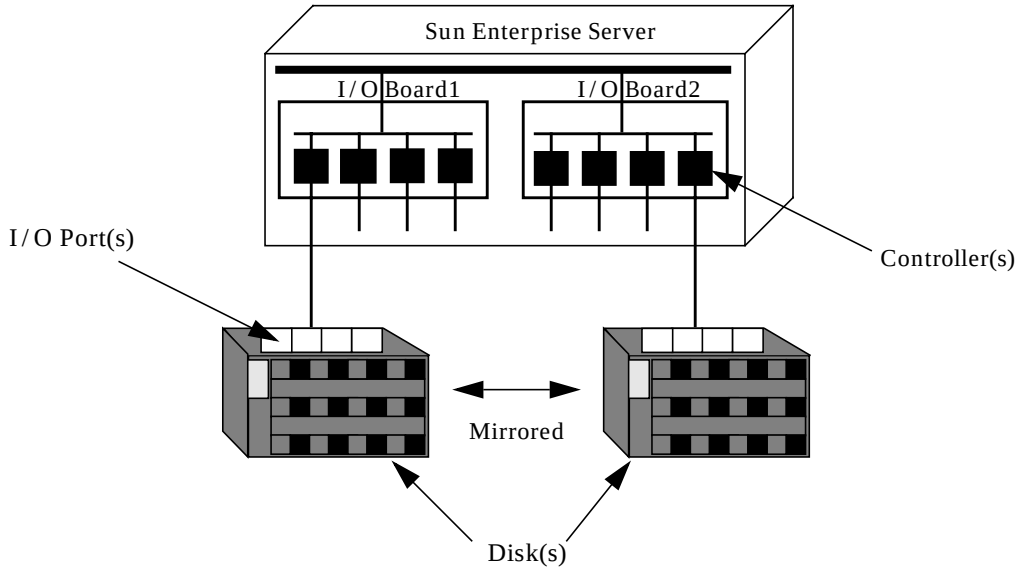
Today, systems should continue to run even in the event of a disk, link, power, or other failure. The main purpose of mirroring any disk is to keep the data accessible when a component fails. If the system disk is properly configured and can be hot swapped, the server does not have to be rebooted when the system disk fails.

In this article, I will examine ways to mirror your system disk, and I will take a look at some of the characteristics of the hardware that is involved. The first step in configuring for availability is to design the system without any single points of failure (SPOF). Many papers and articles talk about every component except the system disk. This is unwise, since you cannot run a server without the system disk. The system disk should always be a part of your high availability plan.

Hardware Configuration

When configuring hardware for high availability, every component involved in booting, accessing the root filesystem, and accessing the swap volume should support two separate paths. In addition, all components involved in mirroring your system disk must be designated as hot swap components. Keeping these things in

mind, a Sun Enterprise™ Server (such as the Sun Enterprise™ 450 or the Sun Enterprise 10000) should support dual I/O paths for the boot/root/swap devices shown in the following diagram.



In this figure, the boot/root/swap disk and mirror have separate I/O buses, I/O boards, I/O controllers, I/O links, and I/O ports. Of course, they also exist on separate disks.

Software Configuration

There are several software alternatives from which to choose when configuring boot, root, and swap to be highly available. The current product offerings are:

- Solstice DiskSuite™

  Solstice DiskSuite requires that you create state database replicas. A state database stores the Solstice DiskSuite configuration and state information. At least four database replicas are recommended, and this translates into four available partitions. The remaining system slices are then configured as single (one-way) concatenations/stripes, preparing them to become a submirror. A second disk is sliced in the same manner as the root disk. This disk is also placed

into a single concatenation/stripe, preparing it to be the second submirror. Before the system is rebooted, a one-way mirror is created from the first submirror. Then, the second submirror is attached. This creates a mirror of the original boot disk.

If a failure occurs on one side of the redundant hardware, the system is able to continue running on the hardware which is unaffected. You can replace the failed hardware while the system is operating. Solstice DiskSuite can continue to operate regardless of it's own majority consensus algorithm (which requires 51 percent of the total database replicas to be available for the system to reboot). Once you have replaced the broken hardware, you simply resynchronize the effected mirror.

- Sun Enterprise Volume Manager™

  Sun Enterprise Volume Manager (which is technically equivalent to Veritas Volume Manager) requires that your root disk be encapsulated. This imposes some restrictions on the type of filesystems you can place on the root disk. Sun Enterprise Volume Manager also requires two unassigned zero length slices for configuration and state information, known as the private and public regions. With Sun Enterprise Volume Manager, you must install the redundant disk as a new disk, which discards any previous data. You will have to reboot twice to enact the necessary system changes. You then mirror the root disk to the newly installed second disk. The mirrored disk contains Veritas Volumes which cannot be converted to a Solaris™ filesystem.

  If a failure occurs on one side of the redundant hardware, the system is able to continue running on the hardware which is unaffected. You can replace the failed hardware while the system is operating, and then take the appropriate steps to resync the root disk.

- Solstice DiskSuite with Alternate Pathing (AP)

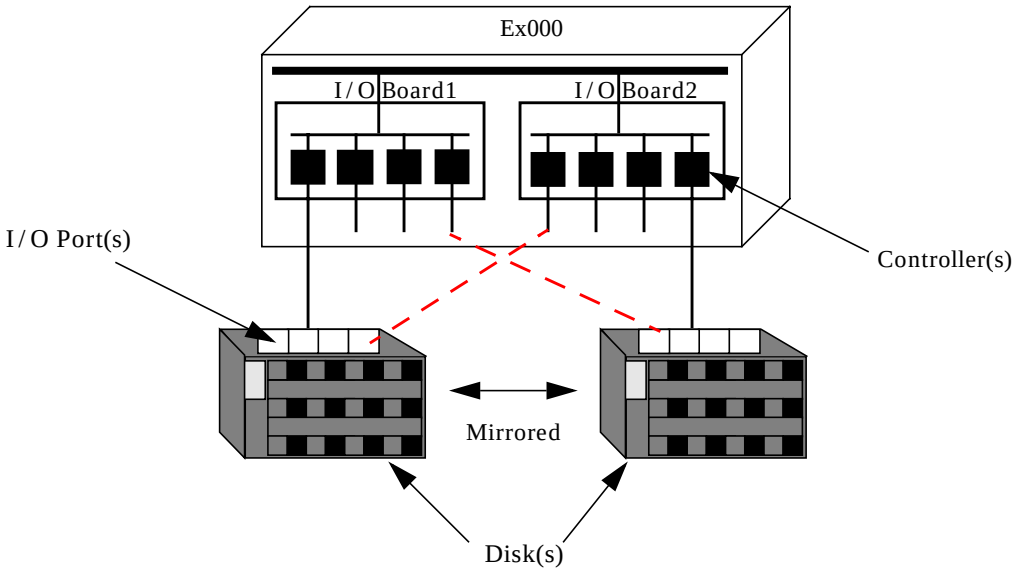
  If you use AP in conjunction with Solstice DiskSuite, the process of setting up and mirroring the boot/root/swap device is essentially the same as it would be without AP. One difference is that you must first configure AP. Then you must layer the Solstice DiskSuite configuration "on top of" AP by specifying AP metadevice names (rather than specifying physical device names). Also, you must set aside additional disk slices for the AP state database replicas. One of the benefits of using AP is that you can dynamically switch between the physical paths that are used to access your system disk.

- Sun Enterprise Volume Manager with Alternate Pathing (AP)

  To use Sun Enterprise Volume Manager with AP, you should first configure AP. Then use Sun Enterprise Volume Manager to encapsulate the appropriate AP metadevices in order to mirror the system disk. You must also set up the AP state databases replicas. It is usually not a good idea to place AP state database replicas on the root disk before you encapsulate the root disk with Sun Enterprise Volume Manager (since you will have to move or remove those database replicas if you

unencapsulate the root disk at any time in the future). One of the benefits of using AP is that you can dynamically switch between the physical paths that are used to access your system disk.

The last two options, above, use AP in conjunction with a volume manager. AP provides controller redundancy, and the volume manager provides data redundancy. The main purpose of AP is to help protect against I/O controller failures (and to support Dynamic Reconfiguration (DR), which is beyond the scope of this article). For disk controllers, an automatic switch operation occurs whenever a path failure is detected during normal operation. For network controllers, you must manually perform the switch operation if a failure occurs (with the current release of AP). Normally, when you add AP into the mix, two additional controllers and links would be added to the configuration. This doubles the number of available data paths, as displayed using red dashed lines in the following diagram.



It requires care to successfully recover from a failure while the system is operating. Recovery procedures should be well defined and understood. If you replace any hardware that involves data I/O, the affected data bus must be in a quiescent state. This is important because system applications may share the same data path that the system disk uses.

For more detailed information on configuring a volume manager and using Alternate Pathing in conjunction with a volume manager, please refer to the “Guide to High Availability: Configuring boot/root/swap,” Prentice Hall, July 1999, ISBN #0-13-016306-6. The book is available through http://www.sun.com/books, amazon.com fatbrain.com or Barnes & Noble bookstores.

Author's Bio: Jeannie Johnstone Kobert

Jeanie is a member of the Enterprise Engineering staff at Sun Microsystems. She joined Sun in 1997, with over nine years of UNIX experience. Her previous work experience with Intelligent Storage controllers brings a data-Guide center approach to high availability, which is the core of this BluePrint.