



Education

Deduplication
Methods for Achieving Data Efficiency

Devin Hamilton, Data Domain
Principal Systems Engineer, DMF DPI Co-Chair

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.

Data Deduplication Implementation Overview

Deduplication has become a very popular topic in the industry because of the potentially large reduction in cost and increase in efficiency it offers. Deduplication technologies are being promoted at various points within the storage network including source deduplication, deduplication of data in transit, and deduplication at the storage destination. Deduplication technologies are also being promoted in all tiers: backup, archiving, and primary storage.

Each of these storage use cases represents a unique set of challenges. Implementing any deduplication technology has major implications for scale, performance, and functionality. Deduplication also has long term legal and compliance implications for records management. This session will review various deduplication technologies available and the implications of each.

Tutorial Learning Objectives

Upon completion of this tutorial, the attendee should understand:

- Differences between various deduplication strategies
- Where, when, and how to utilize deduplication to boost storage efficiency in one's own environment
- Beneficial effects of implementing these technologies in backup and archive

Tutorial Discussion Topics

- Definition
- Deduplication 101
 - ◆ Where it Happens
 - ◆ General Concepts, How it Works
- Deduplication Simplified
- Deduplication Ratio
- Applications
- Benefits
- Choosing a Solution
- Use Cases



**Check out SNIA Tutorial:
Introduction To Data
Protection**

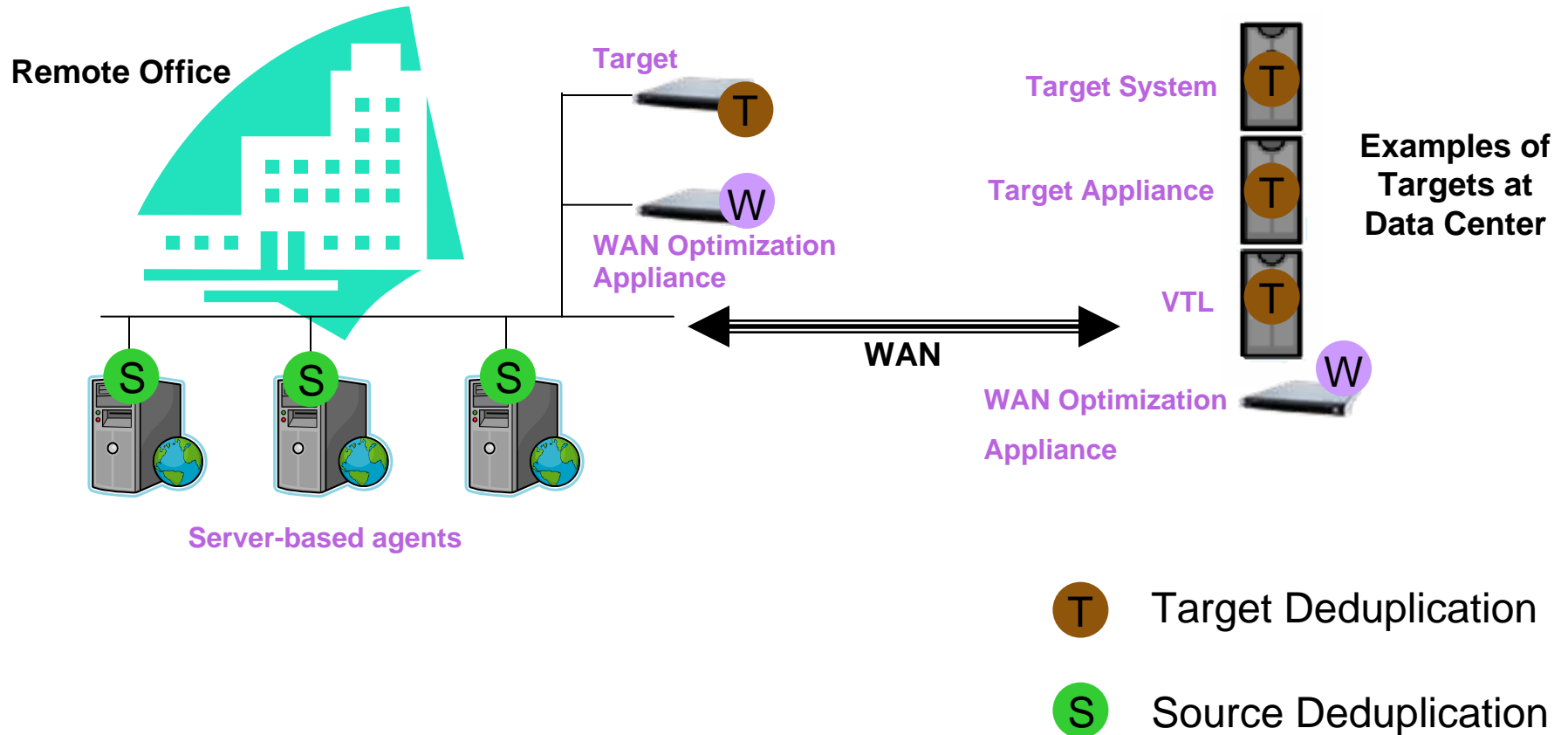
Definitions

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, 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.

Where Deduplication Happens



Benefits for Storage

	Current Challenges	Benefits of Deduplication
Workload	<p>Data growth is exponential</p> <p>Backup, Archive, Disaster Recovery all consuming massive storage resources</p>	<p>Significant reduction in storage footprint</p> <p>Greater amount of data stored in much less physical space</p>
Cost Focus	<p>Rising costs of capacity requirements</p> <p>Storage costs limiting growth</p>	<p>Dramatic reduction in cost per Terabyte of data stored</p> <p>Keep more data longer before buying more storage</p> <p>Reduced footprint, power, cooling</p>
Performance	<p>Aging architectures not keeping up with backup size / window</p> <p>Time to recovery not meeting Recovery Time Objective (RTO)</p>	<p>Benefit from backing up to disk or VTL</p> <p>High throughput to disk or VTL</p> <p>Online access to data</p>
Data Safety	<p>Data health is often linked to storage cost</p>	<p>Enables low cost high integrity storage of data</p> <p>Enables low cost offsite replication</p>

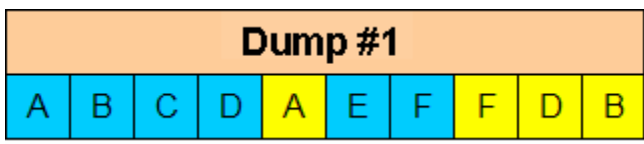
General Concepts



- Data Compression and Data Deduplication are different (potentially complimentary)
- Deduplication results will vary significantly amongst application, method, and environment of usage
- Data stored is available for recall or access regardless of the effects of deduplication
- Use of unique reference pointer instead of sending or storing multiple copies or versions of original data
- Deduplication has many uses including backup, primary storage, disaster recovery, archive and WAN optimization.

Deduplication – How it Works

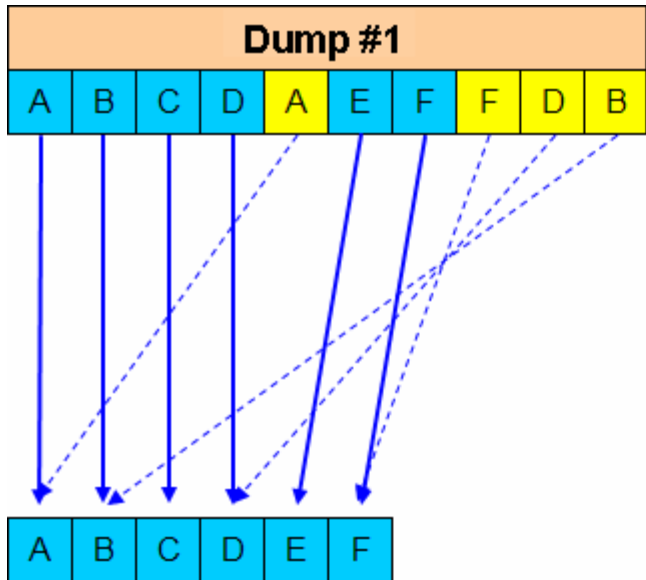
- Evaluate Data
- Identify Redundancy
- Create or Update Reference Information
- Store or Transmit Unique Data Once
- Recall (read) or Reproduce Data

Deduplication Simplified



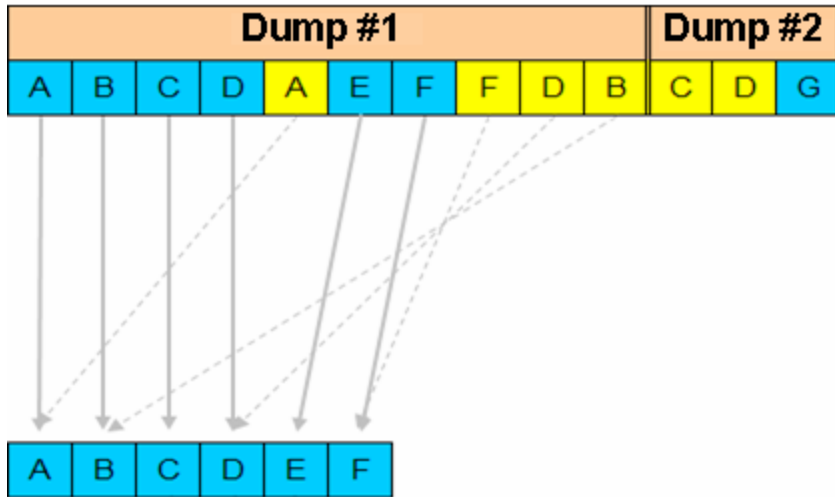
 = new unique data
 = repeat data

Deduplication Simplified



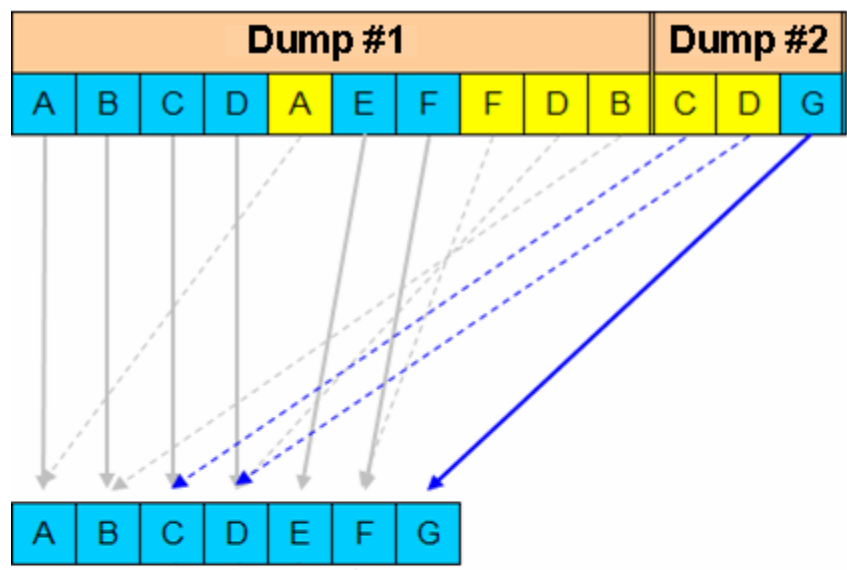
- = new unique data
- = repeat data
- = pointer to unique data segment

Deduplication Simplified



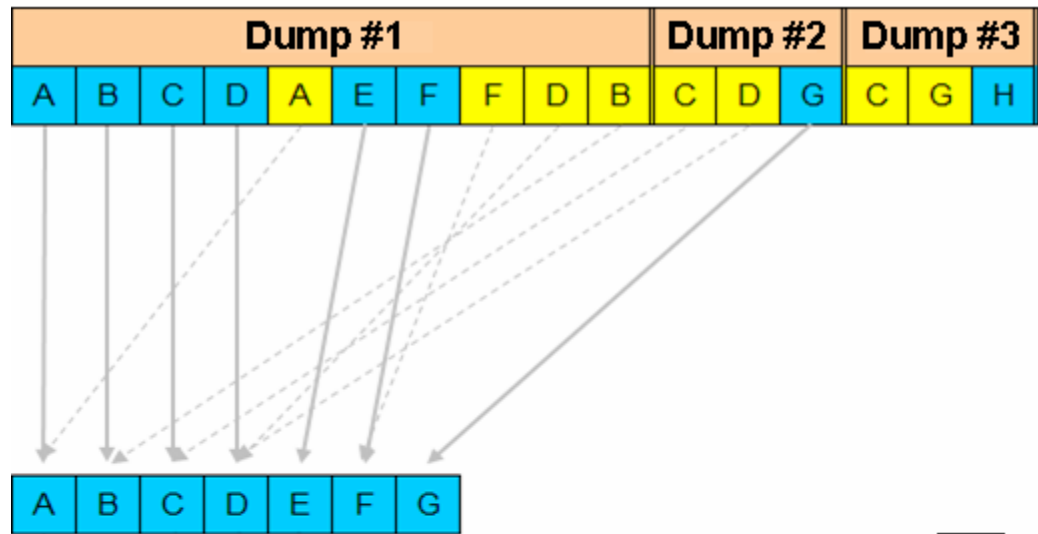
- = new unique data
- = repeat data
- = pointer to unique data segment




Deduplication Simplified



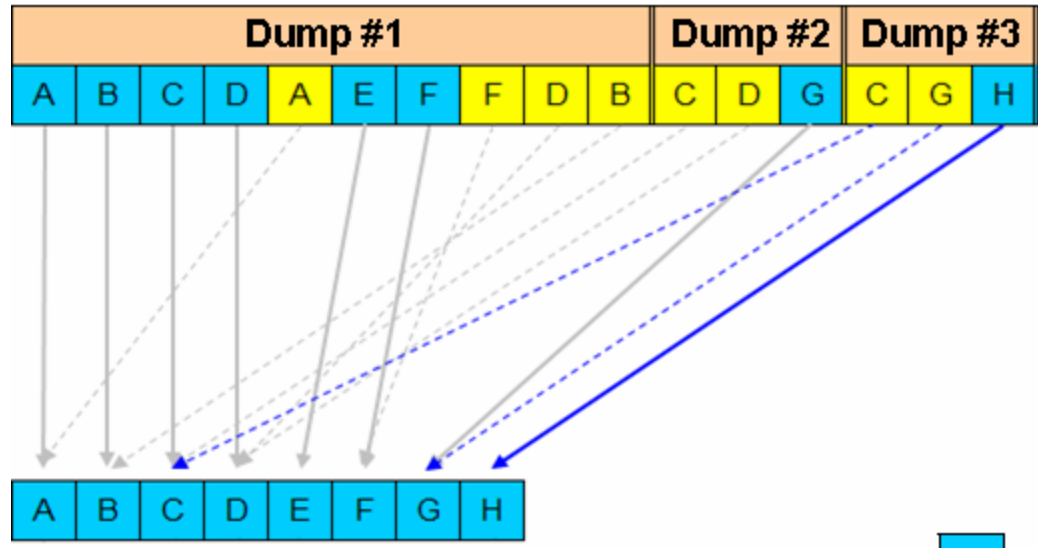
- = new unique data
- = repeat data
- = pointer to unique data segment

Deduplication Simplified



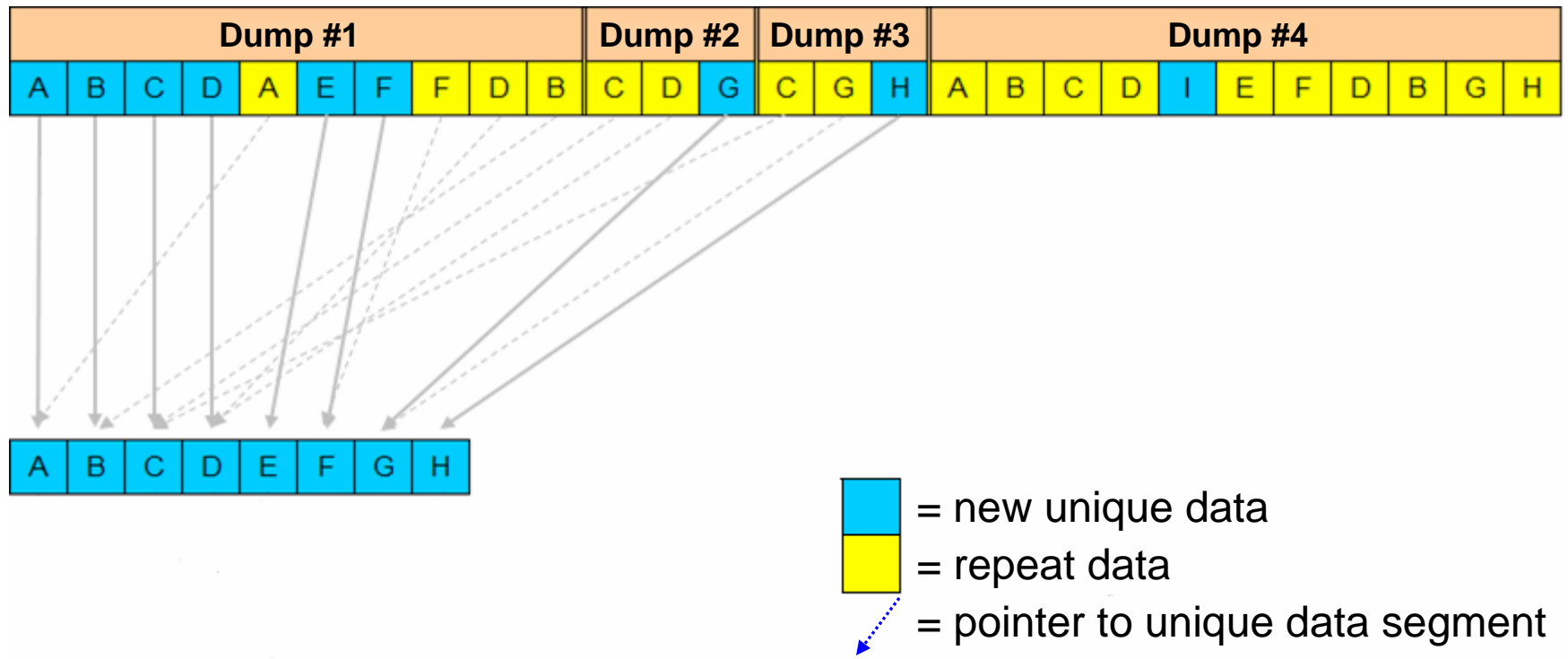
 = new unique data
 = repeat data
 = pointer to unique data segment

Deduplication Simplified

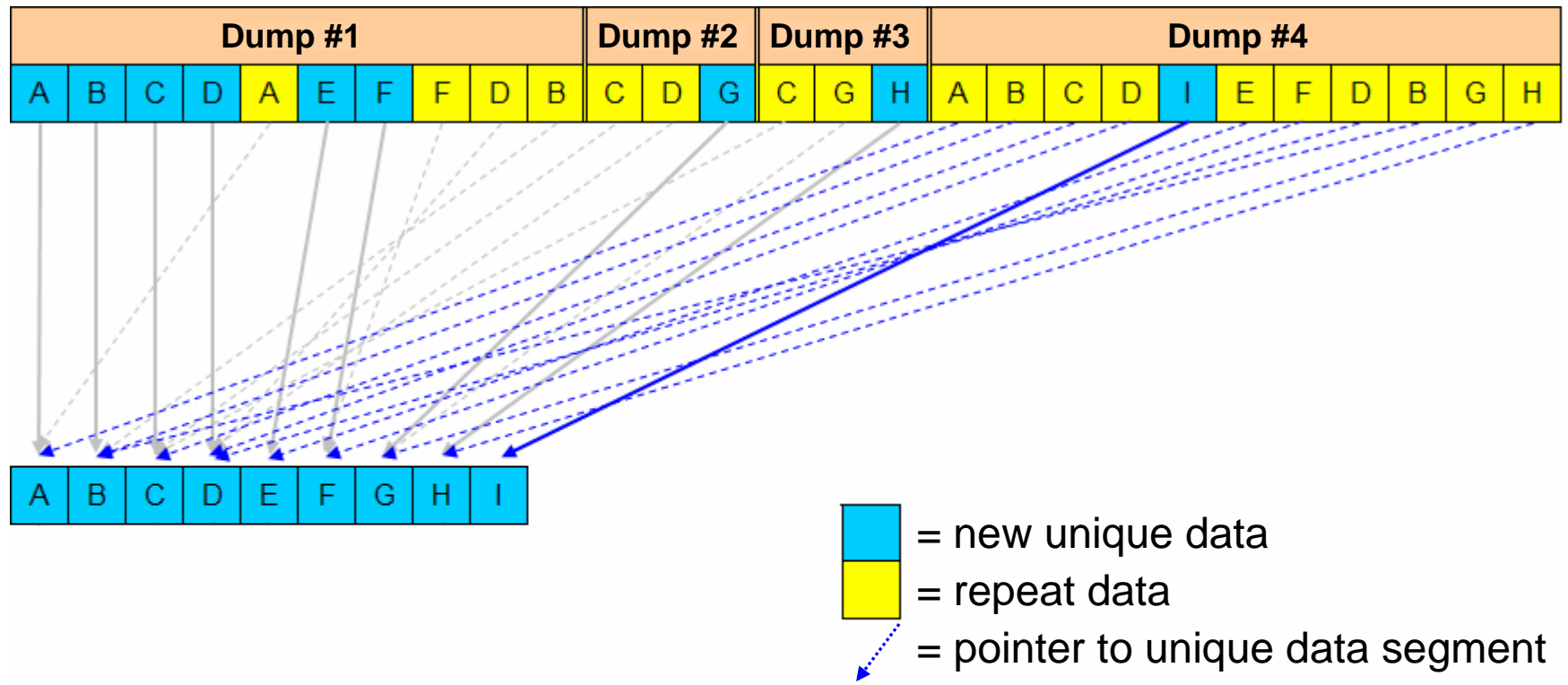


= new unique data
 = repeat data
 = pointer to unique data segment

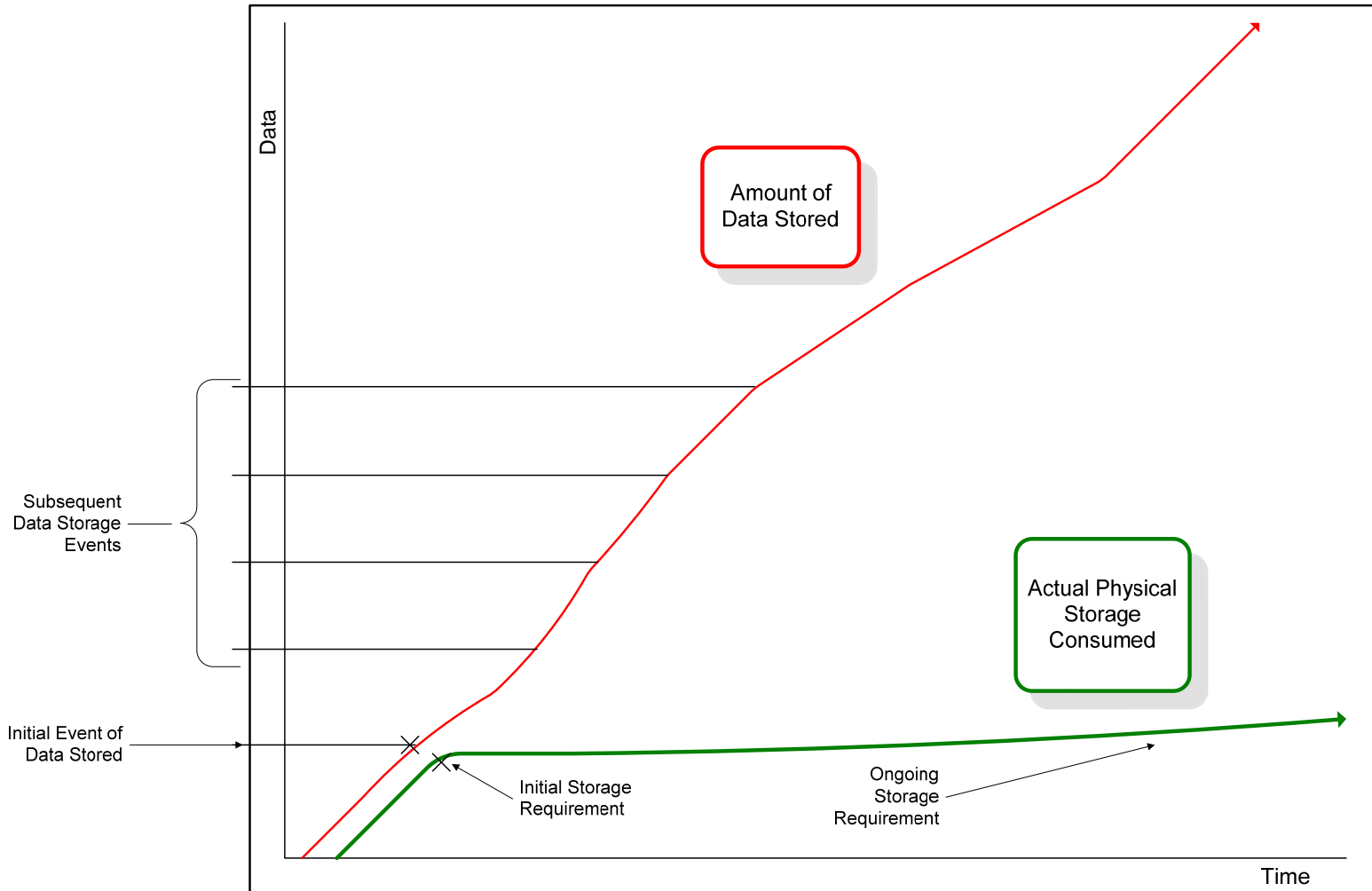
Deduplication Simplified



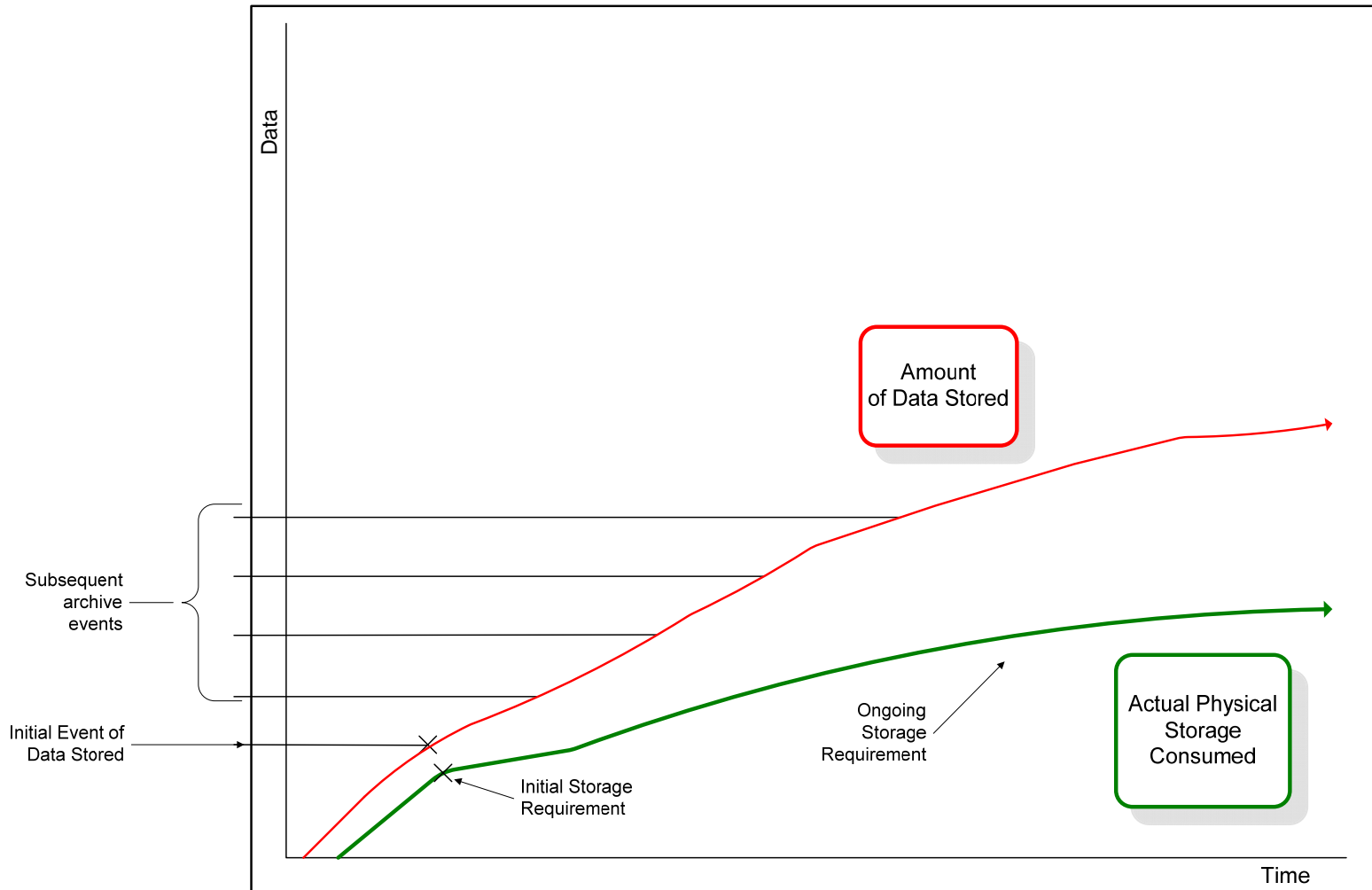
Deduplication Simplified



Deduplication of Backup Data



Deduplication of Archive



Recalling Stored Data from Deduplication

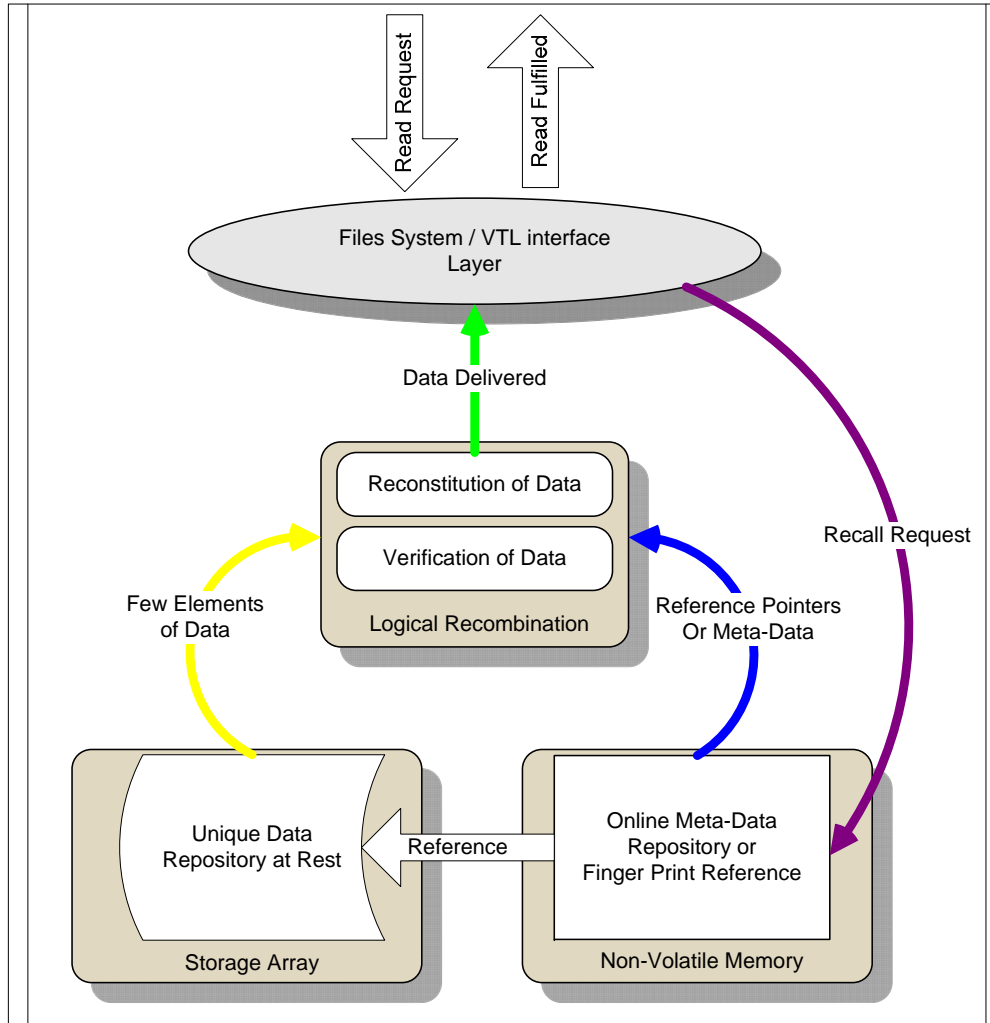
➤ Common questions:

- ◆ If deduplicated, will restore be slow?
- ◆ What happens to the redundant data?
- ◆ When recalling data (read) how does it all come back?
- ◆ Can I get back data from a day ago, a week ago, 6 weeks ago?
- ◆ What happens when a failure occurs?

➤ All great questions!

- ◆ Restore speed may vary
- ◆ Redundant data is discarded only after some reference is captured to its existence
- ◆ All data comes back by recombining unique elements of data
- ◆ You can get back anything you stored!

Recalling the Data



**Example
generic
deduplication
system
recalling data**

What to Consider

- What are the factors that will impact your results?
- Effects of different data types
- Effects of bandwidth and latency
- Math behind deduplication
- 50% deduplication = 2:1 reduction
- Application types will deduplicate differently
 - ◆ Database, image, email, home directory, CRM
- Varied deduplication results – nightly vs. average
 - ◆ 20:1, 50:1, higher? Average 5-10x, higher?

Estimating the Deduplication Ratio

- What are the factors that will impact your results?
 - ◆ Sources of redundant data
 - › Distributed copies
 - › Derivative works
 - › Temporal duplication (e.g. backup)
 - ◆ Type of deduplication
 - › Fixed or variable segment
 - ◆ Backup specific factors
 - › Retention depth
 - › Full vs. incremental backups
 - › Organic change rate of data
 - › Organic growth of data
- “Your mileage may vary!”

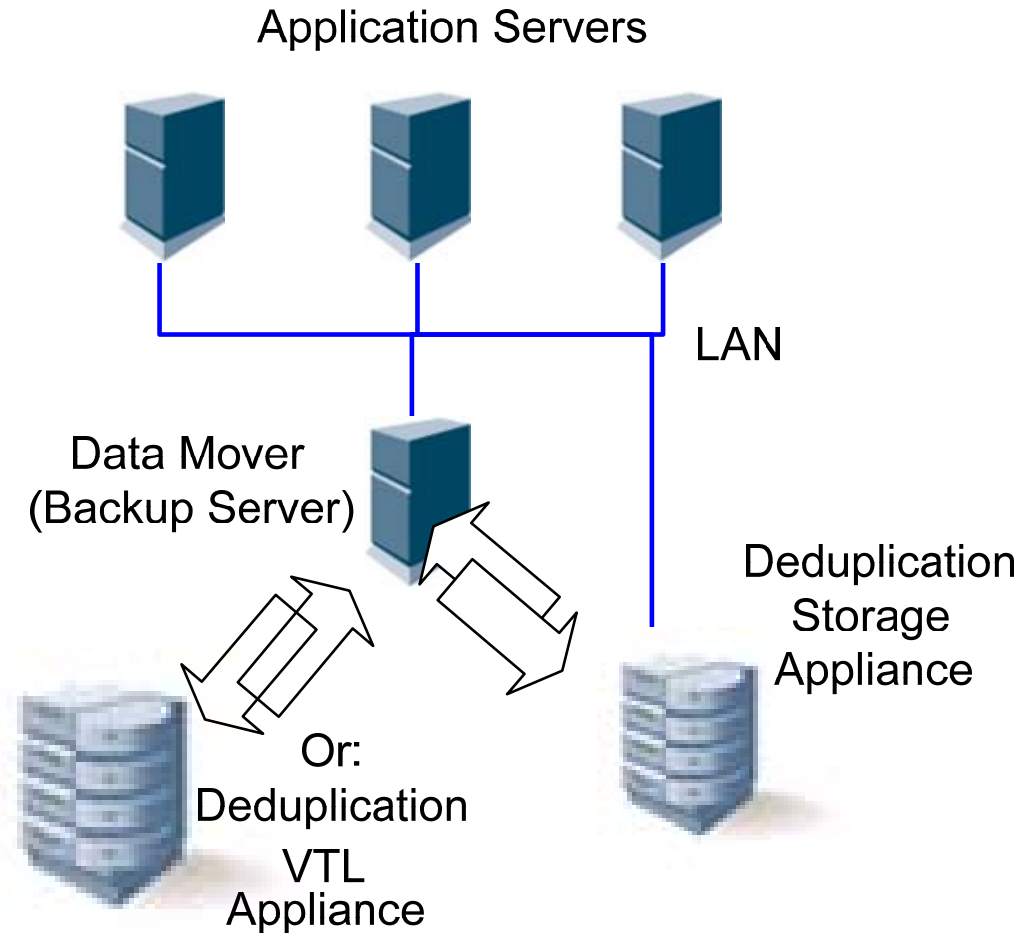
Design Approach

- Appliance
- Software
- Gateway
- System
- Grid Storage

Appliance 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 real-time (in-line) or post-process
- May include self-contained or gateway configurations
- May include other methods of data reduction in addition to deduplication, such as compression
- Allows global deduplication across all clients
- Scale of deduplication space may vary by implementation

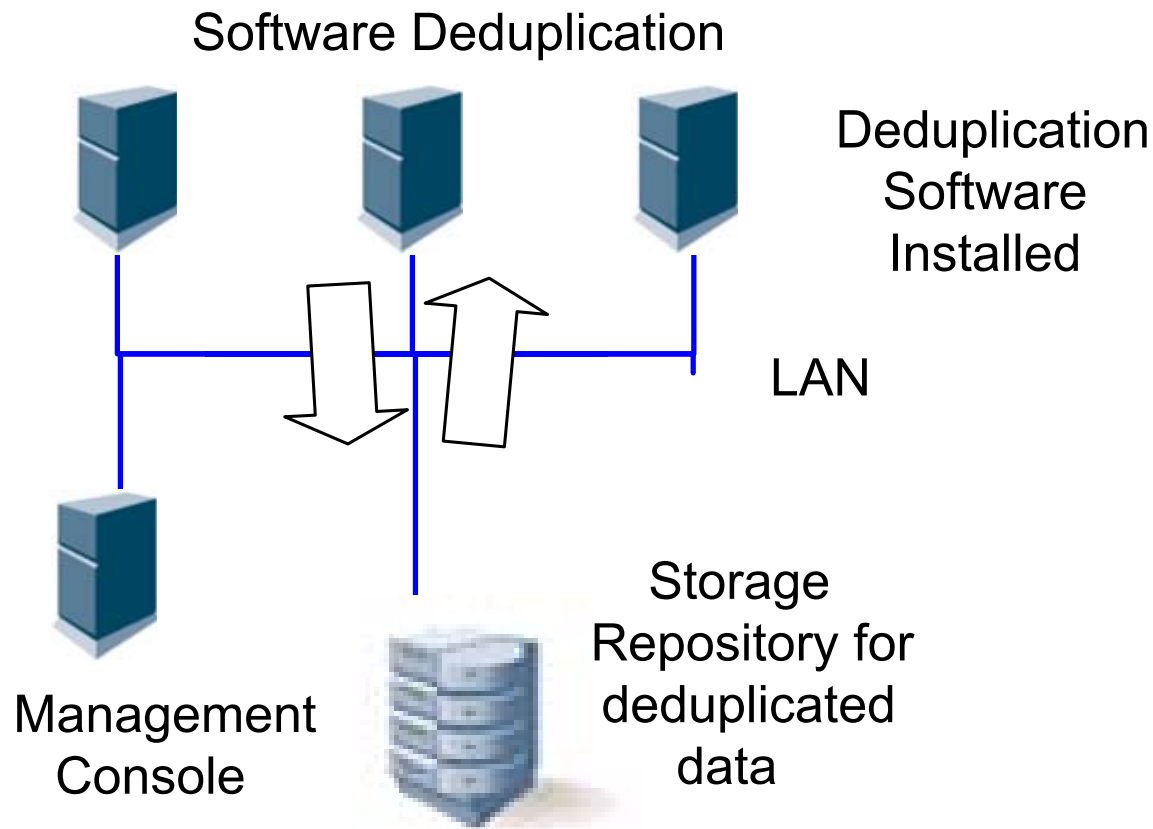
Appliance Deduplication Example



Software 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
 - ◆ Allows 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

Software Deduplication



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, ex. 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, example 4KB to 16KB

➤ Method chosen may affect deduplication results

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

Applications for Deduplication

➤ Near-line Storage

- ◆ Low cost tiered storage devices – targets for inactive data at rest

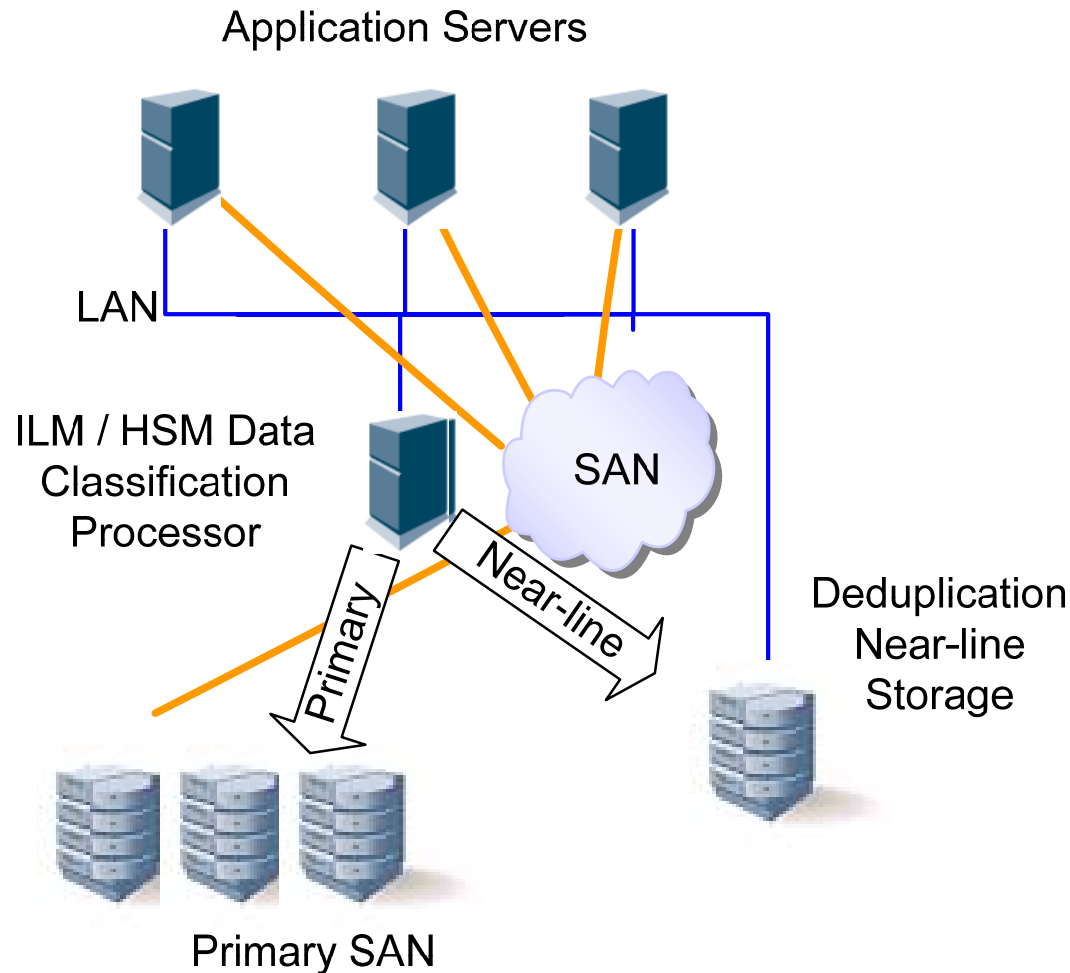
➤ Backup and Recovery

- ◆ Backup to disk efficiently with long retention - recoverability

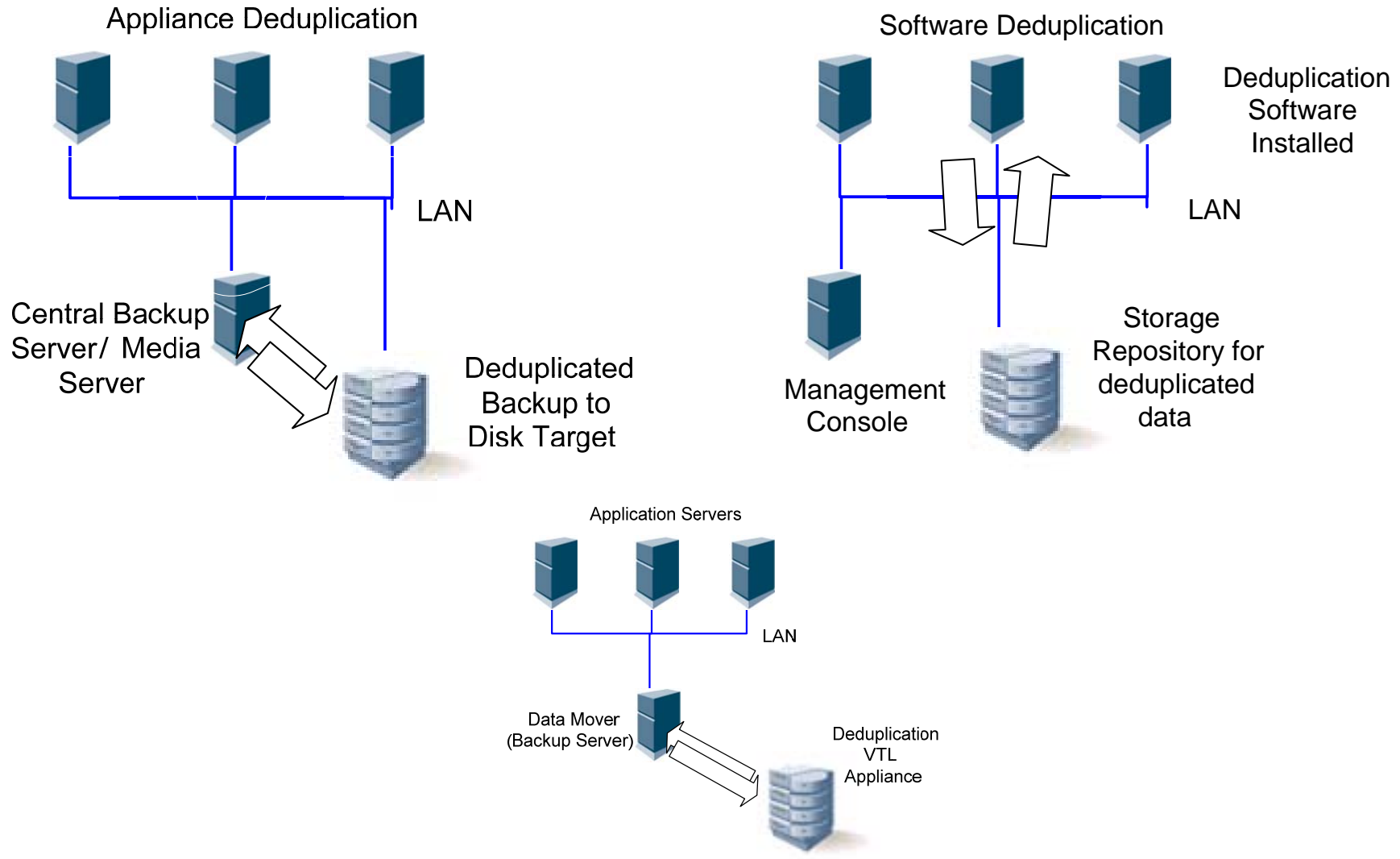
➤ Archiving

- ◆ Retention depth and very low cost disk utilization for archive requirements

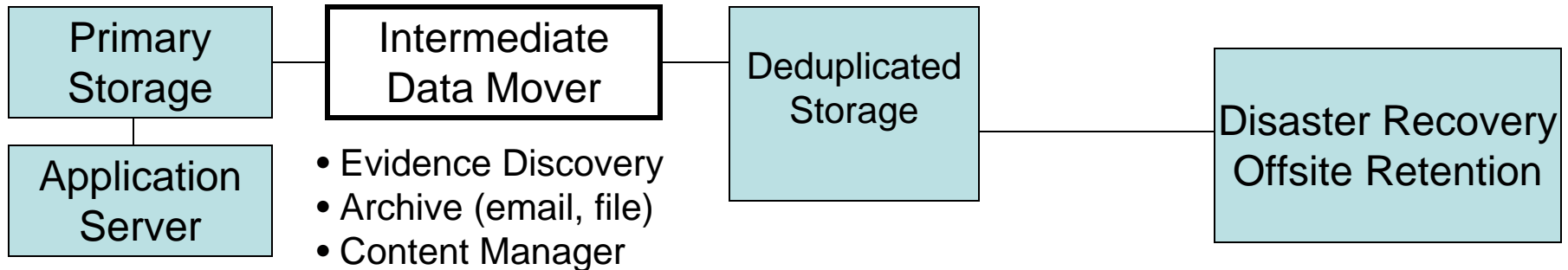
Deduplication for Near-line Use



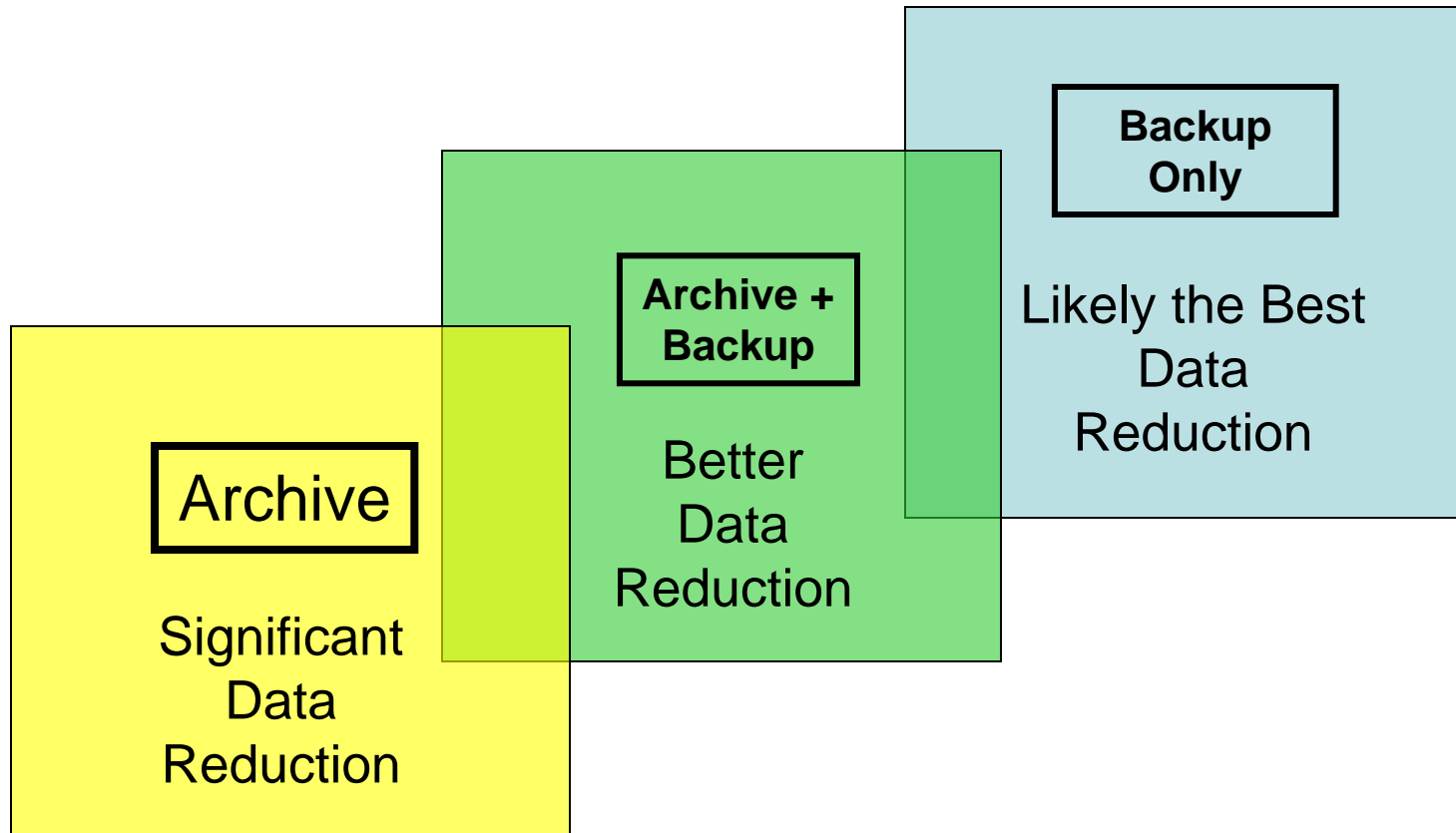
Deduplication for Backup and Recovery



Deduplication of Archive



Combined Application Benefits



Deduplication for Data Movement

➤ Disaster Recovery

- ◆ Replicate all data after deduplication for bandwidth efficiency
- ◆ Replication of deduplicated data meeting offsite requirements
- ◆ Offsite archive

➤ Bandwidth Optimization

- ◆ Increasing WAN efficiency
 - › More applications per pipe
- ◆ Branch Center Consolidation
- ◆ Backup Centralization

Potential Benefits of Deduplication

- Reduced storage cost
- Green Solution
 - ◆ Reduced footprint, power, and cooling
- Reduced network bandwidth consumption
- Reduced management time / spend (opex)
- Improved data recoverability
- Reduced dependency on tape

Potential Benefits Examined

- **Retain Data Longer / Store More Data / Send more Data**
 - ◆ In less physical space
 - ◆ Using less bandwidth
 - ◆ Consuming less time
- **Recovery / Restore from Disk**
 - ◆ Fast random access to data
- **Disaster Recovery**
 - ◆ Replication efficiency
 - ◆ Securing data offsite
 - ◆ Recovering post disaster event quickly
- **Cost savings**
 - ◆ Drive down storage and bandwidth costs
 - ◆ Hardware and software reductions
 - ◆ Management simplicity
 - ◆ Space, Power, Cooling
- **Environmental**
 - ◆ Green Solutions

Enabling Disaster Recovery

- Deduplication allows efficient local backup
- Deduplication enables efficient replication of data
- Recovery at remote DR location
- Recovery consists of standard backup / restore practices
- Deduplication allows smaller footprint for disaster recovery implementation further reducing costs
- User realizes savings in bandwidth and storage costs
- IT environment easier to manage
- Greater level of automation in data protection practices

How Much Can You Protect for DR?

Example:

T3 @ 40% Use: 200GB / Day

**Full weekly,
incremental daily:
200GB Protectable**

Full Being Replicated



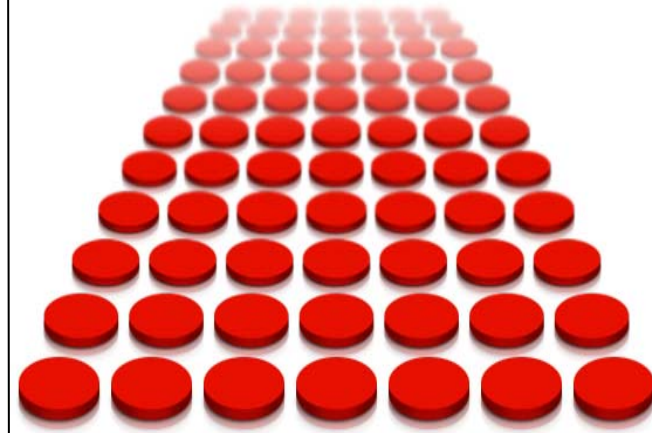
**Incrementals-only:
Over WAN
4TB Protectable**

- 5% change daily
- 200 GB / 5%



**Deduplication:
20 TB Protectable**

- 5% change daily
- 5x deduplication: 1%
- 200 GB / 1%



Choosing a solution with Deduplication

- Storage or network or both?
- Remote office solution?
- Central Solution?
- Backup to disk?
- Software or hardware?
- Long term use case?

➤ Some Applicable Use Cases for Deduplication:

- ◆ Typical production backup environment
- ◆ E-mail archive / legal discovery (EED)
- ◆ Branch office consolidation
- ◆ Disaster recovery
- ◆ Application acceleration
- ◆ Storage for or backup of Virtual Machines
- ◆ High density small footprint readily accessible storage away from primary disk
- ◆ Utilitarian data storage for test and development, code tree, & versioning

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

**Devin Hamilton
Larry Freeman
Michael Fishman
Shane Jackson
Jason lehl
Rory Bolt
Mike Dutch
Matt Brisse
Gideon Senderov
Jeff Porter**