



Education

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

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



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

$$\% = \frac{\text{Bytes In} - \text{Bytes Out}}{\text{Bytes In}}$$

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

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

Space Reduction Ratio & Percent

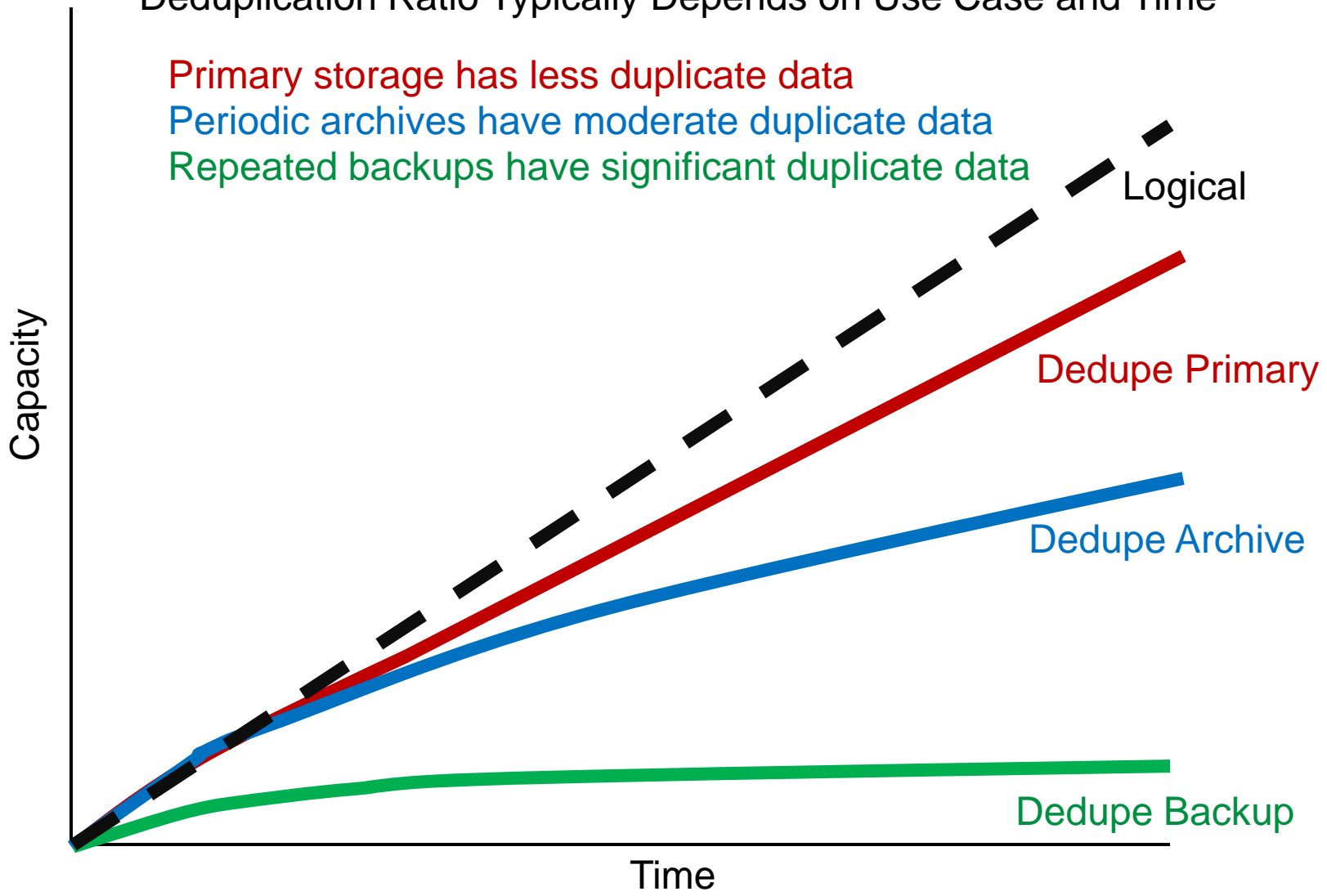
Space Reduction Ratio (In:Out)	Space Reduction Percent
2:1	$1/2 = 50\%$
5:1	$4/5 = 80\%$
10:1	$9/10 = 90\%$
20:1	$19/20 = 95\%$
100:1	$99/100 = 99\%$
500:1	$499/500 = 99.8\%$

- Ratios can meaningfully be compared only under the same set of assumptions
- Relatively low space reduction ratios still provide significant space savings

Capacity Savings over Time

Deduplication Ratio Typically Depends on Use Case and Time

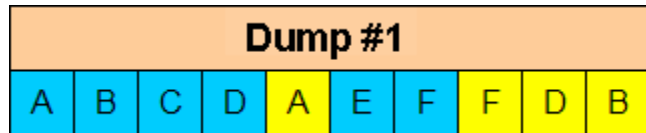
- Primary storage has less duplicate data
- Periodic archives have moderate duplicate data
- Repeated backups have significant duplicate data





- 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

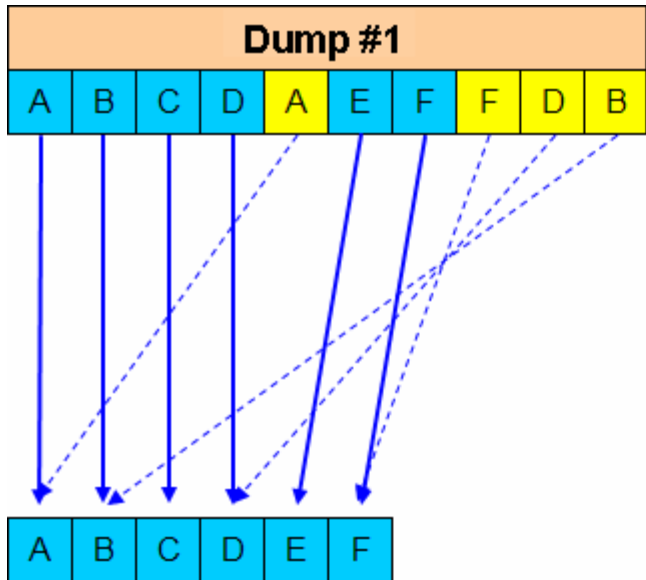





Data Deduplication Simplified



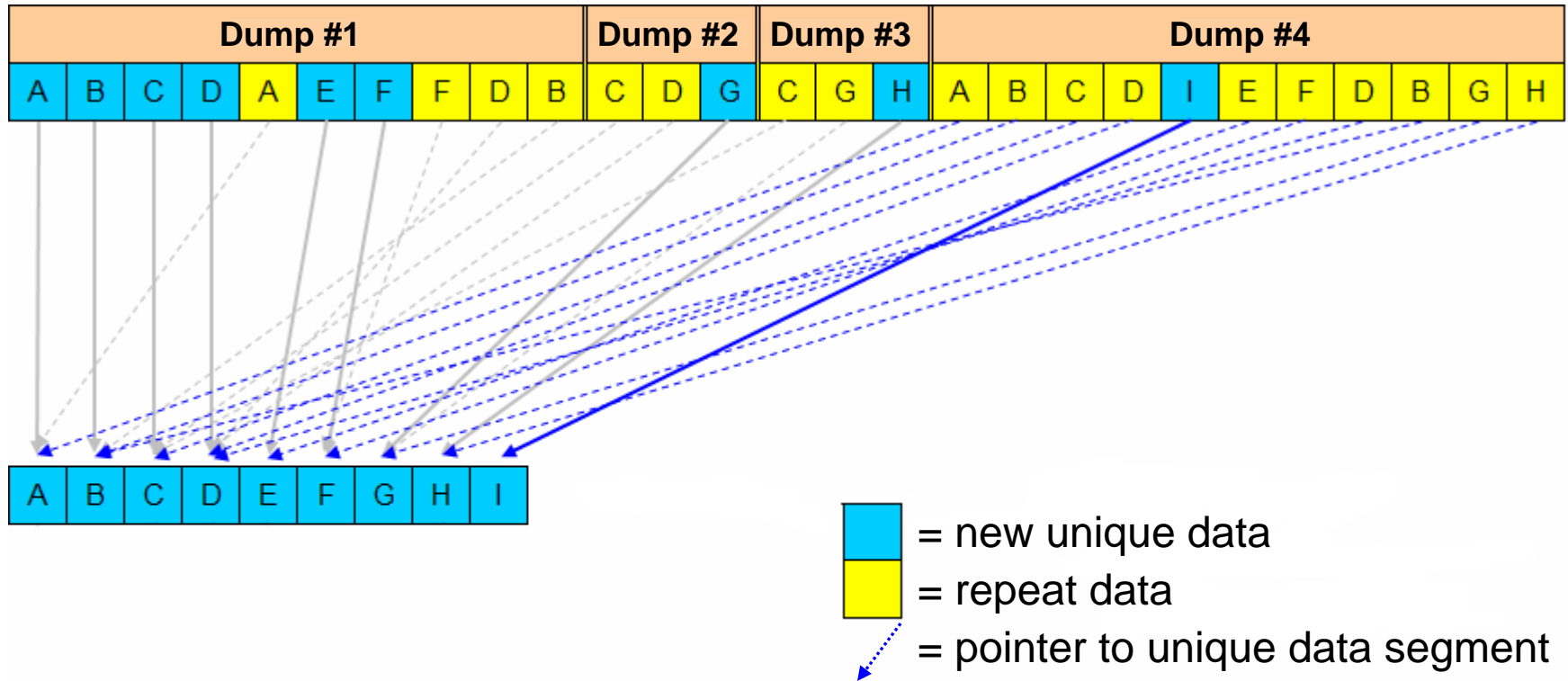
 = new unique data
 = repeat data

Data Deduplication Simplified

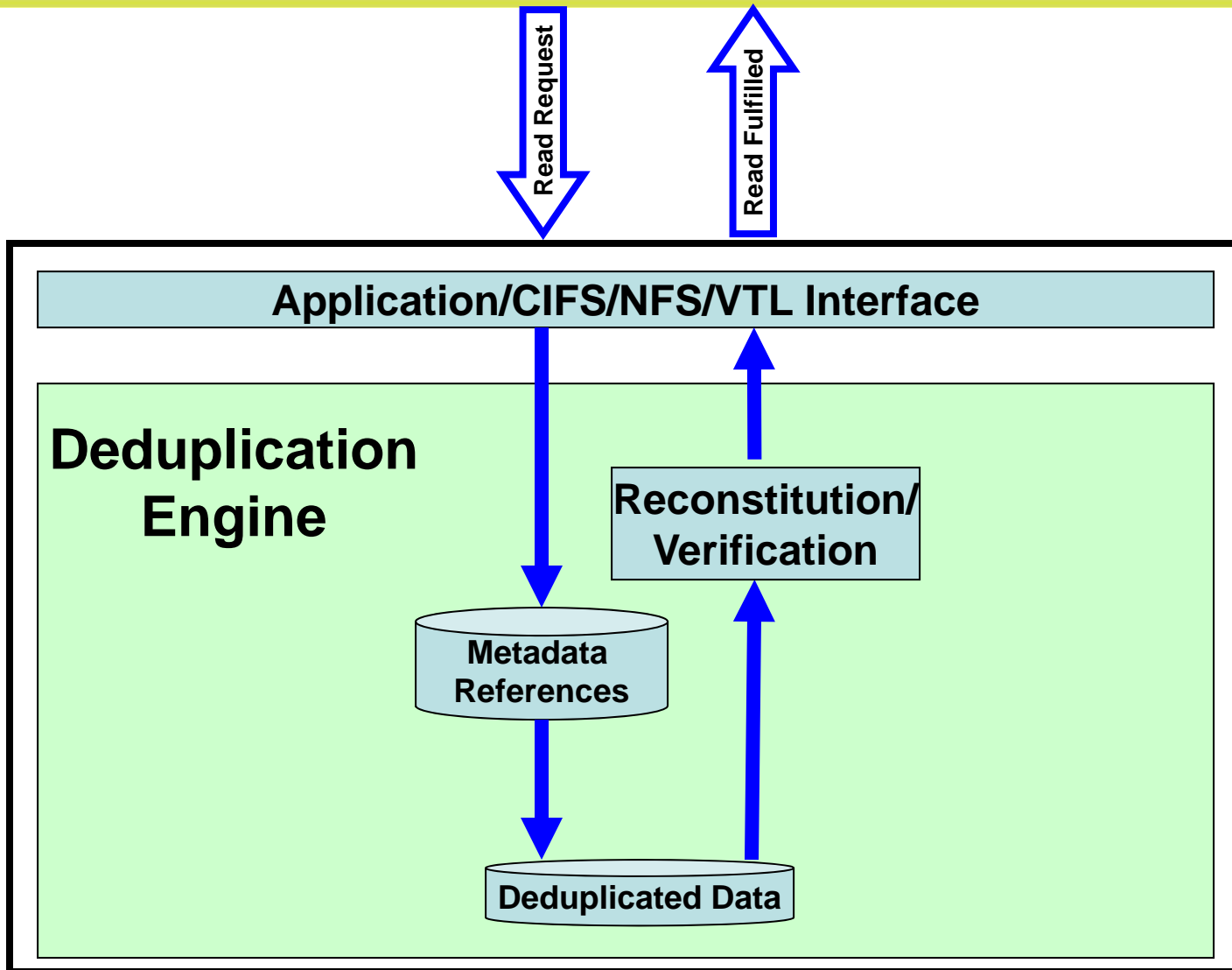


-  = new unique data
-  = repeat data
-  = pointer to unique data segment

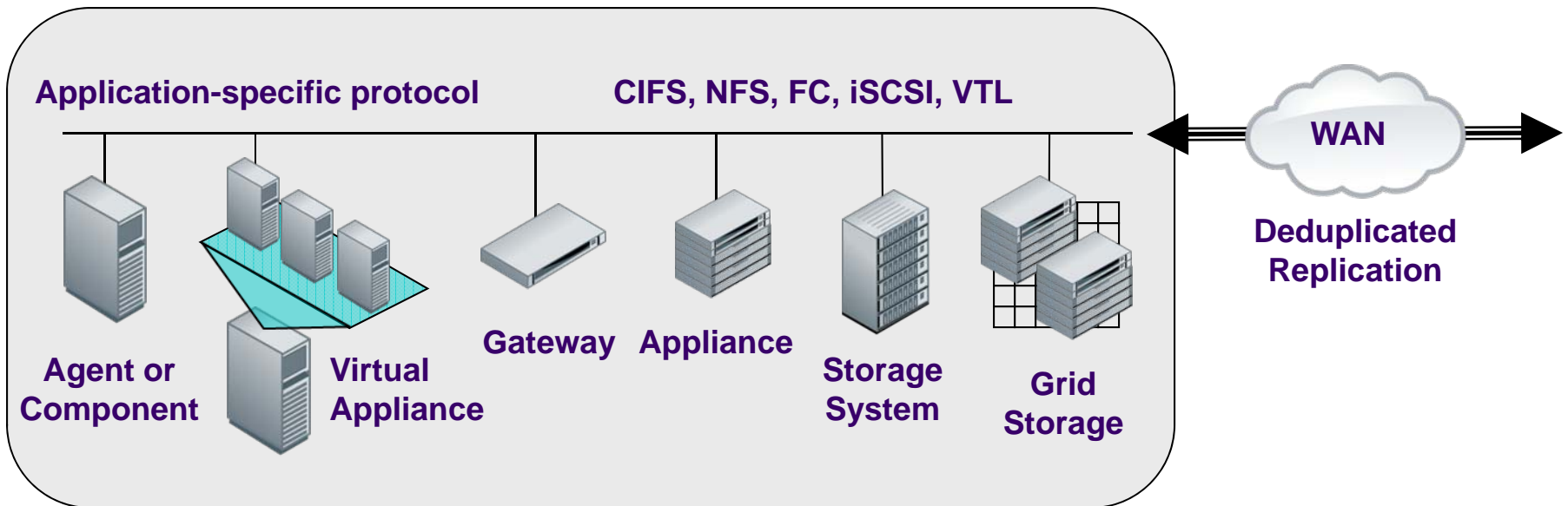
Data Deduplication Simplified



Reading Deduplicated Data



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

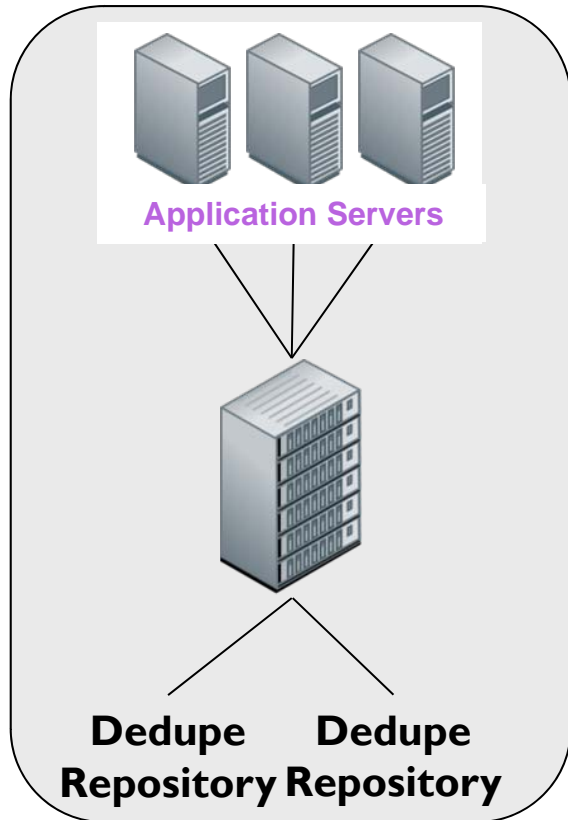
- ◆ Evaluation of data uses a variable length window to find duplicate data in stream or volume of data processed
- ◆ Provides variable granularity, e.g. ranges 4KB-12KB, 32KB-96KB

➤ Method Chosen May Affect Deduplication Results

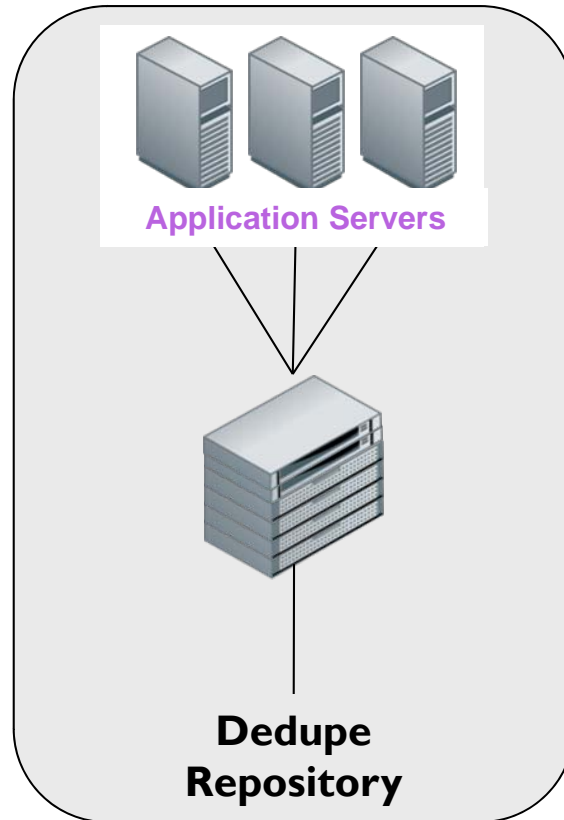
- ◆ Effects observed will vary by method
- ◆ Segmentation may not apply to all deduplication

Data Deduplication Scope

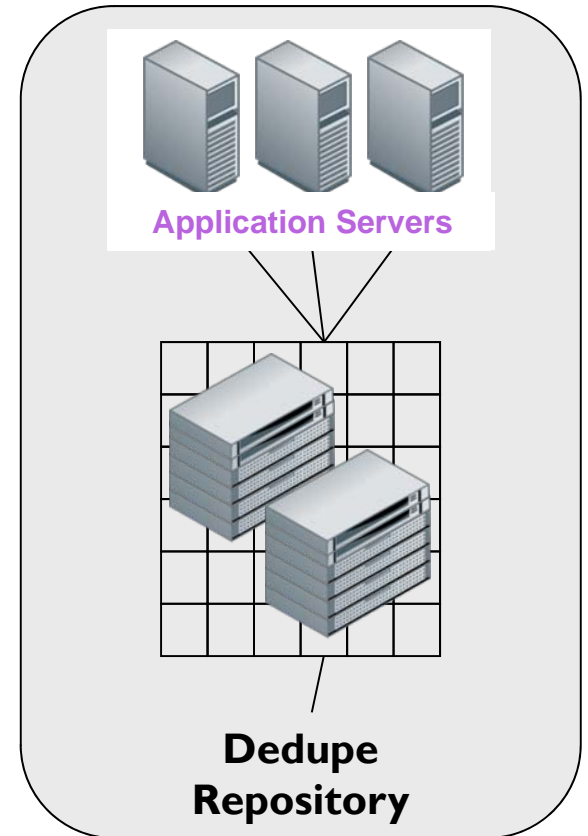
**Multiple Repositories
Per Storage Controller**



**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

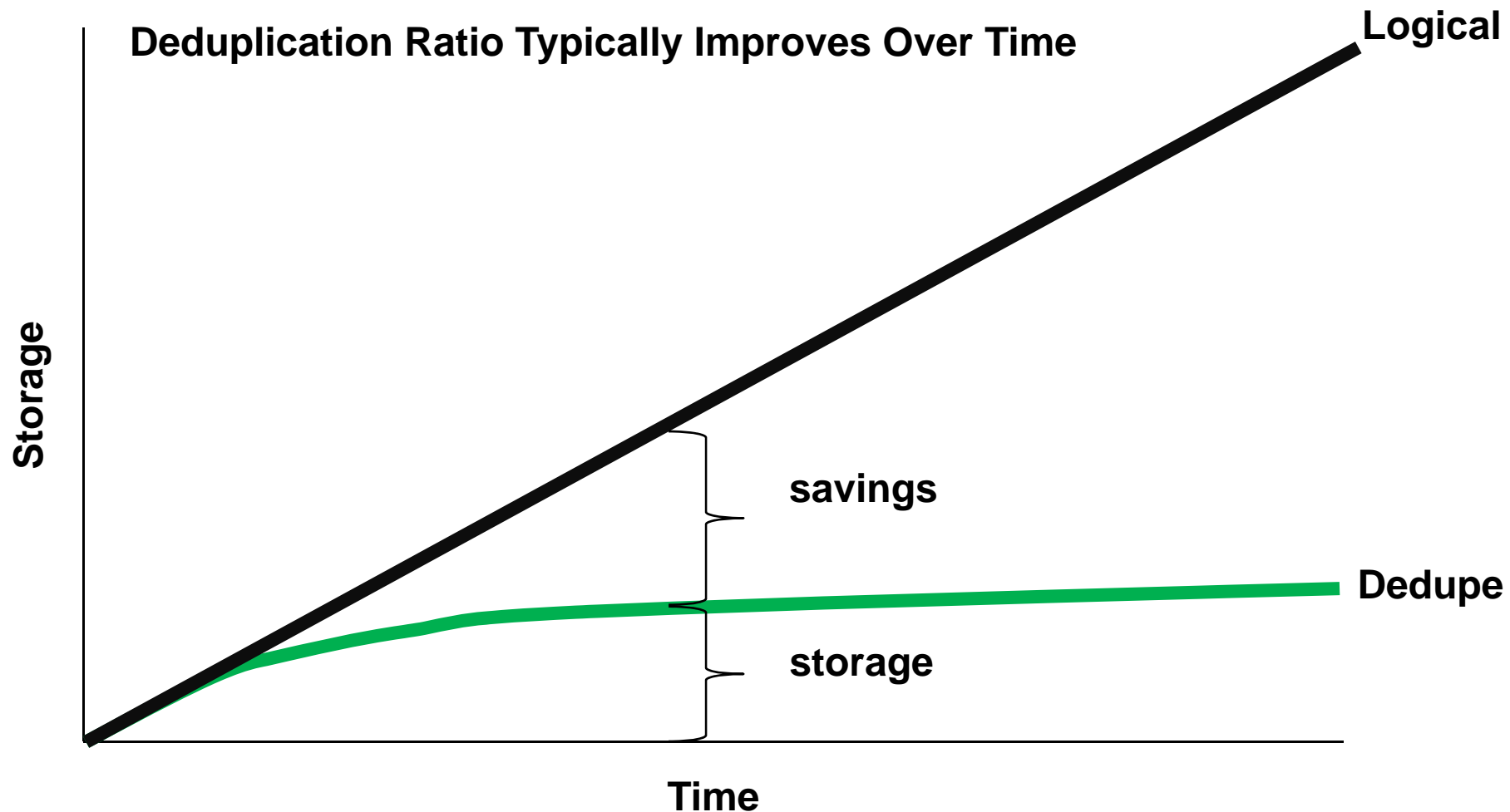
Factors associated with higher data deduplication ratios	Factors associated with lower data deduplication ratios
Data created by users	Data captured from mother nature
Low change rates	High change rates
Reference data and inactive data	Active data, encrypted data, compressed data
Use of full backups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

www.snia.org/forums/dmf/knowledge



Understanding Data Deduplication Ratios

Backup: Capacity Savings Over Time



➤ Disaster Recovery

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

➤ Bandwidth Optimization

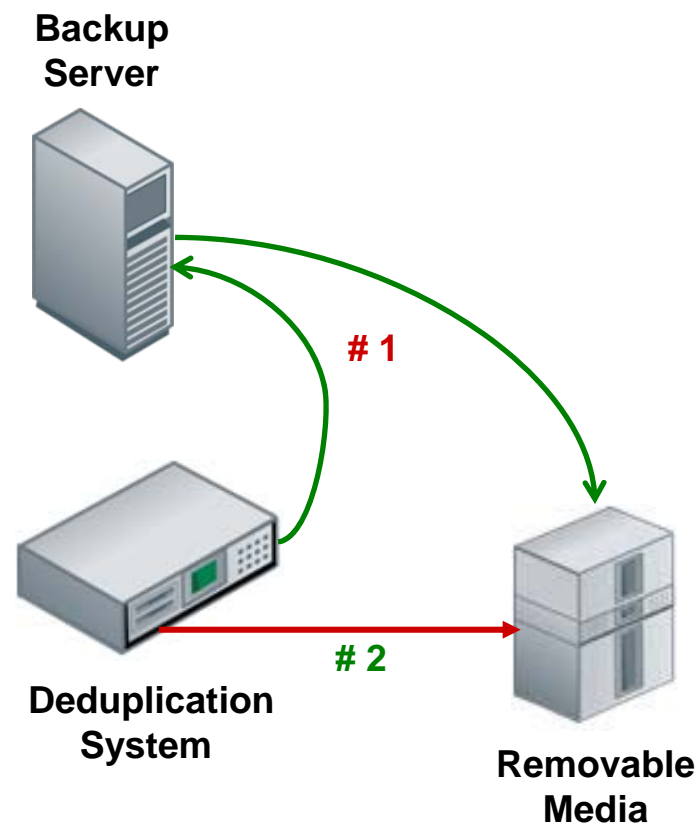
- ◆ Increasing WAN Efficiency
 - › Transfer more information per pipe
- ◆ Support Remote Office Protection
- ◆ Enable Backup Centralization
- ◆ Consolidate Physical Tape Creation



**Check out SNIA Tutorial:
Deduplication's Role in
Disaster Recovery**

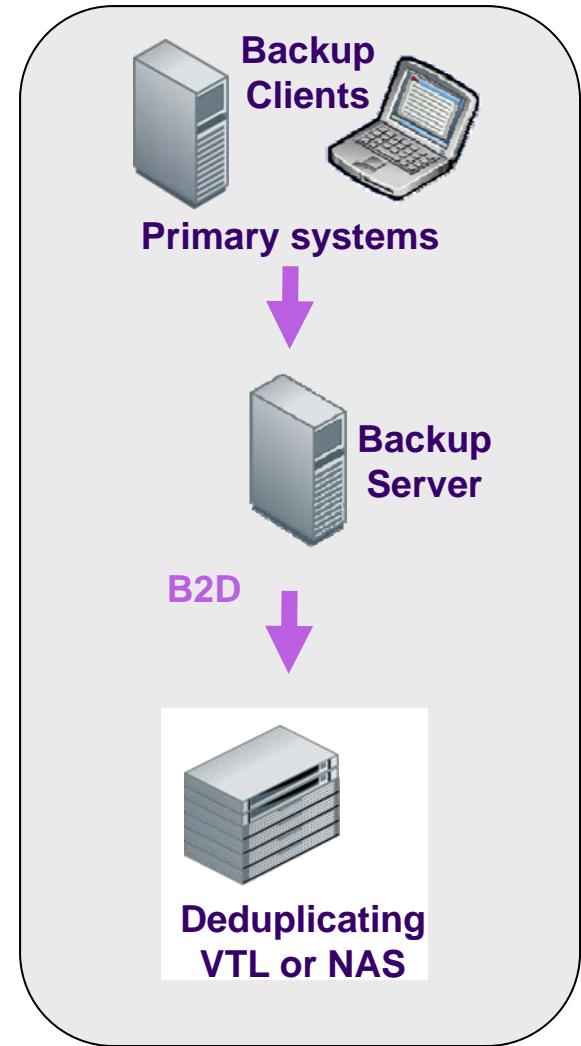
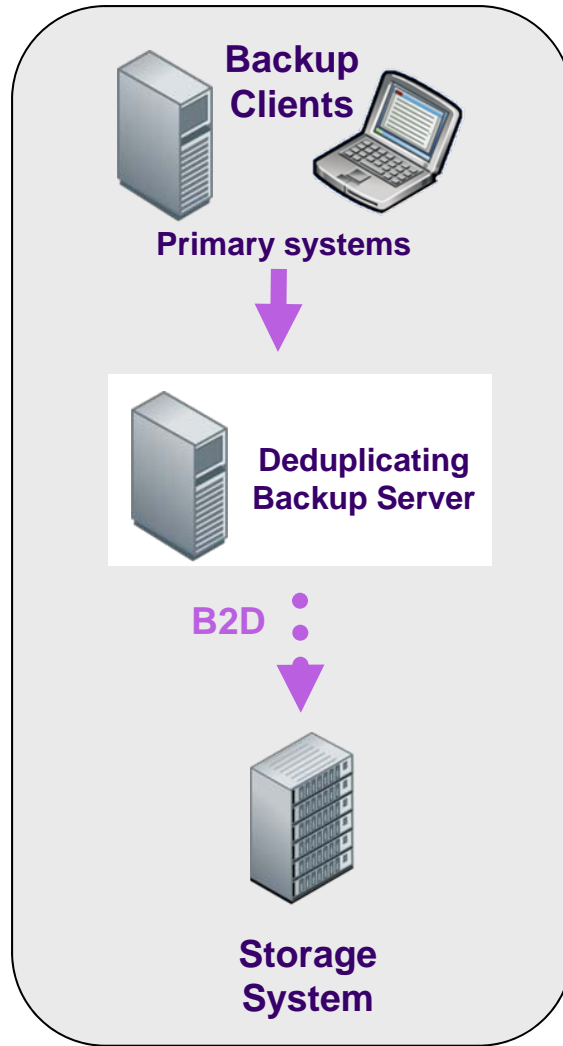
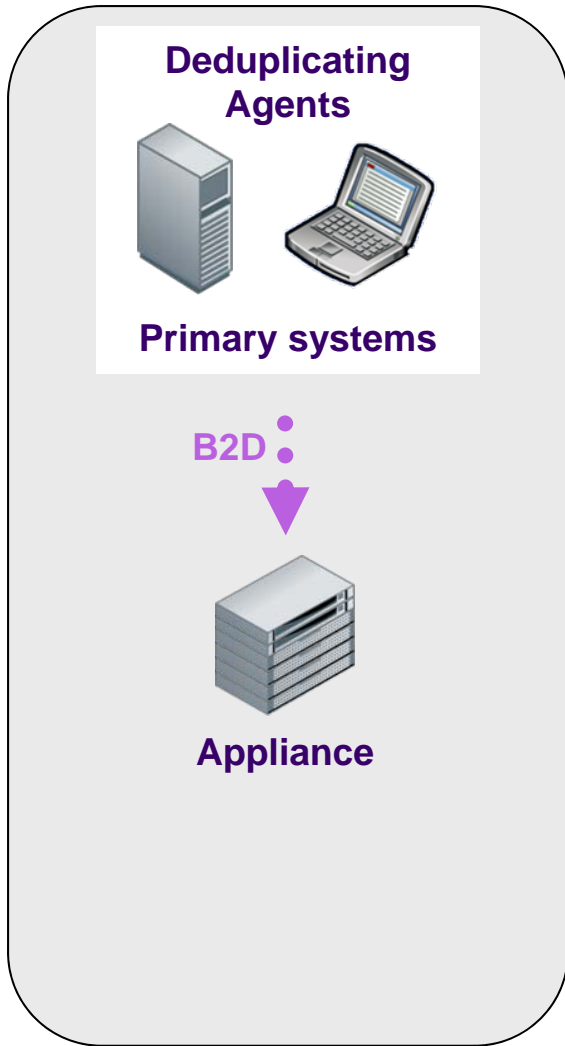
Backup: Removable Media Creation

- 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

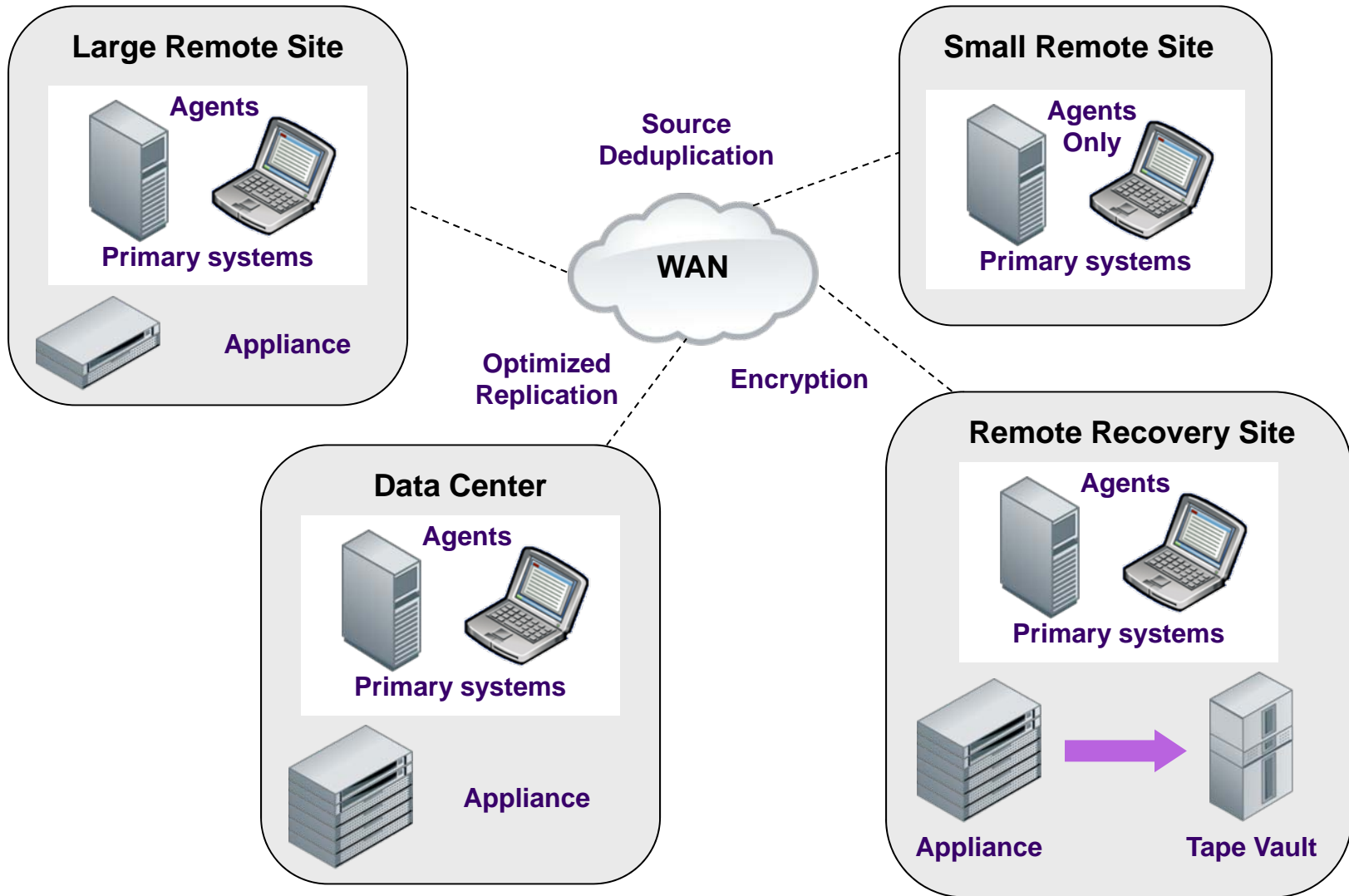


Deduplication within Secondary Storage Use Cases

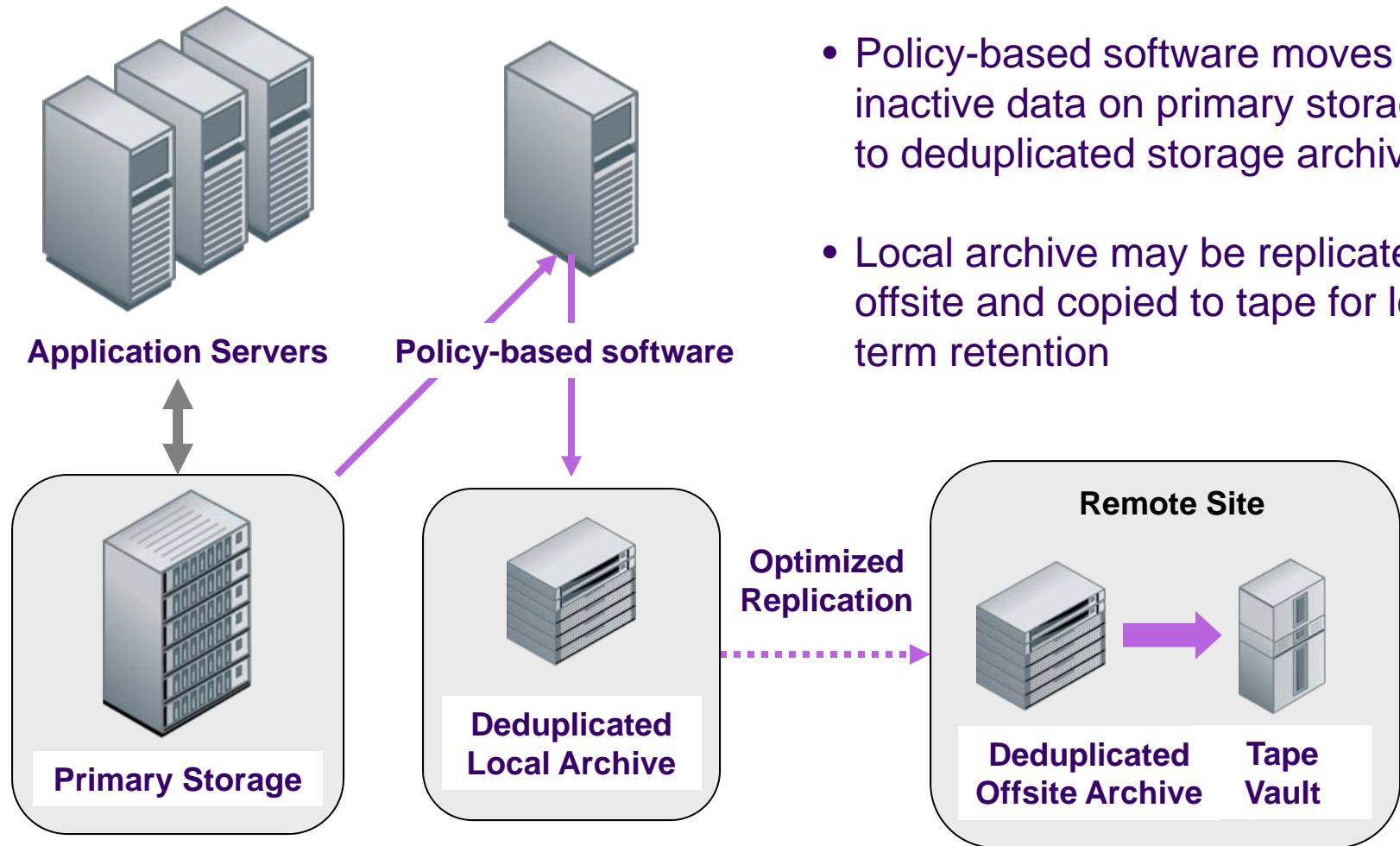
Deduplication for Backup and Recovery



Backup Remote Office Source Deduplication



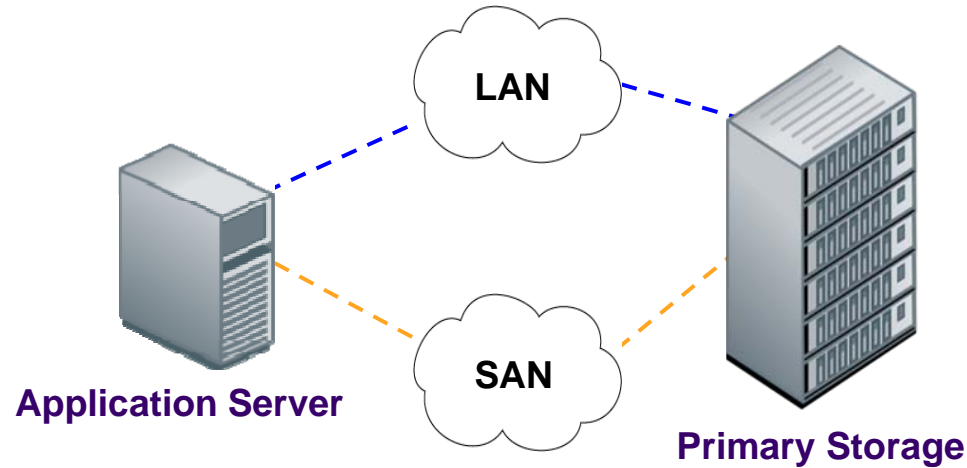
Deduplicated Archive Repositories



- 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

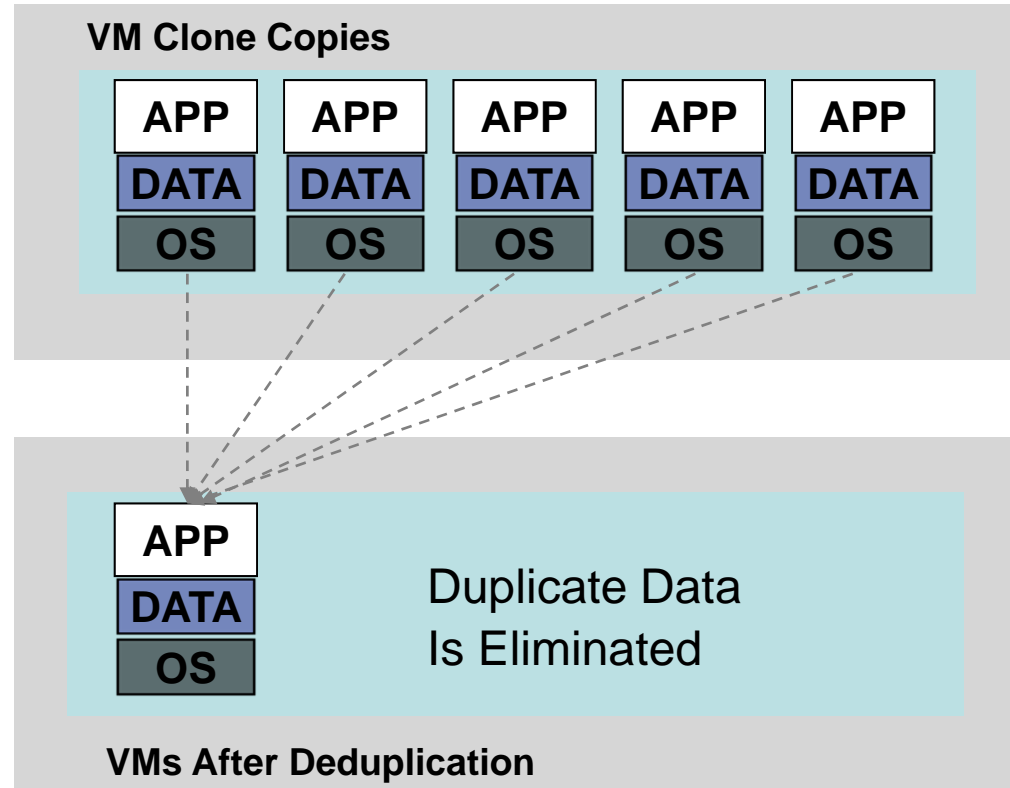
Deduplication of Primary Storage Use Cases

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

- Balance the tradeoff between savings and performance impact
- Examples of Active Data
 - ◆ Unstructured data
 - ◆ Structured data
 - ◆ Virtual Machines



- 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/dpco