The material contained in this tutorial is copyrighted by the SNIA.

Member companies and individual members may use this material in presentations and literature under the following conditions:

- Any slide or slides used must be reproduced in their entirety without modification
- The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

This presentation is a project of the SNIA Education Committee.

Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.
This tutorial has been developed, reviewed and approved by members of the Data Management Forum (DMF)

- The DMF is an industry resource to those responsible for the accessibility and integrity of their organization’s information
- The DMF focuses on the technologies and trends related to Data Protection, ILM and Long-term digital information retention

<table>
<thead>
<tr>
<th>DMF Workgroups:</th>
<th>Information Lifecycle Management Initiative (ILMI)</th>
<th>Long-term Archive and Compliance Storage Initiative (LTACSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Protection Initiative (DPI)</td>
<td>Developing, educating and promoting ILM practices, implementation methods, and benefits</td>
<td>Addressing the challenges of retaining, securing, and preserving digital information for the long-term</td>
</tr>
<tr>
<td>Defining best practices for data protection and recovery technologies such as Backup, CDP, Data deduplication and VTL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Abstract

Data deduplication is a capacity optimization technology that is being used to dramatically improve storage efficiency. This technical session will:

- Review various data deduplication methodologies
- Identify the factors that influence space savings
- Provide scenarios where data deduplication is used
Agenda

- Overview
- How Data Deduplication Works
- Scenarios
- Q & A
Space Reduction Terminology

**Data Deduplication** is the replacement of multiple copies of data—at variable levels of granularity—with references to a shared copy in order to save storage space and/or bandwidth.

**Subfile Data Deduplication** is a form of data deduplication that operates at a finer granularity than an entire file or data object.

**Single Instance Storage** is a form of data deduplication that operates at a granularity of an entire file or data object.

**Compression** is the encoding of data to reduce its storage requirement - deduplicated data can also be compressed.
Space Reduction Ratio & Percent

Space Reduction Ratio of Bytes In vs Bytes Out

\[ \text{Ratio} = \frac{\text{Bytes In}}{\text{Bytes Out}} \]

Space Reduction Percent of Bytes In vs Bytes Out

\[ \% = 1 - \left( \frac{1}{1 - \text{Ratio}} \right) \]
### Space Reduction Ratio & Percent

<table>
<thead>
<tr>
<th>Space Reduction Ratio</th>
<th>Space Reduction Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:1</td>
<td>1/2 = 50%</td>
</tr>
<tr>
<td>5:1</td>
<td>4/5 = 80%</td>
</tr>
<tr>
<td>10:1</td>
<td>9/10 = 90%</td>
</tr>
<tr>
<td>20:1</td>
<td>19/20 = 95%</td>
</tr>
<tr>
<td>100:1</td>
<td>99/100 = 99%</td>
</tr>
<tr>
<td>500:1</td>
<td>499/500 = 99.8%</td>
</tr>
</tbody>
</table>

- Ratios can meaningfully be compared only under the same set of assumptions.
- Relatively low space reduction ratios provide significant space savings.
Data Deduplication – How it Works

- Evaluate Data
- Identify Redundancy
- Create or Update Reference Information
- Store and/or Transmit Unique Data Once
- Read and/or Reproduce Data
Data Deduplication Simplified

Dump #1
A B C D A E F F D B

= new unique data

= repeat data
Data Deduplication Simplified

- New unique data
- Repeat data
- Pointer to unique data segment
Data Deduplication Simplified

<table>
<thead>
<tr>
<th>Dump #1</th>
<th>Dump #2</th>
<th>Dump #3</th>
<th>Dump #4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D</td>
<td>A E F F</td>
<td>C D G C</td>
<td>A B C D</td>
</tr>
<tr>
<td></td>
<td>F D B</td>
<td>G H</td>
<td>I E F D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B G H</td>
</tr>
</tbody>
</table>

- Blue = new unique data
- Yellow = repeat data
- Arrow = pointer to unique data segment

Understanding Data Deduplication
© 2009 Storage Networking Industry Association. All Rights Reserved.
Reading Deduplicated Data

Application/CIFS/NFS/VTL Interface

Deduplication Engine

Reconstitution/Verification

Metadata References

Deduplicated Data
## Design Approach

<table>
<thead>
<tr>
<th>Component</th>
<th>Hardware (e.g., chip or card) integrated into a larger system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gateway</td>
<td>A dedicated data deduplication engine that must be combined with a storage system</td>
</tr>
<tr>
<td>Appliance</td>
<td>A dedicated deduplication engine integrated with a storage system</td>
</tr>
<tr>
<td>Storage System</td>
<td>A general purpose storage system with data deduplication capabilities</td>
</tr>
<tr>
<td>Grid Storage</td>
<td>A storage system that can scale independently without constraints to physical attributes</td>
</tr>
<tr>
<td>Software</td>
<td>Includes application agents, virtual appliances, or storage software</td>
</tr>
</tbody>
</table>
Multiple deployment examples are illustrated
Specific deployments selected based on customer situation
Source or Target

Source Deduplication
- Identifies duplicate data at the client
- Transfers unique segments to a central repository
- Separate client and server components

Target Deduplication
- Identifies duplicate data where the data is being stored
- Stores unique segments
- Standalone system

Considerations
- Neither approach enables a greater or lesser space savings
- Scope of data deduplication may vary by implementation
 Inline or Post-Process

❯ Inline Deduplication
  ➢ Data deduplication performed before writing the deduplicated data

❯ Post-Process Deduplication
  ➢ Data deduplication performed after the data to be deduplicated has been initially stored

❯ Considerations
  ➢ A product may implement both methods
  ➢ A product may provide methods to control when particular data is deduplicated
  ➢ May impact replication, usable capacity, scalability, etc.
Fixed or Variable Size Segment

- **Fixed Length Segment Deduplication**
  - Evaluation of data includes a fixed reference window used to look at segments of data during deduplication process
  - Provides fixed granularity, e.g. 4KB, or 8KB, or 128KB

- **Variable length Segment Deduplication**
  - Evaluation of data uses a variable length window to find duplicate data in stream or volume of data processed
  - Provides variable granularity, e.g. Average 4KB or 32KB

- **Method Chosen May Affect Deduplication Results**
  - Effects observed will vary by method
  - Segmentation may not apply to all deduplication
Data Deduplication Benefits

Data Deduplication can help organizations:

- Satisfy ROI/TCO requirements
- Manage data growth
- Increase efficiency of storage and backup
- Reduce overall cost of storage
- Reduce network bandwidth
- Reduce operational costs including:
  - Infrastructure costs required space, power and cooling
  - Movement toward a greener data center
- Reduce administrative costs
Data Deduplication Scope

- Multiple Repositories Per Controller
- Single Repository Per Controller
- Single Repository Shared by Multiple Controllers

• System capacity varies independently from the scope
Applications for Deduplication

- **Backup and Recovery**
  - Backup to disk efficiently with long retention – recoverability
  - Replication for offsite data movement

- **Archive Repository**
  - Long-term retention and preservation

- **Primary Storage**
  - Lower physical capacity required for storage of active data
Backup: What to Consider

- Factors that will Impact your Results:
  - Different applications or data types
  - Bandwidth and latency
  - Policies and methodologies
  - Data protection overhead
  - Compression and encryption

- Deduplication Scope

- Deduplicated Data Resiliency

- Scalability
  - Capacity
  - Performance
## Backup: Factors Impacting Space Savings

<table>
<thead>
<tr>
<th>Factors associated with higher data deduplication ratios</th>
<th>Factors associated with lower data deduplication ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data created by users</td>
<td>Data captured from mother nature</td>
</tr>
<tr>
<td>Low change rates</td>
<td>High change rates</td>
</tr>
<tr>
<td>Reference data and inactive data</td>
<td>Active data, encrypted data, compressed data</td>
</tr>
<tr>
<td>Applications with lower data transfer rates</td>
<td>Applications with higher data transfer rates</td>
</tr>
<tr>
<td>Use of full backups</td>
<td>Use of incremental backups</td>
</tr>
<tr>
<td>Longer retention of deduplicated data</td>
<td>Shorter retention of deduplicated data</td>
</tr>
<tr>
<td>Wider scope of data deduplication</td>
<td>Narrower scope of data deduplication</td>
</tr>
<tr>
<td>Continuous business process improvement</td>
<td>Business as usual operational procedures</td>
</tr>
<tr>
<td>Smaller segment size</td>
<td>Larger segment size</td>
</tr>
<tr>
<td>Variable-length segment size</td>
<td>Fixed-length segment size</td>
</tr>
<tr>
<td>Format awareness</td>
<td>No format awareness</td>
</tr>
<tr>
<td>Temporal data deduplication</td>
<td>Spatial data deduplication</td>
</tr>
</tbody>
</table>

[www.snia.org/forums/dmf/knowledge](http://www.snia.org/forums/dmf/knowledge)
# Backup: Influence of Backup Methodology

<table>
<thead>
<tr>
<th>Method</th>
<th>Day 1 (Bytes in)</th>
<th>Day 2 (Bytes in)</th>
<th>Day 3 (Bytes in)</th>
<th>Day 4 (Bytes in)</th>
<th>Day 5 (Bytes in)</th>
<th>Day 6 (Bytes in)</th>
<th>Day 7 (Bytes in)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Changed files</td>
<td>40 GB</td>
<td>1 GB</td>
<td>2 GB</td>
<td>3 GB</td>
<td>1 GB</td>
<td>2 GB</td>
<td>3 GB</td>
</tr>
<tr>
<td>Bytes out</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily Full</td>
<td>100 GB</td>
<td>100 GB</td>
<td>100 GB</td>
<td>100 GB</td>
<td>100 GB</td>
<td>100 GB</td>
<td>100 GB</td>
</tr>
<tr>
<td>Daily Differential</td>
<td>100 GB</td>
<td>10 GB</td>
<td>20 GB</td>
<td>30 GB</td>
<td>40 GB</td>
<td>40 GB</td>
<td>50 GB</td>
</tr>
<tr>
<td>Weekly Full + Daily Differential</td>
<td>100 GB</td>
<td>10 GB</td>
<td>20 GB</td>
<td>30 GB</td>
<td>40 GB</td>
<td>40 GB</td>
<td>50 GB</td>
</tr>
<tr>
<td>Weekly Full + Daily Incremental</td>
<td>100 GB</td>
<td>10 GB</td>
<td>10 GB</td>
<td>10 GB</td>
<td>10 GB</td>
<td>20 GB</td>
<td>20 GB</td>
</tr>
</tbody>
</table>

Backup: Influence of Backup Methodology

Understanding Data Deduplication
© 2009 Storage Networking Industry Association. All Rights Reserved.
Backup: Deduplication for Data Movement

Disaster Recovery
- Replicate all Data after Deduplication for Bandwidth Efficiency
- Meet Offsite Requirements without Physical Transport

Bandwidth Optimization
- Increasing WAN Efficiency
  - Transfer more information per pipe
- Support Remote Office Protection
- Enable Backup Centralization
- Consolidate Physical Tape Creation
Create Long-term Removable Media Storage for Compliance and Archive

Different Data Path Approaches

- (#1) Path through backup server
- (#2) Path direct from deduplication system to removable media storage
Deduplication within Secondary Storage Use Cases
Deduplication for Backup and Recovery

- Deduplicating Agents
- Backup Clients
- Primary systems
- Storage System
- Backup Server
- Deduplicating Backup Server
- B2D
- Appliance

Understanding Data Deduplication
© 2009 Storage Networking Industry Association. All Rights Reserved.
Backup Remote Office
Source Deduplication

Large Remote Site
- Agents
- Primary systems
- Appliance

Small Remote Site
- Agents Only
- Primary systems

Data Center
- Agents
- Primary systems
- Appliance

Remote Recovery Site
- Appliance
- Tape Vault
Deduplicated Archive Repositories

- Policy-based software moves inactive data on primary storage to deduplicated storage archive
- Local archive may be replicated offsite and copied to tape for long term retention
Deduplication of Primary Storage Use Cases
Primary Storage with Deduplication

- Bring the benefits of deduplication to primary storage
- Supports networked storage (SAN & LAN)
- Storage efficiency supports replicating more data
- Space/performance tradeoff
Virtual Machine Storage

- Balance the tradeoff between savings and performance impact

- Examples of Active Data
  - Unstructured data
  - Structured data
  - Virtual Machines

VMs After Deduplication

Duplicate Data Is Eliminated

VM Clone Copies

APP
DATA
OS

APP
DATA
OS

APP
DATA
OS

APP
DATA
OS

APP
DATA
OS
Please send any questions or comments on this presentation to SNIA: trackdatamgmt@snia.org

Many thanks to the following individuals for their contributions to this tutorial.
- SNIA Education Committee

Data Deduplication and Space Reduction (DDSR) Special Interest Group
- Matthew Brisse
- Daniel Budiansky
- Mike Dutch
- Michael Fishman
- Larry Freeman
- Devin Hamilton
- Jason Lehl
- Shane Jackson
- Gene Nagle
- Thomas Rivera
- Tom Sas
- Gideon Senderov

It’s easy to get involved with the DMF!
- Find a passion
- Join a committee
- Gain knowledge & influence
- Make a difference

www.snia.org/forums/dmf
The specific backup application determines where deduplication is performed.
Deduplicated Primary Storage

- Space/Performance tradeoff
- File Services
  - Home directories
  - Tech Pubs
  - Email
  - Fixed content
- Replicated databases
- Test and application development
- Source code version control system
- Virtualized environments
Storage Tiering with Deduplication

Tier 1: Primary Storage

Tier 2: Deduplicated Storage

Application Servers
File Tiering with Deduplication

NAS Client

Files and Stubs

Primary Storage

Shared Files

Secondary Storage

NFS/CIFS

Migrate

A

B

C

D

E

A

B

B

C

D

A

E

D

E
Backup Remote Office Target Deduplication

Remote Site
- Primary systems
- Appliance

Target Deduplication

Remote Recovery Site
- Primary systems
- Appliance ➔ Tape Vault

Data Center
- Primary systems
- Appliance

WAN

Optimized Replication

Encryption

Understanding Data Deduplication
© 2009 Storage Networking Industry Association. All Rights Reserved.
Deduplication within the Backup Cloud

Application and Data on Customer Premises

Deduplicated Data

Deduplication performed by the backup client

Application and Data on Customer Premises

Backup as Service within the Cloud

Data

Deduplication Target within the Cloud