TeraMerge® Enables
Single Level Synthetic Backup Consolidation
for the
Lite Version of the True incremental Backup System®

(Reduces the Impact of Backups on Networks)

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® single level synthetic backup consolidation, as used in the Lite Version of the True incremental Backup System® (TiBS). 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. There is another white paper titled "TeraMerge® Enables Multiple Level Synthetic Backup Consolidation".}

Introduction

The last decade saw an explosion in hard drive capacities, enabling data servers, network workstations, and roaming clients to carry more information than ever before. As a result, the once benign process of keeping information backed up has become an IT mission.

This paper presents a powerful strategy for managing the backup function that significantly minimizes the impact on data networks and backup clients, while reducing the risk of data loss due to media failure. This new design allows backup administrators to re-define their backup plan to achieve higher reliability and greater backup efficiency. This technology returns IT resources to their intended purpose and improves the utility of the backup system hardware.

Traditional Backups

Most traditional backup solutions use a combination of four types of backups: Full, Differential, Cumulative Incremental, and Incremental (True incremental). There has been some confusion in the past about the accepted definitions. These terms have often been used improperly and interchangeably by different companies. The following definitions have been accepted by industry leaders and government agencies such as the National Institute of Standards and Technology (NIST).

Full (level 0)
In a full or epoch backup, all of the data on a client system is backed up. This can be across the network or to direct attached storage. The data backed up in a full backup is not just file data, it is also meta data such as directories, folders, filenames, file attributes, access control lists and other security information.

Differential (level 1)
A differential backup will backup the files that have been modified or created since the last full backup. Full restores typically require two backup volumes, the last full backup and the latest differential. Data that has changed and remains on the client system since the last full backup, but hasn't changed since the previous differential, will be backed up again. Periodic network full backups are suggested. (A differential backup is a specific type of cumulative incremental backup.)

Cumulative Incremental (level 1+)
A cumulative incremental backup will backup the files that have been modified or created since the previous lower level backup (see TeraMerge® Enables Multiple Level Synthetic Backup Consolidation white paper). In a two level backup strategy, as discussed in this paper, a cumulative incremental backup is identical to a differential backup (since the previous lower level backup is a full).

Incremental (True incremental)
An incremental backup will only backup the files that have been modified or created since the most recent backup (at any level). This is much less data than any of these other types of backup (usually a single day). Each restore requires the last full backup as well as every incremental backup volume since then, in sequence. Periodic network full backups are still suggested.

Limitations with Traditional Two Level Backups:

A typical two level backup strategy will take a periodic network full (level 0) backup weekly, bi-weekly or monthly depending on the size of the data, the available network bandwidth, and the perceived cost associated with loss of data.

A periodic daily backup will take either the differential (cumulative incremental) or incremental (True incremental) changes from the backup clients. Each strategy has its own limitations.

Differential (Cumulative Incremental)

A daily differential (level 1) backup will backup all the data modified or created since the last full backup. Full restores typically require two backup volumes, the last full backup and the latest differential.

The limitation with this approach concerns each successive daily (level 1) backup, which will typically be larger than the previous, in a near linear growth. An increasing amount of data will be backed up repeatedly and redundantly. Data that has changed and remains on the client system since the last full backup, but hasn't changed since the previous daily differential, will be backed up again. This is an inefficient use of network bandwidth and other IT resources. Most importantly, a new network full backup will eventually be needed to mitigate this data growth.

Incremental

A daily incremental backup will backup all data modified or created since the most recent backup (at any level). This will reduce the daily network and backup client loads.

The primary limitation with this approach is that restores can be a very slow process. Each restore requires the last full backup as well as every incremental backup since then, in sequence. Two weeks after a full backup, a restore request could need more than a dozen tapes. A large number of incremental tapes increase the probability that a bad tape or a bad spot on a tape could result in data loss. If the backup system stores data on disk then storage sizes and cost will continue to escalate. In order to mitigate the increasing number of tape volume required for restore, periodic full backups over the network are still suggested.

The backup server for this type of system must be able to efficiently determine exactly what backup volumes to read and what data to send to the client. If not, an additional burden is placed on the client during the restore process, because many copies of the same file may be unnecessarily transferred. Deleted files could be restored, with possible security and regulatory implications. The system administrator or user would need a mechanism to identify and cull these files.

> **Co-Location of Day to Day Backup Volumes**
>
> Some products will attempt to keep related backup volumes together on a single tape. This approach reduces the tape complexity for restores, but requires free space to be left available on tape cartridges to append future data. This also increases the exposure if a single tape fails (many days worth of backup data could be lost at once). Tape Mirroring is often suggested to mitigate this risk, but that further increases the cost associated with unused tape space and inefficient use of expensive tape library hardware.

Problem Summary

Differential (Cumulative Incremental)

- Periodic network full backups are suggested
- Repeated and redundant sending of incremental data
- Near linear growth of level 1 data that must be backed up daily over the network
- Inefficient use of network bandwidth and IT resources
- Heavy network and client loads
- Deleted files can be restored
- Restores can transfer many copies of the same file
- Greater exposure to data loss due to tape failure

Incremental

- Periodic network full backups are still suggested
- Slow restore time
- Increasing number of backup volumes or tapes needed for a restore
- Greater exposure to data loss due to tape failure
- Deleted files can be restored
- Restores can transfer many copies of the same file
- Inefficient use of network bandwidth and IT resources
- Heavy network and client loads

Engineering a Better Backup Solution – A True incremental Approach

Desired Requirements for Teradactyl® to Implement:
- Create a one-time full backup from a backup client over the network
- Eliminate the need for periodic full backups over the network
- Periodically back up only new or modified files (True incremental forever)
- Eliminate the need to take copies of files that have not changed and have already been backed up
- Generate a new full backup volume from data stored entirely on the backup system without accessing the network or the backup client (a synthetic backup consolidation)
- Reliably recover the last synthetic consolidated backup state in the event of a backup client media failure

TeraMerge® – The True incremental Backup System® (TiBS)

The patented TeraMerge® synthetic backup consolidation process was developed to implement the key features of a True incremental Backup System®.

Recent incremental data (data since the last full backup was generated) is mirrored in a backup server disk cache. Data restores can occur from the cache, reducing tape load requests and restore times. Restore of a failed client partition from the last backup may only require one full tape read if all incremental data can come from the cache.

TiBS merges the most recent True incremental changes from a backup client with cumulative incremental data from the backup server cache to create a new Synthetic Cumulative Incremental Backup volume. The backup server reuses files that would normally be resent by a backup client in a traditional two level differential backup.

New Synthetic Full Backup volumes are generated by merging the last full backup with cumulative incremental data from the backup server disk cache and carefully omitting older files that have been deleted or updated on the client system. This TeraMerge® synthetic backup consolidation is performed with no interaction with backup clients. This eliminates the periodic full backup load from the data networks and backup clients.

The resulting Synthetic Full Backup volume contains the same data as one that would have been taken from the client at that point in time using a traditional two level backup. All merging processes occur on the backup server. Backup clients are responsible for sending the latest changes during incremental backups and the backup server confirms the list is complete. The backup server reuses files that would normally be re-sent by the backup client in a traditional two level backup to generate new Synthetic Cumulative Incremental Backup volumes and complete Synthetic Full Backup volumes.

Benefits of TeraMerge® Single Level Synthetic Consolidation

**Elimination of Periodic Network Full Backups**
- TeraMerge® merges the cumulative incremental client data in the disk cache with previous full data on the backup system to produce a new Synthetic Full Backup.
- The new Synthetic Full Backup volume contains the same data as one that would have been taken from the backup client at that time point in time using a traditional two level backup.
- New Synthetic Full Backups are generated with absolutely no interaction with the network or backup clients.

Inherent Data Redundancy/Disaster Recovery Protection
- Each TeraMerge® created Synthetic Full Backup volume is a product of two other volumes (an older full and a cumulative incremental).
- After the first synthetic consolidation there are actually two copies of each full backup (one is a product of two backup volumes).
- This gives an inherent protection against data loss due to media failure, if appropriate data retention policies are maintained.

Reduction of Daily Differential Network Backup Loads
- Traditional two level differential backup has near linear growth in the amount of data that must be backed up over the network.
- TeraMerge® flattens the network data load by eliminating redundant data.
- TeraMerge® reuses data in the backup server cache.
- Only files that have been modified or created since the last True incremental backup are copied over the network.
- This TeraMerge® created Synthetic Cumulative Incremental Backup is a synthetic backup consolidation that contains the same data as a level 1 differential backup, but with significantly less network load.

Traditional Network Load
vs.
TeraMerge® Network Load

■ Level 1 Data Size ■ TeraMerge Network Size
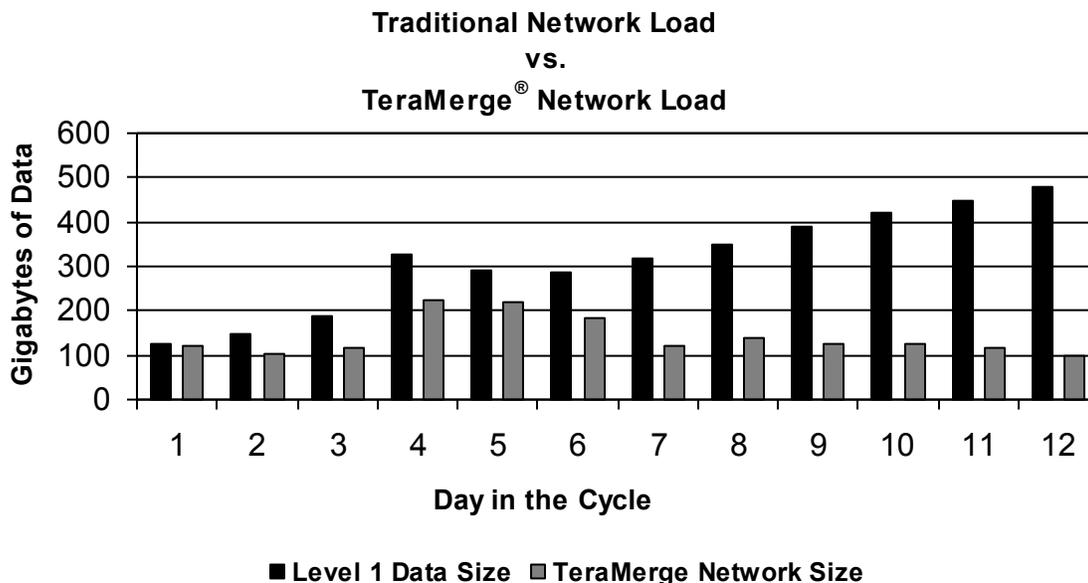
Backup Server Disk Cache Processing
- Allows the reuse of data in the disk cache for synthetic backup consolidations. This significantly decreases the network load.
- Allows faster restores. Often only a single full volume tape read is necessary, if differential data can be taken from the disk cache.
- Serves as a staging area to write to tape devices. This allows much more efficient data streaming to tape, resulting in faster writes with higher compression rates**.**
- Separates the network backup function from the tape write function. Network backups can continue even if there are hardware problems with a tape library or drive.
- Backup consolidations can continue on schedule if the network becomes unavailable.
- The backup server disk cache can also be recovered from tape.

Estimating TeraMerge® Efficiency

Most sites can reduce the network and client load for the backup function by 75%-95% (4 to 20 times). The efficiency for each site depends on several key factors:

Current Schedule (C)
Frequency of full backups and how many levels are in the current schedule are both important factors. In general, sites that generate full backups more frequently benefit the most by eliminating full network backups, while sites that use less aggressive schedules will enjoy other benefits including reduced impact on IT resources.

Average Traditional Data Rate (R)
This is the average over a backup cycle of the size of differential data taken on each backup (from the Network Load Chart above). It can be estimated by adding up the tape utilization for differential backups taken between two full backups and dividing by the number of differential backups taken in the cycle.

Average TeraMerge® Network Rate (N)
This is the average of the TeraMerge® network size data for each True incremental backup (from the Network Load Chart above). This value is typically found to be 1% -2% of the total data size.

A Simple Model for Predicting the Efficiency of TeraMerge® versus a Traditional Level Backup

Let: C = Number of days in a backup cycle (includes one day for full backup)
R = Average daily data rate, as a percent of total data
N = Average TeraMerge® daily data network rate, as a percent of total data

Then the network utilization for a single backup cycle would be:

	Traditional	TeraMerge®
Day 1	100	N
Day 2	R	N
Day C	R	N

$$\text{Total} = 100 + (C - 1) * R \qquad \text{Total} = C * N$$

$$\text{Efficiency} = 1 - \frac{C * N}{100 + (C - 1) * R}$$

If we use the data from the Network Load Chart above ($C = 14$, $R = 3.06\%$, $N = 1.18\%$), then

Efficiency = 88.2%
Reduction in Network Bandwidth = 8.4 times

Summary

The capabilities of TeraMerge® synthetic backup consolidation shifts the paradigm for backup management by permanently removing the burden of periodic full backups from networks and backup clients. Without the performance trade-offs imposed by traditional level backup products, backup administrators are free to rethink their backup policies and procedures. TeraMerge® allows for low network and backup client loads with increased reliability and improved user response times. Using TeraMerge® technology, sites can efficiently centralize backups over existing data networks and take advantage of new high capacity tape drives and disk drives.