UNDERSTANDING DATA DEDUPLICATION

Tom Sas – Hewlett-Packard
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 Protection and Capacity Optimization (DPCO) committee which any SNIA member can join for free.

The mission of the DPCO is to foster the growth and success of the market for data protection and capacity optimization technologies.

2010 goals include educating the vendor and user communities, market outreach, and advocacy and support of any technical work associated with data protection and capacity optimization.
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
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 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.

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

Subfile Data Deduplication is a form of data deduplication that operates at a finer granularity than 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

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

% = 1 - \left( \frac{1}{1 - \%} \right)

% = \frac{\text{Bytes In} - \text{Bytes Out}}{\text{Bytes In}}
### Space Reduction Ratio & Percent

<table>
<thead>
<tr>
<th>Space Reduction Ratio (In:Out)</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 still provide significant space savings.
Understanding Data Deduplication

Primary storage has less duplicate data
Periodic archives have moderate duplicate data
Repeated backups have significant duplicate data

Capacity Savings over Time

Deduplication Ratio Typically Depends on Use Case and Time

Logical

Dedupe Primary

Dedupe Archive

Dedupe Backup
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 Deletion/Space Reclamation
Data Deduplication Simplified

Dump #1

= new unique data

= repeat data
Data Deduplication Simplified

Dump #1

= 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</td>
<td>B</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>D</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>E</td>
<td>C</td>
</tr>
<tr>
<td>D</td>
<td>E</td>
<td>F</td>
<td>D</td>
</tr>
<tr>
<td>A</td>
<td>E</td>
<td>F</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>F</td>
<td>D</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>D</td>
<td>H</td>
<td>D</td>
<td>H</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>D</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>E</td>
<td>C</td>
</tr>
<tr>
<td>D</td>
<td>E</td>
<td>F</td>
<td>D</td>
</tr>
<tr>
<td>A</td>
<td>E</td>
<td>F</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>F</td>
<td>D</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>D</td>
<td>H</td>
<td>D</td>
<td>H</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>D</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>E</td>
<td>C</td>
</tr>
<tr>
<td>D</td>
<td>E</td>
<td>F</td>
<td>D</td>
</tr>
<tr>
<td>A</td>
<td>E</td>
<td>F</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>F</td>
<td>D</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>D</td>
<td>H</td>
<td>D</td>
<td>H</td>
</tr>
</tbody>
</table>

- **Blue** = new unique data
- **Yellow** = repeat data
- **Blue Arrow** = pointer to unique data segment
Reading Deduplicated Data

Application/CIFS/NFS/VTL Interface

Deduplication Engine

Reconstitution/Verification

Metadata References

Deduplicated Data

Read Request

Read Fulfilled
Design Approach

- 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
- Reduces network traffic during backup and replication

Target Deduplication
- Identifies duplicate data where the data is being stored
- Stores unique segments
- Standalone system
- Reduces network traffic during subsequent replication
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. ranges 4KB-12KB, 32KB-96KB

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

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

- System capacity varies independently from the scope
Applications for Deduplication

- **Primary Storage**
  - Lower physical capacity required for storage of active data

- **Data Protection**
  - Backup to disk efficiently with longer retention and recoverability
  - Replication for offsite data movement, disaster recovery and business continuity

- **Archive Repository**
  - Long-term retention and preservation
Backup: What to Consider

❖ Capacity Optimization
  ❖ Data type
  ❖ Operational Policies
  ❖ Design and Implementation

❖ Performance Optimization (bandwidth and latency)
  ❖ Technology

❖ Resiliency
  ❖ Clustering
  ❖ Grid
  ❖ Replication
## 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>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: Capacity Savings Over Time

Deduplication Ratio Typically Improves Over Time

Storage vs. Time

- Logical
- Dedupe
- savings
- storage
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

Check out SNIA Tutorial: Deduplication’s Role in Disaster Recovery
Backup: Removable Media 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
  - Primary systems
    - B2D
    - Appliance

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

- Backup Clients
  - Primary systems
  - Backup Server
    - B2D
    - Deduplicating VTL or NAS
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

Application Servers ➔ Policy-based software ➔ Local Archive ➔ Remote Site

Primary Storage → Deduplicated Local Archive

Deduplicated Offsite Archive ➔ Tape Vault
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
Q&A / Feedback

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

Matthew Brisse
Don Deel
Mike Dutch
Larry Freeman
David Hill
Judy Leach

Gene Nagle
Richard Reitmeyer
Thomas Rivera
Tom Sas
Gideon Senderov

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

www.snia.org/dpc0