Sun StorEdge™ Network Data Replicator 3.0/3.0.1

Software Architecture Guide

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303-4900 U.S.A.
650-960-1300

Part No. 816-3792-10
December 2001, Revision A

Send comments about this document to: docfeedback@sun.com

Please
Recycle

Contents

Preface

This document describes the software architecture and operation for the Sun StorEdge™ Network Data Replicator (Sun SNDR) Versions 3.0 and 3.0.1 software. The intended audience includes Sun support engineers and system administrators.

How This Book Is Organized

See TABLE P-1 for package differences between Versions 3.0 and 3.0.1.

Chapter 1 describes the basic architecture of the Sun SNDR software.

Chapter 2 describes the replication and synchronization modes.

Chapter 3 describes the Sun SNDR software's bitmap management and behavior during synchronization.

Chapter 4 describes miscellaneous topics such as the I/O grouping, using the software in a cluster, and using the software with the Sun StorEdge Instant Image software.

Using UNIX Commands

This document might not contain information on basic UNIX® commands and procedures such as shutting down the system, booting the system, and configuring devices.

See one or more of the following for this information:

- *Solaris Handbook for Sun Peripherals*
- AnswerBook2™ online documentation for the Solaris™ operating environment
- Other software documentation that you received with your system

Typographic Conventions

Typeface or Symbol	Meaning	Examples
`AaBbCc123`	The names of commands, files, and directories; on-screen computer output.	Edit your `.login` file. Use `ls -a` to list all files. `% You have mail.`
`AaBbCc123`	What you type, when contrasted with on-screen computer output.	`% su` `Password:`
AaBbCc123	Book titles, new words or terms, words to be emphasized. Command-line variable; replace with a real name or value.	Read Chapter 6 in the *User's Guide*. These are called *class* options. You *must* be `root` to do this. To delete a file, type `rm` *filename*.
[ ]	In syntax, brackets indicate that an argument is optional.	`scmadm` [`–d` *sec*] [`–r` *n*[:*n*][,*n*]...] [`–z`]
{ *arg* | *arg*}	In syntax, braces and pipes indicate that one of the arguments must be specified.	`sndradm -R b {p | s}`

\	At the end of a command line, the backslash (\) indicates that the command continues on the next line.	`atm90 /dev/md/rdsk/d5 \`
		`/dev/md/rdsk/d1` `atm89 \`
		`/dev/md/rdsk/d5 /bitmaps/map2 \`
		`ip sync`

Shell Prompts

Shell	Prompt
C shell	*machine-name*%
C shell superuser	*machine-name*#
Bourne shell and Korn shell	$
Bourne shell and Korn shell superuser	#

Related Documentation

Note – You can use the *Sun StorEdge Network Data Replicator 3.0 System Administrator's Guide, Sun Cluster 3.0 U1 and Sun StorEdge 3.0 Software Integration Guide,* and *Sun StorEdge Network Data Replicator 3.0 Configuration Guide* with the Sun SNDR Version 3.0.1 software.

For the latest version of storage software documentation, go to:

`http://www.sun.com/products-n-solutions/hardware/docs/Software/`

Application	Title	Part Number
Man pages	`sndradm`(1M) `dscfg`(1M) `file`(1M) `fwcadm`(1M) `pkgadd`(1M) `pkgrm`(1M) `scmadm`(1M) `svadm`(1M)	N/A
Release	*Sun StorEdge Network Data Replicator 3.0.1 Release Notes*	806-7513
	Sun Cluster 3.0 U1 and Sun StorEdge Software 3.0 Release Note Supplement	816-2136
	Sun StorEdge Instant Image 3.0.1 Release Notes	806-7678
Sun Cluster with Sun StorEdge software	*Sun Cluster 3.0 U1 and Sun StorEdge Software 3.0 Integration Guide*	816-1544
Installation and user	*Sun StorEdge Instant Image 3.0.1 Installation Guide*	806-7675
	Sun StorEdge Network Data Replicator 3.0.1 Installation Guide	806-7514
	SunATM 3.0 Installation and User's Guide	805-0331
	SunATM 4.0 Installation and User's Guide	805-6552
	Sun Gigabit Ethernet FC-AL/P Combination Adapter Installation Guide	806-2385

Application	Title	Part Number
	Sun Gigabit Ethernet/S 2.0 Adapter Installation and User's Guide	805-2784
	Sun Gigabit Ethernet/P 2.0 Adapter Installation and User's Guide	805-2785
	Sun Enterprise 10000 InterDomain Networks User Guide	806-4131
System administration	*Sun StorEdge Network Data Replicator 3.0 System Administrator's Guide*	806-7512
	Sun StorEdge Instant Image 3.0 System Administrator's Guide	806-7677
	TCP/IP and Data Communications Administration Guide	805-4003
	System Administration Guide, Volume 3 (for the Solaris 8 operating environment)	806-0916
	Sun StorEdge Fast Write Cache 2.0 System Administrator's Guide	806-2064
Configuration	*Sun StorEdge Network Data Replicator 3.0 Configuration Guide*	806-7550
	Sun StorEdge Instant Image 3.0 Configuration Guide	806-7676
	Sun Enterprise 10000 InterDomain Network Configuration Guide	806-5230

Package Differences

TABLE P-1 shows the package differences between the versions in **boldfaced** text.

TABLE P-1 Sun SNDR Version 3.0 and 3.0.1 Package Differences

	Sun SNDR Version 3.0	Sun SNDR Version 3.0.1
Marketing Part Number	NWDRS-300-99Y9 (English) NWDRS-300-99YS (Localized)	Same as Version 3.0
Software Media	Sun SNDR and core services software part number 724-6969-01 (English) part number 724-7033-01 (Localized)	Sun SNDR and core services software part number 724-6969-**02** (English) part number 724-7033-**02** (Localized; available 30 to 60 days after English release)
Operating Environment	Solaris 7 8/99 (also known as Update 3) Solaris 7 11/99 (Update 4) Solaris 8 Solaris 8 6/00 (also known as Update 1) Solaris 8 10/00 (Update 2) Solaris 8 01/01 (Update 3) Solaris 8 04/01 (Update 4)	**Solaris 2.6 05/98 with these patches:** **105181-28 - kernel super patch** **106639-06 - rpcmod** Solaris 7 8/99 (also known as Update 3) Solaris 7 11/99 (Update 4) Solaris 8 Solaris 8 6/00 (also known as Update 1) Solaris 8 10/00 (Update 2) Solaris 8 01/01 (Update 3) Solaris 8 04/01 (Update 4)
Build Version	Base build 3.0.28	Same as Version 3.0
Patches Installed During Install Process	None	**111945-02 - Storage Cache Manager** **111946-02 - Storage Volume Driver** **111947-01 - Instant Image (if it is also installed with the Sun SNDR software)** **111948-02 - Sun SNDR software** **112046-01 - Solaris 2.6 compatibility**

TABLE P-1 Sun SNDR Version 3.0 and 3.0.1 Package Differences *(Continued)*

	Sun SNDR Version 3.0	Sun SNDR Version 3.0.1
Installation Updates	None.	**Core services software CD includes updated files for these packages:** • **core - `probe_script`** • **`SUNWnvm` - postinstall** • **`SUNWscmu` - preinstall and postinstall** • **`SUNWspsvu` - postinstall** **Sun SNDR software CD includes updated files for these packages:** • **`SUNWrdcu` - postinstall**
Supporting Software	Any TCP/IP network transport software such as SunATM™ or Gigabit Ethernet transports	Same as Version 3.0
Sun Cluster 3.0 Support	The Sun StorEdge Versions 3.0 and 3.0.1 services software *is not cluster-tolerant* in the initial release of the Sun Cluster 3.0 software. The software is not expected to fail over or fail back when a Sun Cluster logical host fails over and fails back.	
	The Sun StorEdge Version 3.0 services software installed with patches is cluster-aware in a two-node, Sun Cluster 3.0 Update 1 software environment.	The Sun StorEdge Version 3.0.1 services software as installed is cluster-aware in a two-node, Sun Cluster 3.0 Update 1 software environment.
	It can coexist with the Sun Cluster 3.0 U1 environment and fails over and fails back as the logical host containing the software product fails over and fails back. A Sun Cluster aware product can then be made highly available by utilizing the High Availability framework that Sun Cluster provides.	It can coexist with the Sun Cluster 3.0 U1 environment and fails over and fails back as the logical host containing the software product fails over and fails back. A Sun Cluster aware product can then be made highly available by utilizing the High Availability framework that Sun Cluster provides.
Servers	Sun Enterprise™ server models 2x0 through 4x0 Sun Enterprise server models 3x00 through 10000 Sun Fire™ server models 3800, 4800, 4810, and 6800	Same as Version 3.0

TABLE P-1 Sun SNDR Version 3.0 and 3.0.1 Package Differences *(Continued)*

	Sun SNDR Version 3.0	Sun SNDR Version 3.0.1
TCP/IP Connection Hardware	The Sun SNDR software requires a TCP/IP connection between the primary and secondary server. A dedicated TCP/IP link is not required. Each server must have the proper ATM or Ethernet hardware installed to support the TCP/IP link. The Sun SNDR software operates over any TCP/IP networking technology but has been qualified only on 10, 100, and 1000 Mbit Ethernet and ATM 155 and 622 technologies.	Same as Version 3.0
Documentation	*Sun StorEdge Network Data Replicator 3.0 Release Notes* **Sun StorEdge Network Data Replicator 3.0 Installation Guide* **Sun StorEdge Network Data Replicator 3.0 System Administrator's Guide* **Sun StorEdge Network Data Replicator 3.0 Configuration Guide* * On product CD media	*Sun StorEdge Network Data Replicator* **3.0.1** *Release Notes* **Sun StorEdge Network Data Replicator* **3.0.1** *Installation Guide* **Sun StorEdge Network Data Replicator 3.0 System Administrator's Guide* **Sun StorEdge Network Data Replicator 3.0 Configuration Guide* ***Sun Cluster 3.0 U1 and Sun StorEdge Software 3.0 Integration Guide*** ***Sun Cluster 3.0 U1 and Sun StorEdge Software 3.0 Release Notes*** * On product CD media

Accessing Sun Documentation Online

A broad selection of Sun system documentation is located at:

`http://www.sun.com/products-n-solutions/hardware/docs`

A complete set of Solaris documentation and many other titles are located at:

`http://docs.sun.com`

Ordering Sun Documentation

Fatbrain.com, an Internet professional bookstore, stocks select product documentation from Sun Microsystems, Inc.

For a list of documents and how to order them, visit the Sun Documentation Center on Fatbrain.com at:

`http://www.fatbrain.com/documentation/sun`

Sun Welcomes Your Comments

Sun is interested in improving its documentation and welcomes your comments and suggestions. You can email your comments to Sun at:

`docfeedback@sun.com`

Please include the part number (8xx-xxxx-xx) of your document in the subject line of your email.

CHAPTER 1

Overview

This chapter contains the following topics:

The Sun SNDR Software Architecture

Note – In this document, a volume is a raw disk partition or a volume created by a volume manager.

The Sun SNDR software is a remote replication facility for the Solaris™ operating environment. It is a Sun StorEdge service (as is the Sun StorEdge Instant Image software).

The Sun SNDR software enables you to replicate disk volumes between physically separate primary and secondary sites in real time. To transport data, the Sun SNDR software uses any Sun network adapter that supports TCP/IP.

It is designed to be active during normal application access to the data volumes and continually replicates the data to the remote site. Think of the Sun SNDR software as mirroring software which operates at the volume level on storage attached to two or more hosts that communicate using TCP/IP.

You can update the data on the secondary volume by issuing a command to *synchronize* the primary and secondary volumes. You can also restore data from the secondary volume to the primary volume by issuing a command to *reverse resynchronize* the volumes.

The Sun SNDR software uses volume sets that you define. A volume set consists of a primary volume residing on a local site and a secondary volume residing on a remote site. The volume set also includes a bitmap volume on each site to track write operations and differences between the volumes.

You can use RAID volumes as part of your Sun SNDR software strategy. Volumes can be any RAID level. The RAID levels of volumes in a volume set do not have to match.

Network Protocols and TCP/IP Connection

The Sun SNDR software requires a TCP/IP connection between the primary and secondary hosts. The software communicates across this network using kernel Remote Procedure Calls (RPCs). See FIGURE 1-1.

Each host must have the proper ATM or Ethernet hardware installed to support the TCP/IP link. The Sun SNDR software operates over any TCP/IP networking technology but has been qualified only on 10, 100, and 1000 Mbit Ethernet and ATM 155 and 622 technologies. A dedicated TCP/IP link is not required.

Although the Sun SNDR software is most likely to be used with SunATM link-level interfaces, the Sun SNDR software can be used with any Sun-supported link-level interface that is TCP/IP-capable, such as Gigabit Ethernet, Gigabit Ethernet Fibre Channel, and others.

When using ATM, ensure that the configuration supports TCP/IP by using either Classical IP or LAN Emulation. For more information on configuring the SunATM interface for these protocols, refer to the *SunATM Installation and User's Guide*.

The Software in the Sun StorEdge Services Stack

The architecture of the Sun SNDR software in the kernel I/O stack is shown in FIGURE 1-1.

The Sun StorEdge services are implemented as layered pseudo-device drivers in the Solaris kernel I/O stack. As shown in FIGURE 1-1, the Sun SNDR software resides in the network storage control kernel module `nsctl` framework above the volume manager or the storage device driver and below the file system.

These drivers rely on the `nsctl` framework to support this layering and to provide runtime control. The Sun SNDR software is implemented as an `nsctl` I/O filter module, enabling it to be integrated with other Sun StorEdge services. By being in the data path, the Sun SNDR software transparently provides remote replication capabilities.

This architecture makes the Sun SNDR software independent of a volume manager or file system. However, implementation below the file system results in limitations on the software. In particular, the Sun SNDR software does not have file system-specific information and cannot perform file system level replication. See "About Replicating File Systems" on page 5.

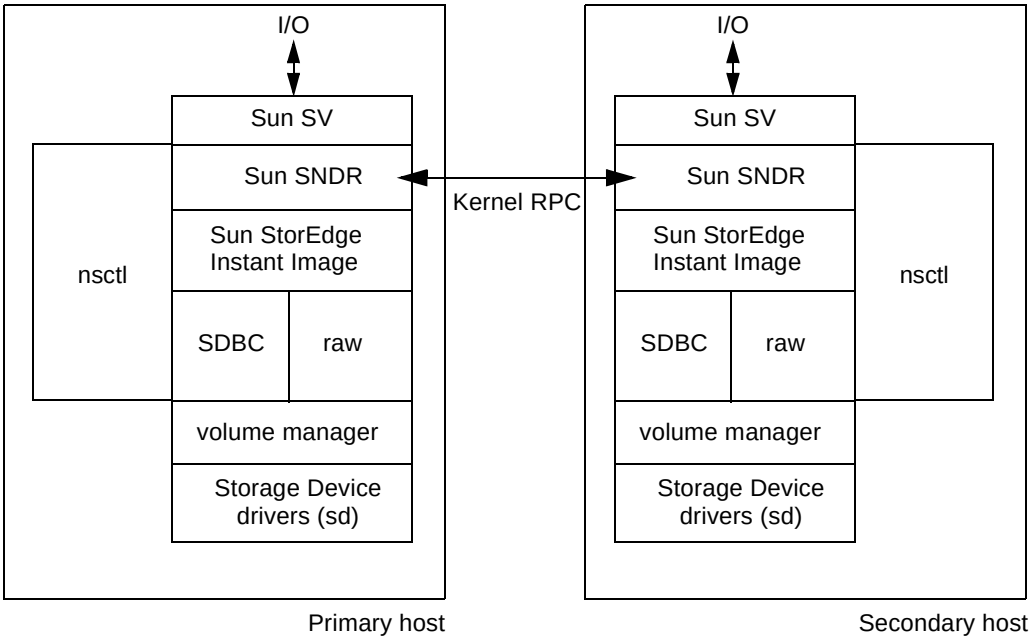
Kernel RPCs allow the software to make procedure calls on other Sun SNDR hosts across a network:

1. A Sun SNDR process sends a request to a Sun SNDR process on another host, which sends a reply to the originator.

2. The kernel RPC facilities route and authenticate these messages through ports registered in the kernel by the Sun SNDR participants on each host. Messages are packaged for transport by using the External Data Representation (XDR) library routines.

Data flows to the Sun SNDR driver from user-layer applications on the primary host through the Sun StorEdge SV layer. Sometimes user-layer applications reside above the file system. Other times these applications run in Data Base Management Systems (DBMS) that might read and write directly to raw partitions. In any case, I/O commands process the data to its destination on the storage device.

The I/O commands targeted to Sun SNDR volumes are intercepted by the SV driver and routed through the Sun StorEdge I/O stack before being passed on to the storage device driver or the volume manager. The SV driver is a very thin layer in the I/O stack and operates by interposing onto the DDI entry points to the underlying device driver. The I/O commands originating in user space are intercepted at the top of the Sun StorEdge Service I/O stack. The SV driver routes them through the stack and feeds them back to the storage device driver or volume manager at the bottom of the stack. Data also flows in the opposite direction, from the storage back to user space.



1. I/O commands and data enter and exit the Sun SNDR software through the Sun StorEdge Storage Volume (SV) driver software.
2. Mediated by `nsctl`, the data flows through Sun SNDR (and optionally the Sun StorEdge Instant Image software) and the Storage Device Block Cache (SDBC) drivers to its destination on the storage array or in user space.

FIGURE 1-1 Sun SNDR Software in the Sun StorEdge Services Stack

About Replicating File Systems

If a file system is replicated, remember that the Sun SNDR software is not a file system replicator but a volume replicator. When you replicate a Sun SNDR volume that contains a file system, the secondary host volume gets an exact copy of the bits on the primary host volume, including any file systems on that volume.

When replicating, the primary host file system is mounted. Do not mount the file system on the secondary host until you are ready to fail over to that site. Changes appear on a replicated file system volume only after a remount.

Also, a file system on secondary host can be mounted only in read-only mode while the Sun SNDR volume set continues to replicate. Once secondary host volumes are placed into logging mode, the file system can be mounted for read / write operations.

Host Relationships and Logging

The relationship to the origin of the user-layer application I/O determines whether one volume is configured as the primary volume and another is configured as the secondary volume.

User-layer applications on the primary host generate the I/O intercepted by the SV driver. The primary host is defined as the machine that hosts:

- User-layer applications
- The storage containing the primary volume

Because of its relationship to the origin of the user layer application I/O, the primary host is also referred to as the *local host*. As a practical matter, the host where user-layer application I/O originates must be configured as the primary host.

The *secondary host* is defined as the machine that hosts the storage containing the secondary volume. The secondary volume of a volume set cannot be mounted on the secondary host; it must be unmounted. The I/O on the secondary volume is mediated entirely by the Sun SNDR software. The secondary host is also referred to as the *remote host*.

This relationship is the normal operating mode for the Sun SNDR software, except when the volume set is placed in *logging* mode and I/O from user-layer applications has been enabled on the secondary host. For example, this situation occurs when the network link between the two hosts attached to the mirrored volumes breaks or is disconnected. In logging mode, the Sun SNDR software stops actively copying data between the two systems and tracks the I/O on both systems using the volume set's bitmap volumes. See also "Logging" on page 26.

The SUNWnvm 3.0 Package and Fast Write Cache

Note – You cannot use the Sun StorEdge Fast Write Cache (FWC) product (all versions, including the SUNWnvm Version 3.0 software) in any Sun Cluster environment because cached data is inaccessible from other machines in a cluster. To compensate, you can use a Sun caching array.

The Sun StorEdge core services Versions 3.0/3.0.1 CD contains the Sun StorEdge SUNWnvm Version 3.0 software package. This package is intended for those users whose systems include Version 2.0 of the Sun FWC hardware and software product and who wish to continue using the Sun FWC product. FWC Version 2.0 is incompatible with the Sun StorEdge Versions 3.0/3.0.1 services software.

The Sun StorEdge services Installation Guides contain more information about removing Version 2.0 and installing the SUNWnvm 3.0 package.

The SUNWnvm 3.0 package and NVRAM boards reduce the frequency of disk I/O access by caching the written data blocks in nonvolatile memory and then destaging the cached data to disk asynchronously.

Terminology

TABLE 1-1 Sun SNDR Terminology

Asynchronous replication	Asynchronous replication confirms to the originating host that the primary I/O transaction is complete before updating the remote image. That is, completion of the I/O transaction is acknowledged to the host when the local write operation is finished and the remote write operation has been queued. Deferring the secondary copy removes the long distance propagation delays from the I/O response time.
Fast Resynchronization	See Update synchronization.
Fast reverse synchronization	See Reverse synchronization.
Full synchronization	Full synchronization performs a complete volume-to-volume copy, which is the most time-consuming of the synchronization operations. In most cases, a secondary volume is synchronized from its source primary volume. However, restoration of a failed primary disk might require reverse synchronization, using the surviving remote mirror as the source.
Logging	Mode where a bitmap tracks writes to a disk, rather than a running log of each I/O event. This method tracks disk updates that have not been remotely copied while the remote service is interrupted or impaired. The blocks that no longer match their remote sets are identified for each source volume. The Sun SNDR software uses this log to re-establish a remote mirror through an optimized update synchronization rather than a complete volume-to-volume copy.
Primary or local: host or volume	The system or volume on which the host application is principally dependent. For example, this is where the production database is being accessed. This data is to be replicated to the secondary by the Sun SNDR software.
Replication	Once a volume set has been initially synchronized, the software ensures that the primary and secondary volumes contain the same data on an ongoing basis. Replication is driven by user-layer application write operations; Sun SNDR replication is an ongoing process.
Reverse synchronization	An operation used during recovery rehearsals. Logging keeps track of test updates applied to the secondary system during the rehearsal. When the primary is restored, the test updates are overwritten with the blocks from the primary image, restoring matching remote sets.
Secondary or remote: host or volume	The remote counterpart of the primary, where data copies are written to and read from. Remote copies are transmitted without host intervention between peer servers. A server might act as primary storage for some volumes and secondary (remote) storage for others.

TABLE 1-1 Sun SNDR Terminology *(Continued)*

Synchronization	The process of establishing an identical copy of a source disk onto a target disk as a precondition to the Sun SNDR software mirroring.
Synchronous replication	Synchronous replication is limited to short distances (tens of kilometers) because of the detrimental effect of propagation delay on I/O response times. The completion of I/O operations is only acknowledged after the local write and the remote write operations have both finished.
Update synchronization	Update synchronization copies only those disk blocks identified by logging, reducing the time to restore remotely mirrored sets.

In FIGURE 1-1, SDBC is the software that provides caching functionality. If the system contains the FWC hardware (NVRAM boards), the cache is placed in write-behind mode. In write-behind mode, writes are copied to NVRAM and acknowledged. Later, the write blocks are destaged from host memory. This scheme lowers the latency for small writes, provides for write cancellation, and allows small sequential writes to be coalesced into a larger single write to disk. Read caching is provided in systems with or without FWC hardware.

With the `SUNWnvm` 3.0 package, the Sun SNDR software performance can be improved because of the decreased latency in issuing the write and receiving the acknowledgment. The performance improvements are apparent during synchronous replications, as I/O acknowledgments are confirmed by the FWC driver and the I/O disk access latency is eliminated at both the local and remote sites.

CHAPTER 2

Replication and Synchronization Modes

The chapter contains the following topics:

Replication Modes

The Sun SNDR software supports two modes of data replication: *synchronous replication* and *asynchronous replication*. The replication mode is a user-selectable parameter for each Sun SNDR volume set. (Use the `sndradm` enable command and select the volume set's `sync` or `async` parameter. Use the `sndradm -R m` command to change the replication mode thereafter.) The volumes can be updated synchronously in real time or asynchronously using a store-and-forward technique.

Typically, a primary volume is first explicitly copied to a designated secondary volume to establish matching contents. As applications write to the primary volume, the Sun SNDR software replicates changes to the secondary volume, keeping the two volumes consistent.

In the event of planned or unplanned outages, the Sun SNDR software maintains per-device bitmap volumes that are marked to indicate changed blocks with a granularity of 32 Kbytes per segment. This technique allows for optimized resynchronization by allowing the Sun SNDR software to resynchronize only the blocks that have changed.

Synchronous Replication

In synchronous mode, a write operation is not confirmed as complete until the remote volume has been updated. Synchronous mirroring forces the Sun SNDR software to wait until an acknowledgement of the receipt of the data is received from the secondary by the primary before returning to the application. The application is not acknowledged until the write at the secondary site is complete.

The advantages of synchronous replication are that it is more reliable and can help reduce the risk of data loss; one disadvantage might be an increase in response time, especially for large data sets or long distance replication

Data Flow

1. The application issues a write to the file systems or raw device on the primary site.
2. Write goes into the Sun SNDR software layer where a bit is set in the bitmap for the data that is being requested to be written.
3. Data is written to the local disk.
4. The Sun SNDR software on the primary site sends data to the Sun SNDR software on secondary site to be replicated.
5. Data is received by the software on the secondary site.
6. The software on the secondary site issues a write request of data.
7. Data is written to the disk on the secondary site.
8. Once the write is committed to the disk or stored on NVRAM on the secondary site, the Sun SNDR software on the secondary site receives an ACK (acknowledgment).
9. The software on the secondary site sends the ACK to the Sun SNDR software on the primary site.
10. The Sun SNDR software on the primary site receives the ACK.
11. The Sun SNDR software on the primary site clears the bit in the bitmap.
12. The Sun SNDR software on the primary site informs the application that a write has been committed.
13. The application issues the next write request.

Asynchronous Replication

In asynchronous mode, a write operation is confirmed as complete before the remote volume has been updated. Asynchronous mirroring allows the Sun SNDR software to return to the host as soon as the write has been completed on the primary volume and been placed on a per-volume queue for the secondary site. The software queues local writes for later transmission to the remote host.

Subsequently, the secondary site receives the queued requests in the order that they were queued. Once the I/O has been completed at the secondary site, notification is sent to the primary. The remote image is updated after the I/O complete signal is sent to the local host.

The advantages of asynchronous replication are that it provides fast response and has the least impact on the response time of the primary application. Here, the long-distance network pipe becomes the bottleneck, forcing local writes to be queued for later transmission. The disadvantage is that there is a possibility of more data loss at the secondary site after primary site or network failures.

Data Flow

1. The application issues a write to the file systems or raw device on the primary site.
2. Write goes into the Sun SNDR software layer on the primary site where a bit is set in the bitmap for the data being requested to be written.
3. Data is written to the local disk.
4. The Sun SNDR software on the primary site puts the data in the queue of data that needs to be transferred.
5. The application on the primary site is informed of the write completion.
6. Data from the queue is transferred on a FIFO basis from the primary site to the secondary site.
7. Data is received by the Sun SNDR software on the secondary site.
8. The Sun SNDR software on the secondary site issues a write request of data.
9. Data is written to the disk on the secondary site.
10. Once the write is committed to the disk on the secondary site, the Sun SNDR software on the secondary site sends the ACK to the Sun SNDR software on the primary site.
11. A bit is cleared in the bitmap if the reference count is zero (0). See “Reference Counts” on page 34.

Things to Consider Before Using Asynchronous Replication

Queue size - When planning an asynchronous strategy, consider the size of the queue of data that needs to be transferred. If this queue becomes full, the Sun SNDR software is designed to throttle back the I/O to avoid an out-of-memory condition in the kernel. If this condition occurs, the I/O latency will match the drain rate of the queue and applications requesting I/O on the volume will slow. See "Performance Considerations" on page 36 for a discussion of potential performance bottlenecks.

Write ordering - Consider if it is important that the remote writes are applied in the order in which they were posted by the source application. When two separate processes (threads) running asynchronously to each other attempt to copy their respective volumes to remote disks, there is no guarantee that the relative order will be preserved. A single thread of update is required somewhere in the replication process to support order-dependent writes.

Write-order dependencies often arise with the replication of database management systems (DBMS), where the system extends across multiple volumes that must be in a self-consistent state to ensure system integrity. If volume replication must be done asynchronously for performance reasons, Sun SNDR I/O groups may be employed to ensure order-dependent writes. See "I/O Groups" on page 36.

Alternatively, points of consistency might be obtained following short periods of quiesced I/O to ensure that all records have been posted remotely. Replication is usually stopped at that time for secondary access. The Sun StorEdge Instant Image software can be combined with the Sun SNDR software when using this approach. See "Sun SNDR Software with Sun StorEdge Instant Image Software" on page 38.

Synchronization Modes

Note – See Chapter 3 to read about bitmap behavior during synchronization.

When a volume set is enabled using the `sndradm -e` command, you must initially synchronize the primary and secondary volumes in the set (use the `sndradm -E` command if the volumes are already identical). Using one of the Sun SNDR synchronization modes ensures that the primary and secondary volumes contain the same data and that they are identical at a clearly defined time. Synchronization is driven by the software through the `sndradm` command and progresses to completion.

Once a volume set has been synchronized, the software ensures that the primary and secondary volumes contain the same data on an ongoing basis through *replication*. Replication is driven by user-layer application write operations; Sun SNDR replication is an ongoing process.

The Sun SNDR software synchronizes data in a *forward* (from the primary to the secondary) or *reverse* (from the secondary to the primary) direction.

The Sun SNDR software synchronizes data in five modes:

- Full synchronization
- Fast resynchronization (also known as an update)
- Reverse synchronization
- Fast reverse synchronization (also known as a reverse update)
- Logging

The Sun SNDR software provides two methods for synchronization after a scheduled or unscheduled link failure:

- Automatic, where synchronization occurs automatically when the link is reestablished
- Manual, where synchronization requires operator intervention

When Not To Resynchronize

Resynchronization is discouraged if the Sun SNDR software interruption is the warning of a larger rolling disaster. It is best to maintain the target site in a dated-but-consistent state, rather than risk a disastrous interruption that leaves the target site inconsistent and difficult to return to full integrity. This scheme is why the auto-synchronization option is disabled by default.

For example:

1. Two volumes are synchronized and then in replicating mode when a link failure is detected.
2. Both volumes revert to logging mode and bits in the bitmap volumes are marked as dirty.
3. The link returns online and a resynchronization is started: the bitmap volumes are logically OR'ed.
4. One of the volumes fails.

The target volume is now in an inconsistent state because it now is a mix of data written from the two volumes before the second (volume) failure and data which has not yet been synchronized. In this instance, the fact that one of the volumes is now unrecoverable makes the target difficult to return to full integrity.

Full Synchronization

Full synchronization starts a full copy operation from the primary volume to the secondary volume. It also enables replication concurrently from the primary volume to the secondary volume; any new writes to the primary volume are also replicated to the secondary volume. After the operation is complete, the Sun SNDR software maintains the normal replicating mode for the volume: either synchronous or asynchronous replication.

Note – The volumes may be made identical using other methods, not just a Sun SNDR full synchronization. When network latencies justify it, you can perform the initial synchronization of an volume set by backing up a source or primary volume on magnetic tape on one site, then restoring the volume from the tape on the other site. During the intervening period (that is, the period between when the backup is completed and the restore is started), place the source or primary volume in logging mode.

Make sure that the backup copy is a physical copy (for example, by using the `dd`(1M) command) and not a logical copy (for example, one made using the `tar`(1M) or `cpio`(1M) commands). The copies must have identical blocks, not just identical files.

Data Flow

FIGURE 2-1 shows the full forward synchronization process.

1. The Sun SNDR software on the primary system (`host1`) requests disk blocks from the active primary volume. The data might already be resident in the primary system data cache, or might require a local disk access.
2. The Sun SNDR software transmits the disk blocks, with destaging instructions, over the connection to a cache region on the secondary system.
3. The Sun SNDR software on the secondary system updates its remote volume and acknowledges the update to the primary system.

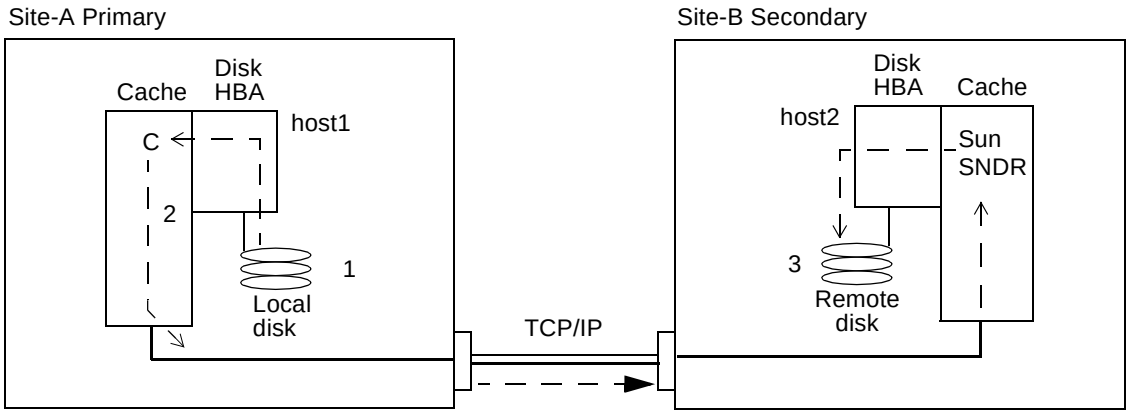


FIGURE 2-1 Full Synchronization (Volume-to-Volume Copy)

Fast Resynchronization (Update)

During fast resynchronization mirroring, the Sun SNDR software initiates replication of only the changed primary site volume data to the secondary site, based on the bitmap. Only the blocks marked dirty in the bitmap are copied to the target volume.

After the mirroring is complete, the Sun SNDR software maintains the normal replicating mode. The software can also be placed in logging mode. See "Logging" on page 26.

The Sun SNDR software resynchronizes the secondary volume from the primary volume. It updates the secondary volume according to the changes based on logs maintained while replication was stopped. It also enables concurrent replication between the primary and secondary volumes; any new write operations to the primary volumes are also replicated to the secondary volumes.

If a remote copy interruption lasts numerous hours and the updates are widespread, logging and fast resynchronization provide diminishing returns. As time passes, the proportion of bits set to true in the bitmap volumes of a volume set might reach 100 percent. The overhead of logging, coupled with fast resynchronization, must then be balanced against that of a full synchronization without intervening periods of logging.

Logging and fast resynchronization serve as a built-in safety net should one of your replication processes be disturbed. The software monitors the network connections between the primary and secondary hosts. Link failures and remote system failures are detected by the transport interface and are passed to the Sun SNDR software.

Data Flow

FIGURE 2-2 shows an update resynchronization using an ATM link from the primary system to its secondary system, when the secondary volumes are stale from the interruption.

1. The Sun SNDR software on `host1` examines a bitmap from the primary and secondary hosts for the Sun SNDR software-managed volumes affected by the interruption.
2. The Sun SNDR software on `host1` requests the blocks that were updated during the interruption from the up-to-date volume. The data might already reside in the `host1` data cache or on the local disk.
3. The Sun SNDR software on `host1` transmits the update blocks 3R to `host2` Sun SNDR software using the SunATM™ hardware connection.
4. The Sun SNDR software on `host2` refreshes its stale replicated image with the updated blocks and acknowledges the action to `host1`.

5. The Sun SNDR software revises the bitmap to track the remote update.

All steps repeat until the remote replicated image is up-to-date.

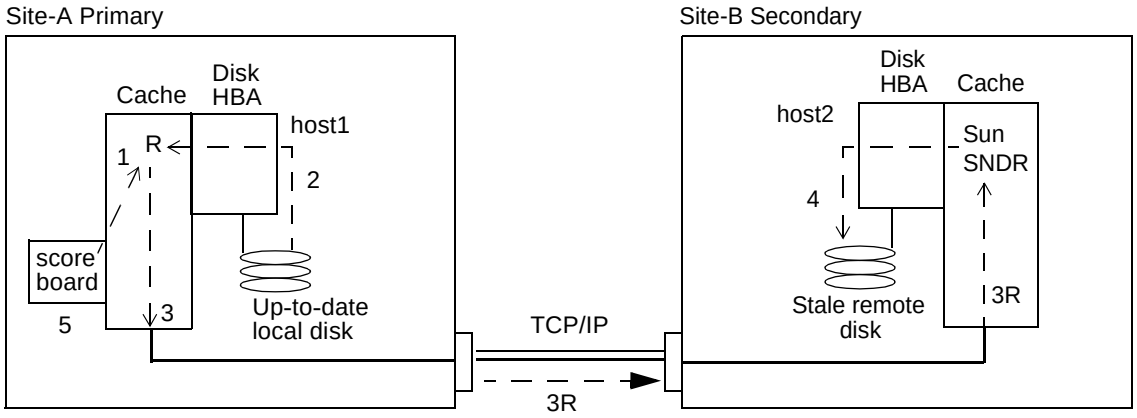


FIGURE 2-2 Update Synchronization of a Secondary Volume Set

Reverse Synchronization

Caution – Do not start the primary application (such as a database application) that writes to the volumes until the full reverse copy operation finishes executing.

During reverse synchronization mirroring, the Sun SNDR software replicates the volume data at the secondary site to the primary site, using either normal full synchronization mirroring or fast resynchronization mirroring.

It starts a full reverse copy operation from the secondary volume to the primary volume. It also enables concurrent replication from the primary volume to the secondary volume; any new writes to the primary volume are also replicated to the secondary volume.

Data Flow

This command starts reverse full synchronization, as the secondary volume on `host2` is resynchronizing the new primary volume on `host1`. FIGURE 2-3 shows the full reverse synchronization process.

1. The data might already be resident in `host1` data cache, or it might require a secondary disk access. If so, the Sun SNDR software on `host1` requests blocks from the up-to-date secondary volume on `host2`.
2. The Sun SNDR software on `host2` transmits the cache blocks 2R over the intersite fiber link to a Sun SNDR software region on `host1` with destaging instructions.
3. The Sun SNDR software on `host1` updates its disk.

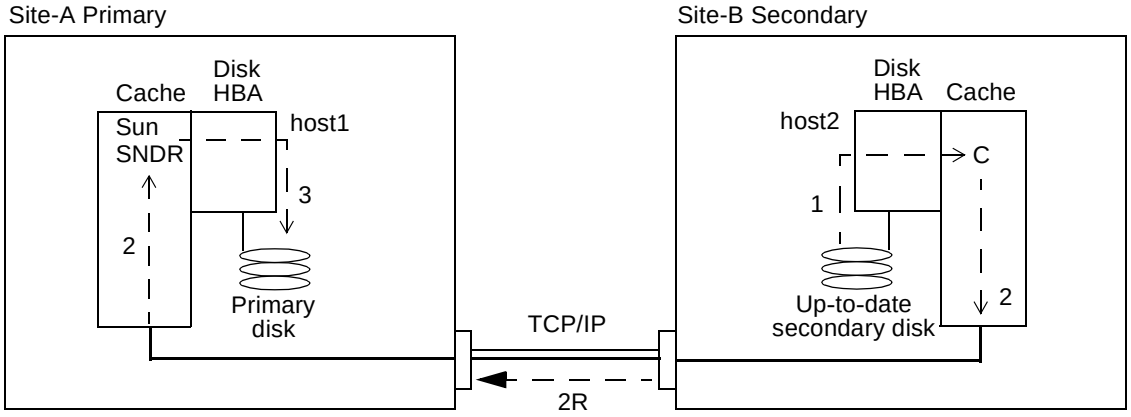


FIGURE 2-3 Reverse Full Synchronization

Fast Reverse Synchronization (Reverse Update)

During fast reverse synchronization mirroring, the Sun SNDR software compares the bitmaps between the primary and secondary sites and replicates only the changed blocks from the secondary site to the primary site.

It resynchronizes the primary volume from the secondary volume. It updates the primary volume according to the changes based on logs maintained while replication was stopped. It also enables concurrent replication between the primary volume and secondary volumes; any new write operations to the primary are also replicated to the secondary volumes.

Data Flow

FIGURE 2-4 shows a reverse update resynchronization from the secondary system to its primary system.

1. The Sun SNDR software on `host1` retrieves the secondary bitmap 1R from `host2` for one of the Sun SNDR software-managed volumes affected by the interruption.
2. The Sun SNDR software on `host1` requests the blocks updated during the interruption from the up-to-date secondary volume of `host2`. The data might already be resident in `host2`'s data cache, or it might require secondary disk access.
3. The Sun SNDR software on `host2` transmits the updated blocks 3R to `host1` Sun SNDR software region of cache using the intersite link.
4. The Sun SNDR software on `host1` refreshes its stale image with the updated blocks.
5. The Sun SNDR software on `host1` revises the bitmap to track the remote update.

All steps repeat until the primary volume is up-to-date.

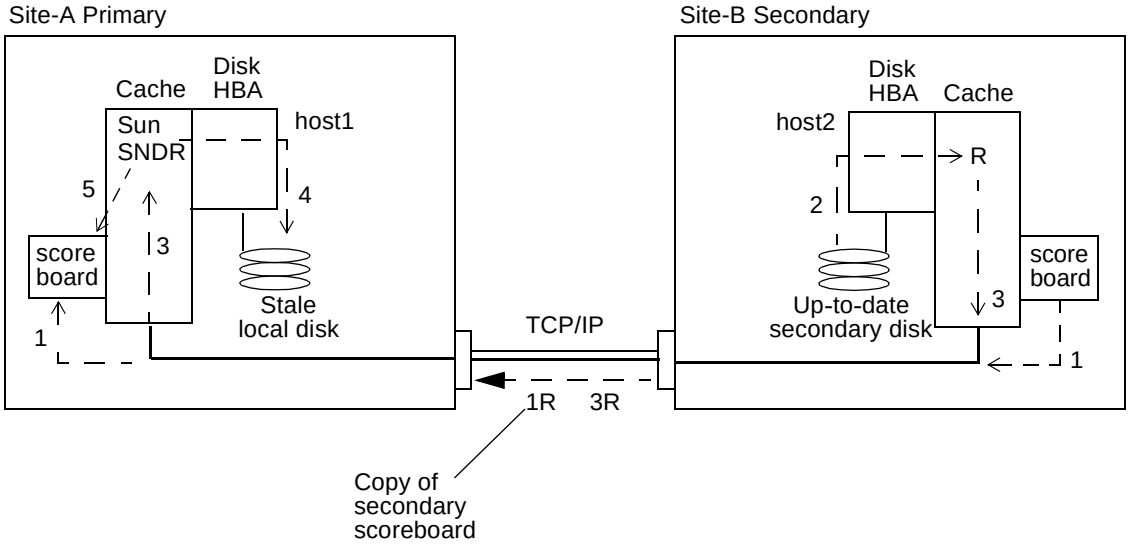


FIGURE 2-4 Reverse Update Synchronization

Logging

Note – With synchronous and asynchronous replication, the Sun SNDR software automatically switches to logging mode if there is a break in the network or if the primary site is down.

During logging, the Sun SNDR software only updates the bitmaps at the primary site; no replication occurs. At a later time, the bitmaps at the primary and secondary sites are compared and the changed blocks in the primary site volume are mirrored by fast resynchronization to the secondary site.

If all volume sets in an I/O group are replicating (meaning that the secondary volumes contain a valid point-in-time copy of the corresponding primary volumes), when one volume set enters logging mode, all other sets in the I/O group will enter logging mode automatically. This scheme ensures that the secondary volumes will contain a valid point-in-time copy.

You can use logging to save on telecommunications or connection costs. The risk, however, is the costs incurred by increased data loss if the primary is lost. If you lose the primary, you do not have the data at the secondary that was written to the primary during the period of logging.

You can also perform logging on the secondary site before a failover. You can then update the primary site using a reverse sync or reverse update sync command.

Note – To resume Sun SNDR software operations after using the `sndradm -l` logging command, use the `sndradm -m` command to perform a full resynchronization or the `sndradm -u` command to perform an update resynchronization. Note also that, when issued from the secondary host, the `sndradm -l` command does not work on the secondary volume for any volume that is currently synchronizing.

Example Replication Scenarios

This section describes three example scenarios:

- “Multihop Replication” on page 27
- “One-to-Many Replication” on page 28
- “Many-to-One Replication” on page 29

Multihop Replication

The Sun SNDR software allows multihop volume sets.

In a multihop set, the secondary host volume of one volume set can be the primary host volume of another volume set. FIGURE 2-5 shows one primary and one secondary host volume; the secondary host volume B becomes the primary host volume A1 to the secondary host volume B1.

Multi-hop configurations can become very complex and the use and administration of multihop sets must be carefully considered.

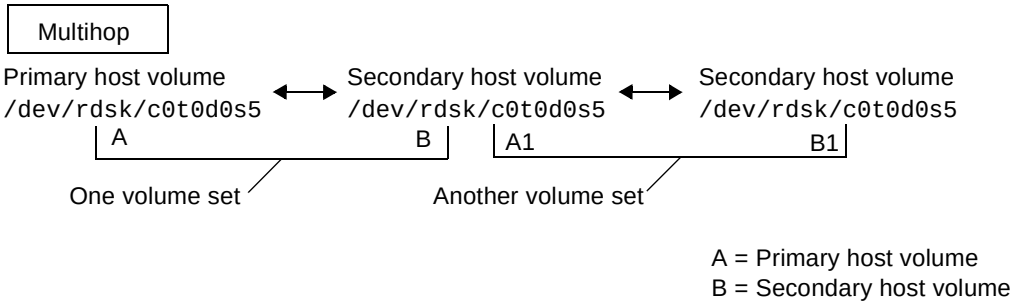


FIGURE 2-5 Multihop Volume Sets

For example, consider what happens if resynchronization operations for every volume set in the multihop chain are performed in synchronous mode. The I/O proceeds along each link of the chain and the I/O acknowledgment will not be confirmed until the last link is reached, at which point the process will complete.

If both sets were configured to replicate synchronously in the example in FIGURE 2-5, the I/O acknowledgment from B1 would be received at A1; then the acknowledgment at B would be received at A. In a multihop configuration where every set in the chain is configured to replicate synchronously, the I/O latency at the primary node (assuming a forward replication) is the combined latency of every link and disk access along the chain.

Conversely, when volume sets are part of an multihop configuration where all sets replicate asynchronously, the contents of any given non-primary volume is unpredictable with respect to its neighbor until the resynchronization completes on all nodes.

These examples are for illustration only, however. The Sun SNDR software does not place any restrictions on the configurations between sets along the chain and a mix of synchronous and asynchronous sets is most useful.

As an another example, configure the A+B volume set as a synchronous SNDR set running over a dark fiber in the same room (that is, ensure a consistent copy of the volume without adversely affecting performance on the primary site). Make the A1+B1 volume set an asynchronous set, running across a network to a remote location (that is, replicate the volume to a remote location at a comparatively fast rate by performing the replication asynchronously because of the high network latencies).

As described below in "Sun SNDR Software with Sun StorEdge Instant Image Software" on page 38, multi-hop configurations can be expanded and the performance of these configurations improved when Sun StorEdge Instant Image software is coupled with the Sun SNDR software.

One-to-Many Replication

In a one-to-many volume set, you can replicate data from one primary volume to many secondary volumes residing on one or more hosts. One primary and each secondary site volume is a single volume set (each volume requires its own unique bitmap volume). When you perform a forward resynchronization, you can synchronize one volume set or all volume sets; in this case, issue a separate command for each set. You can also update the primary volume by using a specific secondary volume. FIGURE 2-6 shows one primary and three secondary host volumes and therefore three volume sets: A and B1, A and B2, and A and B3.

When one-to-many replication is performed in synchronous mode, I/O from the primary is sent to the first secondary in the configuration A+B1. The software waits for the I/O acknowledgment before starting to send the I/O to the second secondary volume in the configuration (B2). (In the Sun SNDR 3.1 release, writes will be queued and processed in parallel, and this wait for the acknowledgment from the preceding secondary will be eliminated.) This pattern is repeated until I/O is acknowledged on the *n*th secondary volume in the one-to-many configuration (B3). In a synchronous one-to-many configuration, the latency at the primary host is the combined I/O latency for every connection to and disk access on the secondary hosts.

When one-to-many replication is performed in asynchronous mode, I/O is queued at the primary host for later transmission and acknowledgment for every secondary host. This scheme allows replication to proceed in parallel during one-to-many asynchronous replications.

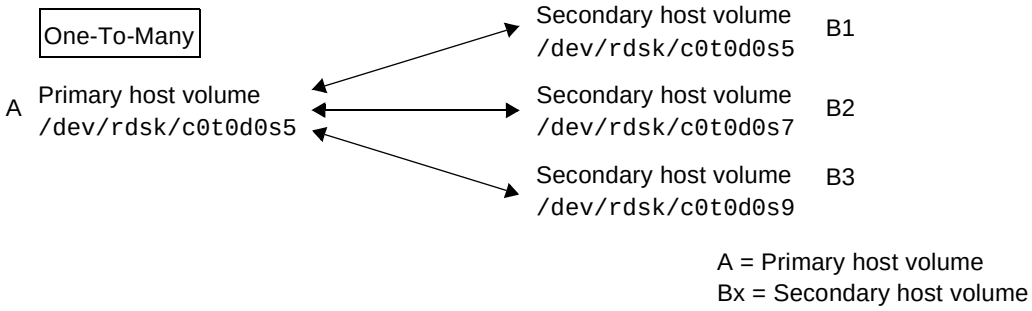


FIGURE 2-6 One-to-Many Volume Sets

Many-to-One Replication

The Sun SNDR software also supports the replication of volumes located on many different hosts to volumes on a single host. The terminology differs from the one-to-many configuration terminology, where the one and the many referred to are volumes. Many-to-one configuration refers to the ability to replicate volumes across more than two hosts through more than one network connection. An example of a many-to-one configuration is shown in FIGURE 2-7.

FIGURE 2-7 shows a simple use of the many-to-one configuration. Host A serves to back up volumes on both Host B and Host C. The Sun SNDR software does not place restrictions on many-to-one configurations, however, and Host A could be configured to be the primary host for some of the replicated volumes and the secondary host for others.

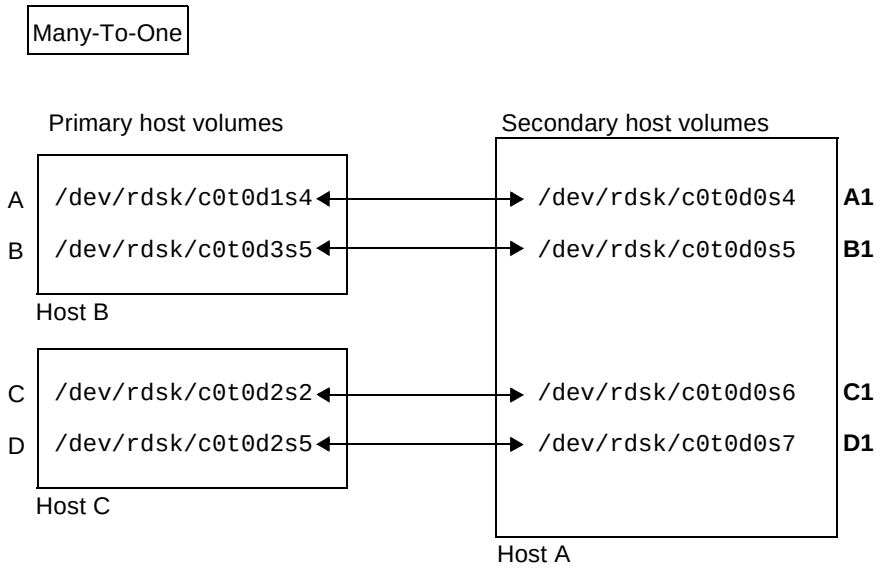


FIGURE 2-7 Many-to-One Volume Sets

CHAPTER 3

Bitmaps

This chapter describes the Sun SNDR software's bitmap management and behavior during synchronization.

Which volume is the source of the data (where the data is copied from) and which volume is the target of the data (where the data is copied to) is important in the context of explaining bitmap management. In a forward synchronization, the primary volume is the source of the data, and the secondary volume is the target. In a reverse synchronization, the secondary volume is the source of the data and the primary volume is the target.

Bitmap Management

The Sun SNDR software maintains a primary and secondary bitmap volume for use during volume replication. It maintains a bit for every 32-Kbyte block of a volume in a volume set. The bit indicates if the data at the block is up-to-date with respect to its replication partner. This technique is known as scoreboarding and the bitmap volume is sometimes referred to as the bitmap, scoreboard, or scoreboard log.

To understand bitmap management, consider how the bitmap is managed in a full synchronization scenario, where every block of storage on the source volume is to be copied to the corresponding block on the target volume.

1. At the beginning of a full synchronization operation, all the bits in the bitmap for the source volume are set to 1. (The source bitmap is not modified in this case.)

   When a bit is set to 1, indicating that the block has not been synchronized, the block is said to be dirty.

2. During replication, as data moves from the source volume to the target, the bits in the bitmap corresponding to the addresses updated are set to 0 and the blocks are said to be clean.

   Conceptually, this replication would proceed in a linear sequence from the start address to the end address. If the only I/O being done in the system were that being done by the Sun SNDR software, the bitmap bits would flip like a line of dominos.

Bitmap Behavior During A Full Synchronization

A synchronization operation does not prohibit I/O from occurring on the source volume. When a write operation destined for an address which has not been copied comes through the I/O stack, the block which is the target of the incoming write is processed concurrently with the synchronization. That is, any new writes to the source volume are also replicated to the target volume.

When the incoming write has completed, the corresponding bit in the bitmap is set to 0. Since the Sun SNDR software checks each bit to see if the block is dirty before copying it, when it does get to this block it will not copy it a second time. After the operation is complete, the Sun SNDR software maintains the normal replicating mode for the volume: either synchronous or asynchronous replication.

When the system is in replicating mode, incoming writes cause the bit corresponding to the block addressed to be set to dirty. The write is then processed and the Sun SNDR software sets the bit to clean. In replicating mode, the Sun SNDR bitmap management is triggered by user-layer application I/O.

Bitmap Behavior During An Update Synchronization

Consider what happens during an update synchronization: a synchronization that does not copy all blocks from the source volume to the corresponding blocks on the target volume.

In an update resynchronization, only the blocks that are marked dirty in the bitmap volume utilized are copied. The only difference in how this synchronization is processed is that the bits in the bitmap volume are not first set to 1. Otherwise, the Sun SNDR software proceeds in the same way as described in “Bitmap Behavior During A Full Synchronization” on page 32, progressing through the bitmap and copying each block if the bitmap indicates that the block is dirty.

Bitmap Behavior During Logging

Bitmap management also occurs during periods when the Sun SNDR software has stopped actively copying data from a source volume to a target volume - that is, when the volume set is in logging mode. (Typically, volume sets are in logging mode when the network between the primary and secondary host is down.)

Unlike the bitmap management during full and update synchronizations, with the Sun SNDR software in logging mode, bitmap management is performed on the bitmap volumes of the primary and secondary volumes in the volume set. Like update synchronization bitmap management, when the Sun SNDR software starts logging, the bits in the bitmap volume are not first set to 1.

When a write request is made on either volume in the set in logging mode, the bit for the block in the bitmap of the corresponding volume is set to 1. This scheme is used because I/O might be permitted on either volume of a volume set when the set is in logging mode.

After logging ends and the Sun SNDR software performs an update resynchronization on the set, the bits in the bitmaps of both of the volumes in the set are logically OR’ed together. This scheme allows the volumes to be resynchronized from either volume in the set.

Reference Counts

For every dirty block being tracked in the bitmap, the Sun SNDR software also maintains a reference count. The reference count indicates the number of I/O requests currently pending (unacknowledged) on a dirty block.

Reference counting is used only during asynchronous operation. As described in this section, once a block had been copied, the bit corresponding to that block is set to 0 (clean).

During asynchronous operation:

1. The reference count increments when more than one process requests a write on a given block of memory before an acknowledgment is received by another process requesting a write on the same block.

2. The reference count decrements when the acknowledgment for the write on the block is received. During asynchronous replication, the bit in the bitmap is not cleared until the reference count reaches 0.

Reference counting ensures data integrity in the case of multiple I/O requests to the same block. During asynchronous operation, these writes exist in the data transfer queue concurrently. Without the reference count, only a single I/O per block could be permitted in the queue at a time.

Under circumstances in which the reference count exceeds one (1), reference counting improves performance. This improvement occurs because reference counts are maintained in memory and the bitmap volume, which is maintained on disk, is not rewritten when subsequent requests on a location previously marked dirty are tracked.

CHAPTER 4

Miscellaneous

This chapter describes the following topics:

Performance Considerations

Several performance considerations exist for the Sun SNDR software. This list is not comprehensive and may not be applicable to all configurations. Consider the following, however, when configuring a system for use with the software:

- Configure bitmap volumes for high performance, particularly the primary bitmap volume. For example, use it in cached arrays or configure it to avoid hot spots (that is, do not put multiple volumes on a single spindle).
- If asynchronous operation is slow, it might be related to the size of the queue of data that needs to be transferred
- Disk speeds on the primary and secondary sites will affect the performance at the primary site when the system operates synchronously.
- The speed and latencies of the network connection affects performance.

I/O Groups

The Sun SNDR software enables you to group volume sets in an I/O group. You can assign specific volume sets to an I/O group to perform replication on these volume sets and not on others you have configured. Grouping volume sets also guarantees write ordering: write operations to the secondary volume occur in the same order as the write operations to the primary volume. This feature is essential in installations requiring you to maintain consistent contents of a group of volumes.

An I/O group is a collection of Sun SNDR software sets that have the same group name, primary and secondary interfaces, and mirroring mode. Mixed groups (those where mirroring modes are asynchronous for one set and synchronous for another set) are not allowed.

By using an I/O group, you can issue a Sun SNDR command that is executed on every member of the group, enabling volume sets to be controlled as a single unit.

I/O group operations are atomic. The change from replicating mode to logging mode is guaranteed to occur on every set in an I/O group and to fail on all the sets if it fails on a single set in the group.

The Sun SNDR software maintains write ordering for volumes in a group to ensure that the data on the secondary volumes is a consistent copy of the corresponding primary volumes. See “Order-Dependent Writes and Volume Set Grouping” on page 37.

Note – The I/O group concept does not matter for synchronous replication; that is, write-ordering is preserved among those volume sets configured as sync.

The auto-resynchronization feature supports the I/O grouping concept. It allows the feature to be enabled or disabled on a per-group basis and controls the resynchronization operation atomically on the group.

I/O grouping has an adverse affect on the Sun SNDR asynchronous operation, as I/O flushing is reduced to a single thread. In this case, consider the size of the data to be transferred since all I/O will be routed through a single queue.

Order-Dependent Writes and Volume Set Grouping

Write ordering is also maintained for *groups* of asynchronously replicating volume sets. (The general definition of write ordering here is that write operations directed to the target volume occur in the same order as write operations to the source.) The group of target volumes is a point-in-time copy of the group of source volumes.

This feature is especially valuable in those cases where you can avoid application requirements that limit operations. For example, a database application might limit partition sizes to no greater than 2 Gbytes. In this case, you might group volume sets to create a virtual large “volume” that preserves write operations. Otherwise, you might risk having inconsistent data by trying to update volume sets individually instead of as a group.

When an application has multiple logical volumes assigned, application data integrity can be maintained by one of the following:

- Specifying that all Sun SNDR software volumes associated with that application are in sync mode
- Using Sun StorEdge Instant Image software to take periodic recoverable point-in-time copies

If you use Sun StorEdge Instant Image software, the remote point-in-time is taken while the application is in the recoverable state. For example, most database applications allow for a hot backup. If a remote point-in-time copy were made of the entire replicated database while the primary was in hot backup mode, then a consistent remote database is available by using the point-in-time copy and the log files taken while the database was in hot backup mode.

Sun SNDR Software with Sun StorEdge Instant Image Software

To help ensure the highest level of data integrity on both sites during normal operations or during fast resynchronization for data recovery, use the Sun StorEdge Instant Image software with the Sun SNDR software.

Enable an Instant Image point-in-time snapshot volume copy of the secondary volume before starting synchronization of a secondary volume from the primary site. A measure of protection is provided by using Instant Image in this scenario.

If a failure occurs during resynchronization, you have a known good copy of usable data; you can resume resynchronization when it is safe to do so. Once the secondary site is fully synchronized with the primary site, you can disable the snapshot or use it for other purposes, such as remote backup or remote data analysis.

Also, you can transfer an Instant Image point-in-time snapshot copy of the primary volume to the secondary site. Applications can remain open and active at the primary site while the copy is being replicated. This scheme works well if the secondary volume is able to be out of sync with the primary volume by some small time delta.

The advantage is that the overhead involved in remotely mirroring the primary data is the snapshot image is mirrored instead. Keeping the secondary site slightly out of sync with the primary also allows the verification of the correctness of the primary data before replicating it to the secondary site.

Sun SNDR Software with Sun Cluster 3.0

Note – Sun SNDR replication within the cluster is not supported; that is, when the primary and secondary hosts reside in the same cluster and the primary, secondary, and bitmap volumes in a volume set reside in the same disk device group.

Sun SNDR volumes can be hosted in a two-node Sun Cluster 3.0 Update 1 (also known as the 07/01 release) environment running most releases of Solaris 8 (Solaris 8, Update 06/00 is not supported). This configuration enables replications to *failover* to another cluster node if the node hosting a Sun SNDR volume crashes.

Failing over involves placing the volumes of the affected node under the control of another node in the cluster and continuing the replication when the new node takes control. This process is automated by Sun Cluster as part of its control of volume management subsystems.

Caution – A primary volume should never be hosted in the same cluster as its corresponding secondary volume.

An important component of successful failover is how the Sun SNDR volumes are configured in a Sun Cluster resource group. A resource group is a grouping of items in a Sun Cluster which are interrelated in such a way as to make it impossible to fail over a single member of the group without failing over all members of the group. That is, members of a resource group are dependent upon one another when a node in the cluster is failed over.

Detailed information about resource groups is available in the Sun Cluster documentation. Also see the *Sun Cluster 3.0 U1 and Sun StorEdge Software 3.0 Integration Guide* for more configuration information.

Information on how resource groups are used in configuring Sun SNDR is described in "Failover" on page 40.

On hosts running the Solaris 8 update 3 operating environment, network multipathing with failover is supported. This feature is provided as part of the standard Solaris operating environment and is transparent to the Sun SNDR software.

Failover

When the node running the Sun SNDR software fails, the Sun Cluster software detects the failure and initiates failover. Conceptually, failover includes restarting processes that were running on the failing node on another node without losing any information. This information is application dependent and outside the control of Sun Cluster. Sun Cluster moves the required file systems, volumes, networking and configuration data.

In the case of the Sun SNDR software, the Sun SNDR hostname, IP address, and control of the volumes being referenced must be moved to the new node. Replication is then restarted at the point that the first node failed. You must configure the Sun SNDR software with a Sun Cluster lightweight resource group consisting of:

- A logical host, which provides the network hook
- A High Availability (HA) storage resource that includes the volume and its associated bitmap volume

Because of its position in the kernel I/O stack, failing over Sun SNDR volumes is similar to failing over a volume manager. The Sun StorEdge services and Sun Cluster software together help ensure that I/O processing on Sun SNDR volumes is enabled at the correct point in the failover process on the new node and that processing on in-transit I/O is completed. The bitmap volumes are used to continue replicating on the new node. *The bitmap volumes in Sun SNDR volume sets running in a Sun Cluster environment must be disk-based, not memory-based.*

If the Primary Volume Is In a Cluster

When the primary volume is hosted in a cluster, the failover hostname and IP address is that of the associated HA Sun SNDR resource group.

Following a failover event, the Sun SNDR failover script initiates an update resynchronization for all affected volume sets if the Sun SNDR autosynchronization feature is enabled for those volume sets. This operation is performed after the failover script has switched over the resource group, as it must be performed after the network switchover has occurred.

If the Secondary Volume Is In a Cluster

When the secondary volume is hosted in a cluster, the failover hostname and IP address is that of the associated HA Sun SNDR resource group.

A failover of a secondary host appears like a network outage to the primary host. Since the secondary host cannot initiate a resynchronization, manual intervention will be required to restart synchronization, unless the SNDR auto-resynchronization facility is enabled.

Both The Primary and Secondary are In a Cluster

This configuration is a special case of the two cases described in "If the Primary Volume Is In a Cluster" on page 40 and "If the Secondary Volume Is In a Cluster" on page 41. It does not impose additional constraints. **Note that the primary volume should never be hosted in the same cluster as its corresponding secondary volume.**

Caution – Sun SNDR replication within the cluster is not supported; that is, when the primary and secondary hosts reside in the same cluster and the primary, secondary, and bitmap volumes in a SNDR volume set reside in the same disk device group.

Sun SNDR Software and Sun StorEdge Fast Write Cache 3.0 (SUNWnvm 3.0 Package)

Note – The Sun StorEdge Fast Write Cache product, all versions, is not supported in any cluster environment.

All versions of the Sun StorEdge Fast Write Cache product are not supported when using the Sun StorEdge Version 3.0/3.0.1 services software in a Sun Cluster 3.0 Update 1 environment because cached data is inaccessible from other machines in a cluster. To compensate, you can use a Sun caching array.

For example, the Sun StorEdge Core Services Version 3.0 and 3.0.1 CDs contain the Sun StorEdge SUNWnvm Version 3.0 software package. This package is intended for those users whose systems include Version 2.0 of the Sun FWC hardware and software product and who wish to continue using the Sun FWC product with Sun SNDR and Instant Image Version 3.0 services software *in a nonclustered environment*.