Configuring Boot Disks

John S. Howard, Enterprise Engineering and
David Deeths, Enterprise Engineering

Sun BluePrints™ OnLine—December 2001

http://www.sun.com/blueprints

Sun Microsystems, Inc.
4150 Network Circle
Santa Clara, CA 95045 U.S.A.
650 960-1300

Part No. 816-3887-11
Revision 1.0 3/7/03
Edition: December 2001

Configuring Boot Disks

Note – This article is the complete fourth chapter of the Sun BluePrints™ book, "Boot Disk Management: a Guide for the Solaris™ Operating Environment" by John S. Howard and David Deeths (ISBN 0-13-062153-6), which is available through www.sun.com/books, amazon.com, and Barnes & Noble bookstores.

This chapter presents a reference configuration of the root disk and associated disks that emphasizes the value of configuring a system for high availability and high serviceability. Although both of these qualities are equally important, the effort to support availability is much simpler than the effort to support serviceability. While you can easily achieve a high level of availability through simple mirroring, the effort involved in configuring a highly serviceable system is more complex and less intuitive. This chapter explains the value of creating a system with both of these characteristics, and outlines the methods used to do so. This chapter also addresses the following topics:

- Principles for boot disk configuration
- Features of the configuration
- Variations of the reference configuration

While the reference configuration reduces downtime through mirroring, the emphasis of this chapter is on easing serviceability burdens to ensure that when a system goes down, it can be easily and quickly recovered regardless of the situation or the staff on hand. While this configuration is useful in most enterprise environments, variations are presented to address a wide variety of availability and serviceability needs. In addition, this chapter is designed for modularity with respect to the other chapters in the book.

While nothing from this point forward in the book requires knowledge of the file system layouts and Live Upgrade (LU) volumes discussed in Chapters 1–3, the reference configuration uses this disk layout, and it may be helpful for you to be familiar with this information. The reference configuration is independent of a volume manager, and you can implement it using either VERITAS Volume Manager

(VxVM) or Solstice DiskSuite software. Despite independence from a specific volume manager, some things are implemented differently with different volume managers. For instance, Solstice DiskSuite software is unlikely to require a contingency disk because it is available on standard Solaris operating environment (OE) boot compact discs (CDs); however, VxVM is not on the boot CDs, and a contingency disk can be an effective way of reducing downtime when the boot image has been damaged.

For information about implementing the reference configuration using VxVM, see Chapter 5 “Configuring a Boot Disk With VERITAS Volume Manager.” For information about implementing the reference configuration using Solstice DiskSuite software, see Chapter 7 “Configuring a Boot Disk With Solstice DiskSuite Software.” Note that some of the procedures discussed in Chapter 5 and Chapter 7 are not obvious and are important even if you do not use the reference configuration.

Configuration Principles

With any architecture, there are trade-offs. The configuration proposed here promotes serviceability and recoverability at the expense of disk space and cost. While this may seem like a substantial trade-off, an investment in simplicity and consistency makes the configuration much safer and faster to recover should a failure occur. With the escalating cost of downtime, a system that you can quickly recover makes up the added cost of installation with the very first outage event. Likewise, a reference configuration that provides consistency throughout the enterprise reduces the likelihood of human mistakes that may cause failures.

In addition, you should consider the impact of having experienced personnel available when configuring and maintaining a system. While you can schedule installations when experienced system administrators who understand volume manager operations are on hand, the true value of an easily serviced and recovered system will be most apparent during an outage when experienced help is unavailable.

The following sections address key design philosophies for the reference configuration. Note that these same philosophies shaped the procedures used to install the boot disks in Chapter 5 and Chapter 7, particularly the choice to use the mirror, break, and remirror process during the VxVM boot disk setup.

Doing the Difficult Work at Installation Time

Setting up the boot disk and related disks with the steps used by the reference configuration presented in this book introduces several tasks on top of the standard procedures. While completing all of these tasks at once can be complicated and can

take more time than performing the default installation, doing so makes things simpler when service is needed. Because installations can be scheduled and controlled, it makes sense to spend a little more time up front to have a configuration that is simple, easy to service, and understood by everyone on the staff.

Striving for Simplicity

The configuration should be simple. Any system administrator with a moderate level of experience should be able to briefly look at the configuration to understand what is going on. There should be few, if any, exceptions or special cases for configuring various aspects of the boot disk.

Creating Consistency in All Things

This is a corollary to simplicity. The more cookie-cutter the configuration is, the more useful an administrator's experience becomes. An administrator who has gone through the recovery of one system, for example, can make that same recovery happen on any other system in the enterprise. Consistency in implementation makes this easier to achieve. In an inconsistent environment, each system poses new problems and a new learning curve that no one wants to tackle during a crisis. Because of this, the reference configuration present a configuration that is flexible enough to be useful in a variety of situations. Both Solstice DiskSuite and VxVM configurations benefit from increased consistency. For example, Solstice DiskSuite metadevice organization can be difficult to understand if an inconsistent naming scheme is used. For VxVM configurations, consistency plays an even bigger role.

Many of the problems in recovering or servicing a VxVM boot device come from the inconsistent configuration produced by the default installation. In a variety of ways, the boot disk is an exception in the world of VxVM. Encapsulating and mirroring the root disk may appear to generate a set of simple, identical disks, but this is not the case. There are several issues that make VxVM's default encapsulation far from ideal. These issues, including the geographic layout of the data, the location of the private region, and the order in which mirrors are attached to `rootdisk` volumes are examined in Chapter 5.

Designing for Resiliency

The reference configuration has designed out the possibility that a single hardware error (or device driver error) could cause an outage. All of the hardware elements that are necessary to support each mirror of the boot device are completely

independent of one another; no single point of failure (SPOF) is tolerated. The examples used to demonstrate our reference configuration use a Sun StorEdge D1000 array in a split configuration as a boot device.

Ensuring Recoverability

The reference configuration applies several layers of contingency to permit easy and rapid recovery. A mirror provides the first level of redundancy, and an additional mirror provides flexibility with backups and an additional level of redundancy. A contingency disk enables recovery even if there are problems with the volume manager setup or software.

To ensure recoverability, it is also important to test the finished configuration to ensure that everything works properly. Later chapters stress the importance of examining configuration changes and verifying proper operation.

Weighing Costs Against Benefits

While disks can be expensive in terms of cost, space, and administrative complexity, allocating an insufficient number of disks can be expensive, too. Although heroic efforts on the part of the system administration staff may be able to solve boot problems, these efforts may involve hours of expensive system administrator time. In addition, as servers become more connected (both to each other and to the lives of the people who use them), availability becomes increasingly important. When a server is unavailable, you might face the added costs of customer dissatisfaction, lost revenue, lost employee time, or lost billable hours. Fortunately, disks are becoming less expensive, and the availability gained by using three or four disks to manage the boot environment (BE) for an important server is usually well worth the price. Over the life of the machine, the cost of a few extra disks may indeed be a very small price to pay. Additionally, the configurations discussed here and in Chapter 5 and Chapter 7 are inherently more serviceable, and events such as upgrades will involve less downtime and less system administration hassle.

Reference Configuration Features

Four disks are used for the boot device and its entourage. The section “Reference Configuration Variations addresses the relative merits of several variations on this design with greater number of or fewer disks.

For VxVM installations, these four disks are the only items to be included in the root disk group (`rootdg`). Any data volumes or file system spaces to be created outside of the core operating system (OS) should reside in other disk groups. Because Solstice DiskSuite software does not partition disks into administrative groups (except in multihost environments), these four disks are not in any sort of separate group in Solstice DiskSuite software configurations.

The disk locations shown in the following table refer to the disk device names used in the examples throughout the book.

TABLE 0-1 Disks in the Reference Configuration

Disk Location	Disk Name
`c1t0d0s2`	`rootdisk`
`c1t1d0s2`	`rootmirror2`
`c2t8d0s2`	`rootmirror`
`c2t9d0s2`	`contingency`

Notice that the disk media name for each disk reflects its function. By providing clear and obvious naming, you can prevent confusion later. If you standardize these names throughout the enterprise, the potential for confusion is even further reduced. Note that `rootdisk` and `rootmirror` are on different controllers. These are the two SCSI host adapters that service each side of the Sun StorEdge D1000 array discussed in Chapter 1 “Partitioning Boot Disks.” Recall that all of the examples in this book use a Sun StorEdge D1000 array in a split configuration. The following paragraphs outline the purpose of each disk.

The *root disk* provides the basis of the BE. It includes the root volume and the `swap` volume. As described in Chapter 1, unless a more secure configuration is required, only one partition should be used to store the root volume (`root`, `usr`, `var`, and so forth). In addition to the root volume, an LU volume can be introduced on or off the boot disk to enable easier patch management and upgrades for the OS.

The *root mirror disk* provides redundancy for the root disk by duplicating all of the boot disk contents. This increases the availability of the system because the BE can still be reached through the root mirror if the boot disk is unavailable. It is important to have the root disk and root mirror on independent paths so the failure of a controller or an array will not adversely affect both of them. The goal of this configuration is to produce a root mirror that is physically identical to the root disk, thereby simplifying serviceability.

The *hot spare* or *additional root mirror* enables an even higher level of availability by acting as a spare for the root disk or root mirror if either fails. This can provide an additional level of redundancy and also reduces the effect of service delays on the redundancy of the system. Because there are only three mirrors in this scenario,

there is still a chance that a controller failure will leave the root disk unmirrored. This can be dealt with by using additional mirrors. An additional mirror is preferred to a hot spare in this situation because there is only one mirrored volume in `rootdg`. The time it would take a hot spare to resync to this mirror would reduce availability when compared to the time it would take to access the additional root mirror. Using a second mirror also allows flexibility because it can be broken off and used as an easy point-in-time backup during a complex service event.

The *contingency disk* allows a final level of protection. The contingency disk is a known-good BE. If certain boot files are accidentally modified or deleted, the boot disk may not boot properly. Since the boot mirror or hot spare mirrors these irregularities, the result is an inability to boot. Because some of these files are checked only at boot time, the problem could be months, or even years old before it is detected. The contingency disk provides a bootable environment, with any necessary volume manager and diagnostic utilities, that is frozen in time and not affected by changes to the boot disk. This enables you to quickly gain normal access to the machine in order to track down and repair the problems with the BE. Contingency disks are not as necessary in Solstice DiskSuite environments because their utilities are available on the bootable Solaris OE CDs.

LU volumes can be configured on one or more of these disks to provide additional options for bootability. If the BE on an LU volume is similar enough to the BE on the boot disk, this could allow the services hosted on the server to be brought up through the BE on the LU disk. Thus, the LU disk allows bootability, and possibly even service availability, even if the boot disk has been accidentally modified so that it cannot boot. If all of the data and applications are stored outside the root disk group, it is much more likely that a non-current disk will support both of these goals. LU volumes on the root disk or contingency disk can be used for this purpose. If these volumes exist on the contingency disk, they should be in addition to the known-good BE, which should be kept static.

Reference Configuration Variations

Obviously, the four-disk reference configuration described here is not ideal for all situations. The ideal environment for such a reference configuration is an enterprise-level computing environment with high-availability expectations. However, you can easily modify the reference configuration to meet a number of needs. In low- or medium-scale environments, or environments with less of an availability concern, the additional cost of a second root mirror, hot spare disk, or contingency disk may not justify the gain in availability. The following paragraphs describe the pros and cons of several variations of this design. Note that it is still a good idea to follow the procedures and suggestions in the rest of the book. For instance, even if several variations of the reference configuration are used in a datacenter, it is good to use

the same installation procedures and common naming conventions on the appropriate disks. Consistency is still the key to allowing system administrators to quickly and effectively service outages on an often-bewildering array of systems.

Although many concerns about boot disk configurations have already been addressed, there are really only two concerns to consider when choosing between variations on the reference configuration: disk failures and bootability failures. Disk failures are essentially random electronic or mechanical failures of the disk. Generally, the only remedy for a disk failure is to replace the disk. Bootability failures often involve human error and occur when the BE is unable to boot because of a misconfiguration or a problem with certain files or disk regions. Because bootability errors often affect the volume manager configuration or are mirrored to the root mirror or hot spare, the existence of those disks does not usually help the problem. While you can mitigate disk failures with root mirrors or hot spares, the remedy for bootability failures involves restoring the BE or booting from the contingency disk.

In a high-availability environment, it is essential that the restored BE or contingency disk has the programs, files, and patches to support the necessary services. Without a contingency disk, you can use any of the following methods to restore bootability:

- If you used Solstice DiskSuite software as the volume manager, you can boot from the Solaris OE installation CDs. Since these CDs contain Solstice DiskSuite software binaries, this provides all necessary Solstice DiskSuite utilities. Because this is usually a fairly easy option, Solstice DiskSuite software installations usually do not require a contingency disk.
- If a recent backup is available, you can use it to restore the boot disk.
- If the boot image was not heavily customized, you can reload it using the same JumpStart image, or by cloning a similar system.
- As a last resort, if good change control documentation is available, you can restore the BE by following the change control documentation; of course, if the change logs are on the boot disk, they will be of little help.

If none of these options are available, it may be extremely difficult and time-consuming to restore the BE to the point that it will support the necessary services. These types of outages are likely to last hours or even days, but could easily have been avoided by implementing any of the plans outlined above.

In systems using VxVM, storing non-OS information outside of `rootdg` alleviates many serviceability issues by eliminating the tendency to have application pieces on the boot disk and by making an alternate boot environment much more likely to support the necessary services. In systems running Solstice DiskSuite software, ensure that the boot disks and non-boot disks are as logically separate as possible.

Implementing Only a Mirrored Boot Disk

In some environments, it may make sense to use a configuration with only the root disk and the root mirror. While this will not achieve the same availability levels as the four-disk reference configuration, it is certainly better than a single-disk (non-mirrored) configuration. The availability level of a system with a mirrored root could vary greatly depending on the speed with which service staff detect and fix failures. It is important to remember that *both* disks need to be monitored. It does little good to have a root mirror if it is not in working condition when the root disk fails.

Having a root mirror generally provides a moderately high level of availability, though it may provide a high level of availability if the time-to-service is small. This availability level assumes that bootability errors are extremely rare, which is likely the case if the boot disk content is relatively static, or if stringent change control is in place. Workstations and machines that have a relatively simple, static configuration (especially where access is restricted) may work well with only a mirrored configuration. However, if the time to service is long, it is a good idea to have an additional mirror or a hot spare.

If occasional downtime is acceptable and the BE can be reinstalled easily, systems may be suited to a simple boot disk plus mirror configuration even if bootability errors are likely to be more common because the boot device is changed frequently or change control is poor. This could be the case for systems with a good backup and restore policy, or for systems that have simple BEs that can be started with JumpStart or reloaded easily. Redundant systems (such as one of a string of front-end web servers) may also be well-suited for this. In the case of redundant systems, a BE can be cloned from a similar system. This is discussed in detail in “Highly Available Services and Boot Disk Considerations” on page 185.

Using Additional Mirrors or a Mirror Plus Hot Spare

Both a hot spare and an additional mirror increase availability; however, the mirror provides better availability because there is no time spent synchronizing after a failure. The advantage of a hot spare is flexibility of which volume it will hot spare. If the only volumes present are on the root disk and root mirror, there is no gain in using hot-sparing over additional mirrors.

Unless there are mirrors in `rootdg` besides the root mirror, hot-sparing does not make sense with VxVM. Because only boot disks should be placed in `rootdg`, a hot spare almost never makes sense in `rootdg` for VxVM.

Since Solstice DiskSuite software does not allow disks to be put into management groups (except in multihosted environments), a hot spare could service disks outside the boot disk and boot mirror. While this could be advantageous to the availability

of other disks, it could be detrimental to the boot disk's availability. It is important to appropriately match the number of hot spares to the number of mirrors and carefully monitor hot spare use so that hot spares are always available.

A boot disk with more than two mirrors works well in most of the same sorts of environments as the simple mirrored boot disk configurations. However, the additional mirror affords increased availability. This is not as important in configurations where the time-to-service is short; but if detecting and fixing problems takes a long time, the additional mirror provides a huge availability advantage over the simple mirror.

This configuration works well in situations where bootability errors are unlikely and service is relatively slow. In some cases, the boot disk may not be monitored at all. If this is the case, an additional mirror or hot spare is especially critical.

Having an even greater number of additional mirrors or hot spares further decreases the likelihood of having disk errors on all disks in the same time window. Additional mirrors or hot spares also provide disk-level redundancy, even in the event of a controller failure. Having two mirrors on two controllers provides data redundancy even if a controller is lost. The availability advantage here is too small to be worth the cost of disks in most situations; however, for configurations with long service times or configurations where availability is of paramount importance, it may be a good idea.

Using Mirrored Boot Disk With Contingency Disk

For environments where bootability failures are common, such as a server supporting a complex set of applications that are heavily tied to the BE, it may be more important to have a contingency disk than an additional mirror. In these types of environments, it is likely that there are lots of people involved in the configuration, making it more likely that disk failures will be detected and fixed. This means that the advantage of an additional mirror is lessened. While it is best for both an additional mirror and a contingency disk to be present, it is not always possible. Given the choice between one of the two, a complex, changing environment probably reaches a better overall availability level with a contingency disk.

As with mirrors, it is possible to have multiple contingency disks. While having contingency disks available on multiple controllers may improve availability, the effect is likely to be negligible, even on systems seeking a very high level of availability. An advantage to multiple contingency disks is the ability to keep one disk updated with the current BE, while keeping the other entirely static. However, this task is probably better relegated to LU volumes, which can manage BEs in a more intuitive way. If you follow the suggestions in Chapter 1, LU volumes could be

available on the boot disk if it is still working, or on the contingency disk. Keeping one or more LU volumes on the contingency disk is relatively easy because today's disks are large enough that the root volume is unlikely to fill even half of the disk.

Note that LU volumes should be used in combination with the contingency disk, not as a replacement for it. Using LU adds some additional complication, so it is still important to have a known-good environment on the contingency disk that is as unaffected by complexity as possible. This includes being unaffected by bugs or misconfigurations involving LU.

LU volumes can serve as a quick-fix in a crisis, but this is not their intended use. It is important to have a contingency disk to fall back on. Since the intent of LU volumes is enabling easy upgrades and sidegrades, that should be their primary use. Using LU volumes as emergency boot media may be possible in some situations, but they lack the fail-safe nature of the contingency disk.

If a bootability error occurs, you can attempt to boot using the most up-to-date LU volume. If the offending change was made after the last update, the disk will boot and should be close enough to the current environment to support the necessary services. If the offending change was made before the last update, the latest LU volume may not boot or provide the necessary services, but the contingency disk or an older LU volume should. Even if the static contingency disk is not current enough to support the necessary applications, having a BE up quickly enables easy access to the underlying volumes and faster serviceability, leading to less downtime.

Summary

This chapter presented a reference configuration that enables high availability and serviceability. Used in conjunction with the file system and hardware suggestions outlined in Chapter 1, and the volume manager setups described in Chapter 5 and Chapter 7, this reference configuration forms the foundation for a simple, consistent, resilient, and recoverable boot disk solution. Variations of the configuration explained the design of the reference configuration, as well as the advantages and disadvantages of different modifications to the configuration. At the very least, this chapter provided an understanding of the importance of optimizing boot disk availability and serviceability.