



Education

Data Deduplication
Methods for Achieving Data Efficiency

Matthew Brisse, Quantum
Gideon Senderov, NEC

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
 - Member companies and individuals may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced without modification
 - ◆ The SNIA must be acknowledged as 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 nor should be construed as legal advice or opinion. If you need legal advice or legal opinion please contact an attorney.
 - The information presented herein represents the Author's personal opinion and current understanding of the 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.**

Data Deduplication Implementation Overview

This technical session will address the nuances of data deduplication and the various design approaches available today. Space savings technologies including data deduplication are used to dramatically improve storage efficiency. The session will address the question of what data deduplication is, how it is performed, and the architectural choices available today. It will address various design approaches including fixed length and variable length segmentation, inline and post processing, source and target, availability and resiliency of deduplicated data, and complementary use of removable media. It will also explore the factors affecting space reduction ratios relative to specific capacity optimization techniques.

- Definitions
- Where and How Data Deduplication Works
- Review Fixed or Variable Segment Approaches
- Review of Different Design Approach
- Removable Media Integration
- What to Expect with Data Deduplication Ratios
- Questions

Definitions

Data Deduplication is the process of examining a data-set or I/O stream at the sub-file level and storing and/or sending only unique data. The definition of "what is a duplicate" is predicated upon the method used to evaluate, identify, track and avoid duplication. The deduplication process includes updating tracking information, storing and or sending data that is new and unique, and disregarding any data that is a duplicate.

Compression is the encoding of data to reduce its storage requirement. Deduplicated data can also be compressed.

Single Instance Storage is the replacement of duplicate files with references to a shared copy.



Check out SNIA
Tutorial:
Green Storage
TWG Status & Plans



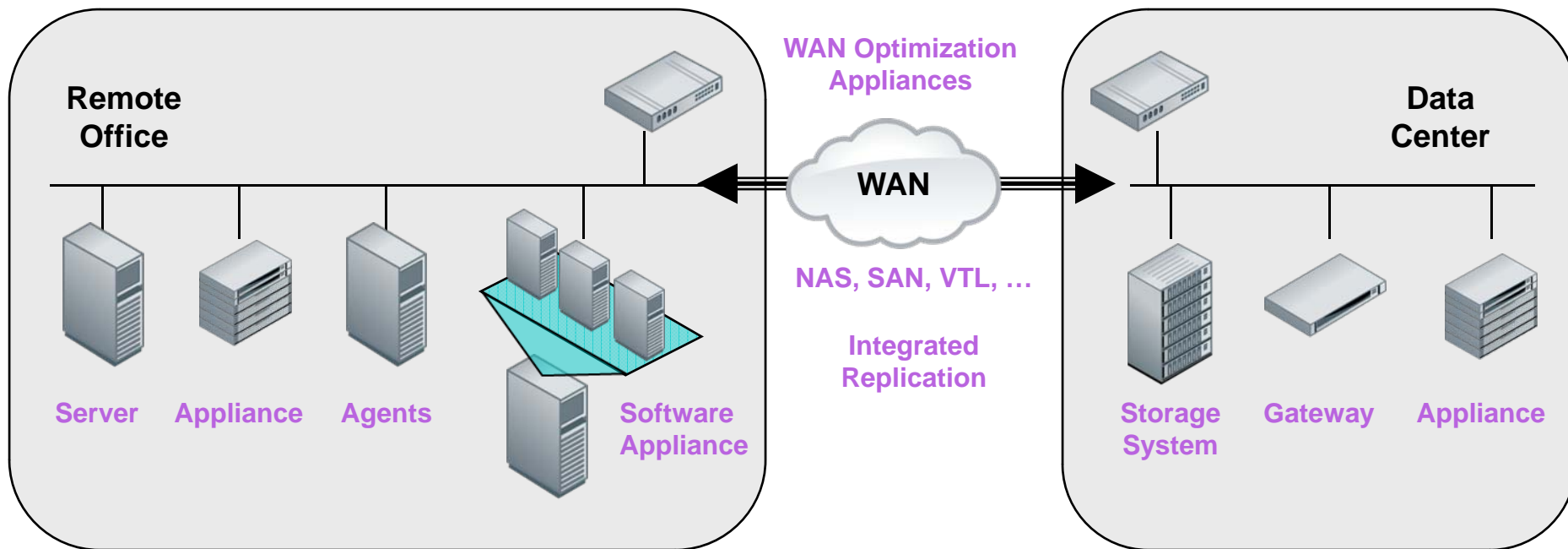
Check out SNIA
Tutorial:
Introduction To Data
Protection



Check out SNIA
Tutorial:
The risk of retention and deletion
in the face of FRCP

Where Deduplication Can Happen

Multiple deployment examples are illustrated... specific deployments are selected based on customer situation



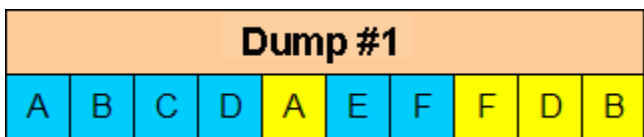
Deduplication – How it Works



➤ How Deduplication Works

- ◆ Evaluate Data
- ◆ Identify Redundancy
- ◆ Create or Update Reference Information
- ◆ Store or Transmit Unique Data Once
- ◆ Read or Reproduce Data

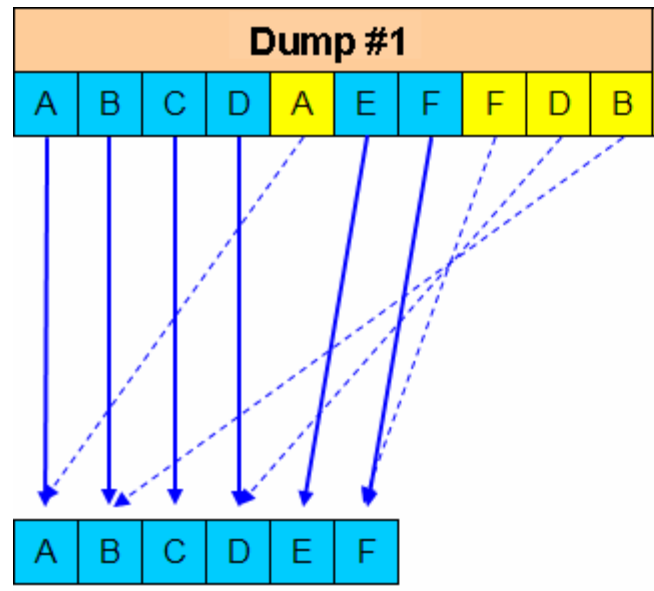
➤ Details To Follow ...




Deduplication Simplified



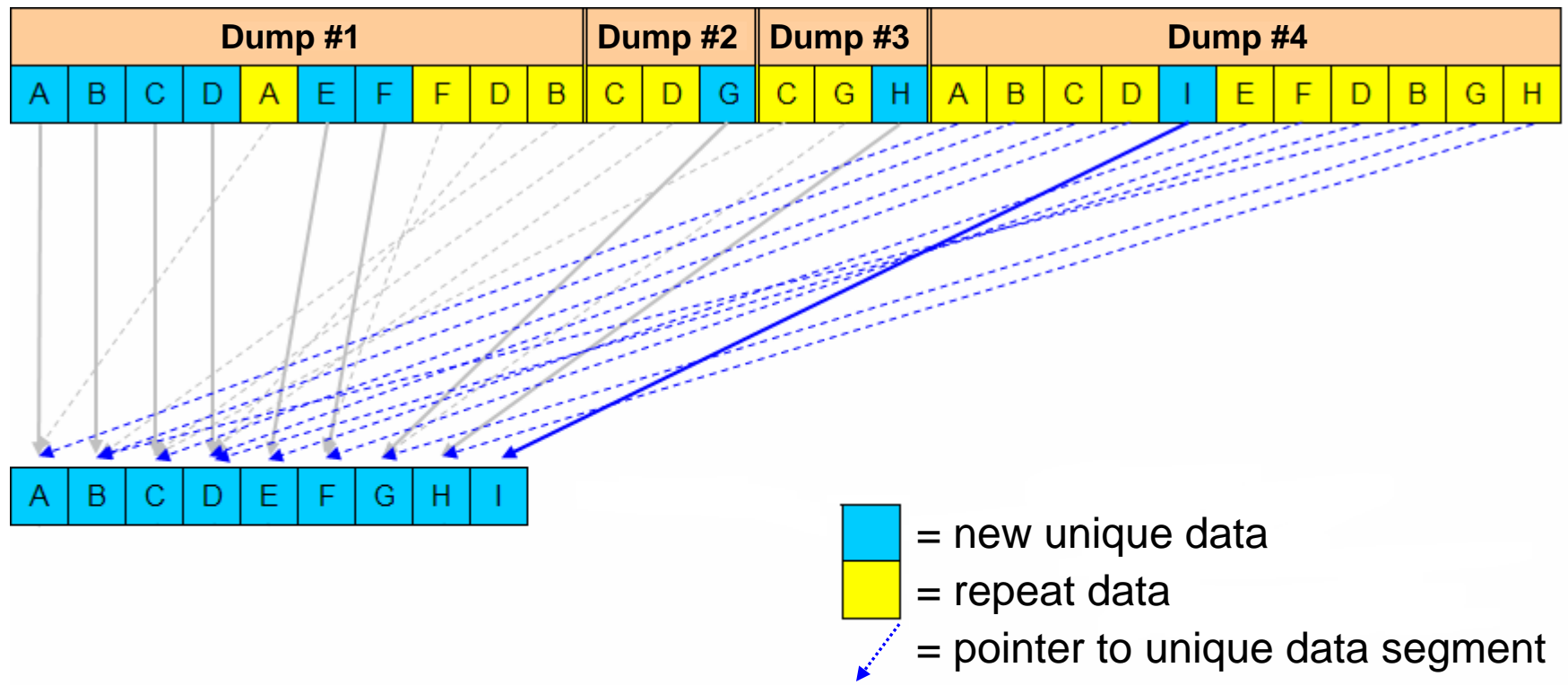
 = new unique data
 = repeat data

Deduplication Simplified

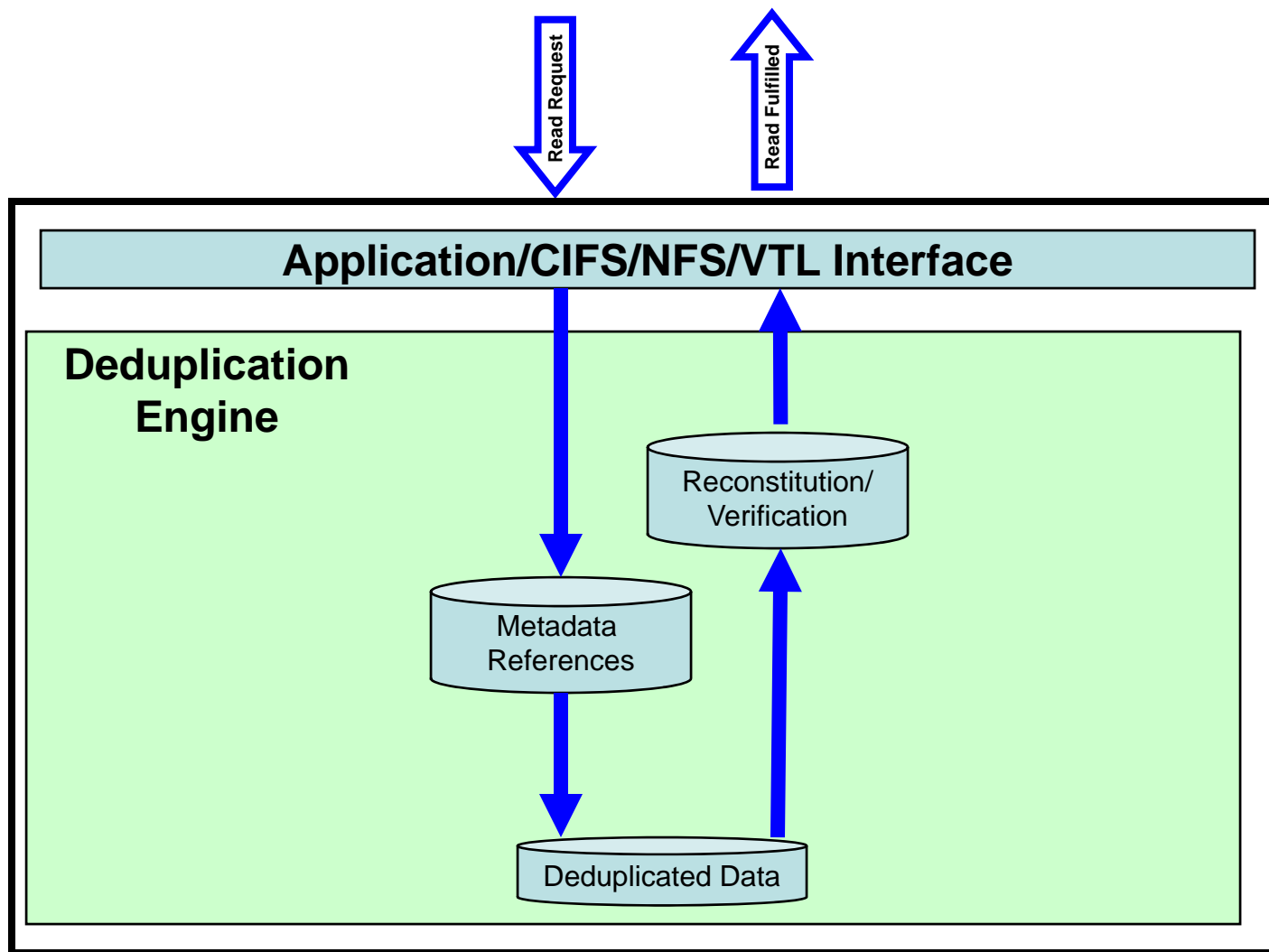


 = new unique data
 = repeat data
 = pointer to unique data segment

Deduplication Simplified



Reading the Data



Fixed or Variable Segment Deduplication

➤ 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

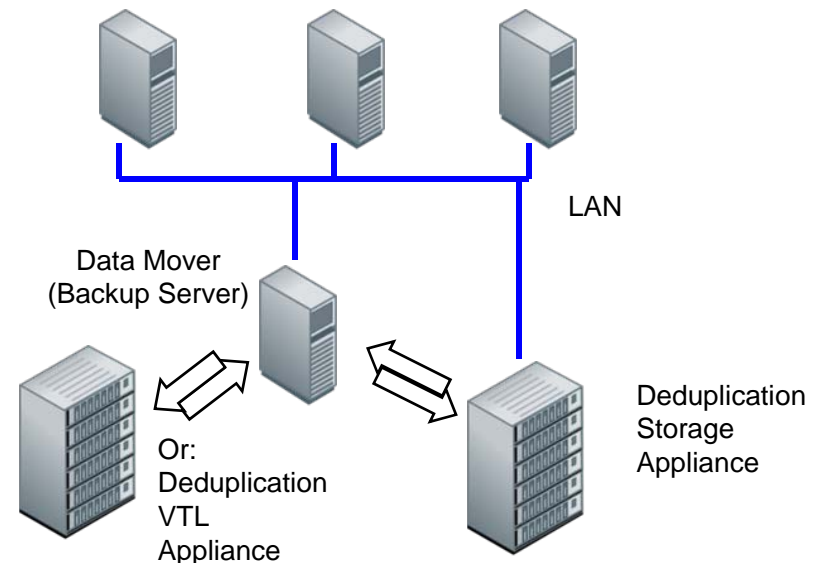
- ◆ Effects observed will vary by method
- ◆ Segmentation may not apply to all deduplication

Design Approach

- **Gateway**
 - ◆ Hardware capable of deduplicating data excluding storage hardware
- **Appliance**
 - ◆ Hardware capable of deduplicating data and/or storing deduplicated data including the storage hardware dedicated for this purpose
- **Storage System**
 - ◆ General purpose storage system that is capable of deduplicating data
- **Grid Storage**
 - ◆ Distributed system capable of deduplicating data across multiple independent components that can scale capacity and performance
- **Agent**
 - ◆ Client-side software capable of deduplicating data for a specific application
- **Software Appliance**
 - ◆ Server-side software capable of storing deduplicated data from application agents
- **Storage Software**
 - ◆ Software capable of deduplicating data and/or storing deduplicated data that is neither an agent nor a software appliance

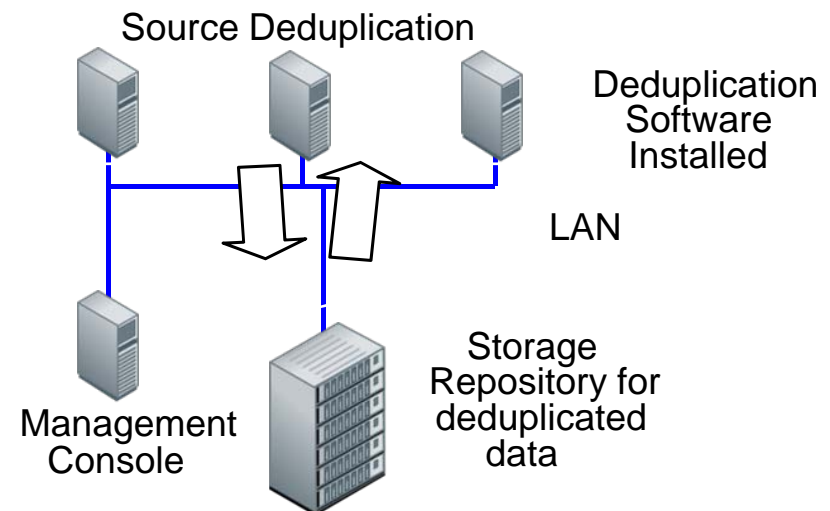
Target Deduplication

- Stand-alone central repository for data
- Could be NAS, SAN, VTL, other
- Could be attached via Ethernet, FC, or both
- Deduplicates data as internal process either in-line, post-process, or both
- May include self-contained or gateway configurations
- May include other methods of data reduction in addition to deduplication, such as compression, or object level differencing
- Allows global deduplication across all clients
- Scale of deduplication space may vary by implementation

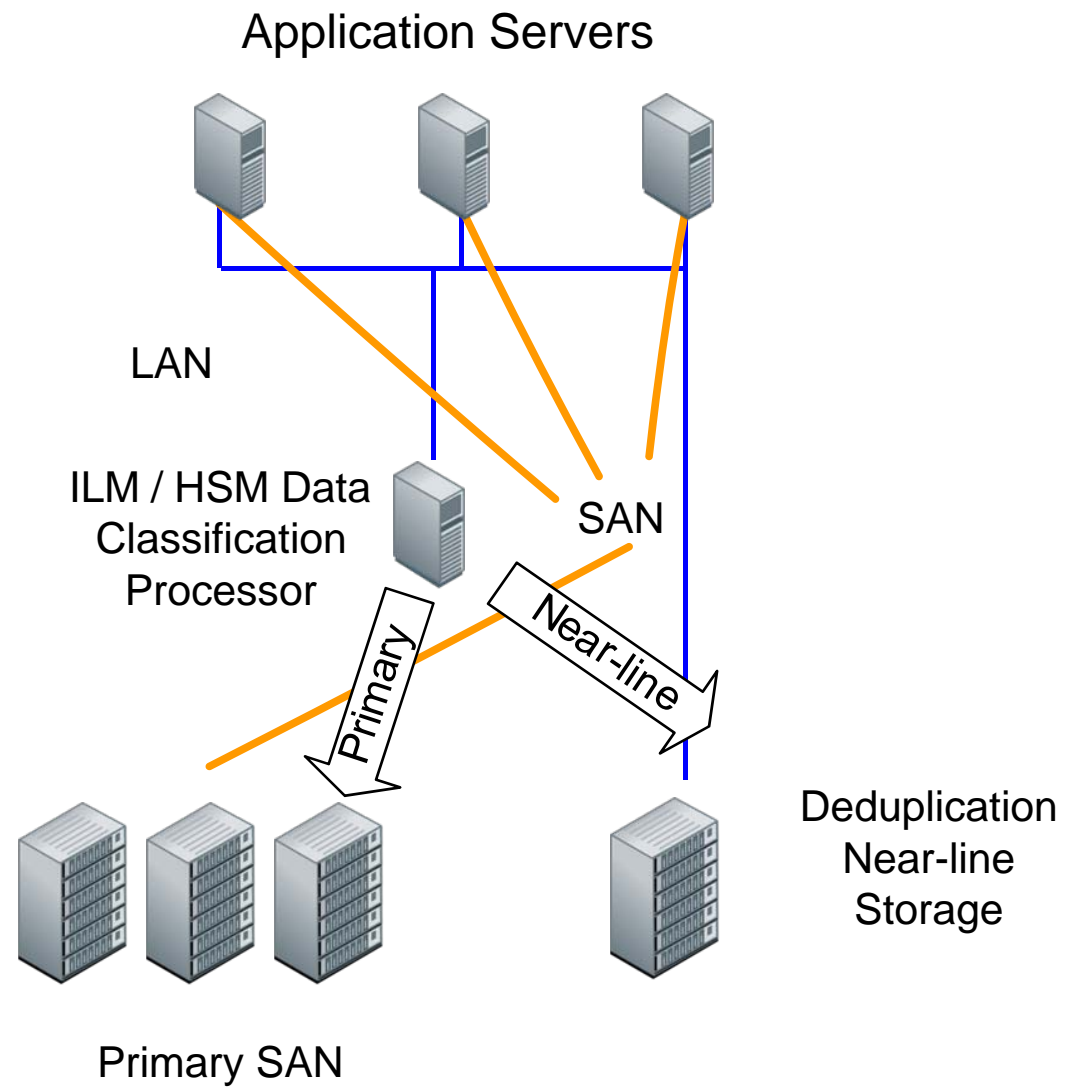


Source Deduplication

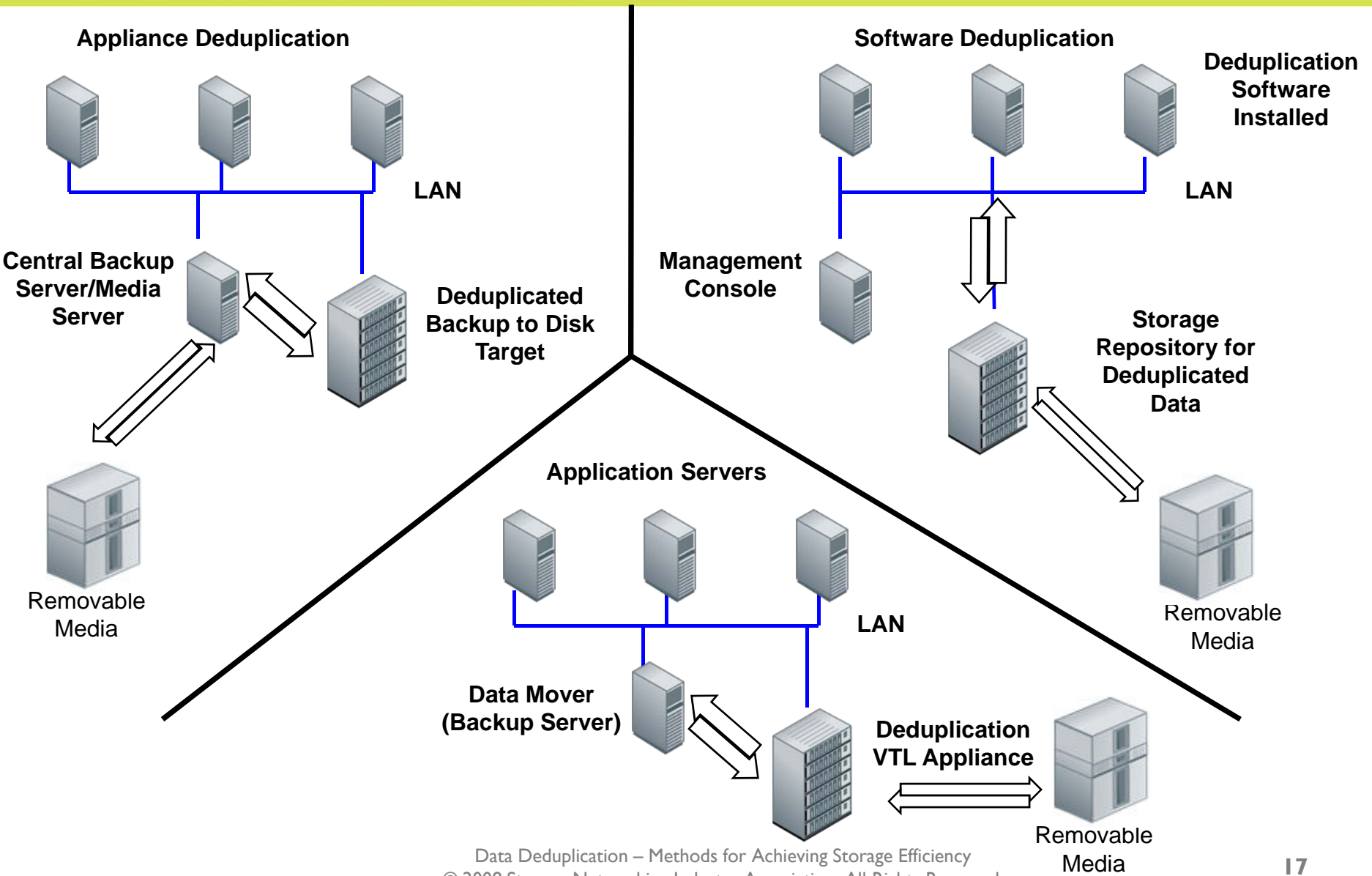
- Client software or agent installed at deduplication point
- Example: Deduplication agent installed on client / server
 - ◆ Agent scans for new or modified data
 - Unchanged data simply referenced again without further processing
 - Allows incremental forever approach while providing full backups
 - Typically shortens backup windows relative to both fulls and incrementals
 - Only unique data sent to central repository
 - Reduction in data transmitted and stored
 - ◆ Could allow global deduplication across all clients
 - ◆ Scale of Deduplication space may vary by implementation
 - ◆ Ideal for branch office applications
 - May not require additional hardware at remote sites
 - ◆ Modest CPU impact
 - ◆ May include other methods of data reduction in addition to deduplication, such as compression



Deduplication for Near-line Use



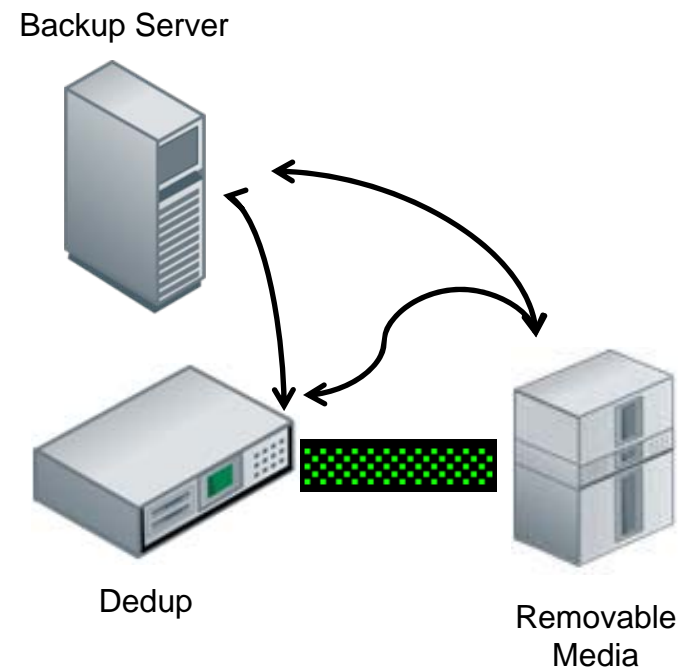
Deduplication for Backup and Recovery



Scope of Data Deduplication

- Data Deduplication for ALL data from ALL sources across the entire solution
 - ◆ Single copy of data across entire solution
 - ◆ Can span multiple solution components
- Global Deduplication
 - ◆ Reference to a repository of deduplicated data spanning multiple computing systems to identify duplicate data.
- Local Deduplication
 - ◆ Reference to a repository of deduplicated data storing data originating from a single storage system
- Can span multiple applications
 - ◆ Backup
 - ◆ Archive
 - ◆ Nearline
- Can span multiple datacenters

- **Deduplication integration with tape**
- **Create long term removable media storage for compliance and archive**
- **Different data path approaches**
 - ◆ Path through backup server
 - ◆ Path direct from Deduplication to removable media storage
- **Two Options for VTL**
 - ◆ Application-specific removable media integration: backup software initiates and manages the entire process
 - ◆ Dedup device manages the removable media environment autonomously



What to Consider

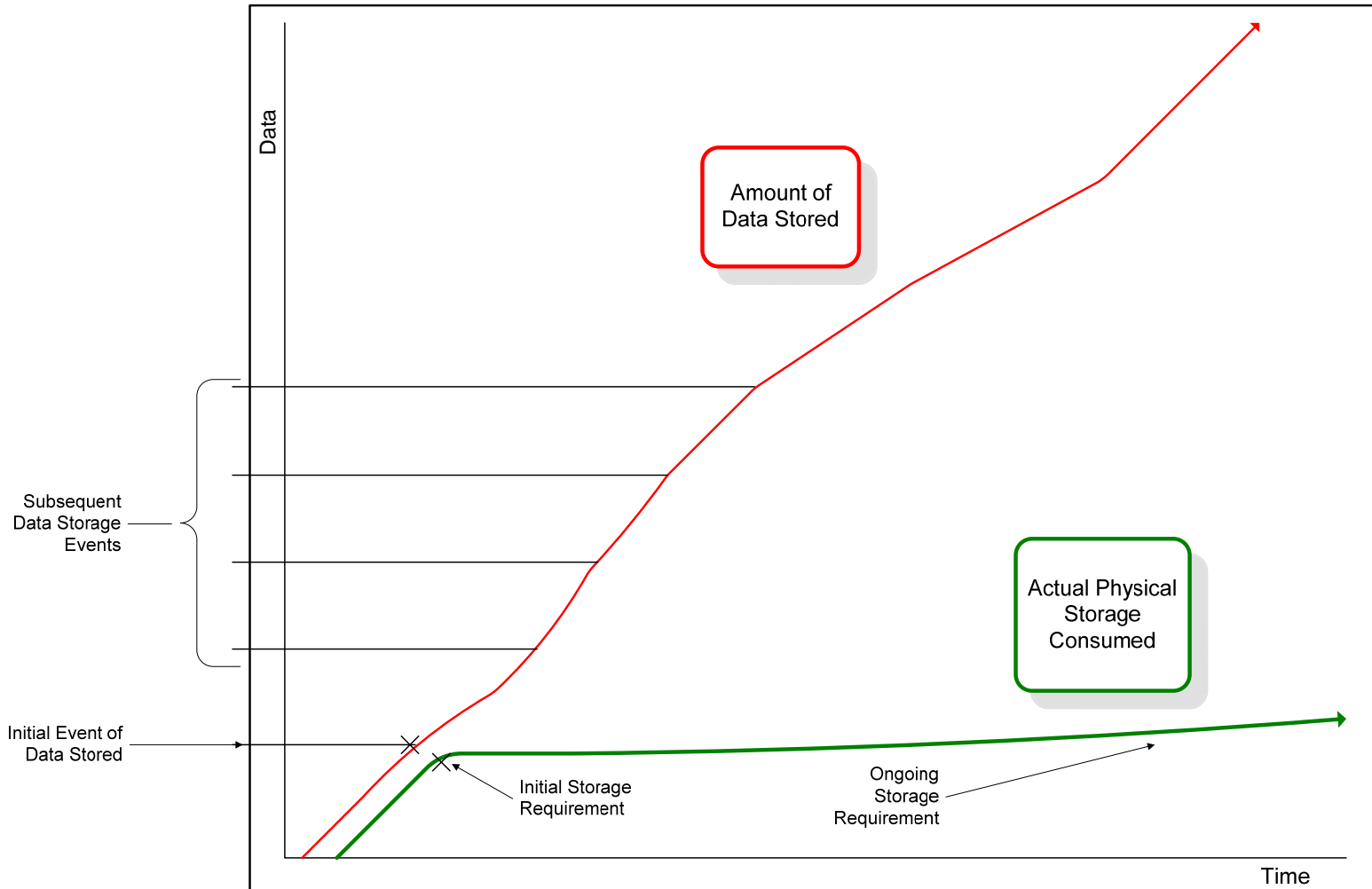
- Factors that will impact your results:
 - ◆ Different applications or data types
 - ◆ Bandwidth and latency
 - ◆ Backup/Archive Policies
 - ◆ Data Protection Overhead
 - ◆ Compression and Encryption
- Global Deduplication and Scope
- Deduplicated Data Resiliency
- Scalability
 - ◆ Capacity
 - ◆ Performance
- Legal Considerations

Estimating the Deduplication Ratio

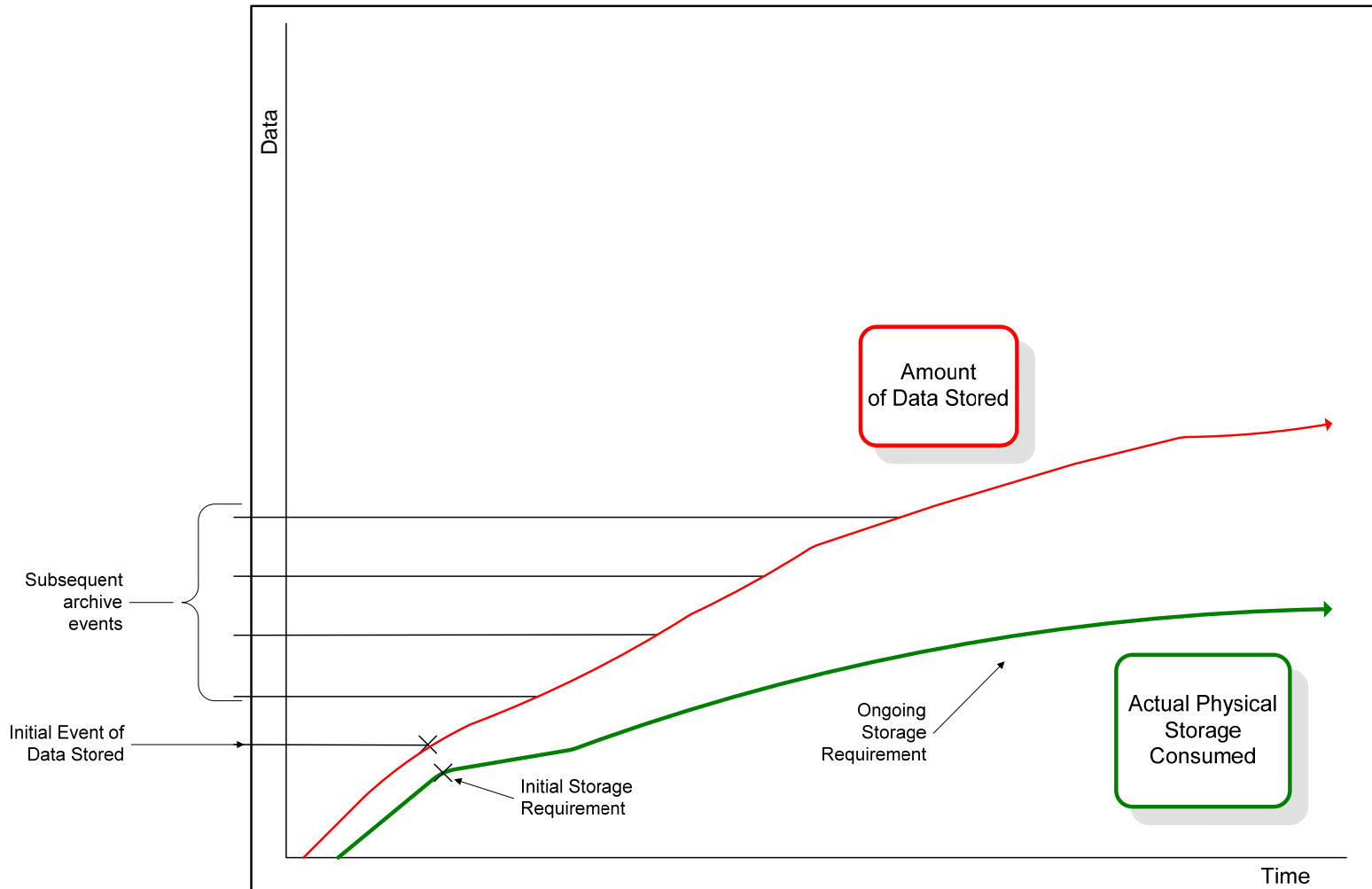
“Your mileage may vary!”

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
Content-aware	Content-agnostic
Temporal data deduplication	Spatial data deduplication

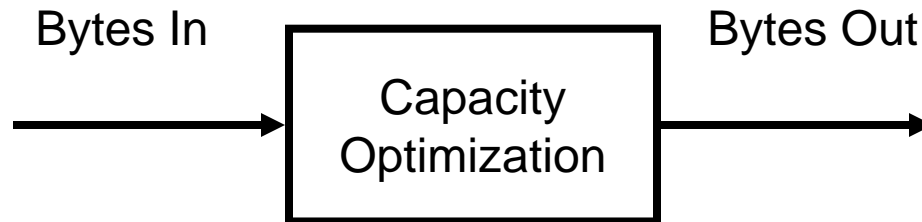
Deduplication of Backup Data



Deduplication of Non-Backup Data



The Space Reduction Ratio



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

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

Example:

Bytes In: (7 Days Full Backup) x (100GB per day) = **700GB**

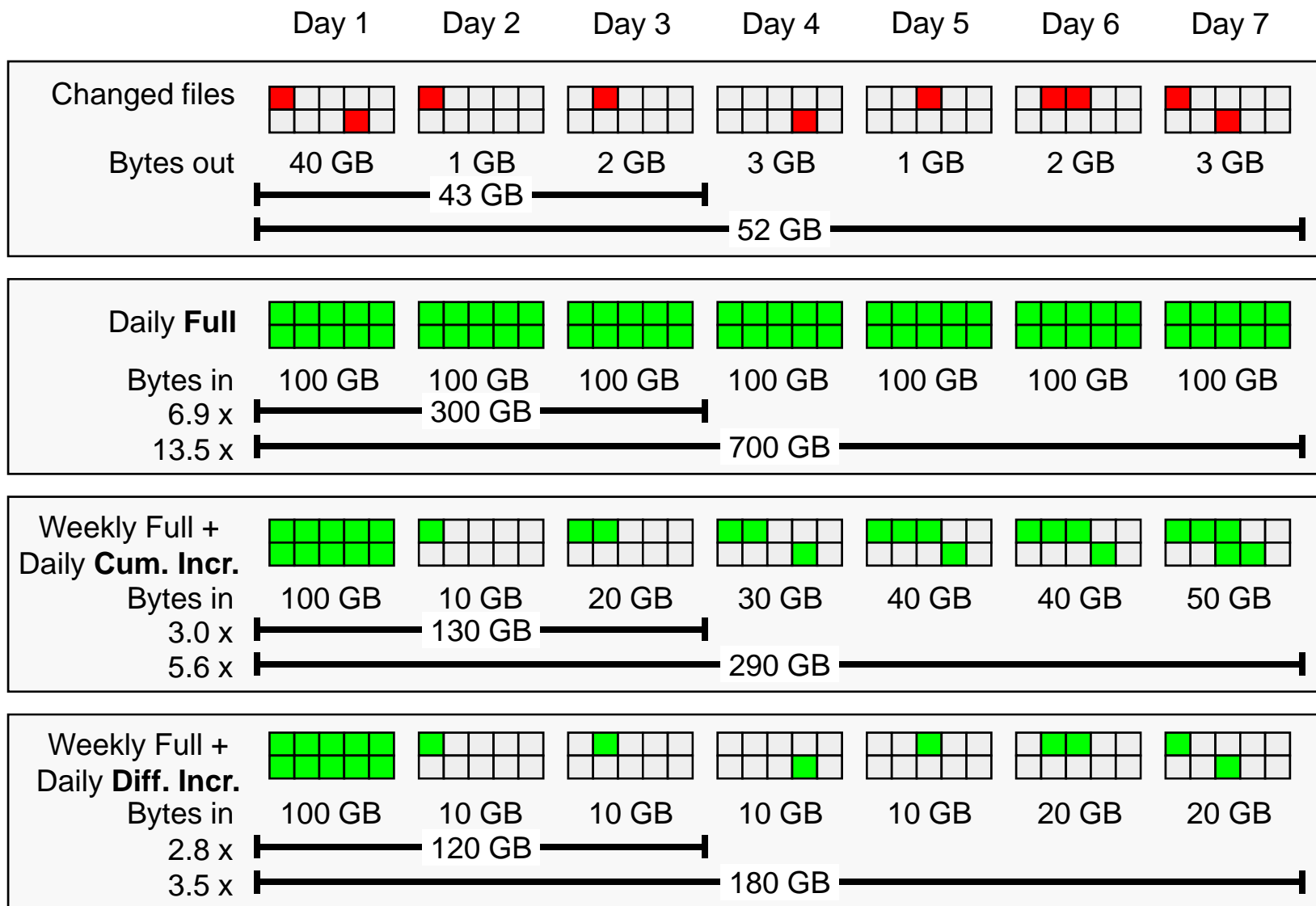
Bytes Out: **52GB** changed data actually stored

→ Space Reduction Ratio	= 700GB / 52GB =	13.5
→ Space Reduction %	= 1 - (1 / 13.5) =	92.6%



Understanding data deduplication ratios

The Deduplication Ratio is influenced by the Backup Methodology



- 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 SIG

**Rory Bolt
Matthew Brisse
Mike Dutch
Michael Fishman
Larry Freeman
Devin Hamilton**

**Jason Iehl
Shane Jackson
Jeff Porter
Gideon Senderov
Jim Shocrylas**