BACKUP

WHITE PAPER | BRENT WELCH | NOVEMBER 2006

TABLE OF CONTENTS

ABSTRACT

Large scale storage systems have demanding backup requirements, and the Panasas Storage Cluster™ helps satisfy these requirements by delivering high throughput to many concurrent backup IO streams to standard backup applications such as Veritas NetBackup™ or EMC® NetWorker™. However, file system performance is just one aspect of implementing an effective backup and restore process. This white paper describes the overall strategy for backup and restore taking into account the tape subsystem, networking, file system workload and administrative policy.

BACKUP OVERVIEW

There are three types of backup considered in this paper: full, incremental and synthetic.

- A full backup is a complete backup of the data at a given time. This is the most complete form of backup and full recovery can be made from this backup tape(s). A full backup tends to put a heavy load on the available resources because of the fact that all the data needs to be transfered from one media to another.

- An incremental backup is where only the changes since the last backup are backed up. There are two forms of incremental backup: cumulative incrementals backup all changes since the last full backup; differential incrementals backup only the changes since the last full or incremental backup. The backup window is much shorter, as they are for the differential backup. With a cumulative incremental scheme, the backup window increases each step away from the full backup. Recovery in a cumulative incremental scheme involves recovery of the last full backup, then recovery of the last cumulative backup. Recovery in a differential incremental scheme involves recovery of the last full backup, then recovery of each of the incremental backups made since the last full. A typical backup schedule might be to do incrementals through the week with a full backup done at the weekend.

- Synthetic backups are a relatively new scheme. Their purpose is to eliminate the full backup window, and to reduce the restore time by generating synthetic full backups offline. After the first full backup, only differential backups are made, which minimizes the backup window. At any time, a synthesized full backup can be created offline by combining the original full and all the differentials. This synthetic full becomes the new baseline without the need for an actual full backup. There is extra work required offline to create the synthetic backup, but this is offset by eliminating or reducing the occurrence of regular full backups.

Each organization's backup requirements are different; the type of backup scheme should be chosen to suit the needs of the organization. All of these backup schemes are a compromise between minimizing the backup window and minimizing the time to restore. However the backup is performed, it will have an impact not only on the storage, but also on the systems using that storage. Backing up data generated by active applications is not without risks. Although the backup and restore software may be able to read the data that the application is working on, that data may not be in a consistent state. This makes that backup data useless for restore purposes because it will put the application in an inconsistent state. A number of backup/restore software vendor address these issues by providing agents that cooperate with the applications that run on the customer's server to make sure that the backups that are being made are consistent and can be used for restores.

BACKGROUND ON BACKUP APPLICATIONS

A backup application has three participants: the scheduler, the tape subsystem, and the file system.

- The scheduler keeps track of when backups occur and maintains a catalog that records on which tapes they reside. It is the overall controller and it generates commands to the other participants to initiate backup or recovery actions. The administrator interacts with the scheduler via a user interface application.

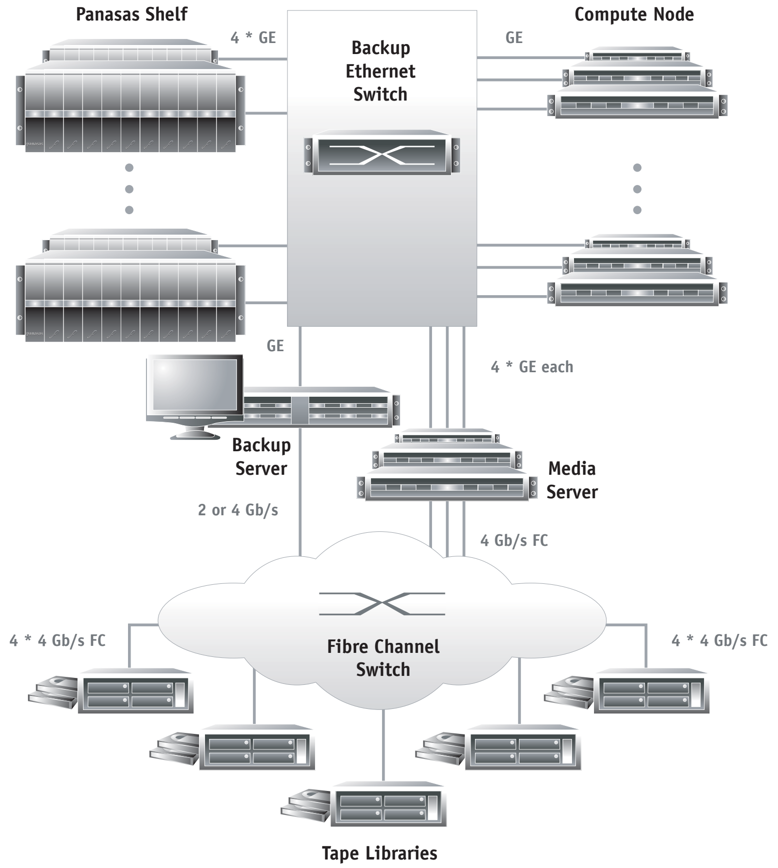
- The tape subsystem consists of one or more tape libraries, each of which can have one or more tape drive units. The *media agent* is a software module that controls the tape subsystem in response to commands from the scheduler.

- The file system stores the data, and must provide a high throughput datapath to the tape subsystem. The *backup agent* is a software module that transmits the data between the file system and the media agent in response to a command from the scheduler.

There is a three-way communication structure between the scheduler, media agent, and backup agent. The most general configurations run these modules on different computer systems that communicate over a network, typically via TCP/IP. The scheduler creates a *control channel* to each backup agent and each media agent. There can be many media agents and backup agents in large configurations. In response to commands from the scheduler, the backup agents and media agents establish pair-wise *data channel* connections. The backup data flows over the data channels, while commands and catalog information flows over the control channels. In some configurations, a media agent and backup agent can be co-located on the same computer system. In this case, the data channel is just a memory buffer.

To create a backup, the scheduler commands the backup agent to begin reading files from the file system and transmit them to a media agent, which in turn writes the files to tape. As the backup agent generates its stream of backup data, it also sends catalog information about the files being backed up to the scheduler. When the media agent changes tapes, or needs operator attention, it also sends status information to the scheduler for use in the catalog and the user interface to the overall backup application.

To restore data, the scheduler uses its catalog information to determine what tapes need to be loaded via the media agent. It commands the media agent to pull the required files off the tape and transmit them to the backup agent, which writes the files to the file system. The diagram below illustrates a standard configuration of connecting Panasas storage to a customer's compute and backup infrastructure. (The backup server has the scheduler while the media agent and backup agent is in media server.)



MEDIA AGENTS AND KEEPING TAPE DRIVES BUSY

A media agent manages tape devices. These software modules are supplied by the backup vendor and are qualified with a broad variety of tape drive equipment. The key to good performance is keeping a tape drive busy. The streaming performance of a tape drive can be quite high: 120 MB/s for DLT, and future generations promise higher speeds.

Technology	Sustained Performance
DLT	120 MB/s
LTO-3	90 MB/s
SDLT	72 MB/s
AIT-4	24 MB/s
DAT	7 MB/s

* 2:1 compression used to determine sustained performance

However, if the media agent cannot feed data that quickly to the drive, then throughput can drop dramatically. The performance issue is that the tape drive cannot cheaply stop and restart. When the drive stops writing, the tape media continues to race past that point until it slows to a stop. To start again, the drive has to rewind slightly to correctly position the write head. The goal of the media agent is to keep its drives busy to maintain their performance.

One common problem is that an individual backup stream from the file system (i.e., the backup agent) is not fast enough to satisfy the needs of the media agent. Especially if there are many small files in the file system, then the overhead of finding each file is high enough that the backup agent may not generate a fast enough stream of file data. If only a single slow stream is available, then the tape drive is under utilized and the time to complete the backup of the storage system.

To address this issue, backup vendors have a multi-stream feature (sometimes called multiplexing) where multiple file backup streams are aggregated together and put onto one tape. The catalog keeps track of this, and there are facilities to de-multiplex streams if administrators need to get a single backup stream onto a single tape. The goal is to configure enough concurrent backup streams, and use multiplexing to keep the tape drives fully utilized.

BACKUP AGENTS

Backup agents scan the file system for modified files, generate a data stream over a data channel to a media server, and transmit catalog information to the scheduler. A traditional backup implementation has one stream per "file system" that in a single server, is a disk partition. The size of a disk partition can range from a few tens of MBs to hundreds of GBs, or more. A LUN on a RAID could be one TB in size or more.

Backup agents run on file servers, or in systems like Panasas, the backup agents run on computer systems called *backup clients*. These are computer systems that run the DirectFLOW™ file system client so they have high performance access to the file system. If the computer system also has access to a tape device, either directly or via a SAN, then it can also run a media agent. During backup, data flows from the Panasas StorageBlades™ to the DirectFLOW file system client, to the backup agent, to the media agent, to the tape device.

A large storage system can have multiple backup clients to achieve scalable backup performance. The ability to scale backup throughput is an advantage of the Panasas storage cluster. In contrast, a traditional file server that runs the backup agent on the server has to share its computing, memory system, and network interface resources with user accesses.

BACKUP PERFORMANCE

This section has some basic rules of thumb for estimating backup application performance. However, there are enough variables in the file system workload, networking configuration, backup client throughput, and tape devices that make accurate predictions difficult. Obtaining optimal backup performance requires tuning the overall system configuration.

The first rule is that multiple concurrent backup streams are required for best throughput. The actual level of concurrency required depends on the file size distribution and the desired performance. A file system filled with mostly small files, means files less than 1 MB, will have low throughput on a single stream. It may require as many as six or eight backup streams against a single shelf to drive the utilization above 50%. Of course, that will cause more interference with on-line activity. Four concurrent streams per shelf with a small file distribution should provide a reasonable balance between backup throughput and on-line activity. A file system with mostly large files (MB or GB) will have a relatively high throughput for a single backup stream. Four concurrent backup streams would place a fairly high load on the system, between 200 and 300 MB/s. In this case, it may be desirable to limit concurrency to one or two streams per shelf to avoid saturating the system with backup traffic.

The second rule is that each backup client has a throughput dependent on its network interfaces and memory system. An older 32-bit system with one single GE network interface may only be able to transfer 50 MB/s between the file system and the media agent. A modern multiprocessor machine with a quad-GE NIC may be able to sustain 300 MB/s, or more. A machine with a 10GE NIC and dual 4 Gb/s FC interfaces may be able to host a backup agent and a media agent and sustain even higher throughputs. Customers can choose to deploy a small number of the latest systems to have the highest throughput per backup client, or they may choose to repurpose older hardware as a small “backup cluster” to achieve the desired throughput.

The final aspect of the configuration is the choice of the tape subsystem. In some configurations, tape is not used at all and the backup target is another file system. This kind of disk-to-disk backup is directly supported by backup applications. Throughput of a tape device depends on its model. The latest DLT tape drives have a throughput of about 120 MB/s compressed, Older tape devices may only have 60, 30, or 15 MB/s of throughput, and future devices promise speeds above 200 MB/s.

Design of an effective backup system requires enough tape devices, enough backup clients, and enough concurrent backup streams to achieve the desired backup throughput. The rules of thumb presented here provide general guidelines that should help design a scalable backup infrastructure. Mapping backup streams to backup clients and media agents will require testing in your environment to determine the effective throughput.

EXAMPLES

Suppose you have a five shelf Panasas system with 500 GB blades providing a total of 25 TBs of storage and a large file size distribution. If the system is about 80% full, this is 20 TB of data. Assume the shelves can deliver 340 MB/s at over 90% utilization, so the five shelf system could deliver 1.7 GB/s. Assume a backup stream can achieve 85 MB/sec, then each shelf would be saturated by 4 streams, and the whole storage system would be saturated by 20 streams. Assume two backup streams are multiplexed onto one tape device, and that there is enough throughput and/or compression that the tape device can absorb the 170 MB/s, then 10 tape devices are required. Assume a media agent can support two drives and has a system throughput of 340 MB/s to match a single shelf. The full backup system is then 5 backup clients and 10 tape devices and has a throughput that matches the file system. If the system were 100% devoted to backup (or restore) and everything was tuned perfectly, it would complete the 20 TB backup in three hours and 15 minutes.

The previous example is actually a better model for a full restore in a disaster recovery scenario. In this case it is reasonable to devote the storage system 100% to the restore operation. Backup is usually configured to impose a more limited load on the system. A more realistic implementation might only achieve 60 MB/s per backup stream, have two streams per shelf multiplexed onto a single tape device, and use two or three backup clients to interface between the file system and tape devices. This would be 600 MB/s of backup throughput, which is about 1/3 the capacity of the storage system. The 20 TB would be backed up in 9.5 hours.

Environments with mostly small files have a harder time achieving high backup throughput. In this case, the throughput is best measured in files/s as opposed to MB/s because the storage system has per file seek overhead. With enough concurrency, a Panasas shelf can deliver between 500 and 1000 “whole-file-operations” each second. This is a cold-cache read of a file that is less than 64K. The rate is a function of the seek performance of the drives. Suppose a 5 TB shelf is 80% full of 20 K files. These files are mirrored and there is a 4 K object descriptor for each mirror, so they occupy 48 K of space and there would be over 80 million files in the system. At 1000 files/second, it would take almost 24 hours to back up and the storage system would be over 90% utilized. This assumes that the backup configuration is perfectly tuned, and that the backup application can accept updates to its catalog at the same rate.

MEDIA ROTATION AND RETENTION PERIODS

A frequently used schedule is utilize to perform full backups each weekend with incremental backups through the week. Usually different tape sets are used for different tape retention periods:

- Daily incrementals may be overwritten after a week
- Weekly full backups may be overwritten after an month
- Monthly full backups bay be overwritten after one year

There are many combinations of retention and overwriting but it is important that complete recovery sets are kept offsite or in a fire-safe vault for disaster recovery purposes.

Keeping backup tapes for long periods should not be confused with archiving. Backup applications do not do a good job of data archiving: little or no meta-data is stored, and there is no refresh or consistency checks of the media.

NDMP

NDMP provides a standard interface between the backup agent, the media server, and the scheduler. It was introduced so that file server vendors could run a standard backup agent on their hardware.

The primary drawback of NDMP is that it lacks some of the more advanced features found in today's backup applications. In particular, the multiplexing feature that puts multiple backup streams onto one tape device is generally not supported. This means that a single NDMP backup stream is bound to a single tape device. As described above, this can cause performance problems because tape drives are not driven fast enough.

Another motivation often associated with NDMP is to be able to confine backup traffic to a separate network for performance reasons. In our experience, the interference between backup traffic and on-line activity occurs at the disk devices, not the network. We assume that each shelf has four trunked GE connections to the backbone network, that the backup clients have two to four GE connections to the backbone, and that the backbone network has adequate performance.

CONCLUSIONS

The ability to backup from a Panasas storage system is no different than backing up from any other storage system. In fact, the Panasas Storage Cluster™ running the standard Symantec or EMC backup software will allow customers to multiplex concurrent data streams to deliver the highest performance to keep tape drives fully utilized. Implementation of an effective backup strategy requires tuning the number of backup streams, backup clients, and tape devices to achieve the desired throughput based on the file size distribution and the load placed on the storage system. The key benefit to Panasas' high performance multi-streaming capability is that customers can backup up more data in smaller backup windows.