Optimizing Your Network Backup Performance: Part II

BY IRA GOODMAN

Optimizing a network backup often turns out to be surprisingly difficult because there are so many variables involved. However, having a powerful, dedicated Master Server and balancing data and devices can help. Also, understanding some simple programs that can be used for both planning and troubleshooting, as well as knowing which software settings can drain resources and effect performance, will allow you to optimize your use of backups.

OPTIMIZING network backup performance is deceptively simple. The variables involved are straightforward. The programs that you can use for troubleshooting are easy to understand and write. The software settings you need to consider are obvious — when you think about them.

Indeed, the difficulty in optimizing backup performance is not in grasping any particular part of the performance problem. The difficulty is that there are so many parts, and they interact in complex ways.

This concluding article explores two important tuning variables that weren't discussed in Part I (***Technical Support,*** April 1998): the choice of device servers and data location. Additionally, this article describes simple programs that you can use to test the throughput of components directly or across a network. Finally, software settings that have a significant impact on performance are discussed.

IDENTIFYING BACKUP SERVERS

If you are distributing backup devices on your network, you will be using several backup servers. In this article, I'll refer to these machines as device servers.

One device server must be designated as the Master Server because it holds the following:

- the backup product
- the schedule, which ensures backups are done regularly and automatically
- the catalog, which lets you track and restore your data

Ideally, the Master Server should be dedicated to backup and restore processing because critical catalog and reporting functions are performed on it. It can also be connected to one or more backup devices, and data from its own disk or from other machines on the network can be backed up to those devices. The backup product on the Master Server routes the data to the appropriate device servers and backup devices according to the job definitions set by the backup administrator.

When choosing a device server, regardless of whether or not it is used as the Master Server, you need to focus on the following:

- CPU characteristics
- memory
- what else is running on the machine

The Master Server has an added requirement — disk space for the backup catalog.

THE BACKUP CATALOG AND PERFORMANCE

Whenever you back up a file, your backup product writes a record in its catalog. You can estimate the amount of space you need by multiplying the maximum number of records you expect to have in your backup catalog by the number of bytes your backup product writes per record.

An extremely efficient backup product writes from 80 to 100 bytes per record, but some products may write two or more times that amount per record. You must calculate carefully and allow room for catalog growth.

Keep in mind that every catalog record is not held indefinitely. You will periodically condense your catalog to discard the records that have expired. If job throughput tests out fine but overall backup time is disappointing, one problem may be the size of

the backup catalog, and you may need to condense it more often.

CPU CHARACTERISTICS

Ideally, device servers should have multiple CPUs, and if you are backing up a substantial amount of data and it's growing rapidly, those CPUs should have a high megahertz rating,

Multiple CPUs are important because they allow multiple processes to run in parallel on the same machine but on different CPUs. This parallelism decreases the load on any single CPU and results in better performance. Parallel processes (also called parallel streams or threads) are important in driving high-speed devices to their maximum with techniques such as "dynamic parallelism."[1]

A high megahertz rating allows every operation on a device server to complete more quickly. This is especially important on the Master Server. Although all the backup-related processing you will do on the Master Server may not be CPU-intensive, you will need the extra speed for a few critical jobs related to cataloging and reporting. For example, a fast CPU normally ensures that all catalog condense operations are done as quickly as possible and that they do not interfere with other processes.

MEMORY

Having sufficient memory on all device servers is important. If you don't have sufficient memory, you will very likely have serious paging problems during backup and restore jobs, causing them to run much more slowly.

OTHER APPLICATIONS

You may experience performance problems if your Master Server also runs other important processes.

Contention between processes will be especially noticeable on the Master Server when you are writing catalog records during a backup process, condensing your catalog, or writing backup-related reports. Severe paging problems could result.

LOOKING AT DATA LOCATION

Ideally, a backup system should be set up with optimal performance in mind so that it can be tuned easily from the first day it is brought online. One key aspect is identifying the largest concentrations of data and deciding how to deal with them.

Even if you have a system in place, it will be worth your time to do the following:

1. Look at your enterprise as a whole and identify the machines with the most data. Backup devices should be directly connected to these machines, so that you avoid moving large amounts of data across the network during backup.

2. Find out where the smaller amounts of data are on your system. Are they held locally? Can you consolidate this data on a fast file server and connect a backup device to this server? This strategy is particularly important if your backup window is shrinking.

3. Test your throughput capabilities wherever you have very large concentrations of data or wherever you suspect you may have a throughput problem. Several techniques for doing this kind of testing are described in the next section. Do you need to invest in larger capacity devices? Should you change from DAT to DLT or DTF tape format? Do you need faster network cards or disks?

4. Plan for future growth. If a new application is being added to a particular machine and you expect it to generate a lot of data very quickly, consider directly connecting a device to that machine.

Data location is critical,especially now that distributed backup processing is common. Complex streaming techniques may need to be used if your data is not balanced correctly and matched to appropriate hardware.

TESTING PERFORMANCE

If you have a backup system in place and have done the theoretical work of balancing the data load among the backup devices and device servers at your disposal, it's probably time to do some systematic testing if you are still unhappy with your performance numbers. The most effective way to do this testing is to isolate components and test their actual throughput.

Some vendors supply utilities for this kind of testing, and the programs can be quite sophisticated and complex, involving many different variables that can be set. However, the theory behind these programs is very simple, and qualified programmers should be able to write their own versions.

THREE HANDY UTILITIES

You can do a great deal of very effective performance testing with three simple utilities:

- disk read
- device write
- network read/write

In these programs, your objective is to produce "pure" benchmark numbers that give a realistic reading of what your hardware and software can actually do.

Here is a skeleton for a typical program:

1. Start a timer.
2. Open.
3. Read or write continuously.
4. Close.
5. Stop the timer.
6. Calculate a throughput result by dividing the amount of data read or written by the time recorded on the timer to do the reading or writing.

Two variables are used to control the programs: the block size of the data read or written and the number of blocks.

The network program is more complex than the programs that do disk read and device write testing because two devices and many other variables are involved. These variables include the number of channels, the number of processes on the network, and whether you are testing local or remote performance.

These performance utilities are normally written in C and do not use TAR or a sophisticated backup product because the aim is to eliminate all backup-related housekeeping and overhead, such as writing catalog entries. You want to see "pure" numbers for throughput in a given situation.

DISK-RELATED TESTING

You can use the "pure" disk read program to do the following:

- check the performance of any server that you think might be causing a bottleneck on your system

1. For an excellent discussion of dynamic parallelism,see "Backup Techniques" by W. Curtis Preston in *Sys Admin* (February 1998),pp. 47 - 54.

- evaluate servers to choose a Master Server
- adjust performance-related settings such as the amount of concurrency per device

You can also use the program to compare raw partition and file-level processing, if you need to make a decision about which type to use.

DEVICE-RELATED TESTING

You can use the "pure" backup device program to perform these functions:

- decide where to attach particular backup devices
- measure how one or more devices perform when they are connected to a specific SCSI card
- test hardware compression settings

The real value of this program is its ability to give you a reasonably reliable reading on exactly what a particular device can do in a given situation that would otherwise be difficult to isolate.

NETWORK-RELATED TESTING

The "pure" network program is ideal for providing an accurate reading of how fast a backup can be written across a network (node-to-node) from a particular server to a particular backup device. By running several copies simultaneously, you can actually simulate a fairly complicated backup job.

If you have the luxury of a quiet system, you can use the program for solid benchmark readings. However, if you are troubleshooting, you probably want to test under conditions that are as realistic as possible.

SOFTWARE SETTINGS

Several software options are very resource-intensive and can affect performance dramatically. Because these options are often set globally, that is, for every job, it's very easy for one of them to be set during installation when it doesn't have to be.

Backup software settings are usually divided into two major categories: source and destination. Following are the performance-related source options that you need to check:

- verification
- network compression
- network encryption

Indeed, the difficulty in optimizing backup performance is not in grasping any particular part of the performance problem. The difficulty is that there are so many parts, and they interact in complex ways.

On less powerful machines, you may need to pay special attention to the "concurrency per device" setting. Decreasing this setting may improve performance by decreasing contention between modules.

You should check these two backup destination options:

- File Checksum
- File Compression

Pay particular attention to these performance-related options if your throughput is suffering only for specific jobs or specific types of jobs. Both options force the backup to do more work, and of course, more work means that backup jobs take more time and use more resources.

VERIFICATION

Verification can take several different forms, and some products let you set "levels" of verification, where each level adds one of the following tasks:

- **Level 1:** The backup reads the tape it has written to "verify" that it can be read.
- **Level 2:** The backup reads as in Level 1, and also checks header values.
- **Level 3:** The backup performs both procedures in the previous levels and compares checksum values. The original checksum is usually generated by setting the File Checksum destination option. If you aren't doing this level of verification, make sure the File Checksum option is off so that you don't waste time doing unnecessary work.

Level 3 verification is usually reserved for backups that are especially critical, that is, for backups of data that can never be recreated or can only be recreated at a considerable cost in time and money. Level 1 is much more commonly used, even though it does add some time to a backup job.

COMPRESSION

Compression often involves a trade-off. A compressed file takes less time to travel across a network and the compressed data takes up less space on a backup media, such as tape. However, the compression process takes time, both CPU and wall-clock, and you shouldn't compress small amounts of data or data that is not very compressible (data without a lot of repeating characters). As a general rule, you should only compress data if you have a fast CPU and a slow network.

ENCRYPTION

Encryption is a security feature that some sites must have. It keeps data safe from prying eyes as it passes over the network and makes files unreadable when they are stored. However, encryption requires very CPU-intensive processing and takes a considerable amount of time.

If security on the network is your only concern, attaching a device directly to the machine holding the data that must be kept secure is often the best solution. Because of the time factor, some backup products choose only to encrypt backup messages to protect the backup image itself from tampering and make the other types of encryption optional.

SUMMARY

Optimizing a network backup often turns out to be surprisingly difficult because there are so many variables involved. They range from hardware mismatches to software settings that add unnecessary work (and time) to backup processing. However, as this concluding article explained, having a powerful, dedicated Master Server and balancing data and devices can help. Also, understanding some simple programs that can be used for both planning and troubleshooting, as well as knowing which software settings can drain resources and effect performance, will allow you to optimize your use of backups.

The best backup software today is very sophisticated and written to optimize performance. If your backup system is also set up to optimize performance, you can

exploit backup product efficiency to its fullest, keeping every kind of network secure cost-effectively. ts

Ira Goodman is a software services manager at Syncsort, Inc. He has 24 years of experience in information processing, and has managed backup product support for almost a decade. He currently manages support for Backup Express, a backup product for networks running UNIX, NT, and NetWare, and advises customers on the setup and maintenance of distributed backup systems and automated libraries.