SNIA

Education

UNDERSTANDING DATA DEDUPLICATION

Thomas Rivera – SEPATON

SNIA Legal Notice

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.

About the SNIA DMF

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

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

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 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

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

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

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

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

Space Reduction Ratio & Percent

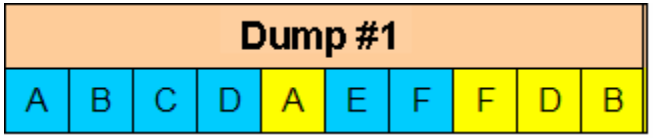
Space Reduction Ratio	Space Reduction Percent
2:1	1/2 = 50%
5:1	4/5 = 80%
10:1	9/10 = 90%
20:1	19/20 = 95%
100:1	99/100 = 99%
500:1	499/500 = 99.8%

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

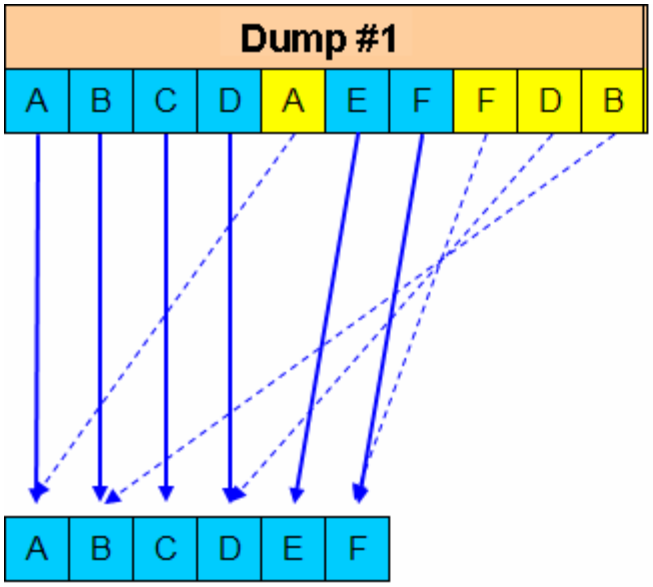
Data Deduplication Simplified



= new unique data

= repeat data

Data Deduplication Simplified

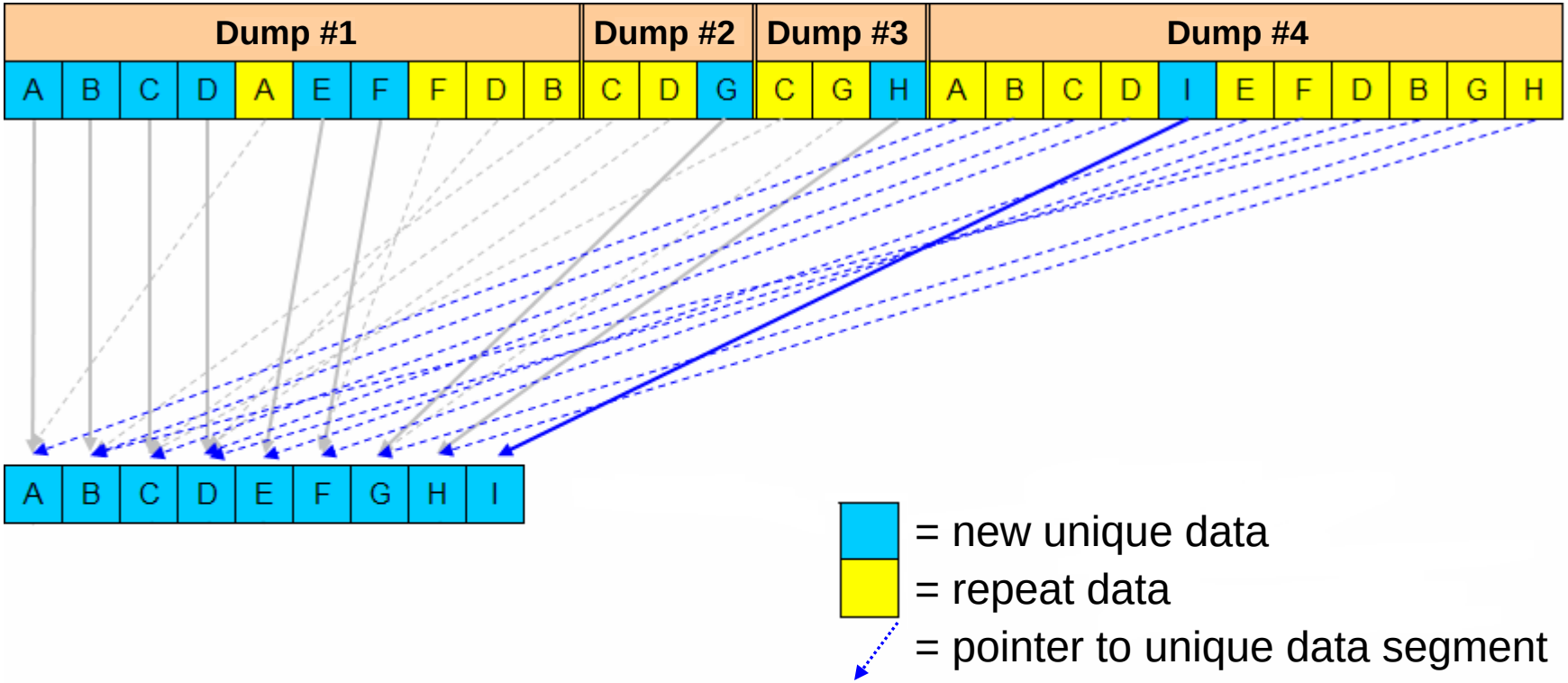


= new unique data

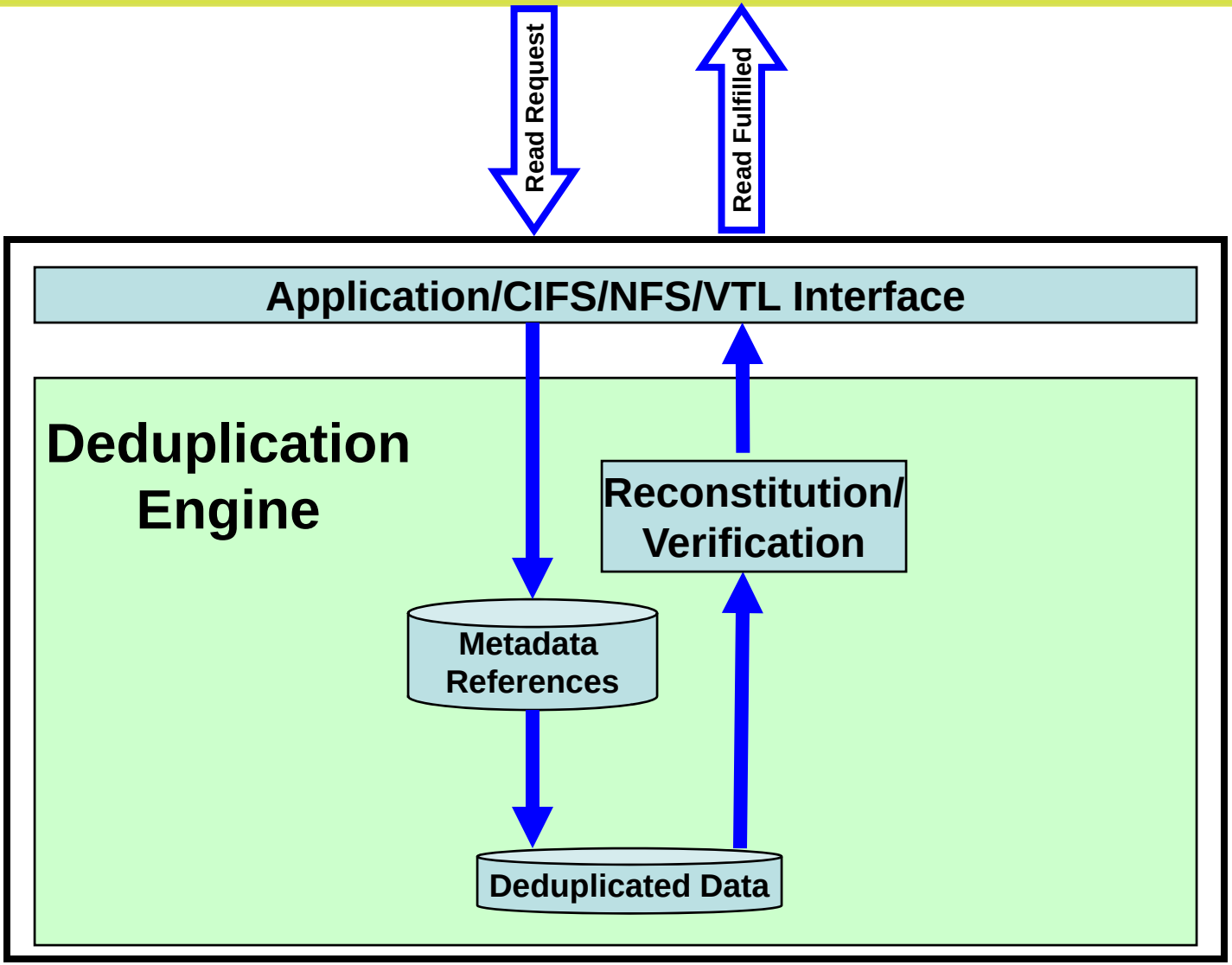
= repeat data

= pointer to unique data segment

Data Deduplication Simplified



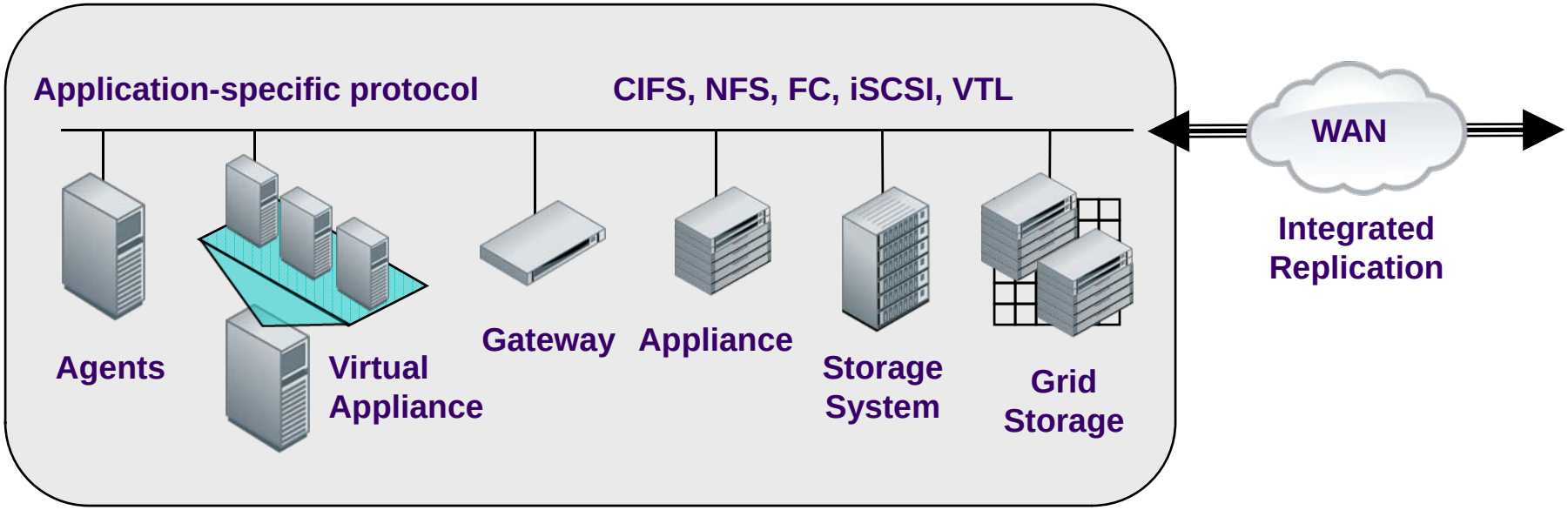
Reading Deduplicated Data



Design Approach

- Component
  - Hardware (e.g., chip or card) integrated into a larger system
- Gateway
  - A dedicated data deduplication engine that must be combined with a storage system
- Appliance
  - A dedicated deduplication engine integrated with a storage system
- Storage System
  - A general purpose storage system with data deduplication capabilities
- Grid Storage
  - A storage system that can scale independently without constraints to physical attributes
- Software
  - Includes application agents, virtual appliances, or storage software

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
- 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.

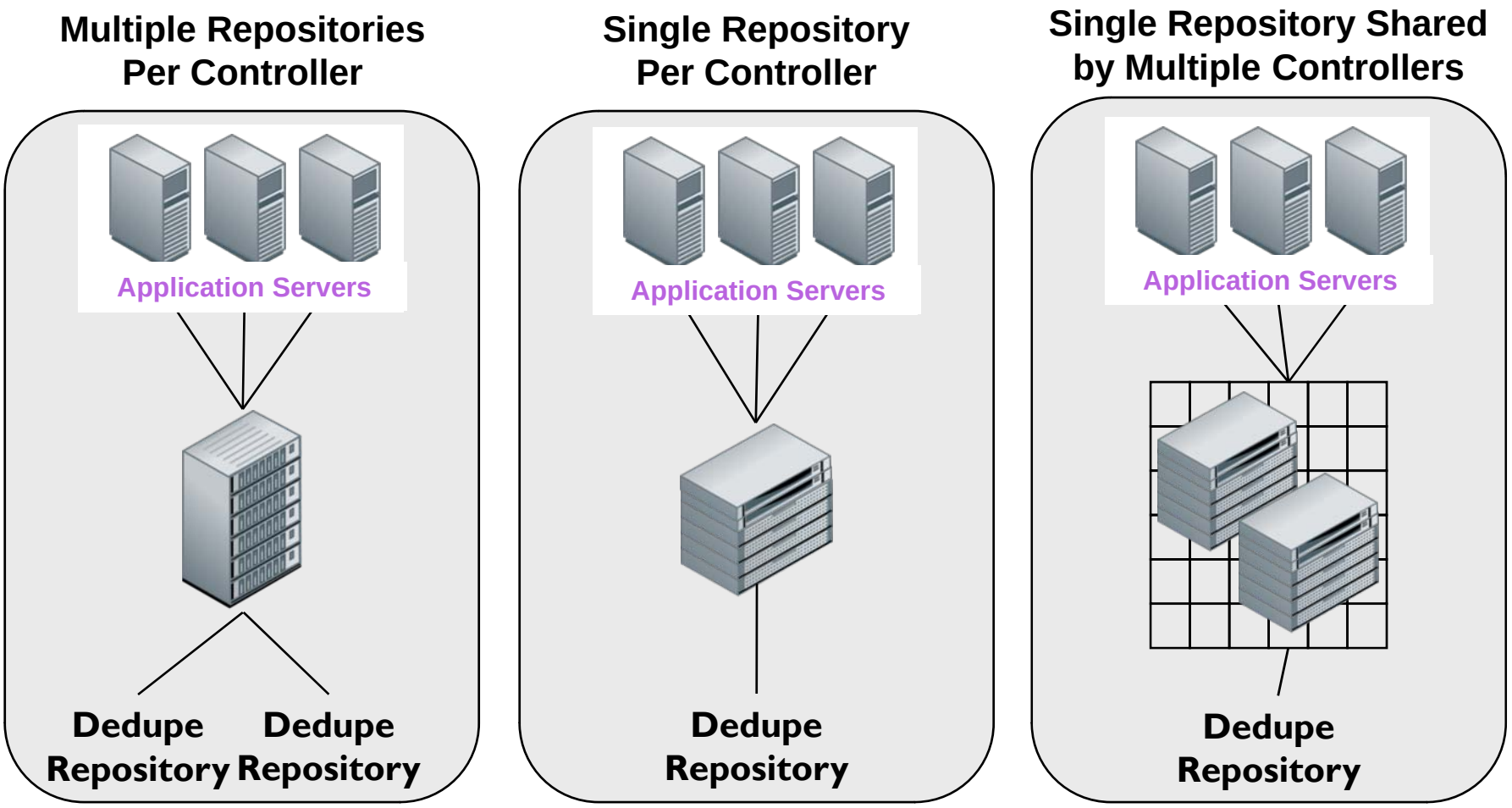
Fixed or Variable Size Segment

- 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



- 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

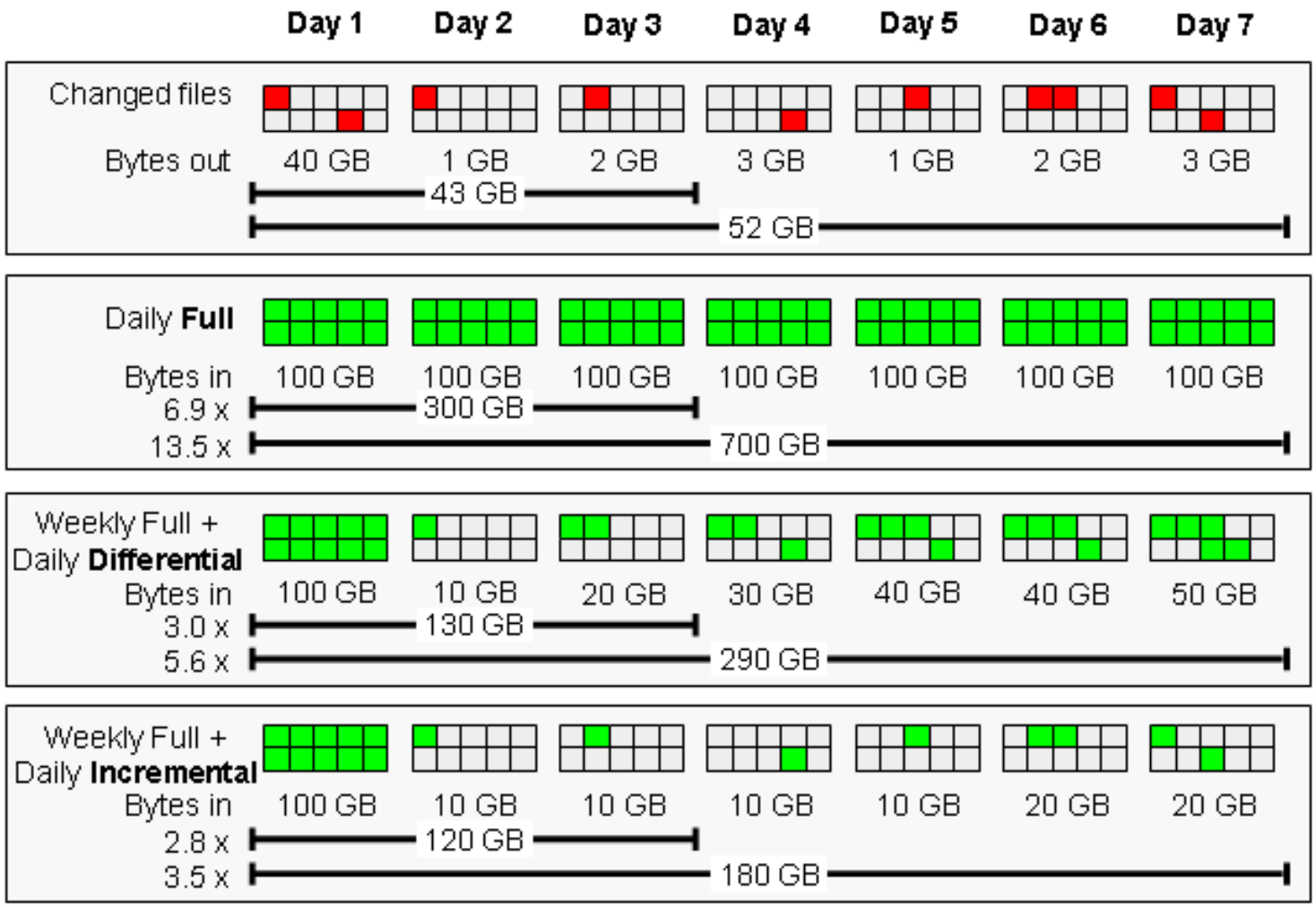
Backup:
Factors Impacting Space Savings

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
Applications with lower data transfer rates	Applications with higher data transfer rates
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

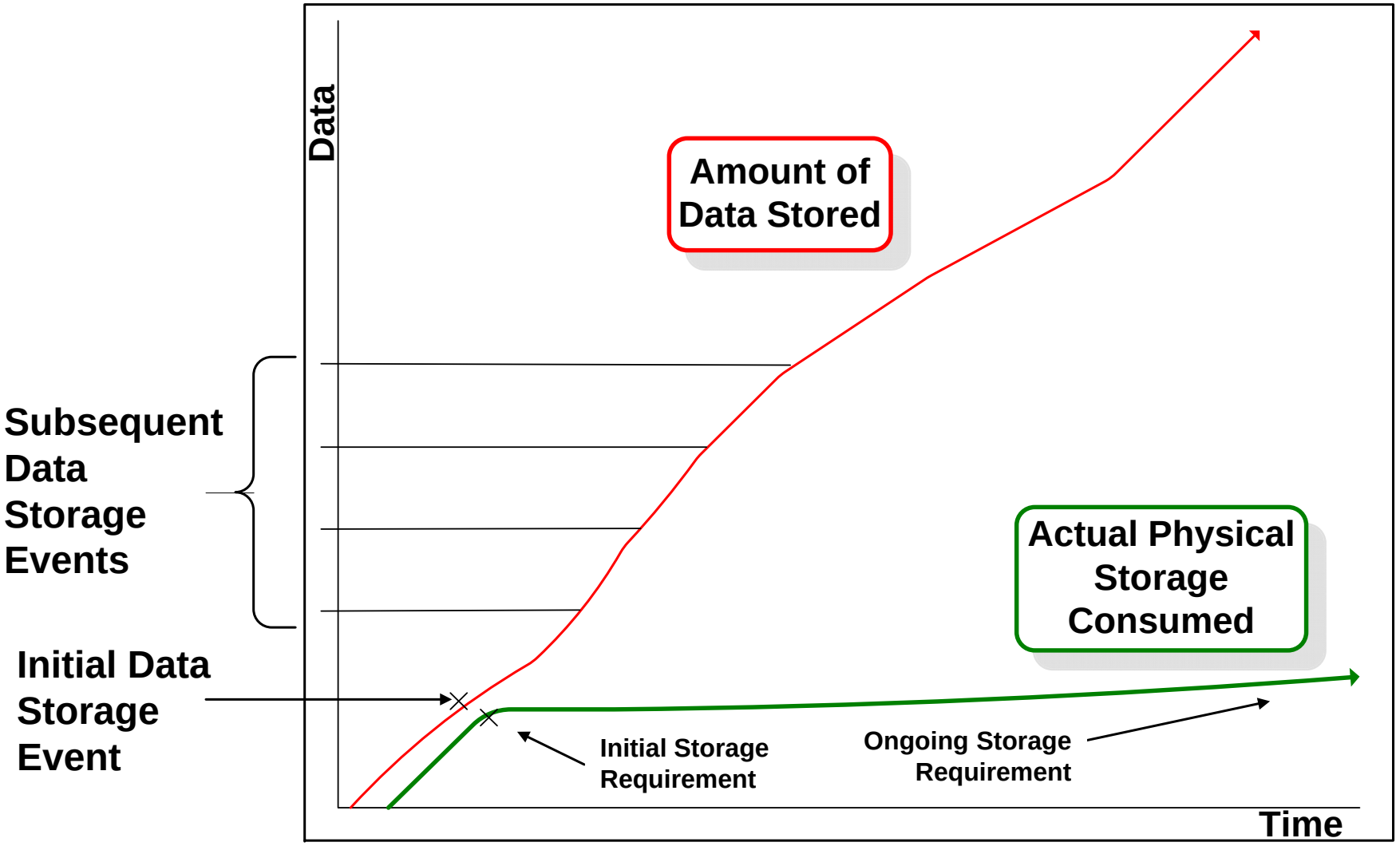
www.snia.org/forums/dmf/knowledge

Understanding Data Deduplication Ratios

Backup:
Influence of Backup Methodology



Backup: Capacity Savings Over Time

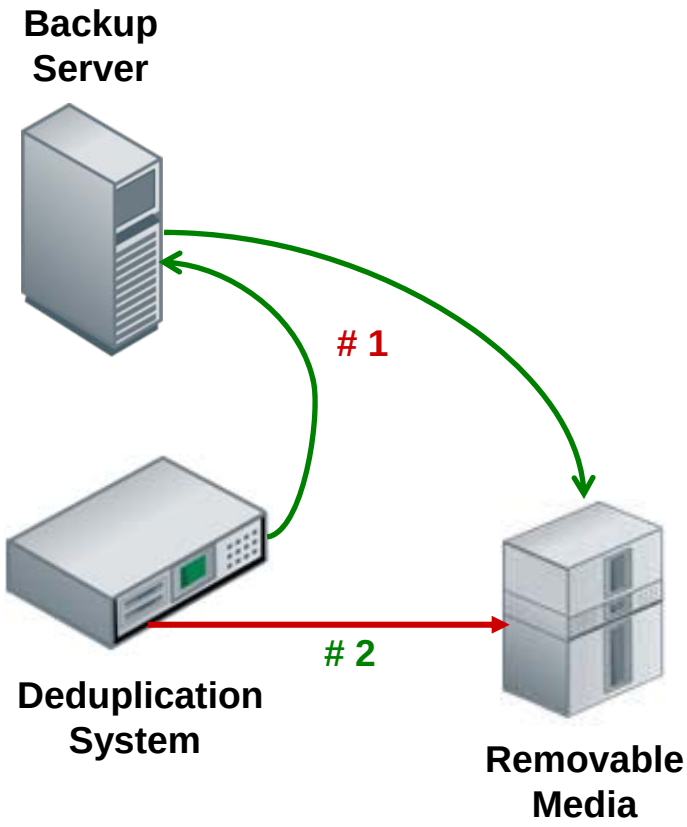


Backup: Deduplication for Data Movement

- 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

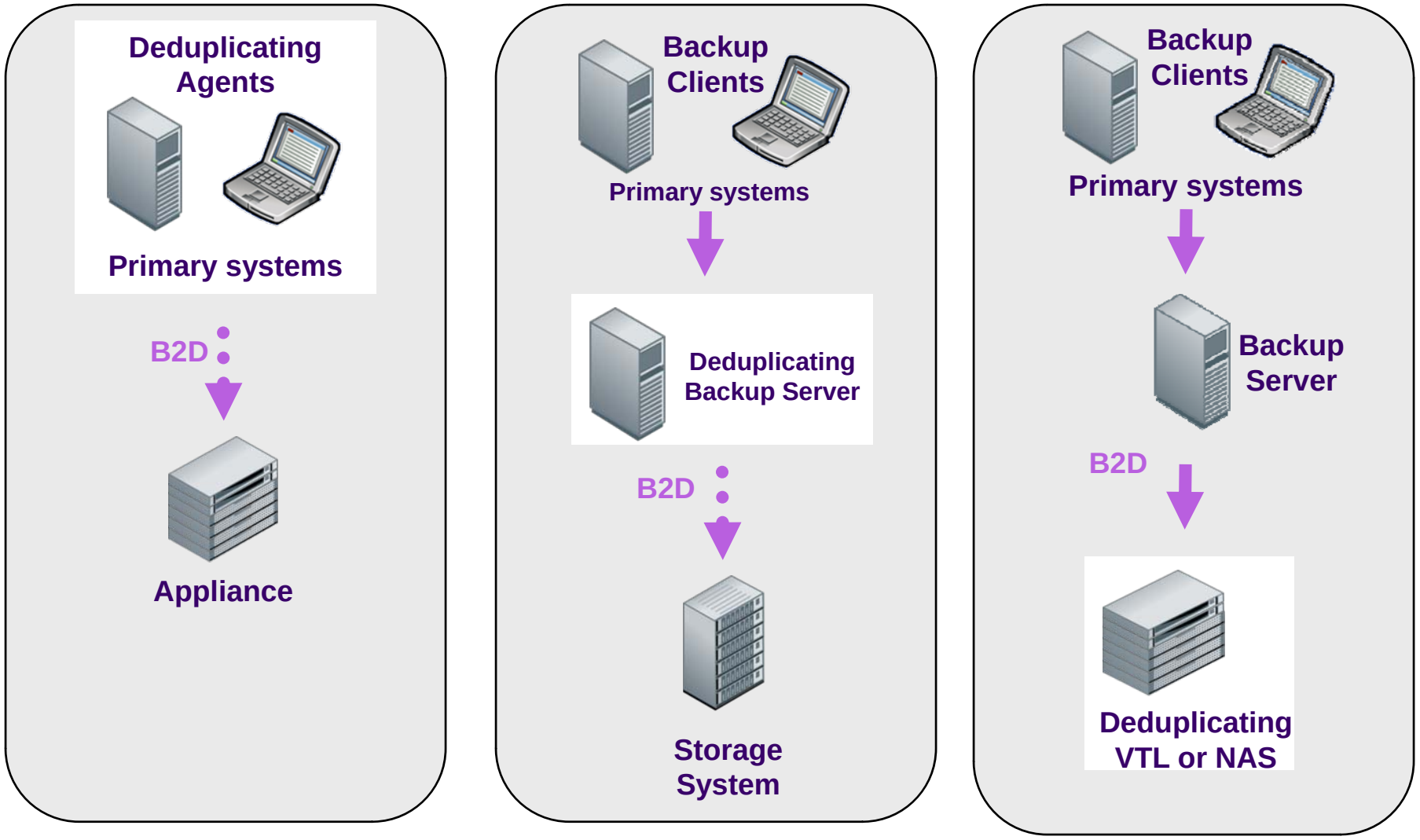
Backup: Removable Media Integration

- 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

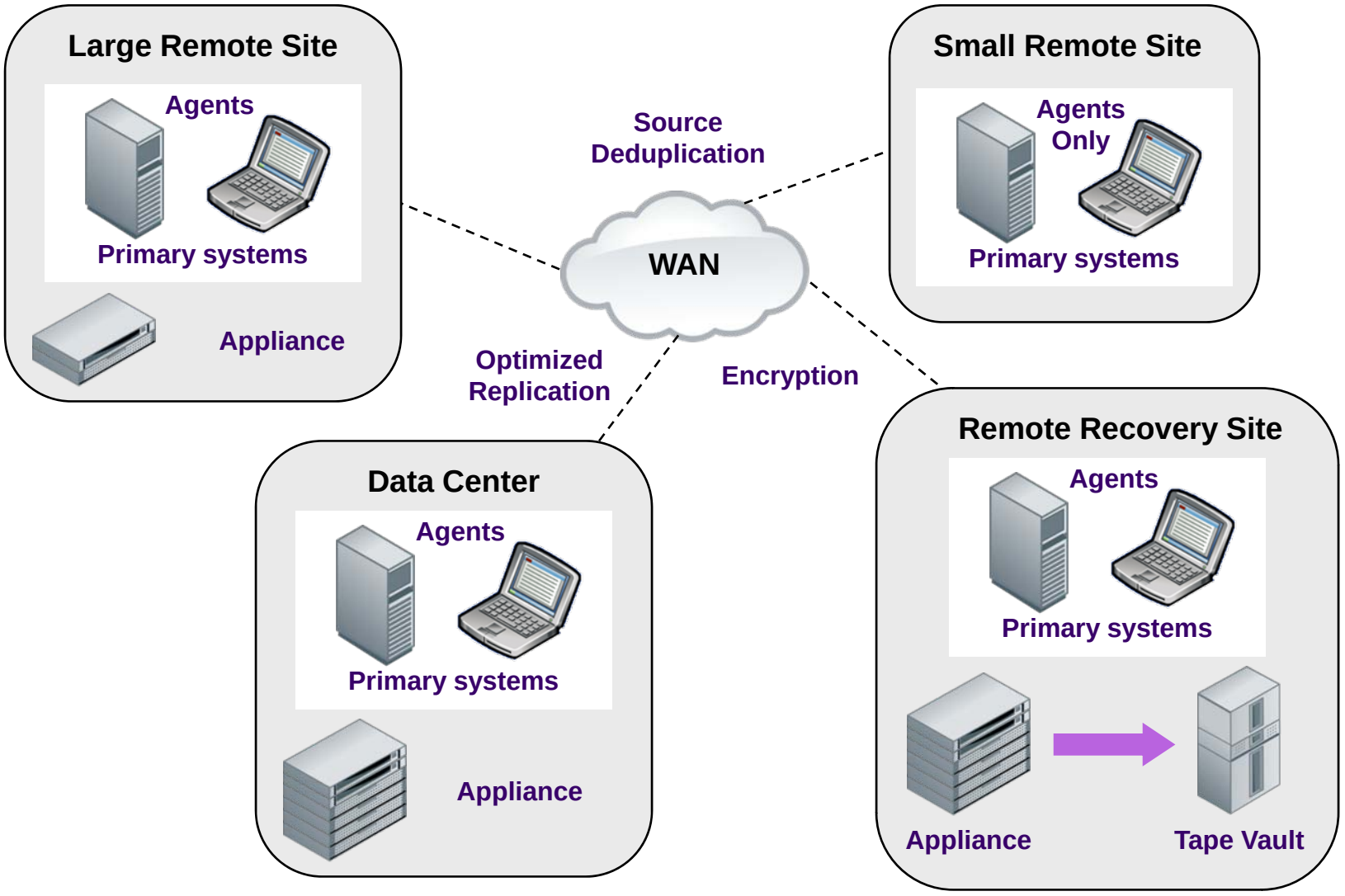


Deduplication within Secondary Storage Use Cases

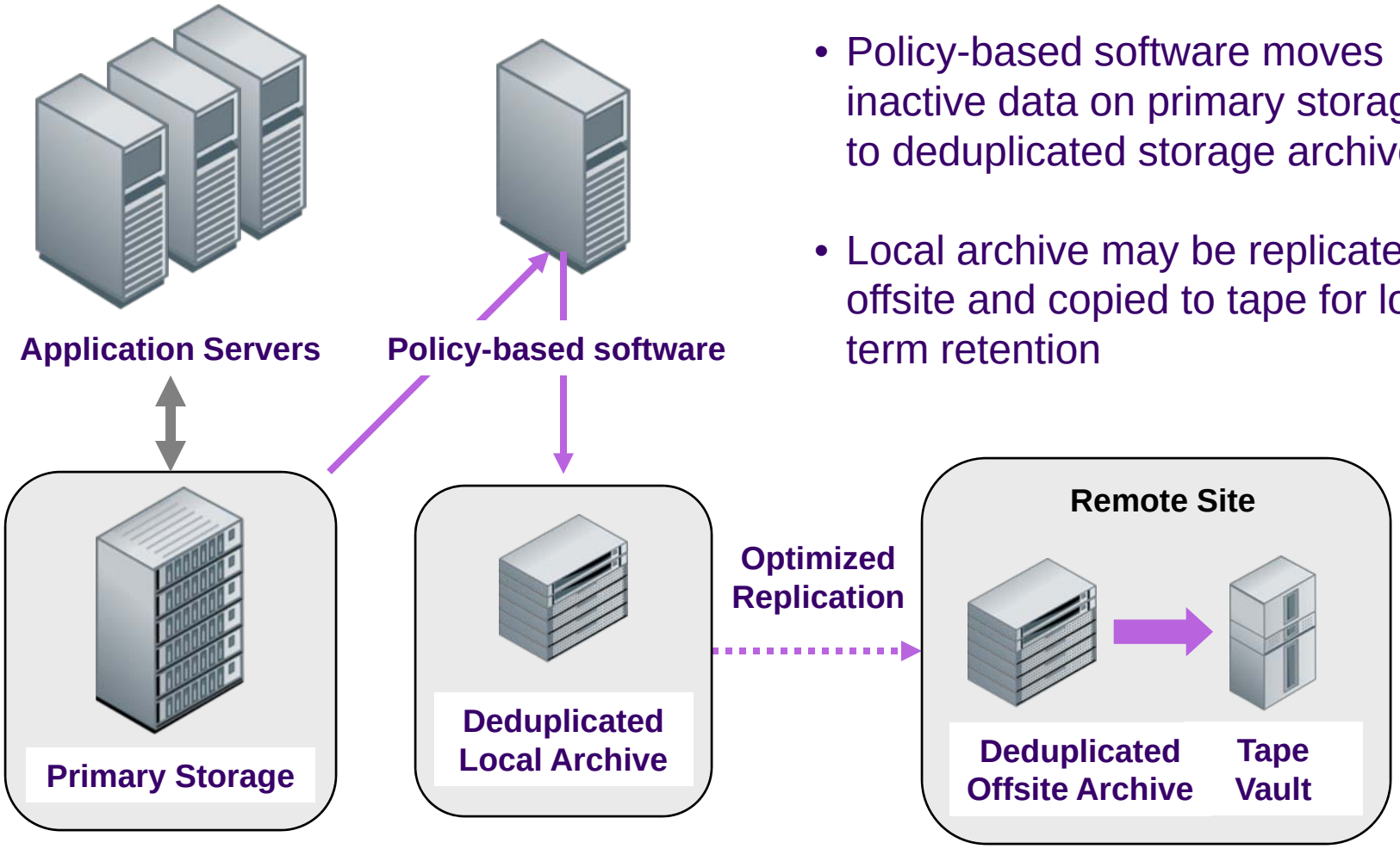
Deduplication for Backup and Recovery



Backup Remote Office Source Deduplication



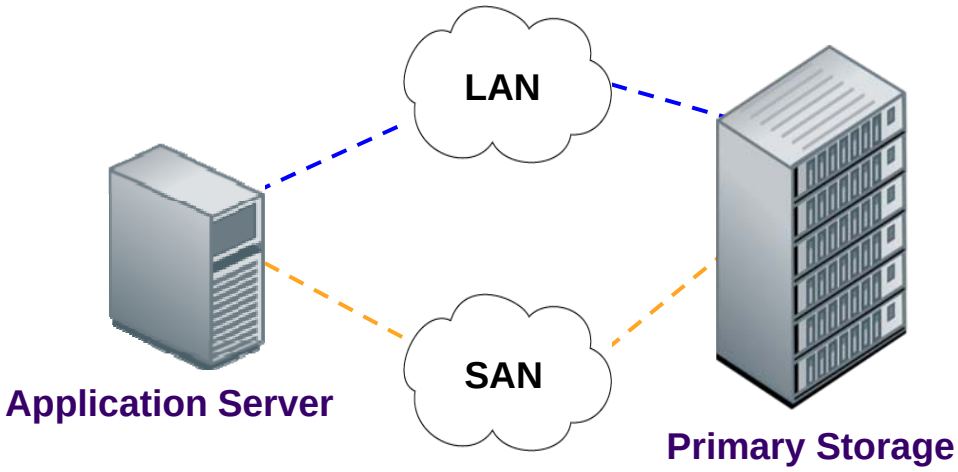
Deduplicated Archive Repositories



- 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

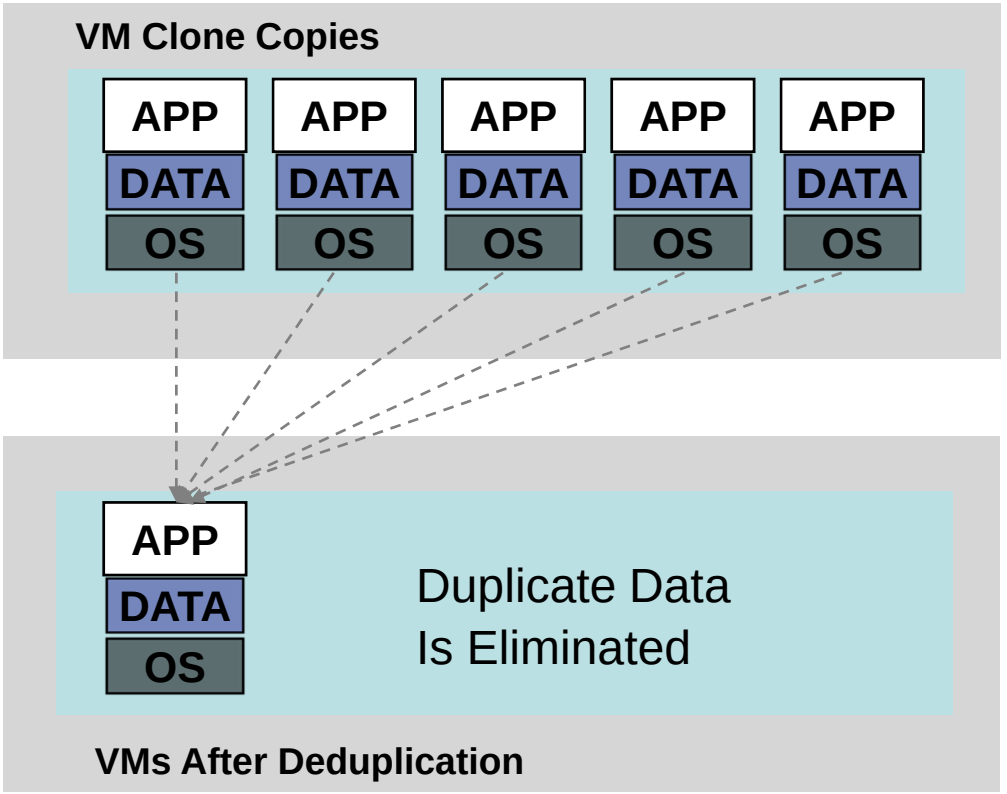
Primary Storage with Deduplication



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

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

Data Deduplication and Space Reduction (DDSR) Special Interest Group

Matthew Brisse
Daniel Budiansky
Mike Dutch
Michael Fishman
Larry Freeman
Devin Hamilton
Jason Iehl
Shane Jackson
Gene Nagle
Thomas Rivera
Tom Sas
Gideon Senderov

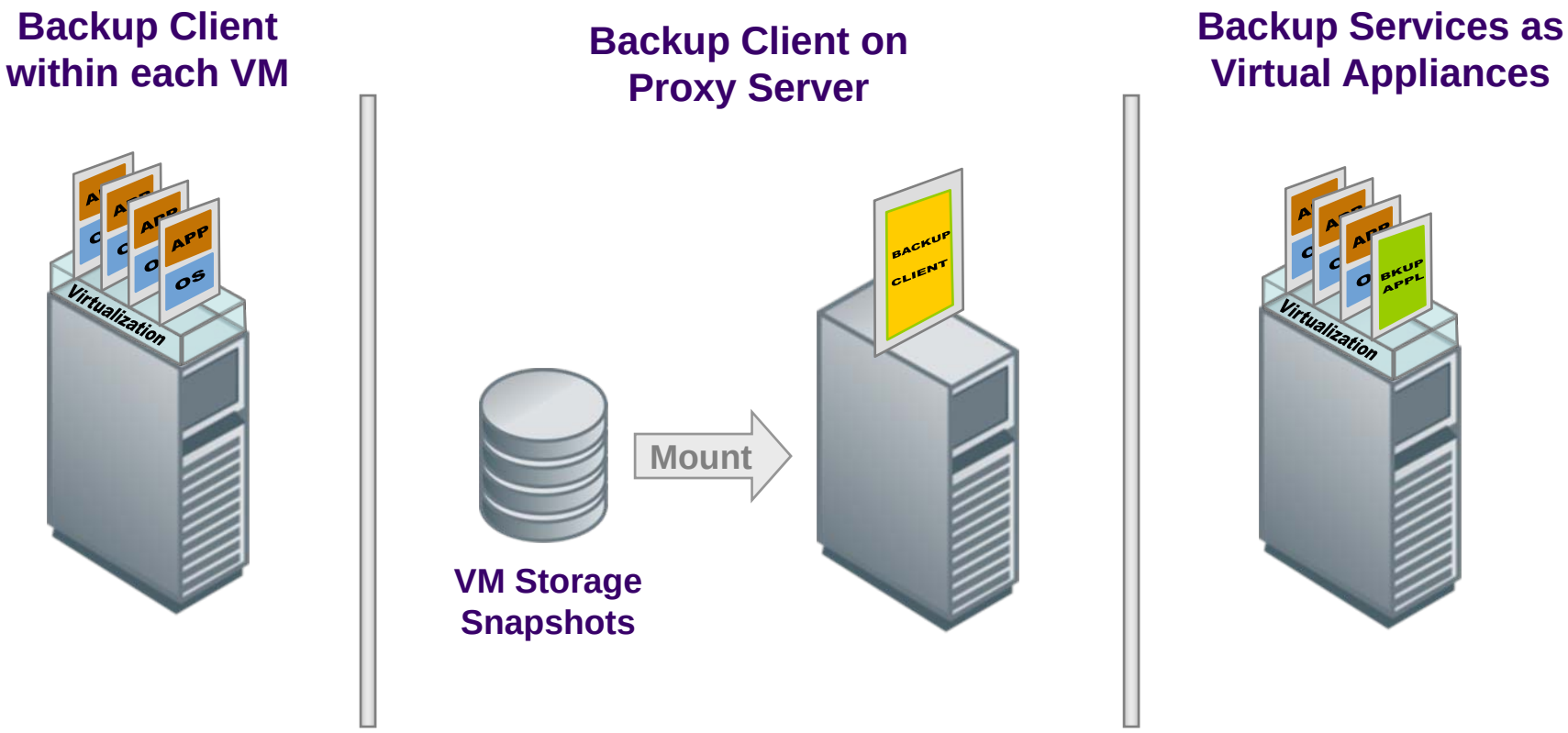
SNIA Data Management Forum

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

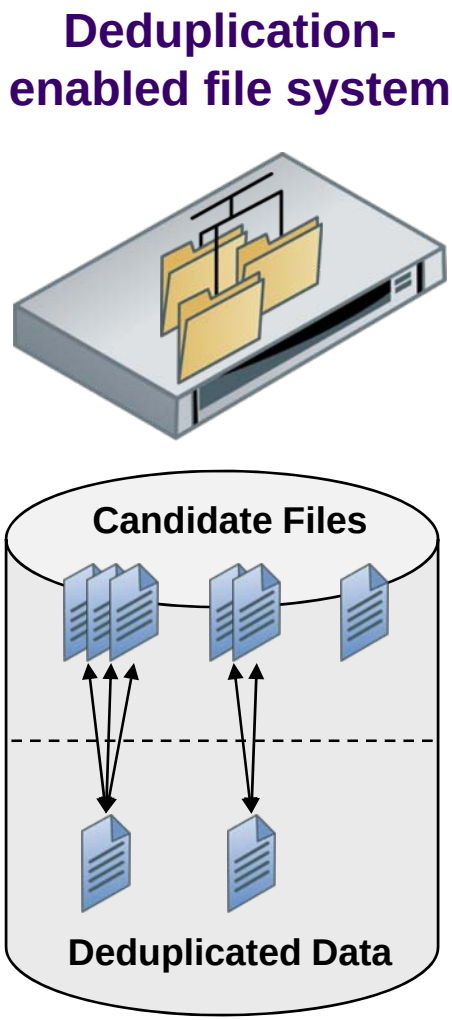
Approaches to VM Backup Details



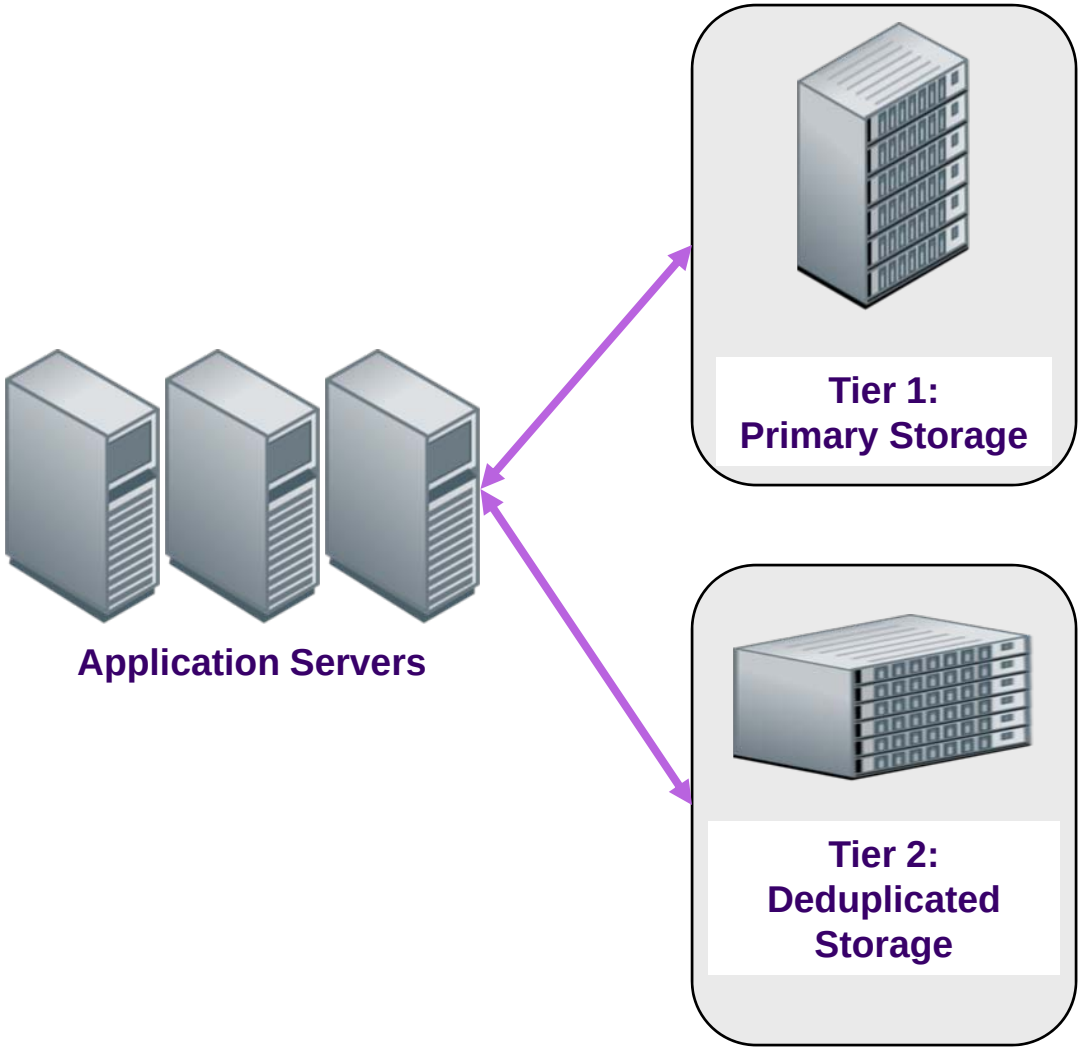
The specific backup application determines where deduplication is performed

Deduplicated Primary Storage

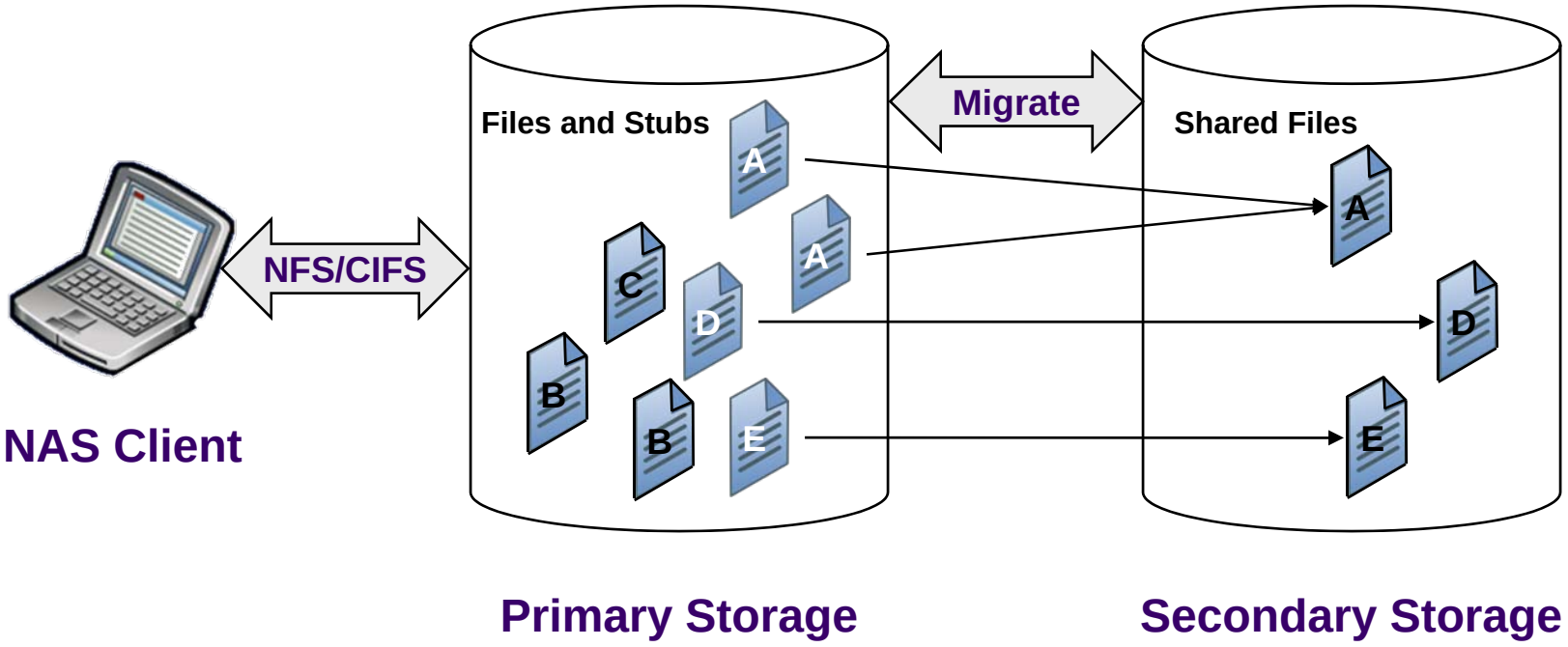
- Space/Performance tradeoff
- File Services
  - Home directories
  - Tech Pubs
  - Email
  - Fixed content
- Replicated databases
- Test and application development
- Source code version control system
- Virtualized environments



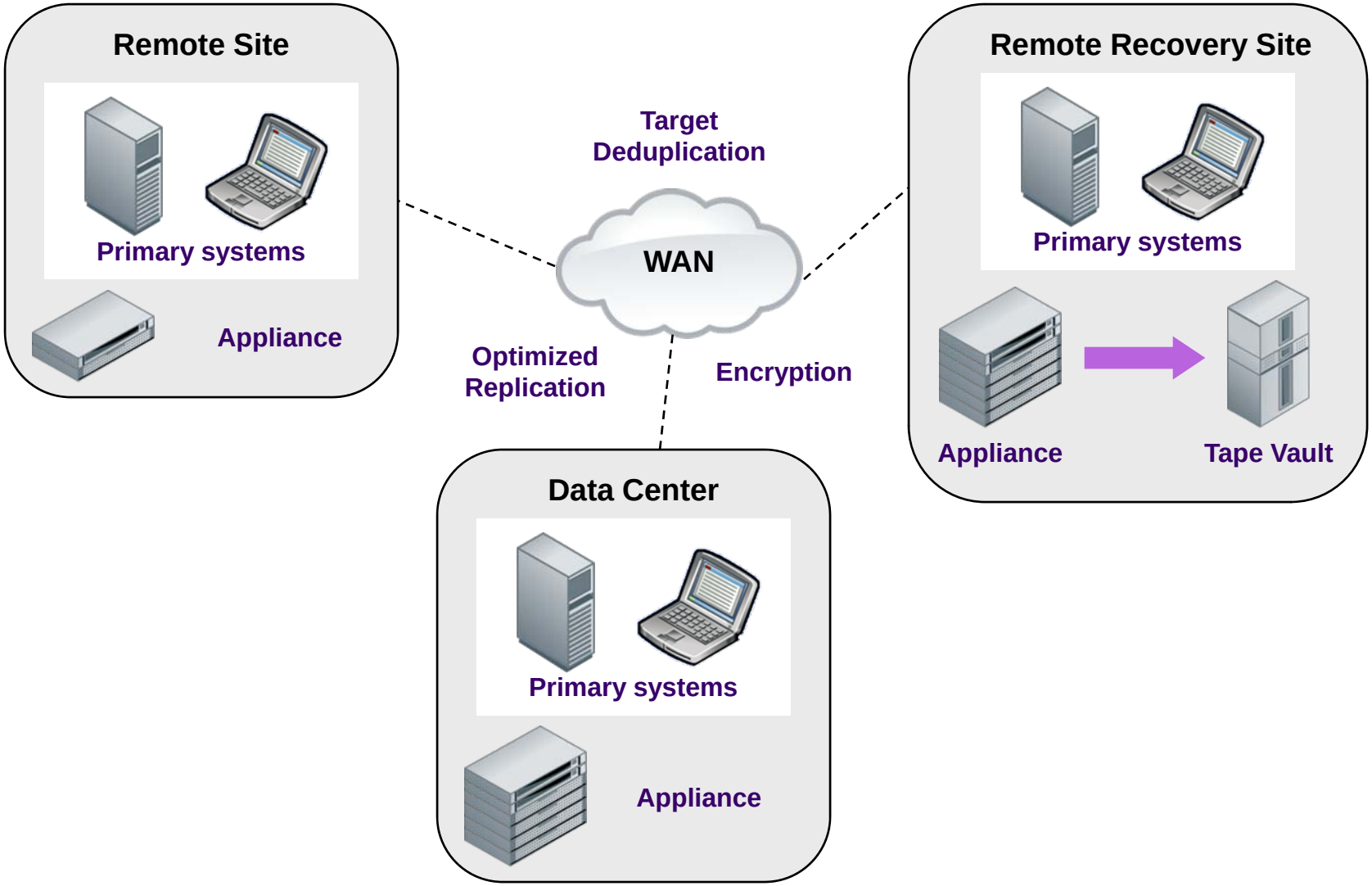
Storage Tiering with Deduplication



File Tiering with Deduplication



Backup Remote Office Target Deduplication



Deduplication within the Backup Cloud

Application and Data on Customer Premises

Deduplicated Data

Deduplication performed by the backup client

Backup as Service within the Cloud

Application and Data on Customer Premises

Data

Deduplication Target within the Cloud