Teradactyl
True incremental Backup System®

TeraMerge®

TeraMerge® Enables
<u>Multiple Level</u> Synthetic Backup Consolidation
for the
Full Version of the True incremental Backup System®

(Increases Backup Server Performance)

Teradactyl LLC.
2450 Baylor Drive SE
Albuquerque, NM 87106
U.S. & International Sales: (505) 338-6000

info@teradactyl.com

www.teradactyl.com

{This white paper studies TeraMerge® multiple level synthetic backup consolidation, as used in the Full Version of the True incremental Backup System® (TiBS). There is also a previous white paper titled "TeraMerge® Enables Single Level Synthetic Backup Consolidation". This technology was first engineered by Teradactyl® for the software release of TiBS 1.0. TeraMerge® was later improved and extended to do multiple level merges for TiBS. Teradactyl has received a patent for TeraMerge® multiple level synthetic backup consolidation.}

Introduction

As computer information systems continue to grow, companies require more efficient strategies for managing backup and recovery processes. Teradactyl introduced TeraMerge® technology to eliminate the constant recopying of data from backup clients over data networks. By recycling data already on a backup server, the True incremental Backup System® (TiBS) generates newly updated synthetic full backup images without contacting backup clients. It also uses disk caching on the backup server to generate cumulative incremental backup volumes, further reducing the impact of the backup function on the rest of the computing environment. This allows administrators to take an original full backup and then only the files that have been modified or created since the last successful network backup.

This technology was originally limited to two backup levels, which required the frequent creation of new synthetic full backup images to keep the backup server cache utilization under control. This paper discusses an extension to TeraMerge® that provides for the creation of multiple backup levels similar to those available in traditional level backup systems. This new exploitation of the merging process dramatically reduces the backup server's workload allowing the same hardware to support an extended amount of data with a reduced tape or media cost.

Traditional Multiple Level Backups

Many backup products implement what is known as a "traditional level backup." The levels are usually defined from zero to nine (0-9). The way backup levels behave can be simple or complex depending on how they are scheduled. Level 0 is defined as the full or epoch backup. All of the data on a client is backed up during a level 0 backup. All higher level backups (level n) will backup either all data changes since the most recent backup at level n or lower (a partial-cumulative incremental backup), or all data changes since the most recent backup at level n-1 or lower (a cumulative incremental backup). For example, a level 2 cumulative incremental backup will backup all data that has been modified or created since the last level 0 or level 1 backup (whichever was most recent).

In this traditional level scheme, there is a classic tradeoff between network and client loads, versus the exposure to potential tape failure in the restore process. Schedules that are more complex offer a significantly reduced network load, but require more tape read requests to restore data completely. The need for more tapes in the restore process increases restore time and the potential exposure to tape failure. In a traditional level backup scheme a single bad tape can lead to lost restore data. Using fewer levels can improve restore time and reliability but at the expense of increased loads on networks, backup clients and backup media costs. Sites that are limited by any combination of the size of data, the rate that data changes, the available backup window, the rate that backup data can be transferred over the available network bandwidth, or backup media costs, will tend to use a 3 level or higher backup scheme.

The Evolution of TeraMerge®

Phase 1 - TiBS was originally designed with four types of backup in mind:

Network Full Backup: The original full or epoch backup, all of the data on a client system is backed up across the network. The data backed up in a full backup is not just file data, it also includes meta data such as directories, folders, filenames, file attributes, access control lists and other security information.

True incremental Backup: A special type of network partial-cumulative incremental backup. A True incremental Backup will only backup the files that have been modified or created since the most recent backup from the client (the most recent backup could be at any level). This is typically a daily backup that only copies the files that have changed since the previous day.

Synthetic Cumulative Incremental Backup: The True incremental data taken from the client is integrated with cumulative incremental data on a backup server disk cache to produce a new Synthetic Cumulative Incremental Backup volume.

Synthetic Full Backup: Periodically, cumulative incremental data in the backup server disk cache is integrated with a previous full backup volume to produce a new Synthetic Full Backup volume.

These four types of backup allow sites to process much of the backup function entirely on the backup server and shrink nightly backup windows significantly. However, new full backups still need to be generated frequently, typically on a weekly or bi-weekly basis. This limits the amount of data that a single backup server can support. If less frequent full backups are required or desired, the backup server disk cache size grows in an almost linear fashion.

Phase 2 - A new type of backup volume was developed to allow TiBS backup-servers to support larger data sizes:

Synthetic Partial-Cumulative Incremental Backup: This new technique allows the data from the backup server cache to be copied to backup media, creating an intermediate synthetic backup volume between full and True incremental level backups. Once the new Synthetic Partial-Cumulative Incremental Backup volume is created, data can be removed from the disk cache. This can be done multiple times allowing the frequency of full backups to be greatly reduced. Historically there was no merge method to take incremental data, multiple Synthetic Partial-Cumulative Incremental volumes, and a previous full backup to create a new full backup volume. Periodic full network backups are still required. Additionally, each Synthetic Partial-Cumulative Incremental volume adds one more backup volume to the restore process.

Synthetic Partial-Cumulative Incremental Backups do allow full backup frequencies to be decreased to several months, reducing their average daily impact on networks and backup clients. This significantly reduces the backup server load, which results in an overall increase in the amount of data that a single backup server can support.

Phase 3 – The Multi-Level Merge Approach

To increase the amount of data that a single backup server can support, Teradactyl® has extended TeraMerge® to support an arbitrary number of backup levels similar to the way a traditional level backup system processes data.

Synthetic Cumulative Incremental Backup: This new type of backup is based on the Synthetic Full Backup technique. At each intermediate backup level, data accumulated in the cache is merged with previous backup data at that level and written to backup media. If a lower level backup has also been scheduled (e.g. a new Synthetic Full Backup) the data is retained in the cache for the next merge process, otherwise it is removed.

The Synthetic Cumulative Backup gives customers increased flexibility over the original TeraMerge® design. Increasing the number of backup levels beyond the basic full and incremental backups reduces both the workload on the backup server and the amount of space required for storage.

For example, consider a new three level backup approach; Monthly, Weekly, Daily:

An initial Network Full Backup is taken

Each day, a True incremental Backup takes only the changes since the previous day's backup. This True incremental Backup is merged with data in the backup server cache to produce a new Synthetic Cumulative Incremental Backup volume. (This entire process can be referred to as a Network Synthetic Cumulative Incremental Backup).

Each week, the data in the backup cache is merged with previous Weekly data to produce a new Weekly backup volume. The new Weekly backup volume will contain all changes on the backup client since the last Full Backup.

Each month, the cumulative data from the Weekly backup is merged with previous Monthly backup data to produce a new Monthly backup volume. The resulting Monthly full backup contains the same data as one that would have been taken from the target system for that period of time.

Compared with a simple Weekly, Daily two level backup strategy, the three level backup strategy requires full backups to be created at one-fourth the frequency. Because the amount of data that changes on a system is usually a small percentage (typically less than 3% daily), the amount of processing required for Weekly and Daily backups is much lower than that of generating a new full backup volume. The reduction in backup server loading translates directly into lower tape utilization. Because full backups are generated with less frequency, the number of copies of unchanging data is dramatically reduced.

The only disadvantage to adding an additional backup level is that one more backup volume read will typically be required to restore data. Site administrators must make the tradeoff between the additional time required for processing restores and the increase in efficiency of the backup process. Since backups run continuously and restore requests are typically much less frequent, most sites will benefit from an increase in the number of backup levels.

Comparison of Backup Server Loads

To see how multi-level TeraMerge® impacts backup performance, the data from two active sites totaling almost ten terabytes is averaged below. Both sites use a four level backup strategy with Bi-Annual, Monthly, Weekly and Daily backups. Data processing for each backup level on a daily basis is summarized as a percentage of the total data:

Backup Level	Frequency	2 Level	3 Level	4 Level
Bi-Annual	182 days			.6%
Monthly	28 days		3.6%	.8%
Weekly	7 days	14.3%	1.3%	1.3%
Daily	1 day	2.6%	2.6%	2.6%
Average Daily Load		16.9%	7.5%	5.3%

Workload Comparison as a Percentage of Total Backup Data Size

The full backup load for each strategy is estimated as the current full data size divided by the number of days in the full cycle (7, 28, and 182 respectively). The selection of 28 days to represent a month is done to synchronize Monthly and Weekly processing. As a result there are 13 months (364 days) in a year and a new full backup is generated every 6.5 months. This actually results in the generation of 7 monthly backups per full backup cycle in the four level strategy. This extra amount of data processing has been included in these results.

Note that the data load for all three strategies is the same for the Daily backup. This is because once a week, for each strategy, the incremental data is merged and removed from the cache. For the two level strategy, the Weekly full backup represents $1/7^{th}$ of the total data that must be processed each day. Even though only a small percentage of data is changing, the two level strategy must constantly recopy unchanged data. With three backup levels, the impact that full backup processing has on the server's average daily load is reduced by a factor of four and the total workload is reduced by more than fifty percent. Moving to a four level strategy reduces server workloads even further, allowing the same backup server to support over 3 times the amount of data compared with the two level strategy.

Another interesting statistic is the amount of daily work that actually comes off of the network. TiBS tracks the percentage of data generated over the network and the percentage of data that is reused from the backup cache. Both sites we reviewed had a network percentage of approximately 50%. This means that for a one-week daily cycle about 1.3% of the total data size comes from the backup clients each day. An important result of the TeraMerge® process is that any number of backup levels may be configured on the backup server without impacting the workload placed on networks or backup clients.

Inherent Data Redundancy/Disaster Recovery Protection

Each TeraMerge® created synthetic backup volume is a product of two other volumes (an older lower level backup and a cumulative incremental). After the first synthetic consolidation there are actually two copies of each backup consolidation (one is a product of one or more backup volumes). This gives an inherent protection against data loss due to media failure, if appropriate data retention policies are maintained

Offsite Management Considerations

Multiple level merging allows customers to reduce the workload on backup servers allowing a single server to support larger amounts of data. TiBS also provides a mirroring option for offsite considerations. Data is typically written to a pair of tapes at approximately 75% the data rate of unmirrored tape writes. For example, moving from a two level to a four level strategy and mirroring all tapes will still result in a server that can support twice as much data, while creating tapes for offsite and onsite purposes. Even greater time efficiencies and cost savings can be found if tape usage is combined with the Disk Library Interface (DLI).

Summary

The extension of TeraMerge® to support multiple levels of backup provides large, distributed sites with the flexibility to trade off restore times, tape costs, and backup server loading. Data sizes continue to grow and so do related costs. TiBS now supports multiple level backup strategies that allow customers to continue to scale data while controlling the cost of data protection.