TSM 6.1 data deduplication and differences with TS7650G

Oliver Rautenberg
TSM Level 2 Support
Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
## Agenda – Datadeduplication

- **TSM Data Reduction**
  - What is Data Deduplication?
  - TSM Deduplication Overview
  - Deduplication Highlights
  - Data Deduplication within TSM
  - Where TSM Deduplication is Performed
  - Generalized Deduplication Processing
  - Externals: Server Commands
  - Externals: Server Options
  - Considerations for Use of TSM Deduplication
  - Deduplication Example
  - Design Points for Deduplication in TSM 6
  - Comparison of TSM Data Reduction Methods
  - Considerations for Use of TSM Deduplication
  - Client Dependencies
TSM Data Reduction

Client compression
- Files compressed by client before transmission
- Conserves network bandwidth and server storage

Server-side deduplication
- Added in TSM 6
- Conserves server storage
- Improves recovery time as compared to storage on tape

Device compression
- Compression performed by storage hardware
- Conserves server storage

Subfile backup
- Only changed portions of files are transmitted
- Conserves network bandwidth and server storage

Incremental forever
- After initial backup, file is not backed up again unless it changes
- Conserves network bandwidth and server storage

Storage Hierarchy

Incremental forever greatly reduces data redundancy compared to traditional methodologies. TSM has less potential for data reduction via deduplication as compared to other backup products.

Appliance deduplication
- Deduplication performed by storage appliance (VTL or NAS)
- Conserves server storage
Agenda – Datadeduplication

- TSM Data Reduction
  - What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
What is Data Deduplication?

- Data deduplication (often called "intelligent compression" or "single-instance storage") is a method of reducing storage needs by eliminating redundant data. Only one unique instance of the data is actually retained on storage media, such as disk or tape. Redundant data is replaced with a pointer to the unique data copy. For example, a typical email system might contain 100 instances of the same one megabyte (MB) file attachment. If the email platform is backed up or archived, all 100 instances are saved, requiring 100 MB storage space. With data deduplication, only one instance of the attachment is actually stored; each subsequent instance is just referenced back to the one saved copy. In this example, a 100 MB storage demand could be reduced to only one MB.
What is Data Deduplication?

- Data deduplication can generally operate at the file, block, and even the bit level. File deduplication eliminates duplicate files, but this is not a very efficient means of deduplication. Block and bit deduplication looks within a file and saves unique iterations of each block or bit. Each chunk of data is processed using a hash algorithm such as MD5 or SHA-1. This process generates a unique number for each piece which is then stored in an index.
- Find and eliminate duplicate data stored in a TSM storage pool
## Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- **TSM Deduplication Overview**
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
TSM Deduplication Overview

Files A, B, and C have common data. Deduplication reduces the number of unique chunks stored, allowing more objects to be stored on disk for fast access.

Deduplicated disk storage pool stores unique chunks to reduce disk utilization.

Tape copy pool stores A, B, and C individually to avoid performance degradation during restore.

Node | File
--- | ---
Client 1 | A
Client 2 | B
Client 3 | C
Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
Deduplication Highlights

- Disk storage requirement reduced via optional data deduplication for FILE storage pools
- Deduplication processing performed on TSM server and tracked in database
- Reduced redundancy for
  - Identical objects from same or different client nodes (even if file names are different)
  - Common data chunks (subfiles, extents) in objects from same or different nodes
- Post-ingestion (out-of-band) detection of duplicate data on TSM server to minimize impact to backup windows
- Space occupied by duplicate data will be removed during reclamation processing
- Allowed for all data types: backup, archive, HSM, TDP, API applications
- Transparent client access to deduplicated objects
# Deduplication Highlights

- Deployment of new clients or API applications not required
- Legacy data stored in or moved to enabled FILE storage pools can be deduplicated
- Data migrated or copied to tape will be reduplicated to avoid excessive mounting and positioning during subsequent access
- Ability to control number, duration and scheduling of CPU-intensive background processes for identification of duplicate data
- Reporting of space savings in deduplicated storage pools
- Deduplication processing will skip client-encrypted objects, but should work with storage-device encryption
- Native TSM implementation, with no dependency on specific hardware
Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
Data Deduplication within TSM

- Native TSM implementation, with no dependency on specific hardware
- Reduced redundancy for:
  - Identical objects from same or different client nodes (even if names are different)
  - Common data chunks (subfiles, extents) in objects from same or different nodes
- Allowed for all data types: backup, archive, HSM, TDP, API
- Only supported with stgpool using devclass=file
- Server-side Deduplication
  - Deduplication processing performed on TSM server and tracked in database (server-side deduplication)
  - Post-ingestion detection of duplicate data on TSM server to minimize impact to backup windows (out-of-band)
  - No required change to clients or legacy data in storage pools
    - New or Legacy data stored in or moved to enabled FILE storage pools will be deduplicated
  - Transparent client access to deduplicated objects
  - Space occupied by duplicate data will be removed during reclamation processing
Data Deduplication within TSM

- Data migrated or copied to tape will be re-duplicated to avoid excessive mounting and positioning during subsequent access.
- V6.1 client will change to identify meta data vs content.
  - Meta data is unique for each file, even if the file contents are identical. The 6.1 client has been enhanced to send the meta data to the TSM Server before sending any file data which will allow deduplication to be performed against the file data only. This will also allow the TSM Server to perform deduplication against files compressed by the client.
- Deduplication will not be effective for client-encrypted data, but should work for storage-device encryption.
  - The same file encrypted with different keys cannot be deduplicated on the server (not a 100% match). Since encryption is end to end and unless the exact same key is used the chunks will not match.
Data Deduplication within TSM

- Use SHA-1 hashing algorithm
  - Chance of one collision in a 40 PB archive*** (using 4KB / chunk) is $0.5 \times 10^{-28}$
  - The probability of one hard-drive bit-error is about $10^{-14}$
  - Rabin fingerprinting algorithm to determine chunk size
- Entire object hashed as well and used during restore as validation
  - Using MD5 hashing technique
- Reporting of space savings in deduplicated storage pools
- All data within a storage pool will be deduplicated
  - If you don’t want data deduplicated, move it or direct it to an alternate non-dedup pool.
Data Deduplication within TSM

- New storage pool attributes
  - indicate whether the pool is a deduplicated storage pool
  - indicate how many processes to use to find duplicate data
- Deduplication is a 2-phase operation
  - Phase 1 examines the data and identifies common chunks
    - Existing FILE volumes are opened and read (I/O intensive)
    - Data is fingerprinted into chunks (CPU intensive)
    - A SHA digest is calculated on the chunk (CPU intensive)
      - To avoid false-positives, size of chunk and another quick digest are also checked
    - Common chunks are replaced with pointers to the location of the common data (DB updates, LOCK contention)
  - Phase 2 removes the common/redundant chunks
    - Via storage pool reclamation
Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
## Where TSM Deduplication is Performed

<table>
<thead>
<tr>
<th>Approach</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source-side (client-side)</strong></td>
<td>- Deduplication before transmission conserves network bandwidth</td>
<td>- Deduplication consumes CPU cycles on the file/application server</td>
</tr>
<tr>
<td></td>
<td>- Awareness of data usage and format may allow more effective data reduction</td>
<td>- Requires software deployment at source (and possibly target) endpoints</td>
</tr>
<tr>
<td></td>
<td>- Processing at the source may facilitate scale-out</td>
<td>- Depending on design, may be subject to security attack via spoofing</td>
</tr>
<tr>
<td><strong>Target-side (server-side)</strong></td>
<td>- No deployment of client software at endpoints</td>
<td>- Deduplication consumes CPU cycles on the target server or storage device</td>
</tr>
<tr>
<td></td>
<td>- Possible use of direct comparison to confirm duplicates</td>
<td>- Data may be discarded after being transmitted to the target</td>
</tr>
</tbody>
</table>
## When TSM Deduplication is Performed

<table>
<thead>
<tr>
<th>Approach</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In-band</strong></td>
<td>- Immediate data reduction, minimizing disk storage requirement</td>
<td>- May be bottleneck for data ingestion (e.g., longer backup times)</td>
</tr>
<tr>
<td></td>
<td>- No post-processing</td>
<td>- Only one deduplication process for each I/O stream</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- No deduplication of legacy data on the target server</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Out-of-band</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Data must be processed twice (during ingestion and subsequent deduplication)</td>
</tr>
<tr>
<td><strong>Out-of-band</strong></td>
<td>- No impact to data ingestion</td>
<td>- Storage needed to retain data until deduplication occurs</td>
</tr>
<tr>
<td></td>
<td>- Potential for deduplication of legacy data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Possibility for parallel data deduplication processing</td>
<td></td>
</tr>
</tbody>
</table>

Deduplication performed after data ingestion at the target.
Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
Generalized Deduplication Processing

1. Chunk the object
   - Divide object into logical segments called chunks

2. Identify duplicate chunks
   - Hash each chunk to produce unique identifier
   - Compare each chunk identifier with index to determine whether chunk is already stored

3. Eliminate redundant chunks
   - Update index to reference matching chunks
   - Deallocate space for redundant chunks
Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
Externals: Server Commands

- New keywords for DEFINE/UPDATE STGPOOL commands
  - DEDUPlicate = [ Yes | No ]
    - for DEFINE STGPOOL, default is No
  - IDENTIFYPRocess = nn
    - Default number of background processes
    - for DEFINE STGPOOL, default is 1

- New IDENTIFY DUPLICATES command
  - `>>-IDentify DUPlicates--stgpool_name------------------>
  - `-+--------------------------------------------------><
    - `-DURation-----minutes-'  '-NUMPRocess-----number-'`

  - DURATION determines the length of time (in minutes) that the IDENTIFY DUPLICATES process should run
  - NUMPROCESS determines the number of process allowed to run
    - This value overrides the number of processes specified in stgpool definition
    - Value can be set to 0 to gracefully stop all processes for a storage pool
    - The identification processes will utilize a lot of CPU and disk I/O
Externals: Server Commands

Example of QUERY PROCESS output

<table>
<thead>
<tr>
<th>Process Number</th>
<th>Process Description</th>
<th>Status</th>
</tr>
</thead>
</table>

Notice that the process goes idle when not busy. As soon as more data is stored, the process will start identifying the duplicates.
Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
Externals: Server Options

- **DEDUPREQUIRESBACKUP [Yes | No] (default is Yes)**
  - Indicates whether a volume in a PRIMARY DEDUP storage pool can be reclaimed before it is backed up to a non-deduplicated copy pool.
  - If YES, a volume in a dedup primary pool cannot be reclaimed until it has been backed up via STGPOOL BACKUP.
  - If NO, the reclamation criteria remains unchanged (same as previous reclamation criteria).
    - Setting this value to NO could cause unrecoverable data loss in the highly unlikely event an object generates a false-positive match on another extent.
  - This option can be changed dynamically with the SETOPT command.
- **IDENTIFYAUTOstart [Yes | No] (default is Yes)**
  - Indicates whether the duplicate identification processes are to be started automatically when the server starts.
  - If you issue UPDATE STGPOOL DEDUP=YES against a pool that was previously DEDUP=NO, this option controls whether identification processes are started immediately.
  - If this option is NO, use IDENTIFY DUPLICATES command to start/stop the identification processes.
Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
Considerations for Use of TSM Deduplication

- Consider deduplication if
  - Data recovery would improve by storing more data objects on limited amount of disk
  - Data will remain on disk for extended periods of time
  - Much redundancy in data stored by TSM (e.g., for common operating-system or project files)
  - TSM server CPU and disk I/O resources are available for intensive processing to identify duplicate chunks

- Deduplication might not be indicated for
  - Mission-critical data, whose recovery could be delayed by accessing chunks that are not stored contiguously
  - TSM servers that do not have sufficient resources
  - Data that will soon be migrated to tape
Agenda – Data Deduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
**Deduplication Example**

1. **Client1** backs up files A, B, C and D. Files A and C have different names, but the same data.

```
Client1
Vol1
  A  B  C  D
```

2. **Client2** backs up files E, F and G. File E has data in common with files B and G.

```
Client2
Vol1
  A  B  C  D
Vol2
  E  F  G
```

3. **Server** process divides data into "chunks" and identifies duplicate chunks C1, E2 and G1.

```
Vol1
  A1 B1 B2 C1 D1
Vol2
  E1 E2 E3 F1 G1
```

4. **Reclamation processing** recovers space occupied by duplicate chunks.

```
Vol3
  A1 B1 B2 D1 E1 E3 F1
```
# Agenda – Data Deduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- **Design Points for Deduplication in TSM 6**
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
## Design Points for Deduplication in TSM 6

<table>
<thead>
<tr>
<th>Design point</th>
<th>Comments</th>
</tr>
</thead>
</table>
| Server-side | • Avoids need for deployment of client software  
                • Effective for all types of stored data |
| Out-of-band | • Allows deduplication of legacy data in addition to new data  
               • Minimizes impact to backup windows  
               • Concurrent processes to identify duplicate data |
| Index maintained in TSM server database (DB2) | Transactional integrity, scalability, performance, disaster protection |
| Variable-size chunking | • Rabin fingerprinting with awareness of TSM data format |
| Avoidance of false matches | • SHA-1 digest for each chunk  
                                 • Comparison of chunk size  
                                 • MD5 digest for entire data object (checked during restore) |
| Average chunk size 256KB | • Larger chunks require less database overhead  
                                 • Larger chunks reduce the total number of chunks required for given amount of data and therefore reduce collision probability  
                                 • Smaller chunks could improve deduplication ratio |
| Space occupied by redundant chunks recovered during reclamation | • Allows coordinated recovery of space occupied by deleted objects and redundant chunks |
Agenda – Data Deduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
## Comparison of TSM Data Reduction Methods

<table>
<thead>
<tr>
<th></th>
<th>Client compression</th>
<th>Incremental forever</th>
<th>Subfile backup</th>
<th>Deduplication in TSM 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>How data reduction is achieved</td>
<td>Client compresses</td>
<td>Client only sends</td>
<td>Client only</td>
<td>Server eliminates redundant data</td>
</tr>
<tr>
<td></td>
<td>files</td>
<td>changed files</td>
<td>sends subfiles</td>
<td>chunks</td>
</tr>
<tr>
<td>Conserves storage pool space?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Conserves network bandwidth?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Data supported</td>
<td>Backup, archive,</td>
<td>Backup</td>
<td>Backup</td>
<td>Backup, archive, HSM, API</td>
</tr>
<tr>
<td></td>
<td>HSM, API</td>
<td>(Windows only)</td>
<td>(Windows only)</td>
<td></td>
</tr>
<tr>
<td>Scope of data reduction</td>
<td>Redundant data</td>
<td>Files that do not</td>
<td>Subfiles that</td>
<td>Redundant data from any files in</td>
</tr>
<tr>
<td></td>
<td>within same file on</td>
<td>change between</td>
<td>do not change</td>
<td>storage pool</td>
</tr>
<tr>
<td></td>
<td>client node</td>
<td>backups</td>
<td>between backups</td>
<td></td>
</tr>
<tr>
<td>Avoids storing identical files</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>renamed, copied, or relocated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>on client node?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Removes redundant data for files</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>from different client nodes?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Agenda – Datadeduplication

- TSM Data Reduction
- What is Data Deduplication?
- TSM Deduplication Overview
- Deduplication Highlights
- Data Deduplication within TSM
- Where TSM Deduplication is Performed
- Generalized Deduplication Processing
- Externals: Server Commands
- Externals: Server Options
- Considerations for Use of TSM Deduplication
- Deduplication Example
- Design Points for Deduplication in TSM 6
- Comparison of TSM Data Reduction Methods
- Considerations for Use of TSM Deduplication
- Client Dependencies
Client Dependencies

The BA Client will be enhanced, as follows:

1. A new bit in the server function bitmap will be defined. This bit will indicate whether the server can accept a new StreamMarker verb. If the enhanced client sees this bit on in the server's function bitmap, it will separate the meta data from the file data, identifying each with a StreamMarker verb.

2. During restore processing, if the client originally sent StreamMarker verbs to the server during the store operation, the server will send the same verbs back during the retrieve operation (for both classic restore and NQR). However, if the server detects that the client is not at 6.1, it will NOT send the marker verbs because the client won't recognize them. Instead, it will send the data without the markers. In this case, the client will fail automatically as downlevel, because we're going to change the version of the File Decode Block. When a 5.5 or prior client sees the new version in the FDB it receives, it will fail.
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
TS7650G ProtecTIER Overview

- ProtecTIER software resides on TS7650G Deduplication Gateway
- Emulates a tape library unit, including drives, cartridges and robotics
- Uses FC-attached disk array as the backup medium
TS7650G ProtecTIER Overview
Inside ProtecTIER TS7650G
Replication with ProtecTIER
A Simple View of the Backup Process
Impact of Different Speeds
Summary

There are a Variety of Data Reduction Methods
TSM is Extremely Efficient at Data Reduction
Where Can Data Deduplication Occur?
IBM Delivers Data Deduplication Solutions
When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
Inside ProtecTIER TS7650G

New Data Stream

HyperFactor™

Repository

Memory Resident Index

Backup Servers

FC Switch

TS7650G

Disk Arrays

Existing Data

“Filtered” data
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
Replication with ProtecTIER

**Primary Site**
- Backup Server
- ProtecTIER Gateway

**Secondary Site**
- Backup Server
- ProtecTIER Gateway

- Significant bandwidth reduction
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
A Simple View of the Backup Process

Backup Processing Norms

- Backup Server
- Backup Target
- Tape Library
- Truck
- SLA Is Met

Backup process starts

Vault/Off-site process starts

Data is Off-site
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
## Impact of Different Speeds

<table>
<thead>
<tr>
<th>Profile</th>
<th>But</th>
</tr>
</thead>
</table>

- **Receive at 300 MB/s**
- **Post Process @ 100 MB/sec**
- **Backup 6 TB**
  - Takes approximately 6 hours
- **Post process then consumes the next 18 hours!**
- **So, it takes 24 hours to process 6 TB**

- **What about resources in support of vaulting/off-site?**
- **When de-dupe is a post-process it competes for disk resources as any other process?**
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
Summary

- When looking into dedupe based solutions make sure you ask the critical questions:
  - How fast is the dedupe process in an operational environment?
  - If dedupe is done in parallel to ingest, what is the impact on ingest speed?
  - Does capacity scale without impacting performance?
  - How does the solution scale in performance?
  - Does the system need ‘quiet’ times for space management?
  - Will dedupe impact operational/production activities?
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
## There are a Variety of Data Reduction Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Incremental Forever</strong></td>
<td>- Only new or changed files are transmitted</td>
</tr>
<tr>
<td></td>
<td>- Avoids regular full backups required by many backup solutions!</td>
</tr>
<tr>
<td></td>
<td>- Conserves network bandwidth and server storage</td>
</tr>
<tr>
<td><strong>Compression</strong></td>
<td>- Encoding of data to reduce size</td>
</tr>
<tr>
<td></td>
<td>- Typically localized, such as to a file, directory tree or storage volume</td>
</tr>
<tr>
<td></td>
<td>- Also done within storage devices themselves</td>
</tr>
<tr>
<td><strong>Single instance store (SIS)</strong></td>
<td>- Only one instance of a file is retained in the data store</td>
</tr>
<tr>
<td></td>
<td>- Duplicate instances of the file reference the stored instance</td>
</tr>
<tr>
<td></td>
<td>- Also known as redundant file elimination</td>
</tr>
<tr>
<td><strong>Data deduplication</strong></td>
<td>- A form of compression, usually applied to a large collection of files in a shared data store</td>
</tr>
<tr>
<td></td>
<td>- In contrast to SIS, deduplication often refers to elimination of redundant subfiles (also known as chunks, blocks, or extents)</td>
</tr>
<tr>
<td></td>
<td>- Only one instance is stored for each common chunk</td>
</tr>
<tr>
<td></td>
<td>- Duplicate instances of the chunk reference the stored instance</td>
</tr>
</tbody>
</table>

This terminology is not used consistently throughout the industry. In particular, the terms SIS and deduplication are sometimes used interchangeably.
## Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- **TSM is Extremely Efficient at Data Reduction**
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
**TSM is Extremely Efficient at Data Reduction**

**Client compression**
- Conserves network bandwidth and server storage

**Incremental forever**
- Only new/changed files backed up ever
- Conserves network bandwidth and server storage
- Greatly reduces redundancy

**Subfile backup**
- Only changed portions of files are transmitted
- Conserves network bandwidth and server storage

**TSM 6 server deduplication**
- Storage Pool based
- Conserves server storage
- Improves recovery time as compared to storage on tape

**Device compression**
- Conserves server storage

**Comparison of TB Stored**

**Appliance deduplication**
- VTL or NAS
- Conserves server storage

TSM and FastBack’s efficiencies leave less potential for data reduction via deduplication as compared to other backup products.
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
Where Can Data Deduplication Occur?

- Applications like email and content management are building in Single Instance Store and deduplication.
- WAN devices perform deduplication.
- Some deduplication vendors are promoting their appliances for live data as well as a backup target.
- VTLs serve as a target for backup applications and have added in-line and post process deduplication.
- Backup applications like TSM are including server deduplication.
- Some NAS devices perform Single Instance Store or fixed block deduplication on live data or can serve as a target for backup applications.
- Some backup applications can perform client or remote office server deduplication.
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
IBM Delivers Data Deduplication Solutions

Lotus Domino and IBM Content Manager deliver data deduplication

IBM partners for WAN optimization solutions

Future: TSM client and FastBack server deduplication

TSM 6 provides Storage Pool based deduplication

IBM n-series provides Single Instance Store and fixed block deduplication

ProtecTIER provides the industry’s highest scaling and highest performing deduplication solution

Backup Client
Remote Office Server
Source Side
Target Side
SAN
LAN
FC Storage
VTL
Backup Server
NAS Storage
IBM partners for WAN optimization solutions
Agenda – TS7650G (Diligent) ProtecTIER

- TS7650G ProtecTIER Overview
- Inside ProtecTIER TS7650G
- Replication with ProtecTIER
- A Simple View of the Backup Process
- Impact of Different Speeds
- Summary

- There are a Variety of Data Reduction Methods
- TSM is Extremely Efficient at Data Reduction
- Where Can Data Deduplication Occur?
- IBM Delivers Data Deduplication Solutions
- When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?
When Do I Use ProtecTIER vs TSM 6 Built-in Deduplication?

Both Solutions Offer the Benefits of Target side Deduplication:
- Greatly reduced storage capacity requirements
- Lower operational costs, energy usage and TCO (Total cost of ownership)
- Faster recoveries with more data on disk

Use ProtecTIER When:
- Highest performance and capacity scaling are required!
- Up to 500 MB/sec (1GB/s with 2 node) deduplication rates are needed
- Deduplicated capacities up to 25 PB are required
- You desire deduplication be done inline during data ingest
- A VTL appliance model is desired
- Deduplicating across multiple TSM (or other backup) servers

Use TSM 6 Built-in Deduplication When:
- Sufficient TSM server resources can be made available and you desire deduplication operations be completely integrated within TSM
- The benefits of deduplication are desired without separate hardware or software dependencies or licenses (ships with TSM Extended Edition)
- You desire end to end data lifecycle management with minimized data store

Complementary Solutions Today!
Can be used together but don’t deduplicate the same data twice