

Paper #: 4045
Title: Large Scale Backups
Author: Dorne Smith
Hewlett Packard
1900 Garden of the Gods Road
Colorado Springs, CO 80907
Phone: (719) 590-2809

Introduction

Today, computing service providers are faced with a dynamic environment where storage and number of nodes to be managed are growing explosively. Competitive pressures are forcing companies to drastically reduce costs while improving response time to market changes. Technologies come into use quickly, then may be obsoleted almost overnight. At the same time, customers require high availability and reliability. Downtime or data loss can threaten the survival of a business.

This paper is a case study of how systems and networks have been architected to provide a flexible, high performance environment for doing system backup and recovery. It includes information on where performance bottlenecks may occur, and what the limiting factors are for today's most popular storage technologies, such as DDS, DLT, and MO. The emphasis is on how to provide high reliability, high performance, and unattended operation for backup/recovery processes.

Overview

In general, the key to good backup and restore performance is multiple, balanced I/O streams. When backing up systems over the network, this can be accomplished by segmenting the network and placing backup server connections on each major segment. As the number and size of systems to be backed up increases, network segments or subnets can be dedicated to handling backup traffic only. When backing up large systems to locally connected devices, this can be accomplished by architecting the system so that the application data, I/O configuration, and backup data sets are aligned to take maximum advantage of the available system I/O bandwidth.

First, we illustrate a standard configuration for a network backup server and examine three possible methods of organizing network backup topology to provide backups for large numbers of clients. Performance estimates are provided for each method, based on actual production backup data.

Next, we illustrate a large system configuration and examine several factors which affect backup performance. Again, performance estimates are provided based on actual production backup data.

Lastly, we give additional hints on how to back up certain classes of systems which merit special consideration, and conclude with some general comments regarding backup/recovery processes.

Standard Backup Server

Figure 1.1 illustrates our standard configuration for a network backup server.

This configuration consists of:

- > HP E35 computer with 2 full height I/O slots and 10Mb Ethernet LAN.
- > 3 SE SCSI interface cards
- > 1 additional 10Mb Ethernet card
- > 2 160fx MO Jukeboxes for a total of 320Gb of storage
- > 2 48AL 6 tape autochangers OR 2 DLT4000 tape drives
- > Hiback backup software

Each MO jukebox has 4 MO drives, and the storage is logically partitioned by the backup software into 4 - 40Gb storage groups. Each of these groups of media behaves as a virtual 40Gb shared storage device. Using two jukeboxes makes a total of 8 40Gb shared storage devices available, yielding approx 320Gb of on-line storage for each backup server. Each MO jukebox is connected to a dedicated SE SCSI interface.

Half of the storage is allocated for weekly full backups, and the other half is allocated for daily incremental backups. The 48AL 6 tape autochangers or DLT4000 tape drives are used to make weekly copies of full backups for long term storage. Both tape storage devices are connected to a single SE SCSI interface card. Daily incremental backups are retained on the MO media and are not copied to tape. This 50:50 full:incremental storage ratio provides for a 6 day incremental retention period given an average data volatility of 8% per day for the environment. This ratio can be adjusted as necessary for other retention requirements or data volatility.

In actual configurations, the MO storage allocated for full backups is filled to 80-90% of capacity to allow for variations in the size of backups and to provide some space for special backups, etc. If the amount of data stored exceeds this "high water mark", warning mail is sent to backup administrators to alert them that storage is running low and some action needs to be taken.

For planning purposes, it can be assumed that this configuration will provide backups for 160Gb of Unix or NT filesystem data. (Based on an 8% per day data volatility rate and 20% software compression.)

Standard Network Backup Process

Our standard network backup process flow is shown in **figure 1.2**. Full backups of client systems are done on weekends during a 30 hour window beginning at 6:00AM on Saturday and ending at 12:00 noon on Sunday. Daily incremental backups of all data which has changed since the last successful full backup are done during a 12 hour window beginning at 6:00PM and ending at 6:00AM.

Backups are scheduled via cron (UX) or WinAt (NT) on the clients. If an MO storage group is locked by a client doing a backup, other client backups will queue up and wait for the storage to become available. Client backup start times are staggered to avoid large queues.

All network backups are done unattended to the MO jukeboxes, providing extremely high reliability and low operating cost for the process. The process is fully automated, eliminating the requirement for media handling in order to complete the backups. MO media is the most reliable media available for backups. It doesn't wear out (lifetime of media is estimated at 100 years), and is not susceptible to erasure by magnetic fields.

For recoveries, all incrementals and at least one full backup are available "near on line", allowing many recoveries to be completed without media handling.

These features of unattended operation and "near on line" access are especially helpful for remote site management.

One disadvantage to using MO media for backups is that it costs much more than magnetic tape media. Our standard retention time for full backups is six months. The cost of MO media required to provide enough storage to satisfy this requirement would be prohibitive.

Another problem arises from the use of high capacity autochangers for storing backup data. Often, these devices are located in a data center or computer room in close physical proximity to some or all of the systems which are being backed up. Good disaster recovery processes dictate that backup data should be periodically vaulted in a physically secure location some distance away from the computer room or data center.

In order to satisfy the retention and vaulting requirements, we have developed a "hybrid" process in which backups are done unattended to MO jukeboxes as described above, then copied to DDS or DLT tape media for long term storage in a physically secure media vault. Beginning on Monday, jobs are run which copy the previous weekend's full backups to magnetic tape media.

For reporting and monitoring the configuration and status of network backups, a Web interface is provided. As backups complete, the results are ftp'ed to a web server where success/fail summaries and session information are updated hourly. Configuration information, schedules, backup server capacity and utilization are also displayed on the web pages.

Backup administrators check these results Sunday PM and each weekday morning taking corrective action as necessary if backups fail or if the MO storage fills beyond the predefined "high water marks" for utilization.

Network Topologies

Now we look at three network designs for backing up large numbers of small to medium sized clients. Systems which store 20Gb or less after software compression for their full backups are included in this class. We also assume that these backups can be done with the systems and applications fully functional. This is normally referred to as an "on line" or "hot" backup. The performance estimates given are based on measurements taken in actual production environments using the configurations described.

Method 1 - Shared Network

Figure 2.1 shows a shared networking model where the backup traffic shares the LAN with other traffic.

This method is well suited for workstations with 4Gb or less of data (after software compression) to be backed up.

There must be enough available network bandwidth to handle the backup traffic. Normally, backups are done at night and on weekends. It is important to be aware of any other scheduled activities which might be competing for network resources during these time periods.

Network connections to backup servers should be located on network segments serving clients having the most storage. Taking advantage of network segmentation via multiple connections will improve both performance and reliability.

Expect data transfer rates for this method to be in the range of 1Gb - 1.5Gb per hour using a 10Mb LAN. This translates into a backup capacity of 30Gb - 45Gb per LAN connection given a 30 hour weekend backup window. Using the backup server configuration described earlier with two LAN cards gives the capability of backing up 60Gb - 90Gb of data each weekend. Assuming a 20% data compression, this configuration will provide backups for 75Gb - 112.5Gb of uncompressed client data.

These performance estimates are based on two backups running simultaneously on a LAN segment, and a non-backup network utilization level of 20% - 30%.

Method 2 - Dedicated Backup LAN or Segments

Figure 2.2 shows a dedicated networking model where the backup traffic is isolated from other traffic. This model requires installing an additional network card in each client. This additional card is connected to the dedicated backup LAN or LAN segment. LAN card(s) in the backup server(s) are also connected to the backup LAN.

This method is well suited for small to medium size servers located in a computer room or data center. The backup LAN segment wiring can be installed only in this limited physical area, reducing the cost of implementation. This method works well for systems having 20Gb or less of data (after software compression) to be backed up.

Expect data transfer rates for this method to be in the range of 2Gb - 3Gb per hour on a 10Mb LAN segment. This translates into a backup capacity of 60Gb - 90Gb per LAN connection given a 30 hour weekend backup window. Using the backup server configuration described earlier with two LAN cards gives the capability of backing up 120Gb - 180Gb of data each weekend. Assuming a 20% data compression, this configuration will provide backups for 150Gb - 225Gb of uncompressed client data.

These performance estimates are based on two backups running simultaneously on a LAN segment.

Method 3 - Dedicated backup LAN with 100VG link to server

Figure 2.3 shows a networking model designed to maximize the throughput of a 10Mb client LAN environment. In this model, the logical configuration of the backup server is matched to the network topology so that each 10Mb LAN segment maps to a single MO storage group. Since only one backup can be actively storing data to an MO storage group (others may be queued up waiting), this configuration has the effect of dedicating a 10Mb LAN segment to each active backup. (Using a high bandwidth switch with a 100VG link from the switch to the backup server handles multiple 10Mb segments with no speed degradation.)

This method is well suited for small to medium sized servers located in a computer room or data center. The backup LAN segment wiring can be installed only in this limited physical area, reducing the cost of implementation. This method works well for systems having 20Gb or less of data (after software compression) to be backed up.

Expect data transfer rates for this method to be in the range of 2Gb - 3Gb per hour for each LAN segment/storage group. The illustrated configuration consists of 4 storage groups. This translates into a backup capacity of 240Gb - 360Gb of data for a 30 hour weekend backup window. Assuming a 20% data compression, this configuration will provide backups for 300Gb - 450Gb of uncompressed client data.

In order to fully take advantage of the data transfer capabilities of this method, the amount of storage on the backup server should be increased to 600Gb using a model 600fx MO jukebox.

Large Systems

Now we take a look at how to architect backups for large systems. When we say large, we mean systems with configured storage in the range of 60Gb to 500Gb. The key to backing up these systems is to architect the hardware and configure the backup software to obtain multiple, balanced data streams which take maximum advantage of the system's I/O structure. **figure 3.1** shows how this can be accomplished.

Following are a few performance estimates and configuration hints for backing up large systems.

One potentially limiting factor for the speed of a backup data stream is the speed at which the system can deliver data from disk. For file system stores, expect 6Gb - 8Gb per hour. For raw disk stores, expect 10Gb - 16Gb per hour. Using the configuration illustrated in **figure 3.1** with 12 data streams, this yields an aggregate throughput potential of 72Gb - 96Gb per hour for file system stores and 120Gb - 192Gb per hour for raw disk stores.

Another factor is the speed at which the I/O backplane can handle data. For an HP-PB backplane, expect a maximum of 60Gb per hour. For an HSC I/O subsystem, the transfer rate is over 200Gb per hour, and should not generally be a limiting factor.

Another factor is the speed at which the tape drive(s) can accept data. Since these drives have hardware compression built in, their speed is dependent on the compressability of the data being stored. Many large systems have an RDBMS in which most of the data is stored. This data tends to compress very well. A compression ration of 4:1 is typical, and is used in the estimates given here. For DDS-II drives, expect 6Gb - 8Gb per hour. (Note: This is a close match to the system's ability to source data for filesystem stores.) For DLT drives, expect 10Gb - 15Gb per hour. (Note: DLT figures are based on results obtained using a SE SCSI interface)

Another factor is the speed at which the SE SCSI bus can handle data. Expect a maximum of 15Gb - 17Gb per hour.

In general, if systems are architected as shown in **figure 3.1**, we get the following results when running actual data stores:

Aggregate thoughput for filesystem stores:

- > 4 data streams - 20Gb - 24Gb per hour (DDS-II or DLT)
- > 12 data streams - 60Gb - 65Gb per hour (DDS-II or DLT)
(SE SCSI limit)

Aggregate throughput for raw stores:

- > 4 data streams - 20Gb - 24Gb per hour (DDS-II)
40Gb - 50Gb per hour (DLT)
- > 12 data streams - 60Gb - 65Gb per hour (DDS-II or DLT)
(SE SCSI limit)

These values assume a 4:1 hardware compression ratio. It is also assumed that the backup data sets are configured to provide balanced data streams from independent sets of disks. (No single disk device is included in more than one data stream, and each data set includes the same amount of data to be stored.)

Additional Hints & Special System Considerations

System .vs. Application Data

For many large systems, especially those having a large application database, it is convenient to define two separate backups; one for the system/platform, and one for the application. Very often, system backups can be done using a network backup process, and application backups can be done using the method described for large system

backups. This approach makes sense where support for the platform is provided by one group or organization, and support for the application is provided by a separate group. It also has some advantage in disaster recovery or system replication for testing, development, etc. If the data has been partitioned so that applications are on a separate backup, application data recoveries can easily be done onto an alternate system platform without overwriting system parameters such as network configuration, hostname, etc.

Data Warehousing Systems

These systems are characterized by large amounts of disc storage (Often 200Gb or greater), and low data volatility. Updates are normally done via periodic batch jobs. The data normally resides in an RDBMS which must be stored/recovered when it is in a consistent state as a monolithic unit. It is often practical to do only weekly full backups of these systems. The application/RDBMS can be shut down while the backup runs (usually 4 - 6 hours), then restarted after the backup completes. Since recovering individual database files would not yield a consistent state, (and some RDBMS systems actually use raw disk storage), raw disk backups are an option worth considering.

OpenMail Systems

Large OpenMail systems are characterized by very large numbers of very small files. The entire application filesystem must be stored and recovered as a monolithic unit. Individual database file recoveries are not generally useful. Filesystem stores are not practical due to the extended time required for completion. (It can take days!) Short periods of application downtime nightly are normally acceptable.

For these systems, we have provided mirrored disk to allow maximum application availability. Backups are performed via raw disk stores in order to attain reasonable store/restore times. The process works as follows:

- > Shut down the OpenMail application.
- > Split mirrors.
- > Bring application back up. (Downtime: 30 min or less)
- > fsck split disks.
- > Do raw disk backups. (approx 2 hours for 40Gb)
- > Merge mirrors.

This process is done nightly, providing a full backup of the system each time.

OLTP Systems

These systems normally have data managed by a DBMS with transaction logging active. Backups must normally be done on line using DBMS utilities. Recoveries require roll forward of logs since the last backup. For these systems, it is often practical to do platform/system backups via the network, and handle application data backups separately. Transaction log backups must normally be done as part of the application backups.

For small to medium size systems (20Gb or less), it may be practical to do checkpoints to system disk using DBMS utilities, then capture this data and transaction logs as part of the system backup. Using this method, the DBMS storage area is excluded from the backups.

File Servers

These systems are characterized by large numbers of files and frequent user file recoveries. For this type of system, network backups to MO work well. The time required to completely store/recover a system is sometimes a concern, but is offset by the fact that data lost or corrupted due to a hardware failure can be restored without restoring the complete system. It is helpful to dedicate enough MO storage to these systems to hold two or three weeks of full backups. This makes user file restores fast and easy.

Comments: Bottlenecks, Performance, and Philosophy

For filesystem stores, large numbers of very small files will cause serious performance problems. If backups are going very slowly, and the CPU is heavily utilized for system calls, this could be an indication that filesystem overhead is the bottleneck. If this is the case, investigate the possibility of using raw disk stores.

Faster storage devices are not an automatic cure for backup performance issues. The key to fast backups is multiple balanced data streams. The ability of the system to source data is often the bottleneck even when using DDS-II drives. If this is the case, replacing DDS-II with DLT will not increase backup speeds. DLT is a good choice for cases where backup software or utilities do not allow configuring multiple data streams. DLT also offers higher reliability and longer media life than DDS-II.

Interleaving is a software technique which is used to provide higher data rates for a single output data stream. It involves multiplexing several "reader" processes into a single "writer" process. This can help utilize higher bandwidth devices such as DLT drives. The disadvantage is that this technique is optimized for backup performance, and recovery performance can be severely impacted. When evaluating performance, it is important to consider both backup and recovery.

Backups tend to expose the raw transfer rates of storage devices. Performance enhancements such as data or inode caching rely on the locality of reference principle, and provide little if any advantage when doing data backup and recovery.

The purpose of backup/recovery processes is to provide insurance against data loss or corruption due to hardware failure, natural disasters, human error, etc. It is tempting to use these processes to provide some form of data archiving by changing retention periods and/or scheduling. Attempting to do this can result in a process which doesn't correctly provide either backup/recovery or archiving. Archiving should be considered separately. Archiving may require different trigger points, data formats, retention, media type, storage locations, security, etc.

When designing backup/recovery processes, extended store/restore times increase the risk of data loss. As a rule of thumb, it should take no longer than 8 - 12 hours to completely store or recover the data on a system. It is important to remember that data recovery is only one part of the total system recovery process. Troubleshooting, hardware repairs, etc. must be taken into account.