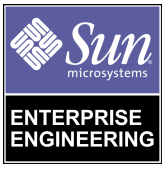
Workgroup Server PCI RAID Solution - The Sun StorEdge™ SRC / P Controller

---

By Don De Vitt - Enterprise Engineering

Sun BluePrints™ OnLine - October 1999



http://www.sun.com/blueprints

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303 USA
650 960-1300 fax 650 969-9131

Part No.: 806-3754-10
Revision 02, October 1999

Please
Recycle

Workgroup Server PCI RAID Solution - The Sun StorEdge™ SRC/P Controller

As with the other installments in this series, this article continues to examine PC interoperability or Workgroup Server related topics. This installment focuses on performance and implementation of the new Workgroup Server internal PCI RAID controller. The official name of this product is the **Sun StorEdge™ SRC/P Controller**. SRC/P stands for Sun RAID Controller for the PCI bus

Because a short article cannot cover every aspect of the SRC/P RAID controller, we will focus on analysis of a popular configuration using the card in the Sun Enterprise™ 450 Workgroup Server. We will cover the following main topics:

- SRC/P basic description
- Several specific configuration options
- Performance considerations in the Sun Enterprise 450 server
- Optimizing the SRC/P controller for data protection of a second disk failure

Background

At the time the Sun Enterprise 450 and the Sun Enterprise 250 Workgroup servers were announced, there were essentially three options for configuring a high performance redundant disk subsystem: an external hardware RAID solution, such as the StorEdge A1000; the internal and external SCSI drives in a software RAID environment such as the Sun Enterprise Volume Manager™ server; or Solstice DiskSuite™ software.

With the introduction of the new Sun StorEdge SRC/P PCI Controller for the Workgroup server, the internal drives of the workgroup servers can be placed into a hardware RAID environment that will improve performance, add redundancy, and off-load the processing of RAID environments from the Sun Enterprise 450 processors to the on-board processor of the SRC/P controller.

While the card is attractive option with the Sun Enterprise 250 workgroup server, this article will focus on the Sun Enterprise 450 performance and configurations.

SRC/P Description

The SRC/P controller is a 64-bit PCI card that fits into 64-bit PCI slots of the Sun Enterprise 450 or 250 servers. The card has three internal 40Mbyte synchronous Ultra SCSI channels supported by three 68-pin SCSI connectors. In addition to the internal connectors, the card also supports three 68-pin VHDCI miniature SCSI connectors to allow external disk packs to be attached. Fully configured with external Sun StorEdge Multipack-12 disk storage units, each SRC/P card can utilize 36 (12x3) Ultra-Wide SCSI drives.

The SRC/P card is capable of supporting RAID levels 0, 1, 5, 1+0 and 5+0 with support hot spares, hot-plug disks, and transparent rebuild, all configurable by the SRC/P Storage Manager GUI program or scripted command-line tools. In addition to off-loading all RAID calculation from the servers CPUs to the controller on the card, the card also has 64 Mbytes of battery backed up memory that allows the card to cache data securely before it can be written to the disk drives. This allows the Solaris™ Operating Environment software to continue disk synchronous operations without waiting for the information to be committed to the disk.

For more specifics on the SRC/P card please refer to:
http://www.sun.com/servers/workgroup/hwraid/

Sun Enterprise 450 and SRC/P Configuration Options

RAID environments trade-offs are always made in three key areas: performance, capacity, and risk. As we discuss the various SRC/P configurations I will highlight these trade-offs so you can weigh them along with the priorities of your system requirements.

Performance Considerations

A fully configured Sun Enterprise 450 workgroup server has five Ultra-Wide SCSI channels. Each of these five channels can support up to four drives installed internally in the system. One channel goes to the Sun Enterprise 450 server's motherboard, which currently must be used to boot the Solaris Operating Environment. (The Solaris Operating Environment cannot currently be booted from SRC/P RAID card.) The remaining four SCSI channels can be attached to regular dual-channel SCSI controllers or the SRC/P RAID controller. To fully utilize all four channels, two SRC/P controllers would be required.

One of the most attractive configurations is to attach as many internal SCSI drives as possible to the SRC/P card to produce RAID volumes. This means that up to 12 drives can be attached internally to one card.

PCI bus Considerations

Before talking about software RAID performance, let's consider which PCI slot to use to support the card. The Sun Enterprise 450 supports considerable PCI bus bandwidth by way of its 10 PCI slots, supported by six PCI buses, implemented by three PCI controller chips. While the traffic going through one SRC/P card can be considerable, it is unlikely that in any real server one SRC/P card could saturate any of the PCI buses it is placed into. Placing other high bandwidth PCI cards on the same PCI bus should however be avoided if possible. Like any high bandwidth PCI card installed into a fully loaded system, it would be best if the SRC/P card were installed into a PCI slot that is attached to a bus shared with as few other PCI cards as possible to maximize the bus bandwidth of the system. The chart below shows the Enterprise 450 PCI slots and the PCI bus and support chip that supports each of these PCI slots.

TABLE 1 Sun Enterprise 450 PCI Bus

PCI Slot Number	PCI Slot type	Chip & Bus Supporting slot	Solaris /dev/rdsk link
1	64 Bit/33Mhz 5Volt	Chip C Bus 2	/pci@6,4000
2	64 Bit/33Mhz 5Volt	Chip C Bus 2	/pci@6,4000
3	64 Bit/33Mhz 5Volt	Chip C Bus 2	/pci@6,4000
4	64 Bit/66Mhz 3.3 Volt	Chip C Bus 1	/pci@6,2000
5	64 Bit/66Mhz 3.3 Volt	Chip A Bus 1	/pci@1f,2000
6	64 Bit/66Mhz 3.3 Volt	Chip B Bus 1	/pci@4,2000
7	64 Bit/33Mhz 5Volt	Chip B Bus 2	/pci@4,4000

TABLE 1 Sun Enterprise 450 PCI Bus

PCI Slot Number	PCI Slot type	Chip & Bus Supporting slot	Solaris /dev/rdsk link
8	32 Bit/33Mhz 5Volt	Chip B Bus 2	/pci@4,4000
9	32 Bit/33Mhz 5Volt	Chip B Bus 2	/pci@4,4000
10	32 Bit/33Mhz 5Volt	Chip A Bus 2	/pci@1f,4000

With six PCI buses, PCI bandwidth or bus contention is rarely an issue with the Sun Enterprise 450. No PCI slot shares its PCI bus with more than two other cards. There are seven 64-bit PCI slots (1-7) that can support the 64 bit SRC/P card. To avoid any possibility of bus contention it would be best to use slots 4, 5, or 6. These slots go to PCI controller chips on their own dedicated bus. These are the best candidates for installing the SRC/P card and avoid any possible PCI bus contention as you continue to upgrade your system. Slot 5 is attached to a controller chip that supports only one other slot on another bus. If you are installing only one SRC/P card, and you have a choice, use Slot 5.

If you have a high bandwidth network card (Gigabit Ethernet or ATM interface) installed in the system, it would be best to place these types of cards in one of these dedicated PCI bus slots (4, 5, 6) as well. Failing to do so may result in bus contention when the system is demanding extremely high bandwidth.

If you already have a working system and want to determine which devices and drives that are supported by which controller on the various PCI buses, the chart can help you decode the output of `ls -l|grep pci` commands of the `/dev` or `/dev/rdsk` directories. In the following example, a previously configured SRC/P RAID 5 is seen by Solaris software as /dev/rdsk/c3t2d0s2. A quick look at the /dev/rdsk directory shows this device pointing to pci@1f,2000. TABLE 1 shows that this is slot 5.

```
eeserver6# cd /dev/rdsk
eeserver6# ls -go|grep c3t2d0s2
lrwxrwxrwx   1      62 Sep 15 13:51 c3t2d0s2 ->
../../devices/pci@1f,2000 pci@1/scsis@4/mscsi@0,0/sd@2,0:c,raw
```

Performance of RAID 5 Disk Subsystems on an Sun Enterprise 450

The Sun Enterprise 450 with a SRC/P card running PC NetLink is a popular server configuration. With respect to performance, the system planner may ask the following questions when planning this kind of server: What configuration options are offered by the SRC/P card? What type of RAID is best for performance? How many disks should the RAID volume have to make sure it doesn't bottleneck first?

We will benchmark a Sun Enterprise 450 server to answer these questions, focusing on the SRC/P card. Much of the information presented here was also used to develop the PC NetLink Sizing Guide available on the Sun web site at `http://www.sun.com/interoperability/netlink/whitepaper/sizing-guide/`

Benchmarks

Benchmarks that measure low level read and write I/O performance of a disk subsystem are extremely useful in planning some types of server functions such as databases. Unfortunately, it is difficult to take these kinds of benchmarks and predict how the disk subsystem performance might affect overall system performance of an Sun Enterprise 450 server running Solaris™ PC NetLink supporting a community of users. The Netbench benchmark from Ziff Davis is a good choice for emulating a large group of end users accessing files on the server via PCs. The benchmark was designed to simulate a high load of users accessing office productivity application files such as word processor files, spreadsheets, and presentations. Each PC client that runs the Netbench benchmark attempts to place as high a load of file operations as possible on the server. Over the period of several minutes they will place a load on the server equivalent to 10-100 PCs doing normal file access. For more information on Netbench, refer to the Ziff Davis Benchmark Operation web site: `http://www1.zdnet.com/zdbop/netbench/netbench.html`

Benchmark Configuration

To measure the effect the disk subsystem has on overall performance we ran the Netbench 6.0 benchmark against a Sun Enterprise 450. As we changed only the disk subsystem we reran the benchmark to see how the performance changed. The configuration we tested was a Sun Enterprise 450 server with 4x400Mhz Processors with one gigabyte of memory and one Gigabit Ethernet connection to the network. The PC clients used were 60 300Mhz Celeron PCs, each with 32Mbytes of memory and a full-duplex 100Mbit Ethernet connection to the network. The network was composed of three Extreme Summit 2 switches interconnected with Gigabit Ethernet.

The Sun Enterprise 450 server ran the Solaris 2.6 Operating Environment with no tuning performed. The Solaris PC NetLink was installed and configured without any performance tuning.

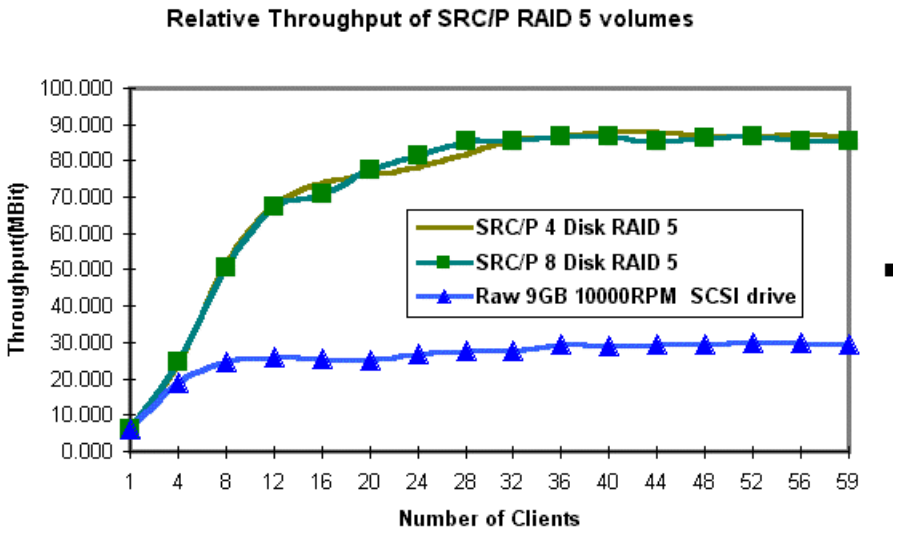
For the purposes of this article we will focus our attention on RAID 5 volumes because they allow for the maximum capacity and still allow for one disk failure redundancy. During each benchmark we also monitored the systems CPU activity to see if the benchmark was limited by CPU performance. The three configurations tested are summarized in the following table.

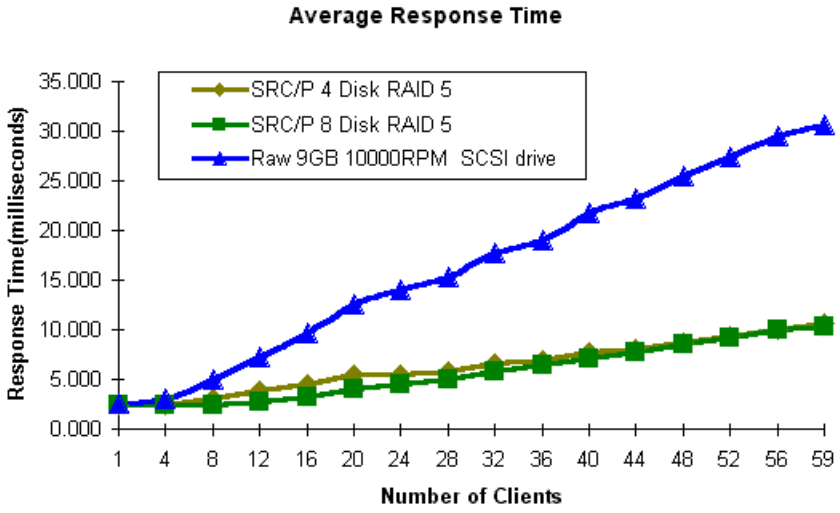
TABLE 2 Configurations Tested with Netbench on a Sun Enterprise 450 Server

Configuration	Note
9 GB 10,000 RPM SCSI drive	Used as a reference. The benchmark was limited by the drive
4 Disk RAID 5 SRC/P controller	Benchmark limited by CPU.
8 Disk RAID 5 SRC/P controller	Benchmark limited by CPU.

The two charts that follow show the total throughput in Mbits/Second and average response time in milliseconds of running the benchmark by changing only the disk subsystem the benchmark was configured to use.

The charts shows that as the number of PC client loads were increased we reached different maximum levels of throughput for each disk subsystem tested.





Both charts show the throughput, and response time, of the server as an ever-increasing load of 100 percent duty cycle PCs were added to the benchmark. The goal of the benchmark, as far as we used it, was to increase the load of PCs performing the benchmark until the server could deliver no further throughput.

Looking at the first chart, let's first contrast the simple RAW SCSI disk performance against any of the RAID configurations. You will see that a RAW SCSI drive limits the benchmark quickly. It took only four PC clients running the benchmark to saturate the drive. While these are high performing 10000 RPM drives, they still were slow (and with no redundancy) when compared to the RAID volumes supported by SRC/P card. As the benchmark added additional clients, the latency response times grew. While the second chart shows average response times, it does not show the maximum response times, which can be considerable.

The benchmarks of the two RAID environments is somewhat anti-climactic. The two volumes tested look almost equal. What can account for this? The Write back cache in the SRC/P card is doing it's job. It allows the Solaris Operating Environment software to immediately return from most operations to the card with little or no wait time. The Solaris driver controlling the card will actually see no delay from the time it issues a write request to the point the card reports the request is completed. The battery backup ensures that even if power is lost to the server, uncompleted disk operations will be finished once power is restored.

During each of the benchmarks the CPU utilization was monitored. It was confirmed that in each case the SRC/P card was able to deliver enough performance to allow the four processors on the Sun Enterprise 450 server to saturate running the

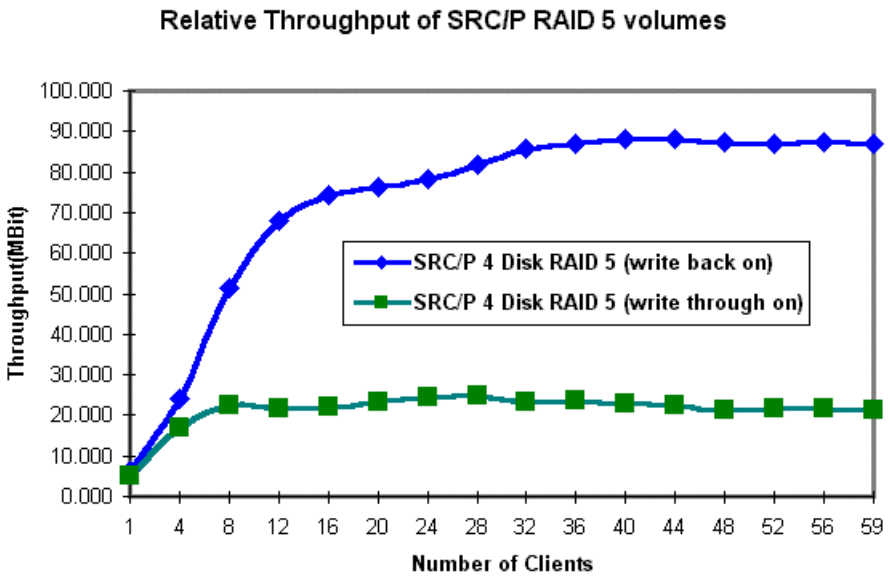
benchmark. Each processor was at 100 percent utilization (high values for `mpstat usr` or `sys` values, not the `wt` or `idl` states) supporting the PC NetLink V1.1 environment while the benchmark was showing its maximum throughput. This means the disk subsystem was not saturated and was not the bottleneck.

This tells us that under similar loading conditions, the system planner can expect the server to deliver maximum performance with any RAID 5 volume consisting of more than four disks. If you need to use small four-disk RAID 5 volumes to manage backups or user groups better, performance should not be an issue for similar loading conditions.

Write Though vs. Write Back

Returning our attention to the battery backed-up write back cache, we will look at the performance of the four disk RAID 5 volume with the write back cache turned off. If the SRC/P card detects that the battery on the card is not charged, it automatically sets the RAID volumes to write through mode. This mode forces all write operations to be committed to disk before the SRC/P card will acknowledge they are completed to the Solaris Operating Environment software. The end result is significantly lower performance until the battery has sufficient charge. If you are installing the SRC/P card for the first time, or the card has not been in a powered system for several days, the battery charge may be too low to support the memory during a power failure. Until the battery is charged you will experience slow performance.

Rerunning the benchmark once more with write through selected results in the following graphs. As expected the performance has dropped considerably.



You can force the write-back to be turned on while the battery is being charged. This is not advisable because it will put the volume at risk if a power failure occurs. It is best to wait for the battery to charge and the write-back cache to be turned on automatically.

Capacity

TABLE 3 shows the relative capacities of four, eight, and twelve disk RAID 5 volumes after configuring, formatting, and creating a UFS file system with the `newfs` command.

TABLE 3 Relative Capacities of Four, Eight, and Twelve Disk RAID 5 Volumes

Type of Disk Volume	Size of Drive Reported by `df -k`	Capacity Compared to Equivalent RAW SCSI Drives	Capacity Efficiency Compared to RAW SCSI
RAW 9 Gigabyte SCSI drive	8705501	1	1 / 1 =100%
SRC / P 4 (9GB) disk RAID 5	26112983	2.999	3 / 4 = 75%
SRC / P 8 (9GB) disk RAID 5	60934332	6.999	7 / 8 = 87.5%
SRC / P 12 (9GB) disk RAID 5	95782979	11.00	11 / 12 = 91.6%

The capacity efficiencies of the RAID volumes were calculated when compared to a RAW SCSI 9GB drive. As you would expect with RAID 5, volume capacities are extremely close to the RAW SCSI disk size (the number of drives in the RAID 5 volume - 1). One disk drive was used to support parity for the RAID 5 volume, so its capacity was not available for the RAID volume.

Only if large volumes and storage capacity efficiency are absolutely required, should you consider making large (more than eight disks) simple RAID 5 volumes. The reason for this is to minimize the risk of a secondary disk failure. This topic is covered separately below. If larger volumes are required, use RAID 5+0 (Parity groups on SRC / P) or used in RAID 0 software RAID environments. The SRC / P Storage Manager GUI will suggest this option whenever you attempt to make a RAID 5 volume of more than six drives. Aside from capacity and performance concerns, another issue is significantly more important to most customers—specifically the risk of a secondary disk failure.

Reducing the Risk of a Secondary Disk Failure

Even with redundant RAID volumes there are periods of time when the volume is vulnerable to a second disk failure. This occurs between the time that the first disk fails and the time the replacement disk has been rebuilt with the data that was on the original failed drive. It is this period of time we want to minimize as much as possible. First let's look at the risk.

Mean Time Between Failure (MBTF)

Disk drives have increased in size over the last few years to 18GB and more. When placed into RAID 5 volumes the size of the total volume becomes extremely large. Losing this amount of data to a second disk failure is a nightmare.

Fortunately the MBTF for individual drives has gone up over the years allowing the risk of having large capacity acceptable. The MBTF of a typical drive shipped by Sun is 1,000,000 hours. These MBTF values are statistical in nature and can lead to a false sense of security. When disks are placed into a RAID environment. The MBTF of a the RAID environment becomes

Total RAID MTBF = (MTBF of 1 Disk) / (Number of disks in the RAID volume)

Redundant RAID environments allow you to continue through the first failure but the MBTF to the second failure is the same equation with one less drive. If a drive fails in a 12-disk RAID 5 environment consisting of 1,00,000 MBTF drives, you are left with 11 drives. The MBTF of this now non-redundant environment is 90909 hours or 10.3 years. If the failure goes unnoticed for long periods of time even this seemingly long MBTF increases the risk considerably that a second error can occur causing full loss of the volume.

There are several steps you can take to reduce the time period a RAID volume is vulnerable to a second disk failure.

When the Volume is Susceptible to a Second Failure

A portion of the time the RAID volume is susceptible to a second disk failure cannot be avoided. This is the time the hardware SRC/P firmware, takes to rebuild the data that was on the original failed drive. This period of time can vary from a few

minutes, to several hours. The length of time this rebuild can take is difficult to predict on an active system. The period of time depends on the size of the drive, the amount of data that was on the drive, and the activity of the system during the rebuild period.

The SRC/P will attempt to rebuild the disk data as quickly as possible, but if it is doing the rebuild while the RAID volume is being used by the system, the rebuild operation can be lengthened considerably. By default, the SRC/P considers the rebuild operation as a low priority operation. A heavily used system that is constantly using the RAID volume that is in the process of rebuilding a disk after a failure can have extremely long rebuild times. In addition to the long rebuild times the system will have degraded performance. The SRC/P controller must synthesize the data lost from the failed drive for every read operation to the disk.

The priority level of the rebuild operation is configurable by way of the SRC/P Storage Manger GUI Options menu. There is a selection to change the Background Task Settings. Raising the priority from the low default setting, allows the rebuild to finish sooner on an active system. If you do raise the priority, expect lower performance of the disk subsystem during the disk failure rebuild times.

Building the Parity on a New RAID 5 Volume

While performing benchmarks for this article, I unknowingly did a benchmark while the SRC/P was building a new RAID 5 volume. After realizing what had happened I reran the benchmark after the rebuild had completed and noticed there was no significant difference in the results. The reason for this is that new volumes have no real data established yet. There will be little performance penalty in using a new volume as its parity is being initialized. However, be aware that until the parity is fully constructed the RAID 5 volume is vulnerable to a single disk failure.

Before allowing the disk to be used in production, wait for the parity build to complete. Monitor the SRC/P SUN Storage Manager GUI to see when the build has completed.

Making a Replacement Drive Available after a Disk Failure

Besides the unavoidable parity rebuild time, the other portion of the time when the RAID volume is vulnerable, is the time it takes the SRC/P controller to see a new replacement to start the rebuild process. This time can be anywhere from 0 seconds to several months if a someone doesn’t notice the disk failure.

Assigning a Hot Spare Drive

Assigning a hot spare drive, via the SRC/P GUI is an effective way to minimize the chance of a second drive failure. A hot spare allows the SRC/P firmware to start rebuilding a RAID volume as soon as a drive failure is detected. If a hot spare is NOT allocated, the period of time when the RAID volume is working without redundancy can be from a few minutes, where there are effective procedures for dealing with the failure, two many days where the failure is not noticed, ignored, or time is needed to allocate a spare. Considerable time can be wasted for the system administrator to detect the failure, allocate a replacement drive of the correct size, and install it into the system. This human response time to install a replacement disk is typically many times the rebuild time, increasing the risk considerably.

Having a spare disk allocated (by way of the SRC/P manager GUI) allows this rebuild of the failed disk data to start immediately, eliminating the human response requirement. Disk prices have also come down over the years. Assigning a spare disk drive is an extremely inexpensive way to reduce the risk of a second disk failure crash that results in the loss of many Gigabytes of data.

To reduce the chance of a second disk failure destroying the RAID volume, assign at least one spare drive to each SRC/P card and make sure you a have a procedure to detect and replace failed drives as soon as possible. If policies are hard to enforce, assign two drives as spares.

Use of Parity Groups

Whenever you define a RAID 5 of more than six drives, the SRC/P Sun Storage Manager GUI will offer to set up parity groups. Parity groups are a mechanism by which large RAID 5 volumes are implemented by way of two RAID 5 volumes striped by RAID 0. This is also know as RAID 50. Parity groups support multiple RAID 5 parity drives, essentially reducing the risk of a second disk failure crashing the volume. If you must create a RAID 5 volume larger than six drives, it is highly advisable to take advantage of this feature. The downside to parity groups is that they do require more disks be used to support the additional parity groups.

Detecting the Hard Disk Failure

It is critical to detect a hard disk failures as soon as it happens and replace the failed drive as soon as possible. The Sun StorEdge SRC/P Controller hardware and SRC/P SUN Storage Manager GUI software have several mechanisms to make sure someone is notified of the failure.

The SRC/P card has its own audible alarm installed directly on the SRC/P card. By default this alarm is turned off. Use the SRC/P SUN Storage Manager GUI to turn the alarm on or off. The alarm is reasonably loud and is an excellent way to grab the

attention of a busy system administrator. If you have never heard the audible alarm before and you forget it's there, it can be quite baffling what the noise is coming out of your server. To help explain the noise to system administrators, who may have never heard the alarm before, it would be useful to place a large sign taped to the system saying "If this system is making an alarm sound, check the SRC/P SUN Storage Manager GUI (/opt/SUNWhwrdg/dptmgr) for a disk failure".

The SRC/P SUN Storage Manager GUI offers a variety of mechanisms to notify users, groups, and other individuals of SRC/P disk failures as well as soft disk errors and recoverable hard disk errors. These events can be sent via email, to groups or users. They can also be sent to event logs, sent to `/var/adm/messages` or to other devices.

Use as many of these features to detect a disk failure as is deemed necessary to guarantee something will be done ASAP. Loosing a large volume to second disk failure, because the system administrator was not aware the first disk failure occurred is a poor reason to loose data. The SRC/P controller and software offer many options to notify you of a problem. Be sure to take advantage of them.

Testing the Disk Failure Notification

Whatever mechanism you choose to detect a disk failure, be sure to test it to make sure it works. Before putting a SRC/P RAID volume in production force a drive failure by using the SRC/P SUN Storage Manager GUI to simulate the failure. double clicking RAID drive icons in the GUI will produce a window that has a Fail Drive Button.

Complex RAID Environments

Once disk drives have been placed into a RAID environment on the SRC/P card, the whole RAID volume appears to the Solaris Operating Environment software as just another SCSI disk. Once several RAID volumes have been created with the appropriate spares, parity groups, and disk failure procedures, they can be members of even larger software RAID environments such as Sun Enterprise Volume Manager software and the Sun Solaris DiskSuite software. Filling an Enterprise 450 with several SRC/P cards and attaching external diskpacks can produce single volumes of considerable size.

The largest single redundant volume (with spares) that can be produced with internal Sun Enterprise 450 server 9 GB disk drives would require eight disks attached to each of two SRC/P controllers. Each controller would support seven disk RAID 5 volumes, with one disk assigned as a spare. When striped with Sun

Enterprise Volume Manager software or Sun Solaris DiskSuite software this would produce a volume of approximately 99.6 Gigabytes. Using external disk packs and multiple SRC/P controllers allows you to produce extremely large volumes.

The maximum number of SRC/P cards that can be placed into a Sun Enterprise 450 in a supported configuration is six cards. Each SRC/P card has three external SCSI buses, each of which can have a Sun StorEdge system 12-pack of 18 Gigabyte drives. This adds up to storage in the neighborhood of 3888 (6x3x12x18) Gigabytes that can be incorporated into a variety of hardware and software RAID environments.

Creating extremely large RAID volumes allows very large databases to be supported, but the more disks you have involved in the RAID volume the higher the risk that a single disk will fail. This makes detecting and replacing failed drives even more critical.

Backup

Using redundant RAID volumes does not imply you don't need to backup. If users accidently delete important data, only backups can restore these files. There are periods after a disk failure when the SRC/P controller needs to rebuild the data of the failed drive. While a second disk failure during this period is unlikely, it is possible, and will happen to someone somewhere. The more systems you manage, the more RAID volumes you have configured the more likely it will happen to you. Nightly backups are the only effective way to minimize the impact of this possibility happening to you.

Author's Bio: Don De Vitt

Don has been on the development teams of almost every software and hardware PC interoperability Product Sun Microsystems has produced over the last 13 years. Don is currently a PC Interoperability specialist within the Enterprise Engineering group and is a member of the PC NetLink engineering team where he has focused on performance related issues.

Don De Vitt started his career as an electrical engineer and has worked in the Automated Test industry (Teradyne Inc.), and PC operating system market (Digital Research from CP/M fame) before coming to Sun Microsystems, Inc.