USING iSCSI MULTIPATHING IN THE SOLARIS™ 10 OPERATING SYSTEM

Aaron Dailey, Storage Network Engineering
Scott Tracy, Storage Network Engineering

Sun BluePrints™ OnLine — December 2005

Part No 819-3730-10
Revision 1.0, 11/28/05
Edition: December 2005

Please
Recycle

TABLE OF CONTENTS

Using iSCSI Multipathing in the Solaris™ 10 Operating System

This Sun BluePrints™ OnLine article describes how to use Internet Small Computer Systems Interface (iSCSI) multipathing in the Solaris™ 10 Operating System (Solaris 10 OS). Implementing iSCSI in a storage solution:

- increases storage availability via fail-over protection
- increases scalability and throughput via link aggregation

This article describes different approaches to implementing multipathing between an iSCSI initiator and an iSCSI target device. It reviews the reasons for multipathing, describes the different approaches that Solaris supports, discusses the trade-offs between those approaches, and provides recommendations for specific configurations.

This article contains the following sections:

About the iSCSI Protocol

The Small Computer Systems Interface (SCSI) is an industry standard set of protocols used for connecting I/O devices (particularly storage devices) to a server or workstation. iSCSI is an Internet Protocol (IP)-based storage networking standard for linking data storage subsystems. By carrying SCSI commands over IP networks, the iSCSI protocol enables users to mount disk devices from across the network onto a local machine. On the local machine, users can access the devices just as they would access any block device. For detailed information, see the iSCSI specification memo submitted by the Network Working Group to the Internet Engineering Task Force (IETF) at:

http://www.ietf.org/rfc/rfc3720.txt

The iSCSI protocol:

- runs across existing Ethernet networks
- uses existing management tools for IP networks
- can be used to connect to Fibre-Channel (FC) or iSCSI Storage Area Network (SAN) environments

An iSCSI network consists of an iSCSI initiator and an iSCSI target. An iSCSI initiator is the client of an iSCSI network. An iSCSI target is any device that responds to client requests from an iSCSI initiator. The iSCSI initiator connects to an iSCSI target over an IP network using the iSCSI protocol. As with SCSI, Logical Units have Logical Unit Numbers (LUNS), where the numbering is per target.

The iSCSI protocol is implemented underneath SCSI and on top of TCP/IP. Multiple TCP connections are allowed per session; these can be optionally used to exploit parallelism and provide error recovery. In addition to SCSI, there are commands to login (connect a TCP session) and logout (TCP session teardown).

iSCSI naming is similar to SCSI—names are globally unique using a worldwide name. The DNS system provides IP name lookups. iSCSI targets may share a common IP address and port, and initiators/targets may have multiple IP addresses.

In iSCSI parlance, a *portal* is any entity that has an IP address, such as a disk array or some other storage resource. In general, a physical network port is synonymous with a portal, such as a network interface controller (NIC) or the Ethernet port in a disk array. Note, however, that a physical port can be associated with multiple IP addresses.

There are two basic types of discovery that the Solaris iSCSI initiator supports:

- **Static configuration** defines an IP address, port, and iSCSI target name for each target.
- **Send target** allows the initiator to query for all supported targets behind a portal.

iSCSI Support in Solaris 10 Release, Update 1

Solaris 10 release, Update 1, includes an iSCSI initiator device driver that enables Solaris to read and write to block storage through any TCP/IP interface, including a standard Ethernet Network Interface Card (NIC).

Potential uses include accessing native iSCSI storage, as well as FibreChannel (FC) storage accessed through bridges. The introduction of iSCSI provides additional options for multipathing. The existing Sun multipathing software solution (MPxIO, also known as the Sun StorEdge Traffic Manager) is still used for certain configurations, and Solaris IP multipathing can aid delivery. Finally, iSCSI includes native multipathing as an optional feature and, as such, it is currently not supported by Solaris, nor is it widely implemented by vendors in their iSCSI products.

Organizations use the new `iscsiadm` command to set up and manage iSCSI devices. For more information, see “Configuring iSCSI Initiators,” in *System Administration Guide: Devices and File Systems*, in the Solaris 10 Product Documentation at:

http://docs.sun.com/app/docs/doc/819-2723/6n50a1n01?a=view

This article focuses on the current Solaris iSCSI implementation that uses standard Ethernet NICs. Although certain features, such as iSCSI TCP Offload Engines (TOEs) and iSCSI boot, might be supported in future Solaris releases, these capabilities are not discussed in this article.

What is iSCSI Multipathing?

In general, *multipathing* is a design pattern for redundancy and automatic fail-over that provides at least two physical paths to a target resource. Multipathing allows for re-routing in the event of component failure, enabling higher availability for storage resources. Multipathing also allows for the parallel routing of data, which can result in faster throughput and increased scalability.

Solaris 10 release, Update 1, supports Solaris network multipathing via:

- **IPMP load spreading**—outgoing network traffic is able to utilize several network interfaces
- **network trunking**—multiple physical network interfaces are treated as one; combining these interfaces, which is accomplished in the TCP/IP stack, allows link aggregation and availability

The Sun multipathing software (MPxIO) allows the merging of multiple SCSI layer paths, such as an iSCSI device exposing the same LUN via two different iSCSI target names (or the same name with different target portal group tags). In addition, one FC path and one iSCSI path can be merged by MPxIO if the target device supports these options. Additional functionality, such as iSCSI Multiple Connections / Session (MC/S), might be supported in future Solaris releases.

Benefits of Multiple Physical Connections

Multiple physical connections between a host and a target provides two major benefits: availability and link aggregation.

Availability

When multiple physical connections exist between host and storage, if one connection is lost, then I/O can be rerouted through other paths. To maximize availability and eliminate single points of failure, the following components in a storage architecture must be made redundant:

- portals on the initiator and target switches
- IP network (various components)
- array controllers

For more information about configuring a high availability network, see *Enterprise Network Design Patterns: High Availability* (Sun BluePrints Online—December, 2003) at http://www.sun.com/blueprints/1203/817-4683.pdf.

Link Aggregation

Multipathing drivers can route data through multiple paths in parallel, resulting in increased throughput between host and storage. Current commodity Ethernet implementations support 1Gb/S full duplex, for an aggregate throughput of 2Gb/S (if transmit and receive traffic is balanced (if, for example, the network supports 1Gb incoming and 1Gb outgoing traffic). Combining multiple links can scale performance.

Solaris 10 OS Multipathing Options for iSCSI Devices

Multipathing iSCSI devices can be implemented at different levels in the Solaris storage protocol stack. The following figure shows the Solaris block I/O stack.

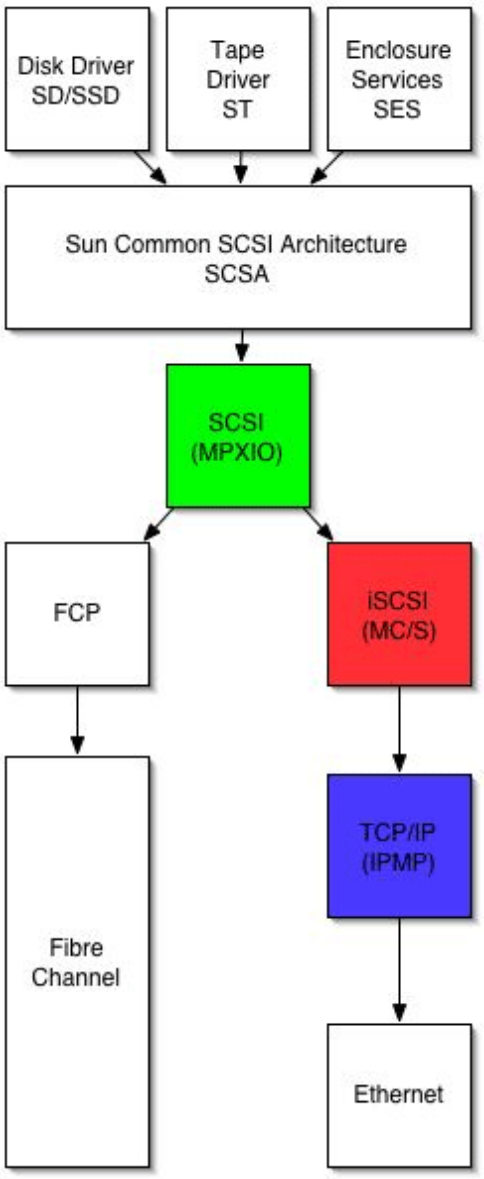

Figure 1. Solaris Storage Stack (Block I/O)

Note – The iSCSI Multiple Connections per Session (MC/S) is currently not supported in Solaris but might be available in a future release.

---

iSCSI is built on the Solaris IP Stack, which includes:

- IP multipathing (IPMP) over TCP/IP

- Above IPMP, iSCSI provides native multipathing using MC/S
- At a higher level (that is independent of the transport layer), Solaris provides multipathing software (MPxIO). Because MPxIO is independent of transport, it can multipath a target that is visible on both iSCSI and FC ports.

Because of their location in the network protocol stack, each multipath solution is useful for different purposes.

IP Multipath

IP Multipath (IPMP) is a native Solaris system facility for network multipathing. Operating at the IP layer in the networking stack, IPMP provides for fail-over and aggregation over two or more NICs. For more information about IPMP, see the Solaris 10 *System Administration Guide: IP Services*, at http://docs.sun.com/app/docs/doc/816-4554.

To implement IPMP, a system administrator selects NICs that are on the same subnet and places them in logical IPMP groups. A daemon (part of the IPMP system) monitors the health of the ports and can be configured to monitor connections to specific iSCSI targets. In the event of a port failure, the other port on the same subnet assumes the same Media Access Control (MAC) address as the failed NIC, and the iSCSI connection continues uninterrupted.

The following figure shows a sample configuration for IPMP.

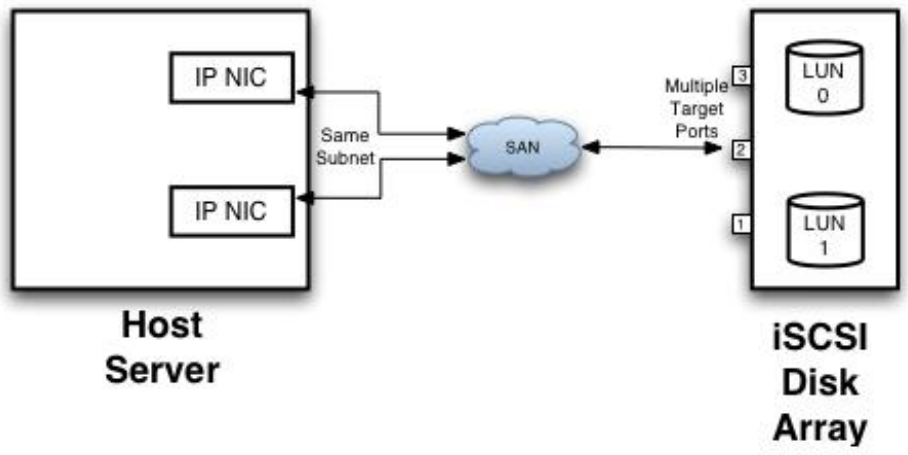

Figure 2. IP Multipathing (IPMP)

IPMP participates in dynamic reconfiguration (DR). On systems that support DR, administrators can replace NICs without disrupting networking traffic. When a NIC is replaced, it is added back to the IPMP group and used thereafter for I/O.

When used in combination with iSCSI, the major limitation of IPMP is that multiple target ports are not multipathed. IPMP enables redundancy between host ports but cannot fail-over to multiple target ports.

iSCSI Native Multipathing

The iSCSI specification addresses the requirement for redundant physical connections. While FC SANs support multiple paths, the iSCSI specification defines what is supported.

In TCP/IP, connections describe communication between two portals. A session is the association between an initiator and target, either of which may have one or more portals. Multiple Connections per Session (MC/S) allows initiator portals to communicate with target portals in an coordinated manner. Target portal and initiator portal redundancy are both supported. Link aggregation is also supported. The following figure shows one configuration that supports MC/S.

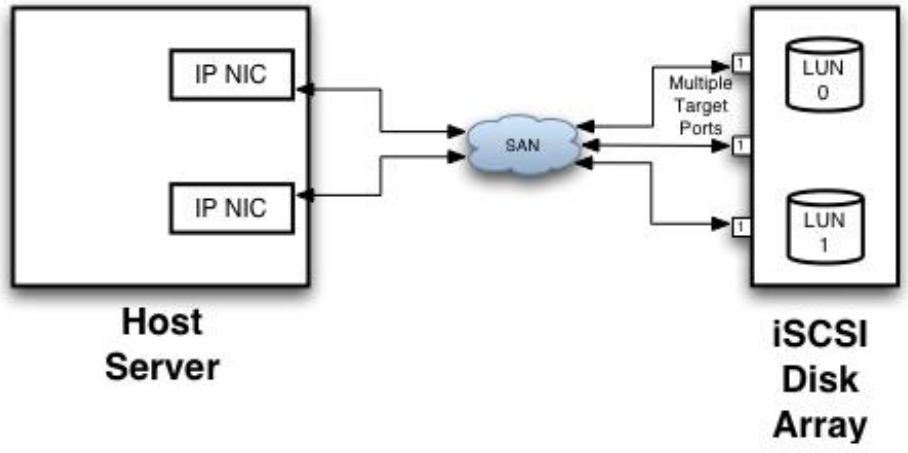

Figure 3. Multiple Connection/Session (MC/S)

MC/S also allows (but does not require) more sophisticated error handling than simply retrying a command. This error recovery allows commands from a failed connection to be recovered quickly by other good connections in the same session. The SCSI layer is not aware of the error.

In general, iSCSI vendors do not yet support MC/S. Therefore, MC/S is not supported in the Solaris 10 release, Update 1, of the Solaris software initiator, but it might be supported in a future release.

Sun Multipathing Software (MPxIO)

MPxIO is a Solaris component that supports multiple physical paths to storage. MPxIO is the current Solaris functionality that supports multiple physical FC connections. Because MPxIO operates above the transport layer (at the SCSI protocol layer), it can support FC, InfiniBand (IB), and iSCSI in certain configurations. For more information about MPxIO, see http://www.sun.com/products-n-solutions/hardware/docs/Software/Storage_Software/Sun_StorEdge_Traffic_Manager/.

FC and iSCSI drivers register logical units (LUNs) with MPxIO. MPxIO matches paths to the same logical unit at the SCSI protocol layer by querying the unique SCSI per LUN identifier from each device. MPxIO collapses duplicate paths to one device so that the target driver and layers above know only of the one device.

The iSCSI initiator driver determines which device(s) to register by examining the SCSI target port identifier of the target. The target port identifier consists of two parts:

- target node name
- target portal group tag (TPGT)

These two parts are concatenated, as shown in the following example target port identifier.

```
iqn.1921-02.com.sun.12432+[1]
```

where the target node name is `iqn.1921-02.com.sun.12432` and the TPGT is `1`.

The iSCSI initiator registers an instance with MPxIO for each LUN for every unique target port identifier.

MPxIO and Multiple SCSI Target Portals IDs

MPxIO might seem to be the ideal solution to the current lack of native iSCSI multipathing support in the Solaris initiator. However, in order for MPxIO to support an iSCSI target, the target must support configuring different SCSI target port identifiers for each portal. One method of doing this, as shown in the following figure, is to allocate portals into multiple target portal groups so that the TPGT makes the target port identifier unique.

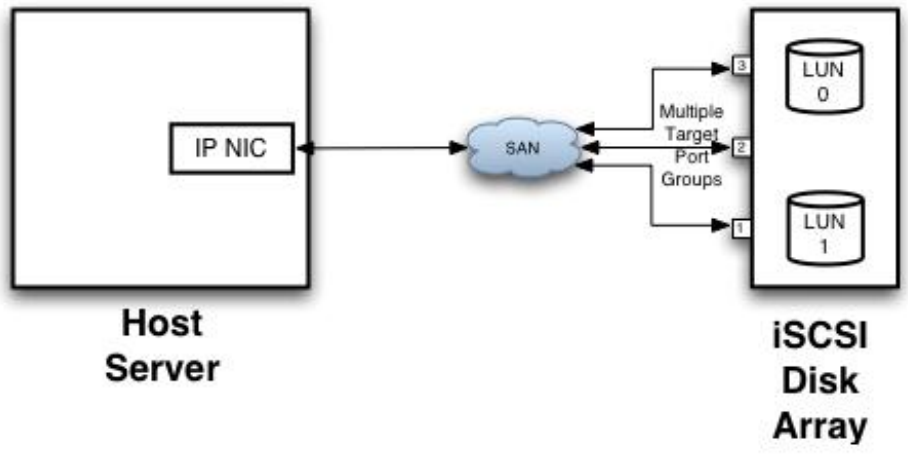

Figure 4. MPxIO with Multiple Target Port Identifiers

Another method is to simply have different iSCSI target names per portal. To create unique names, array vendors can choose either of these approaches.

The target's port configuration determines whether MC/S or MPxIO can be used for multipathing.

- If an iSCSI target supports MC/S, it will present all of its target portals in a single target portal group. With such a target, all target portals form one logical SCSI target port, and the Solaris iSCSI driver therefore registers only one instance of a LUN with MPxIO.
- If an iSCSI target supports MPxIO, it will have different target port groups. Different target port groups force different sessions, so MC/S cannot be used for target port redundancy.

MPxIO with Dual SCSI/FC Bridges

MPxIO can also be used when there are dual iSCSI to FC bridges to a FibreChannel SAN, as shown in the following figure.

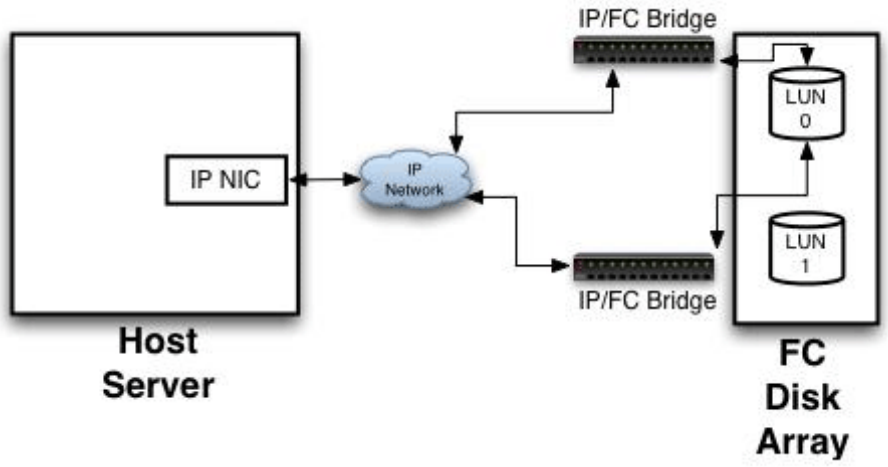

Figure 5. MPxIO with Dual SCSI/FC Bridges

As in the previous example, each LUN has a different target identifier because the iSCSI specification requires unique names for different devices. iSCSI presents both instances to MPxIO, and then MPxIO matches the unique SCSI per LUN identifier, finds that they are identical, and presents one target to the target driver.

MPxIO with Different Transports to the Same Device

Because MPxIO is above the transport layer, MPxIO can support different transports to the same device. In the example configuration shown in the following figure, one LUN appears to the host via FC and iSCSI paths. In this configuration, MPxIO will utilize both paths.

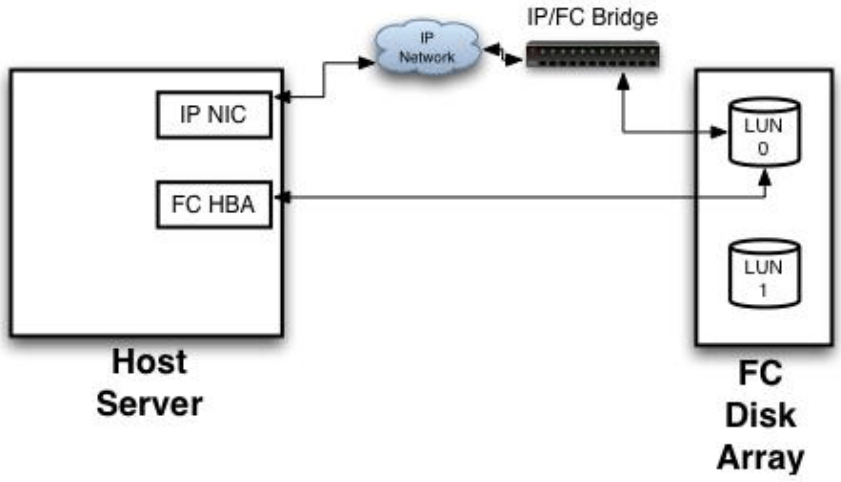

Figure 6. MPxIO with IP/FC Bridge

LUN0 at the disk array appears both to the IP NIC and FC HBA in the host. MPxIO will consolidate the two paths into one and then present it to the target drivers. This is how bridges currently work today. Arrays that support both FC and iSCSI connections natively can use the same mechanism.

Note that, in this configuration, MPxIO performs its default load balancing. For a symmetric access device, this is generally round robin load balancing, so that I/O requests alternate between active links. This is independent of the performance of relative links. Because load balancing is round robin, MPxIO is most useful in configurations in which all links between initiator and target have equal bandwidth and latency.

Summary and Recommendations

Based on the current multipathing options that Solaris supports, consider the following recommendations.

- MPxIO supports target port aggregation and availability in several configurations:
  - Native iSCSI targets, which can present different target port identifiers.
  - Dual iSCSI to FC bridges.
  - Hybrid configurations that connect a target via both iSCSI and FC. Optimal link utilization is achieved when all paths have equal performance capability (and MPxIO will split the load equally).
- IPMP should be used for redundant host ports.

  One recommended base configuration for an iSCSI host is a server with two NICs dedicated to iSCSI traffic. The NICs should be configured using IPMP. Additional NICs should be used for non iSCSI traffic to maximize iSCSI performance.

To maximize reliability, IPMP and MPxIO should both be used. By combining IPMP and MPxIO, redundancy at the host and the target can be achieved, as shown in the following figure.

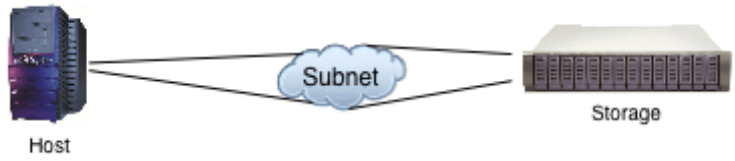

Figure 7. MPxIO and IPMP

Conclusion

In Solaris 10 release, Update 1, Sun provides a software-based iSCSI initiator that allows Solaris computers to connection to iSCSI targets through TCP/IP. Solaris provides multipathing for iSCSI with two existing Solaris components: IPMP and MPxIO.

iSCSI can utilize multiple host NICs by using IPMP. MPxIO, which is already used in FibreChannel SANs, allows using multiple network portals on a target. By implementing the guidelines described in this article, these components together form a robust iSCSI multipathing solution, resulting in increased availability and link aggregation.

References

Publications

- Sun Microsystems, Inc. “Configuring iSCSI Initiators” in *System Administration Guide: Devices and File Systems*, in the Solaris 10 Product Documentation.

  http://docs.sun.com/app/docs/doc/819-2723/6n50a1n01?a=view

- *Solaris Fibre Channel and Storage Multipathing Administration Guide*, Sun Microsystems, Inc. Solaris 10 Product Documentation.

  http://docs.sun.com/source/819-0139/index.html

- Sun Microsystems, Inc. *System Administration Guide: IP Services*, in the Solaris 10 Product Documentation.

  http://docs.sun.com/app/docs/doc/816-4554

- *Internet Protocol Network Multipathing (Updated)*, by Mark Garner (Sun BluePrints™ OnLine—November 2002)

  http://www.sun.com/blueprints/1102/806-7230.pdf

- *Enterprise Network Design Patterns: High Availability* (Sun BluePrints Online—December, 2003)

  http://www.sun.com/blueprints/1203/817-4683.pdf

Web Sites

- Sun Multipathing Software (MPxIO)

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

- Solaris Express 2/05

  http://www.sun.com/software/solaris/solaris-express/

About the Authors

Aaron Dailey

Aaron Dailey is a staff engineer in Storage Network Engineering. Most recently, he has contributed to the Solaris iSCSI initiator. Previously, he worked on the Solaris FibreChannel device driver, and a failover driver for IBM's AIX operating system. Prior to working at Sun, Aaron worked on array controller software at Chaparral Network Storage (now owned by Dot Hill) and Adaptec. He graduated from the University of Virginia with a Bachelor's degree in Computer Science. Aaron lives in Boulder, Colorado and enjoys bicycling.

Scott Tracy

Scott Tracy is a Senior Manager of Storage Network Engineering, working for Sun’s Network Storage Division. He has worked at Sun for over six years. He manages the SAN Software team responsible for software releases for InfiniBand, iSCSI, Fibre Channel, and SCSI connectivity options in the Solaris OS. Previously, he was a manager of Solaris disk and tape driver components, as well as non-Solaris Fibre Channel failover drivers for AIX, HPUX, and MS Windows. Prior to this, he was a kernel developer working on the Solaris disk driver. Before Sun, Scott worked as a driver developer for Adaptec on the Easy CD Creator product and for MCI on database applications. Scott has a BS in Mining Engineering from the Colorado School of Mines, and he currently holds an inactive Certified Public Accountant license in the state of Colorado.

Acknowledgements

The authors would like to thank their colleagues at Sun for their help in the development of this article.

Ordering Sun Documents

The SunDocs℠ program provides more than 250 manuals from Sun Microsystems, Inc. If you live in the United States, Canada, Europe, or Japan, you can purchase documentation sets or individual manuals through this program.

Accessing Sun Documentation Online

The `docs.sun.com` web site enables you to access Sun technical documentation online. You can browse the `docs.sun.com` archive or search for a specific book title or subject at `http://docs.sun.com/`.

To reference Sun BluePrints OnLine articles, visit the Sun BluePrints OnLine Web site at:
`http://www.sun.com/blueprints/online.html`