realtimepublishers.com™

The Definitive Guide™ To

Building Highly Scalable Enterprise File Serving Solutions

PolyServe™

Chris Wolf

Copyright Statement

Chapter 3: Data Path Optimization for Enterprise File Serving

With enterprise file serving, much of the attention concerning availability and high performance is focused on the servers themselves. However, the clients that access data on file servers face many other obstacles and potential bottlenecks along the way. For a client's request to reach a file, it must traverse the network, reach the server, and, ultimately, the server must request the file from its attached storage. This chapter will view the entire landscape of the file serving picture. Although attention will be paid to the servers themselves, you will also see the common problems and pitfalls involved with getting data from storage to a client.

The Big Picture of File Access

Optimizing and protecting file access involves more than just setting up fault-tolerant disks on each file server. Figure 3.1 illustrates a simple example of the objects involved in a client accessing a file from a server.

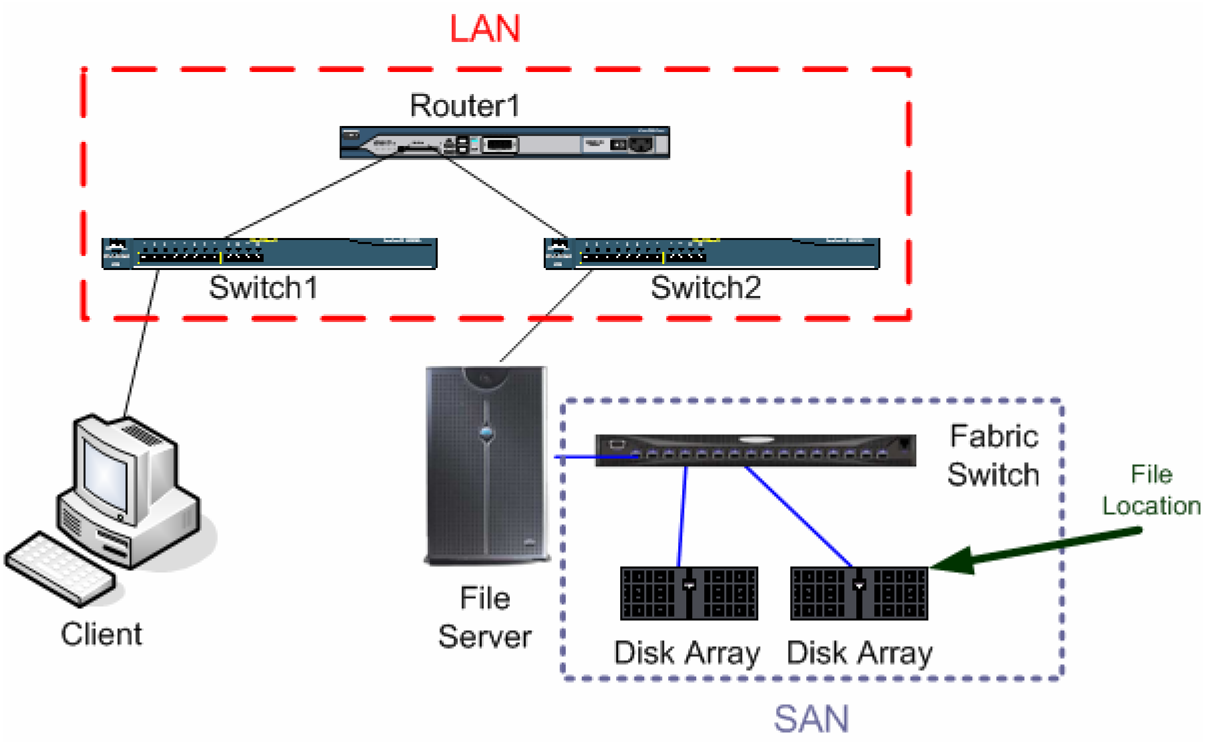


Figure 3.1: Objects in the data path.

For the client to open the file residing on the SAN, the client's request will need to traverse Switch1, Router1, and Switch2 to reach the file server. For the file server to answer the request, the server will need to pull the file from the disk array through the fabric switch on the SAN. In this example, each device between the client computer and the file represents both a potential bottleneck and a single point of failure. A *single point of failure* is any single device whose failure would prevent data access.

Availability and Accessibility

Many administrators tout the fact that some of their servers have been available 99.999 percent of the time during the past year. To make better sense of uptime percentages, Table 3.1 quantifies uptime on an annual basis.

Uptime Percentage	Total Annual Downtime
99%	3.65 days
99.9%	8.75 hours
99.99%	52.5 minutes
99.999%	5.25 minutes

Table 3.1: Quantifying downtime by uptime percentage.

With 99 percent uptime, you have 1 percent downtime. One percent of 365 days yields 3.65 days. If you divided this number by 52 (number of weeks in a year), you would average being down 0.07 days a week. This statistic equates to 1.68 hours (0.07 days × 24 hours) or 1 hour and 41 minutes of downtime per week. If you advance to the 5 nines (99.999) of availability, a server would need to be offline no longer than 5.25 minutes a year. This statistic equates to only 6 seconds of allowable weekly downtime!

Keeping your file servers available is always important. The amount of uptime that is required often varies by organization. For example, if no one will be accessing a file server on a Sunday, it might not be a big deal if it is down for 8 hours. For other shops that require 24×7 access, almost any downtime is detrimental.

When measuring uptime, most organizations simply report on server availability. In other words, if the server is online, it’s available. Although this method sounds logical, it is often not completely accurate. If the switch that interconnects clients to the server fails, the server is not accessible. It might be online and available, but if it is not accessible, it might as well be down.

Thus, uptime percentages can be misleading. Having a server online is only valuable when the data path associated with the server is online as well. To truly deploy a highly available file server, the data path must be highly available as well. For high-performance file serving, the same logic holds true. To meet the performance expectations of the server, the data path must be able to support getting data to and from the server at a rate that—at a minimum—meets client demands.

The next sections in this chapter provide examples of how to ensure availability and performance in the following data path elements:

- Redundant Storage
- SANs
- LAN switches and routers
- Servers
- Power

Let’s start by looking at how to add redundancy and performance to storage resources.

Redundant Storage

Redundant storage can offer two elements that are crucial to high-performance and high-availability file serving. In configuring redundant storage, or Redundant Array of Independent Disks (RAID), you can configure two or more physical disks to collectively act as a single logical disk. This combination can result in both better performance and fault tolerance. In this section, you will see the most common types of RAID configurations as well as their relation to improving enterprise file serving.

RAID Levels

RAID levels are described by number, such as 0, 1, or 5. The most common RAID implementations today are:

- RAID 0
- RAID 1
- RAID 5
- RAID 0+1
- RAID 1+0
- RAID 5+0

Let’s start with a look at RAID 0.

RAID 0

RAID 0 is not considered to be true RAID because it does not offer redundancy. Because of this, RAID 0 is often combined with other RAID levels in order to achieve fault tolerance.

Although not fault tolerant, RAID 0 does offer the fastest performance of all RAID levels. RAID 0 achieves this level of performance by striping data across two or more physical disks. Striping means that data is being written to multiple disks simultaneously. All the disks in what is known as the stripe set are seen by the operating system (OS) as a single physical disk. RAID 0 disk striping is depicted in Figure 3.2.

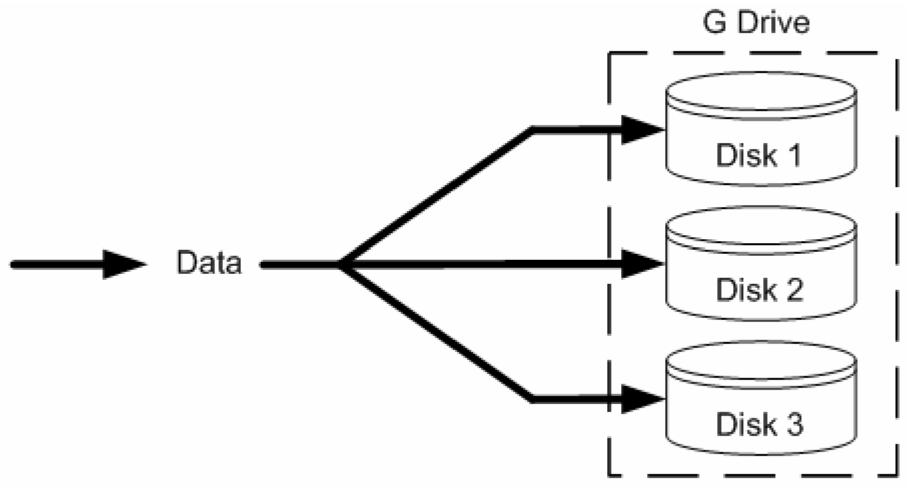


Figure 3.2: RAID 0 operation.

To understand how RAID works, consider the following example. Suppose you wanted to store the word "get" in a RAID 0 array containing three disks. Now picture each disk as being a cup. Since "get" has three letters, a single letter would be stored in each cup. With the letters evenly spaced out, you could theoretically drop all the letters into the three cups simultaneously. This scenario illustrates the advantage of RAID 0—it's fast. The problem, however, is if one of the cups (disks) is lost or damaged, all data is lost.

Because of its lack of fault tolerance, RAID 0 is not considered to be a good fit for file serving. Its raw performance makes it ideal for high-speed caching but makes it risky for storing critical files. As you will see later in this chapter, RAID 0 can be combined with other RAID levels to achieve fault tolerance. This setup can give you the best of both worlds—speed and resiliency. The remaining RAID levels discussed in this section offer fault tolerance at the expense of some of the performance found in RAID 0.

RAID 1

RAID 1 is the first fault-tolerant RAID level, and is available in two forms:

- Disk mirroring
- Disk duplexing

With both disk mirroring and disk duplexing, two or more physical disks provide redundancy by having one or more disks mirror another. Data written or deleted from one disk in the mirror set is automatically written or deleted on all other disks in the set. With this approach, fault tolerance is ensured by having redundant copies of the same data on several disks. The failure of a single disk will not cause any data loss.

The disk mirroring and disk duplexing implementations of RAID 1 differ in how the physical drives are connected. With disk mirroring, the physical disks in the mirror set are connected to the same disk controller. With disk duplexing, the physical disks in the mirror set are connected using at least two disk controllers. Disk duplexing is the more fault tolerant RAID 1 implementation because it eliminates a disk controller as a single point of failure.

When a file is saved to a two-disk RAID 1 array, it is written to both disks simultaneously. Thus, the actual write operation does not complete until the file is finished being written to the slowest disk. The result of this architecture is that the actual performance of the RAID 1 array will be equal to the speed of the slowest disk.

RAID 1 is ideal when you are looking for an easy means to ensure data redundancy. RAID 1 automatically creates a mirror of disks, so you are getting a continuous online backup of data. This setup allows for little to no data loss in the event of a disk failure.

> RAID should not be considered a substitute for backing up data. Although fault-tolerant RAID protects against the failure of a single disk, it does not protect against data corruption or disasters. Because of this shortcoming, regular backups to media stored offsite should still be performed.

The one disadvantage to RAID 1 is that you have to purchase at least twice the amount of disk space for the data you want to store, depending on the number of disks in the RAID 1 mirror. If you are planning to configure two disks to mirror each other, remember that one disk will work exclusively as a backup. For example, a 100GB RAID 1 volume consisting of two physical disks would need a total of 200GB of storage (two disks × 100GB).

RAID 5

RAID 5 operates similar to RAID 0 by striping data across multiple disks. However, it differs in the following ways:

- RAID 5 uses parity to achieve fault tolerance
- RAID 5 requires three or more physical disks (RAID 0 only requires two disks)

With each write, RAID 5 writes a parity bit to one disk in the array. This functionality allows a RAID 5 array to lose a single disk and still operate. However, if a second disk in the array fails, all data would be lost. This loss would result in having to rebuild the array and restore data from backup. In terms of performance, RAID 5 is slower than RAID 0, but outperforms RAID 1.

As RAID 5 uses parity to provide fault tolerance, you must consider the storage of the parity data when sizing a RAID 5 array. Parity writes will effectively take up one disk in the array. Thus, with a three-disk array, two disks would store actual data and the third disk would store the parity bits. If you built a RAID 5 array with three 100GB disks, you would have 200GB of available storage, enabling you to store actual data on 67 percent of your purchased storage. If you add disks to the array, the efficiency of the array improves. For example, with four disks in the array, three disks would store data and one disk would store parity bits, giving you 75 percent utilization of your disks.

RAID 5 has been very popular for enterprise file serving because it offers better speed than RAID 1 and is more efficient. Although it is slower than RAID 0, the fact that it provides fault tolerance makes it desirable.

RAID 0+1

RAID 0+1 arrays provide the performance of RAID 0 as well as the fault tolerance of RAID 1. This configuration is commonly known as a mirrored stripe. With RAID 0+1, data is first striped to a RAID 0 array and then mirrored to a redundant RAID 0 array. Figure 3.3 shows this process.

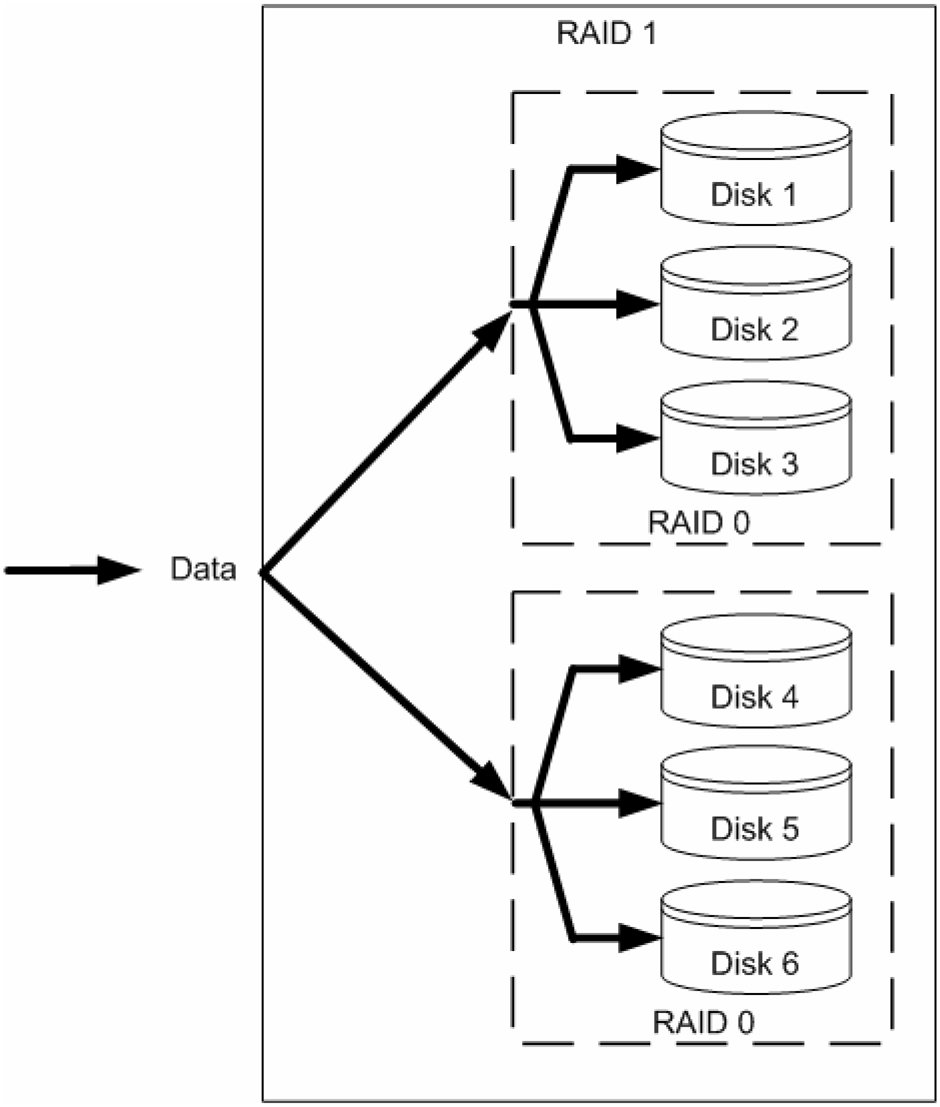


Figure 3.3: RAID 0+1 operation.

RAID 0+1 is configured by first creating two RAID 0 arrays and then creating a mirror from the two arrays. This approach improves performance, but the inclusion of RAID 1 means that your storage investment will need to be double the amount of your storage requirement. Assuming that the illustration that Figure 3.3 shows uses 100GB disks, each RAID 0 array would be able to store 300GB of data (100GB × three disks). As the second RAID 0 array is used for redundancy, it cannot store new data. This setup results in being able to store 300GB of data on 600GB of purchased disk storage.

An advantage to RAID 0+1 is that it offers the performance of RAID 0 and provides fault tolerance. You can lose a single disk in the array and not lose any data. However, you can only lose one disk without experiencing data loss. If you're looking for better fault tolerance, RAID 1+0 is the better choice.

RAID 1+0

RAID 1+0 (also known as RAID 10) combines RAID 1 and RAID 0 to create a striped set of mirrored volumes. To configure this type of RAID array, you first create mirrored pairs of disks and then stripe them together. Figure 3.4 shows an example of this implementation.

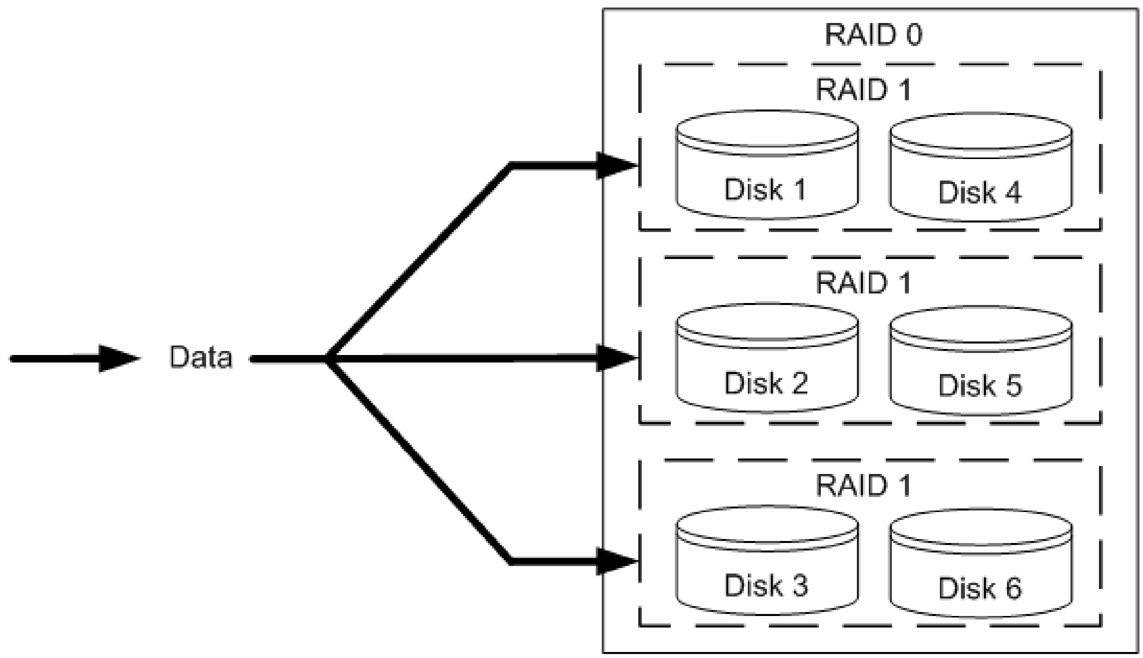


Figure 3.4: RAID 1+0 operation.

Note that the RAID 1+0 configuration is exactly the opposite of RAID 0+1. With RAID 1+0, the RAID 1 arrays are first configured. Then each mirror is striped together to form a RAID 0 array. A major advantage of RAID 1+0 over RAID 0+1 is that RAID 1+0 is more fault tolerant. If there are a total of six disks in the array, you could lose up to three disks without losing any data. The number of disks that can fail is determined by where the failures occur. With RAID 1+0, as long as one physical disk in a mirror set in each stripe remains online, the array will remain online. For example, the array that Figure 3.4 shows could lose disks four, two, and six and remain operational. As long as one disk in each RAID 1 mirror remains available, the array will remain available as well.

RAID 1+0 is similar to RAID 0+1 in terms of storage efficiency. If each disk in the array shown in Figure 3.4 is 100GB in size, you would have 600GB of total purchased storage but only 300GB of writable storage (due to the RAID 1 mirroring). If you're looking for better storage efficiency at the sake of a little speed, RAID 5+0 might be a better option.

RAID 5+0

RAID 5+0 is configured by combining RAID 5 and RAID 0. This array would be configured by first striping data across RAID 5 volumes. This setup is more efficient than RAID 1+0 because only a fraction of your storage investment is lost, instead of half the investment. Figure 3.5 shows this RAID type.

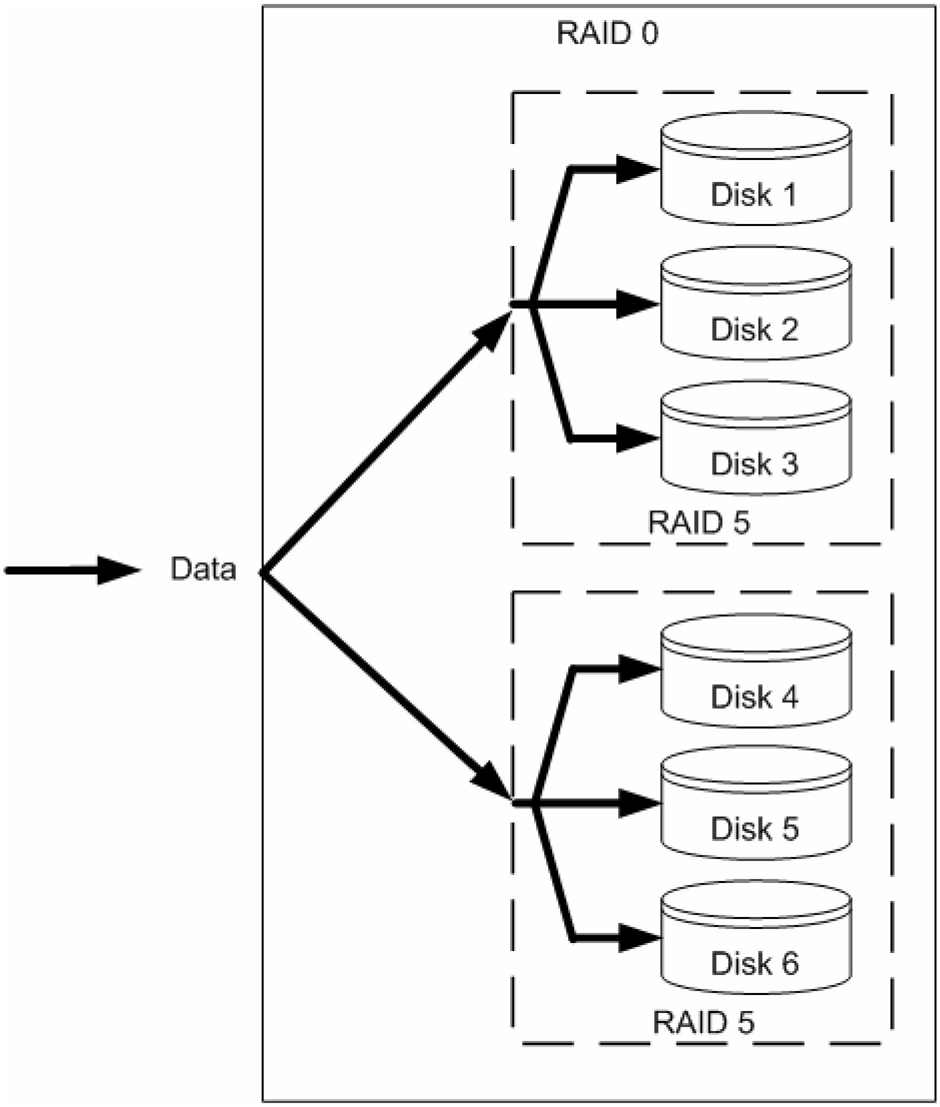


Figure 3.5: RAID 5+0 operation.

Compared with RAID 5, RAID 5+0 provides faster read and write access. However, a problem with RAID 5+0 is that if a drive fails, disk I/O to the array is significantly slowed. Unlike RAID 5, RAID 5+0 is more fault tolerant because it can withstand the loss of a disk in each RAID 5 subarray. With the array that Figure 3.5 shows, both disks one and five could fail and the array would still remain online. If disks four and five failed, the array would go down.

RAID 5+0 sizing is similar to sizing a RAID 5 array, except that a disk in each RAID 5 subarray is used for parity. If the array in the figure had 100GB disks, you would have 600GB of storage space in the array. Of the 600GB, you could write data to 400GB because you give up one 100GB disk in each of the subarrays to parity. As with RAID 5, adding disks to each subarray would provide better storage efficiency.

Hardware vs. Software RAID

As you can see, there are several methods for improving storage performance and fault tolerance for your file servers. When looking to configure each of these RAID levels, you have two general choices—hardware RAID and software RAID. Hardware RAID volumes are set up using a hardware RAID controller card. This setup requires the disks in the array to be connected to the controller card. With software RAID, disks are managed through either the OS or a third-party application. Let’s look at each RAID implementation in more detail.

Hardware RAID

Hardware RAID controllers are available to support all of the most common disk storage buses, including SCSI, Fibre Channel (FC), and Serial ATA (SATA). Hardware RAID is advantageous in that it’s transparent to the OS. The OS only sees the disk presented to it by the RAID controller card. Hardware RAID also significantly outperforms software RAID because no CPU cycles are needed to manage the RAID array. This management is instead performed by the RAID controller card.

Another advantage of hardware RAID is that it is supported by most clustering implementations, whereas most software RAID configurations are not supported by cluster products. To be sure that your configuration is compatible, you should verify that the hardware RAID controller and disk products have been certified by your clustering product vendor.

> ☞ Sometimes hardware alone is not the only compatibility issue. Be sure to verify that your installed hardware is using firmware and drivers that have been certified by your clustering product vendor.

Most of the major RAID controller vendors post technical manuals for their controllers on their Web sites. This accessibility makes it easy to configure RAID controllers in your file serving storage infrastructure.

SCSI RAID

Most of the hardware RAID implementations in production today are in the form of SCSI RAID. Although FC and SATA have gained significant ground in recent years, SCSI still has maintained a large portion of market share. Among the most popular SCSI RAID controller vendors are Adaptec, QLogic, and LSI Logic. Each of these vendors offer products with long-standing reputations and excellent support.

FC RAID

With FC RAID, disk arrays can be configured as RAID arrays and attached directly to a SAN. As SANs have continued to become an integral part of enterprise file serving, FC RAID has risen in popularity. For example, Adaptec’s SANbloc 2Gb RAID solution can allow you to connect FC RAID to a cluster via a SAN and can scale to 112 drives with as much as 16.4TBs of storage. Other vendors that offer FC RAID solutions include Hewlett-Packard, Quantum, Dot Hill, and XioTech.

SATA RAID

SATA has been steadily growing in recent years as a cost-effective alternative to SCSI. SATA offers data transfer rates as fast as 450MBps, depending on the SATA RAID controller and disk vendor. Besides the lower cost, SATA also differs from SCSI in that each disk has a dedicated serial connection to the SATA controller card. This connection allows each disk to utilize the full bandwidth of its serial bus. With SCSI, all disks on the bus share the bandwidth of the bus.

Unlike SCSI, SATA disks are not chained together. Thus, the number of disks in the array will be restricted to the physical limitations of the controller card. For example, the Broadcom RAIDCore 4852 card supports eight ports and all of the popular RAID levels, including RAID 1+0 and RAID 5+0. This controller provided RAID 0 writes at 450MBps and RAID 5 writes at 280MBps during vendor tests.

Many vendors are also developing technologies that allow you to connect SATA disk arrays to an FC SAN. For example, the HP StorageWorks Modular Smart Array (MSA) controller shelf can allow you to connect as many as 96 SATA disks to an FC SAN. This feature gives you the ability to add disk storage to support your file servers on the SAN at a significant cost savings over SCSI.

Software RAID

Software RAID is advantageous in that you do not need a RAID controller card in order to configure it. However, with many organizations deploying clustering technology to support the demands of file serving, software RAID has not been a possibility for shared storage resources in the cluster. There are some exceptions to this rule. For example, Symantec (formerly VERITAS) Volume Manager can set up software RAID that is compatible with some clusters such as Microsoft server clusters. However, most organizations that spend the money to deploy a cluster in the first place don’t try to save a few bucks by cutting corners with RAID.

Having an OS control disks via software RAID can also result in significant CPU overhead. The CPU loading of software RAID often makes it impractical on high-volume enterprise-class file servers. However, some organizations that connect shared cluster storage via hardware RAID will use software RAID to provide redundancy for the OS itself. Having the OS files mirrored across a RAID 1 array can prevent a disk failure from taking a server down. You can also do so with a hardware RAID controller, but if you’re at the end of your budget, you might find using software RAID to protect the OS to be an alternative. As with hardware RAID, you still must use multiple physical disks to configure the RAID array, so breaking a disk into partitions to build a software RAID array is not an option.

Windows OSs natively support software RAID, which can be configured using the Disk Management utility, which is a part of the Computer Management Microsoft Management Console (MMC). Using Disk Management, you can configure RAID 0, 1, and 5 arrays on Windows Server OSs. On Windows client OSs, you can only configure RAID 0 using Disk Management.

With Linux OSs, software RAID 0, 1, and 5 can be configured using the Disk Druid tool during a GUI installation of the OS. If the OS is already installed, you can use the Raidtools package to configure and manage software RAID. Although increasing performance and availability of disks through RAID is an important part of enterprise file serving, there are more elements of the data path that must be considered as well.

Redundant SAN Fabrics

Redundant disks are not of much value if there is only a single path to the disks through a SAN. With this in mind, if you have protected your disk storage through RAID, you should also strongly consider adding redundant data paths between servers attached to the SAN and the storage resources.

Elements of the Redundant SAN

A fully redundant SAN has no single point of failure. An example of a redundant SAN fabric is shown in Figure 3.6.

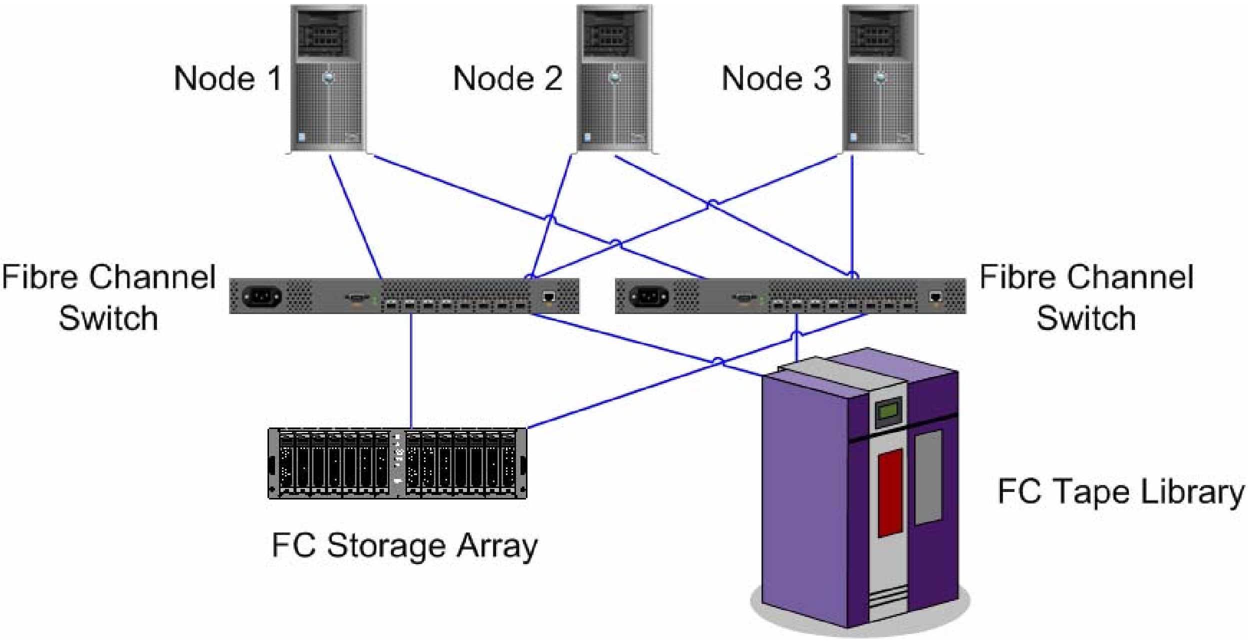


Figure 3.6: Redundant SAN fabric.

In this example, three servers that are part of a shared data cluster all share common storage in a SAN. Fault tolerance begins with redundant FC HBAs in the servers. This setup eliminates an HBA as a single point of failure. Each HBA connects to a separate FC switch. This way, all three servers can withstand the failure of a switch or switch port in the SAN. Finally, the disk array and library in the SAN are also connected to each switch.

Although redundancy adds to the cost, many organizations with SANs have seen redundancy as a necessity. For example, a large hospital recently deployed a non-redundant SAN to connect their file servers to disks resources. The goal was to get better use of their existing storage resources while streamlining backups. However, the administrator's opinion of SANs faded when a switch failed and as a result took down seven servers. With a FC switch serving as the access point to SAN storage, the switch's failure could have devastating consequences.

With a fully redundant SAN fabric, a switch failure will not equate to the failure of several servers. Instead, it will simply be a minor hiccup. However, getting all of this to work is not as simple as just connecting all the devices. Each host OS must be aware of the duplicate paths through the SAN to each storage resource. For this setup to work, you will need to install the multipath drivers for the SAN HBA. You will also need to ensure that the purchased SAN HBAs support multipath.

Managing the Redundant SAN

All of the major SAN switch vendors offer tools to help you manage their products. For example, Brocade's Fabric Manager allows you to manage as many as 200 switches in a SAN. With this product, you can make changes to all switches simultaneously or can make changes to individual switches or even small groups. You can also configure alerting features to alert you if a failure occurs. Other storage vendors have also jumped into the SAN management ring by offering products that collectively manage a variety of SAN hardware devices. Symantec's CommandCentral is an example of software that can manage a diverse collection of storage resources across an enterprise.

There are several products that can assist you in spotting failures on a SAN. How you deal with failures may depend on your IT budget. Some organizations maintain spare parts on hand to quickly resolve failures. This practice could mean having a spare HBA, switch, and FC hard disks. This way, if a failure occurs, you can quickly replace the failed component. Once the failed component is replaced, you can then order the replacement.

☞ Most SAN products have built-in backup utilities. To quickly replace a failed switch and update its configuration, you should perform frequent backups of your SAN switches. Many organizations perform configuration backups before and after each configuration change to a SAN switch. This practice ensures that you will always have the most recent configuration available if you need to replace a failed switch.

Redundant LANs

At this point, you have seen how to improve performance and fault tolerance in the data path from a server to storage. Another element of the data path that is also crucial is the path from the clients to the servers. This path often encompasses the LAN. A simple example of adding fault-tolerant LAN connections to servers is shown in Figure 3.7.

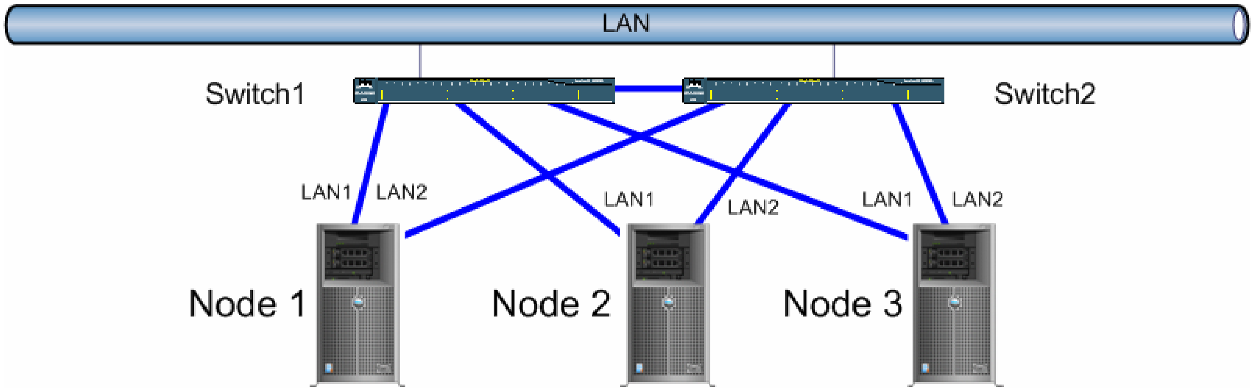


Figure 3.7: Redundant server LAN connections.

The idea of redundant LAN connections is relatively straightforward. As with redundant SAN connections, the redundant LAN illustrated uses a meshed fabric to connect each node to two switches. This approach will make each node resilient to NIC, cable, or switch failure. With this approach, a teamed NIC driver should be installed on each server. This installation will allow the two NICs to be collectively seen as a single NIC and share a virtual IP address.

Aside from meshing server connections to switches, some organizations mesh connections between switches, thus providing for additional resiliency. Figure 3.8 shows this architecture.

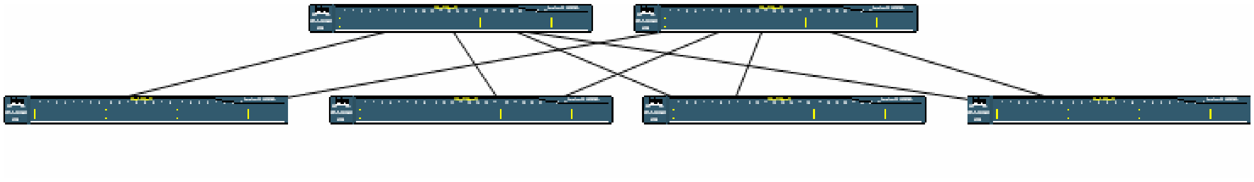


Figure 3.8: Redundant switched LAN connections.

Meshing core and access layer switches can provide fault tolerance for the network backbone, but this also requires additional management. To prevent network loops, Spanning Tree Protocol (STP) will need to be configured on the switches. Loops occur when multiple paths exist between hosts on a LAN. With multiple open paths, it is possible for frames leaving one host to loop between the switches offering the redundant paths while never actually reaching the destination host. Lops can not only disrupt communication between hosts but also flood the LAN with traffic. When configured, STP will dynamically build a logical tree that spans the switch fabric. In building the tree, STP discovers all paths through the LAN. Once the tree is established, STP will ensure that only one path exists between two hosts on the LAN. Ports that would provide a second path are forced into a standby or blocked state. If an active port goes down, the redundant port is brought back online. This setup allows for fault tolerance while preventing network loops from disrupting communication.

One other element of the LAN that can benefit from redundancy is routers. As most hosts on a network route through a single default gateway, failure of a router can shut down LAN communications. This shutdown can be overcome by using routers that support Hot Standby Routing Protocol (HSRP) or Virtual Router Redundancy Protocol (VRRP). Both HSRP and VRRP allow you to configure multiple routers to share a virtual IP address. This functionality provides failover between routers. If one router fails, a second router can automatically assume the routing duties of the first router. Although HSRP and VRRP offer similar functionality, they differ in the fact that HSRP is Cisco-proprietary, while VRRP is an open standard.

Regardless of the protocol used, both HSRP and VRRP can allow you to eliminate a router as a single point of failure. Of course, eliminating the router as a single point of failure comes at the cost of having to purchase and power additional routers for redundancy.

📖 For more information about HSRP, read the Cisco internetworking case study "Using HSRP for Fault-Tolerant IP Routing." This document is available at http://www.cisco.com/univercd/cc/td/doc/cisintwk/ics/cs009.htm.

Redundant Power

With single points of failure eliminated from the LAN, you can focus on power. Power loss, sags, or surges can also wreak havoc on the availability of your file servers. To eliminate these potential problems, both Uninterruptible Power Supplies (UPS) and backup generators can be deployed. Redundant power is no secret in IT circles and has been used for quite some time. If you're managing enterprise-class file servers, odds are that you already have redundant power in place.

In protecting against power failure, the UPS can sustain servers, storage, and network devices for a short period of time. During this period, all devices could be powered down gracefully so as not to corrupt any stored files. In organizations in which availability is crucial, the role of the UPS is usually to maintain servers online long enough for backup generators to start. The backup generator can sustain the critical elements of the network for hours or even days, depending on the number of systems on the network as well as the amount of fuel available to power the generator.

Redundant Servers

Thus far, we have looked at how to add redundancy to power, the LAN, the SAN, and disks. The only aspect of the information system that has been ignored to this point has been the servers themselves. If you have gone this far to protect your file serving infrastructure, you don't want a motherboard failure, for example, to disrupt data access.

Adding redundancy to servers can be accomplished in a few different ways:

- Deploy shared data clusters
- Deploy failover clusters
- Deploy proprietary servers that are fully redundant

Let's start with a look at shared data clusters.

Shared Data Clusters

Shared data clusters have already been fully described in Chapters 1 and 2. They provide full failover support for file servers, allowing a virtual file server entity to move from one physical host to another if a failure occurs. In having this ability, all hardware and software elements of a physical server are eliminated as single points of failure.

In addition to failover support, shared data clusters offer load balancing by allowing multiple nodes in the shared data cluster to simultaneously access files in the shared storage on the SAN. This functionality prevents the performance bottlenecks that are common in other redundant server and clustering solutions. Finally, with shared data clusters running on industry standard x86-class hardware, organizations do not have to fear deploying a proprietary solution when deciding to go with a shared data cluster.

Failover Clusters

Failover clusters, like shared data clusters, offer failover support. If one server's hardware fails, a virtual file server running on that server can simply move to another node in the cluster. All major OS vendors, including Microsoft, Red Hat, and SUSE, offer clustering support with some of their OS products, making them convenient for administrators already familiar with a certain OS.

If performance was not an issue, failover clusters would be an ideal fault-tolerant file serving solution. However, failover clusters lack in the ability to effectively load balance data between hosts. Instead, failover clusters use a shared nothing architecture that allows only one node in a cluster access to a file resource. This setup prevents failover clusters from being able to load balance access to a virtual file server. Instead, access to the virtual file server would have to be provided by one server at a time.

Proprietary Redundant Servers

One final alternative to eliminating servers as a single point of failure is to deploy fully redundant server solutions. These solutions can range in price from tens to hundreds of thousands of dollars. On the low end, companies such as Stratus Technologies offer a server that has fully redundant power, motherboards, CPUs, and storage. At the high end, companies such as Network Appliance and EMC offer fault-tolerant NAS appliances. Although both EMC and Network Appliance share a significant portion of the file serving market, their popularity has been heavily challenged in recent years by companies such as PolyServe that offer fully redundant high-performance file serving solutions that can run on industry-standard hardware.

Eliminating Bottlenecks

In addition to adding availability to the data path by eliminating single points of failure, performance bottlenecks should be a key concern. As with failure points, each element in the data path can represent a potential bottleneck (see Figure 3.9).

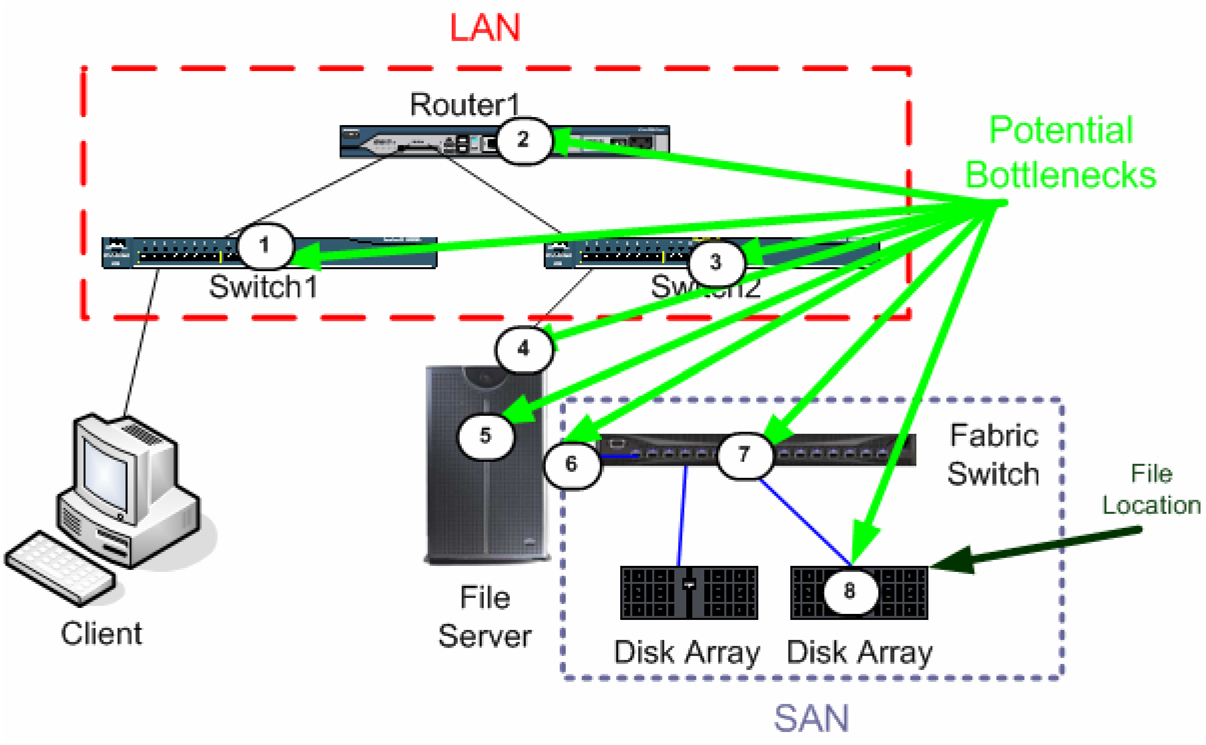


Figure 3.9: Potential data path bottlenecks.

Figure 3.9 points out eight potential bottlenecks in a data path:

1. Client access switch—10Mbps uplink
2. Router—Single router connects all clients to server network segment
3. Server access switch—100Mbps uplink
4. Server NIC—100Mbps NIC
5. Server internal hardware—CPU, RAM, motherboard, and so on
6. Server FC HBA—1Gbps
7. Fabric switch—1Gbps
8. Disk array—Just a Bunch of Disks (JBOD)

Connecting clients to the server LAN through a single "router on a stick" via 10Mbps switches can quickly slow file access performance. The term *router on a stick* refers to a single router that services several LAN segments that are multinetted together. To communicate with each logical network that resides on the multinet, the router will have multiple IP addresses on its network interface that faces the clients.

If a 10Mbps switch connects to the router, you are already faced with all clients having to share a single 10Mbps pipe to access server resources. These bottlenecks could be reduced or eliminated by upgrading the client access switch to 100Mbps and with at least one 1Gbps port to uplink to the router. If several network segments are bottlenecked at the router, you could consider replacing the router with a Layer-3 switch. If needed, a switch with 1Gbps ports could be used. If the server NIC is the bottleneck, it could also be upgraded to 1Gbps or teamed with a second NIC to improve throughput.

At the server, several elements could hurt performance. If there is not enough RAM, too slow of a CPU, or slow hard disks, performance will suffer. The expensive answer to solving server resource bottlenecks is to replace or upgrade hardware. A more scalable solution in the file serving arena is to configure the file servers in a shared data cluster. This solution offers the benefit of load balancing, fault tolerance, and often results in server consolidation.

If you're looking to document a server bottleneck, Windows and Linux offer tools to help pinpoint a problem. On Windows OSs, System Monitor can be used to collect system performance statistics in real time. On Linux, tools such as the Gnome System Monitor or vmstat can allow to you query system performance.

Resource	**Threshold**	**Required Action**
Memory	Committed bytes > Physical RAM Pages/sec > 20	Add or upgrade RAM
Physical Disk	Disk Queue Length > 2 % Disk time > 90%	Upgrade to a faster disk, deploy RAID
Processor	% Processor time > 80% Processor queue length > 2	Upgrade CPU or add an additional CPU
Network	Remains near 100% utilization	Upgrade to a faster NIC or team NICs

Table 3.2: Common performance thresholds.

The SAN also introduces a possible bottleneck site. Arbitrated loop SANs behave like Token Ring LANs and thus provide shared bandwidth for all resources attached to the SAN. Thus, 10 servers attached to an arbitrated loop SAN would have to share its bandwidth. If SAN performance is slow and you are using an arbitrated loop topology, the most effective way to improve performance will be to upgrade to a switched fabric SAN. The same can be said for a 1Gbps SAN. If you have SAN switches and HBAs that support a maximum throughput of 1Gbps, upgrading to a newer 2Gbps or 4Gbps SAN fabric will greatly improve performance.

Finally, the disks themselves could also represent a bottleneck. Upgrading to faster disks is an option; another alternative is to configure the disks as RAID. For example, moving from RAID 1 to RAID 5 could improve performance. Another alternative is to go to RAID 1+0.

Architectural Bottlenecks

Sometimes it’s not the individual pieces that represent the bottleneck; instead it could be the architecture. If the file serving infrastructure is architected poorly, the bottleneck might be the architecture itself. Two typical examples of architectural bottlenecks are single NAS heads and single file servers.

Single NAS Head

As NAS grew in popularity, a big selling point of NAS was that you could consolidate all your file serving resources into a single NAS. With terabytes of available storage, this option seemed like a good idea to many at the time. However, the single NAS head presents severe scalability and performance limitations. The NAS itself represents a single path for LAN clients to access files. Even with teamed NICs, performance scaling is limited. When stuck with a NAS bottleneck, many organizations find themselves adding NAS heads. At first, servers are consolidated, but performance demands must be met by adding NAS boxes. However, adding NAS boxes to handle file serving will likely induce additional management overhead. Thus, you will likely need to reduce user drive mappings as well as restructure backups to accommodate the additional server. If scaling continues, you will need to add another NAS, further compounding the problem.

Single File Server

The single file server represents the same problems presented with single NAS heads. The lone exception to the single file server is that it is not a proprietary solution. However, having a single access point is still an issue. You can add NICs and certainly max out CPU and RAM resources but could still be faced with network throughput bottlenecks in enterprise-class environments. The answer to the performance bottleneck dilemma of both single NAS heads and single file servers can be found in load balancing.

Load Balancing

Traditional load balancing involves balancing a client load between multiple servers. In the traditional load-balancing architecture, each server that participates in a load-balanced cluster maintains its own local storage. Without shared storage, the traditional load-balanced cluster is not a solution for file serving scalability issues. Instead, it is best suited for read-only user-intensive demands such as front-end Web serving or as FTP download servers.

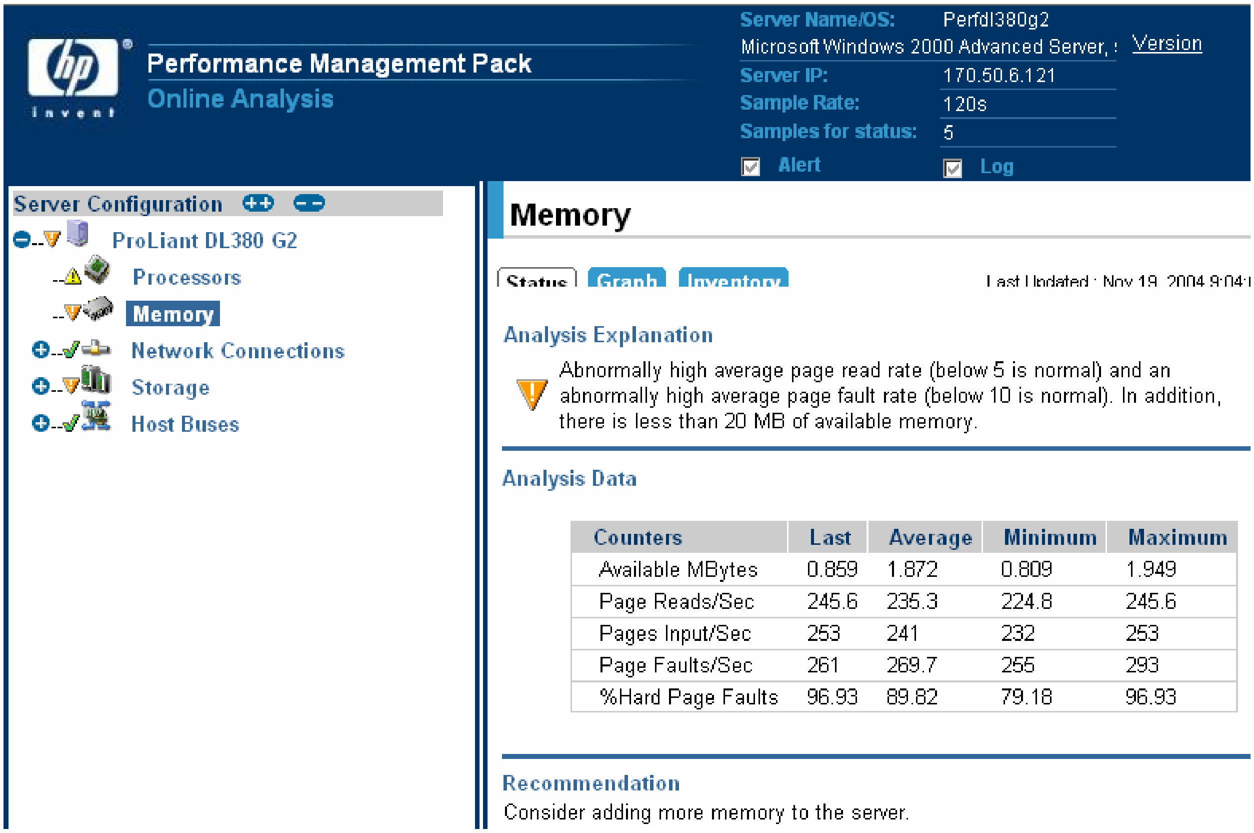
For file serving, the only true way to provide load-balanced read/write access to file system data is through shared data clusters. In the shared data cluster, multiple servers can present a single logical file server application to clients. This setup allows the client load to be distributed across multiple physical servers. This approach can eliminate many of the traditional file serving bottlenecks, including:

- Single network access point
- CPU
- RAM
- Motherboard
- HBA

With a file serving load being distributed across four server nodes, for example, you have four times the amount of server resources to handle client demand. This flexibility can allow you to get more out of your hardware investment and likely extend the life of your servers. In many shops, some servers are over-utilized while others are underutilized. Consolidating file servers to a single shared data cluster will allow you to equally use all file server resources in your organization.

Managing the Resilient Data Path

Building out a resilient data path requires knowledge that crosses several technical boundaries. It’s easy for a storage or server administrator to fail to take LAN performance issues into account. Likewise, it’s easy for a switch and router administrator to disregard SAN issues. To assist administrators in managing performance and fault-tolerance issues, many vendors are developing products that locate and report on server performance. One such product is the HP Performance Management Pack (see Figure 3.10).



Server Name/OS: Perfdl380g2
Microsoft Windows 2000 Advanced Server, :
Server IP: 170.50.6.121
Sample Rate: 120s
Samples for status: 5
☑ Alert ☑ Log

Memory

Last Updated : Nov 19, 2004 9:04:

Analysis Explanation

Abnormally high average page read rate (below 5 is normal) and an abnormally high average page fault rate (below 10 is normal). In addition, there is less than 20 MB of available memory.

Analysis Data

Counters	Last	Average	Minimum	Maximum
Available MBytes	0.859	1.872	0.809	1.949
Page Reads/Sec	245.6	235.3	224.8	245.6
Pages Input/Sec	253	241	232	253
Page Faults/Sec	261	269.7	255	293
%Hard Page Faults	96.93	89.82	79.18	96.93

Recommendation

Consider adding more memory to the server.

Figure 3.10: Discovering a system memory problem with the HP Performance Management Pack.

Tools such as this have grown in popularity because they not only alert you of performance problems but also will recommendations for how to solve them. The Brocade and Symantec tools mentioned earlier in this chapter are ideal for monitoring and managing SAN resources. Other SAN hardware vendors, such as QLogic, offer similar solutions. It’s an easy trap for administrators to invest countless dollars in hardware and then decide to forgo management software in order to save money. The time saved by having software monitor and report on problems within your data path will undoubtedly be worth the cost of the management software.

Many vendors like to tout the fact that they provide an end-to-end solution. However, end-to-end solutions often advertise to solve all problems along a data path, but rarely deliver. When designing a high-performance fault-tolerant data path, be sure to ask yourself or any vendors involved in the project the following questions:

- Is the network path to my file servers fault tolerant?
- Is the SAN path to the file resources fault tolerant?
- Can the planned technologies effectively load balance file access requests?
- How can I monitor and report on LAN bottlenecks and failures?
- How can I monitor and report on SAN bottlenecks and failures?
- How can I monitor and report on server bottlenecks and failures?

With acceptable answers to these questions, you should be ready to enjoy a resilient and fault-tolerant file serving infrastructure.

Summary

Far too often with file serving, administrators focus solely on performance issues from the servers back to storage and all but ignore the remainder of the data path. Hopefully, this chapter has made you aware of all of the aspects of getting a file from a server to a client, and back. With the right architecture in front of and behind your file servers, they should be able to grow and respond to client performance as the needs of your organization evolve. Basing your file serving infrastructure around shared data cluster architecture is the only true way to add fault tolerance and load balancing to file server resources.

As you have seen, all the resources that you need to build a fault-tolerant and resilient file serving data path exist today. Having knowledge of what is available as well as how to use new resources should allow you to build a reliable file serving infrastructure within your organization.

The next chapter will look at building high-performance file serving solutions in both Windows and Linux environments. Chapter 4 will take you through the process of building a high-performance Windows file server and Chapter 5 will explore the process of building a highly available and high-performance Linux file serving solution.

Content Central

Content Central is your complete source for IT learning. Whether you need the most current information for managing your Windows enterprise, implementing security measures on your network, learning about new development tools for Windows and Linux, or deploying new enterprise software solutions, Content Central offers the latest instruction on the topics that are most important to the IT professional. Browse our extensive collection of eBooks and video guides and start building your own personal IT library today!

Download Additional eBooks!

If you found this eBook to be informative, then please visit Content Central and download other eBooks on this topic. If you are not already a registered user of Content Central, please take a moment to register in order to gain free access to other great IT eBooks and video guides. Please visit: http://www.realtimepublishers.com/contentcentral/.