



Education

UNDERSTANDING DATA DEDUPLICATION

Daniel J. Budiansky – Data Domain
Larry Freeman - NetApp

- 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.

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

- Overview
- How Data Deduplication Works
- Scenarios
- Q & A

Space Reduction Terminology

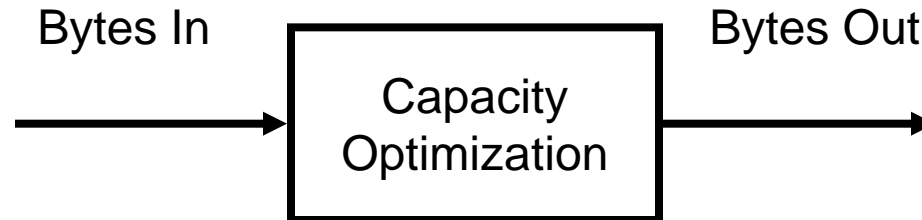
Data Deduplication is the replacement of multiple copies of data—at variable levels of granularity—with references to a shared copy in order to save storage space and/or bandwidth

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

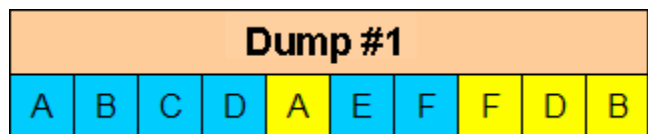
Space Reduction Ratio	Space Reduction Percent
2:1	$1/2 = 50\%$
5:1	$4/5 = 80\%$
10:1	$9/10 = 90\%$
20:1	$19/20 = 95\%$
100:1	$99/100 = 99\%$
500:1	$499/500 = 99.8\%$



- 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

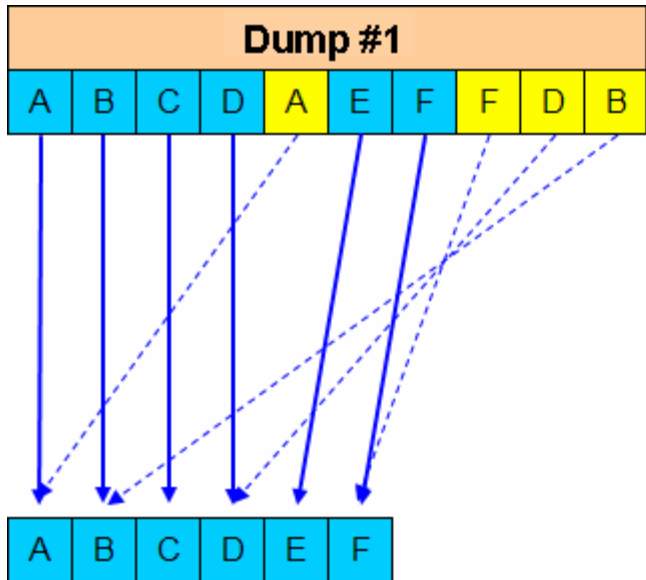
- Evaluate Data
- Identify Redundancy
- Create or Update Reference Information
- Store and/or Transmit Unique Data Once
- Read and/or Reproduce Data




Data Deduplication Simplified



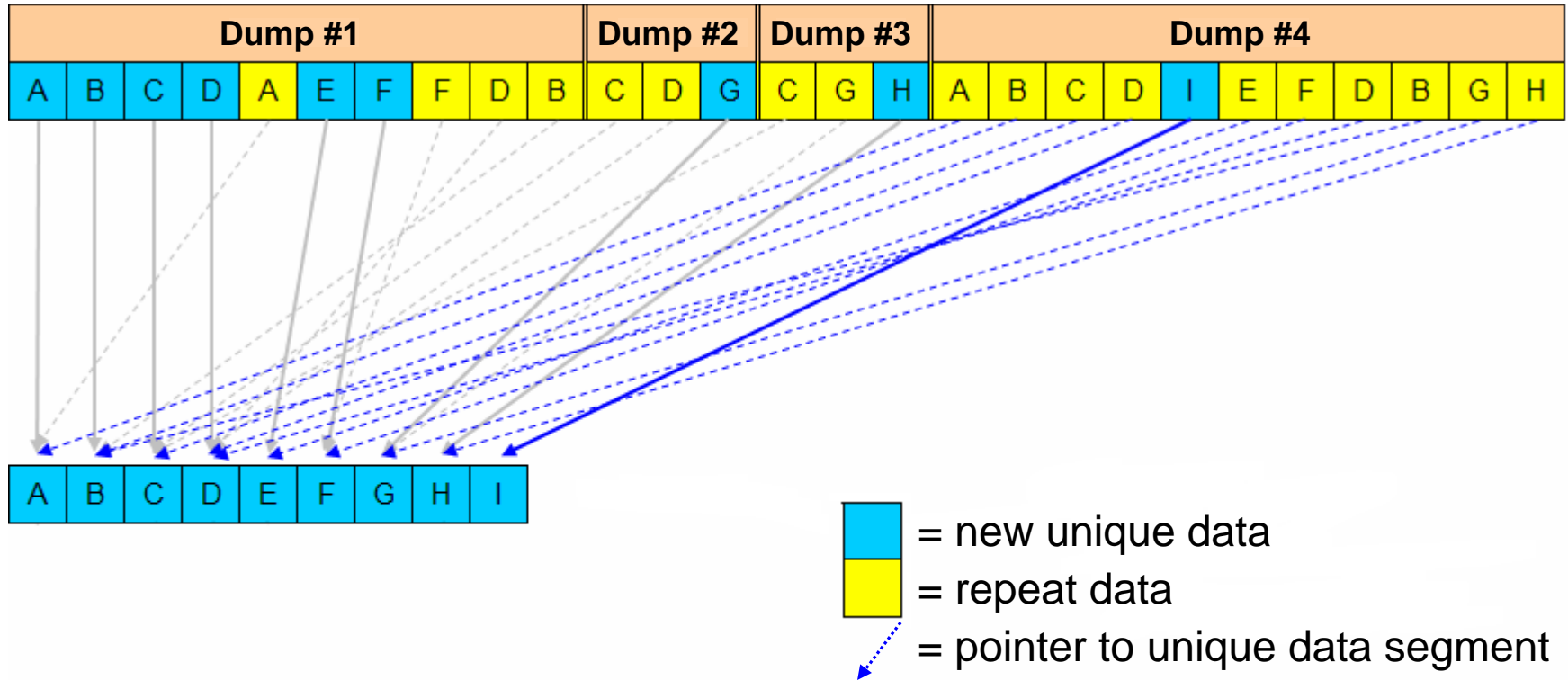
 = new unique data
 = repeat data

Data Deduplication Simplified

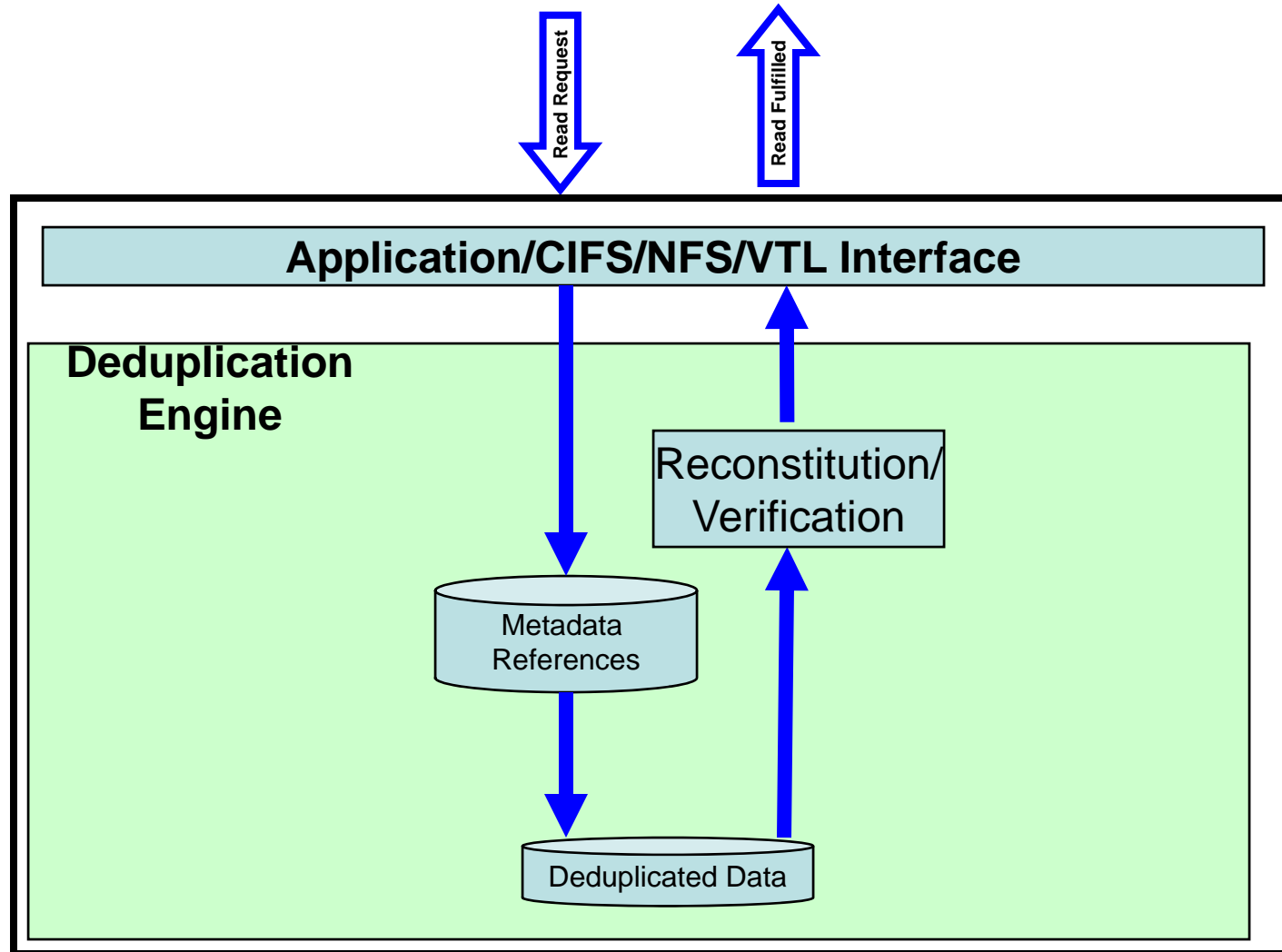


-  = new unique data
-  = repeat data
-  = pointer to unique data segment

Data Deduplication Simplified



Reading Deduplicated Data



Design Approach

➤ 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

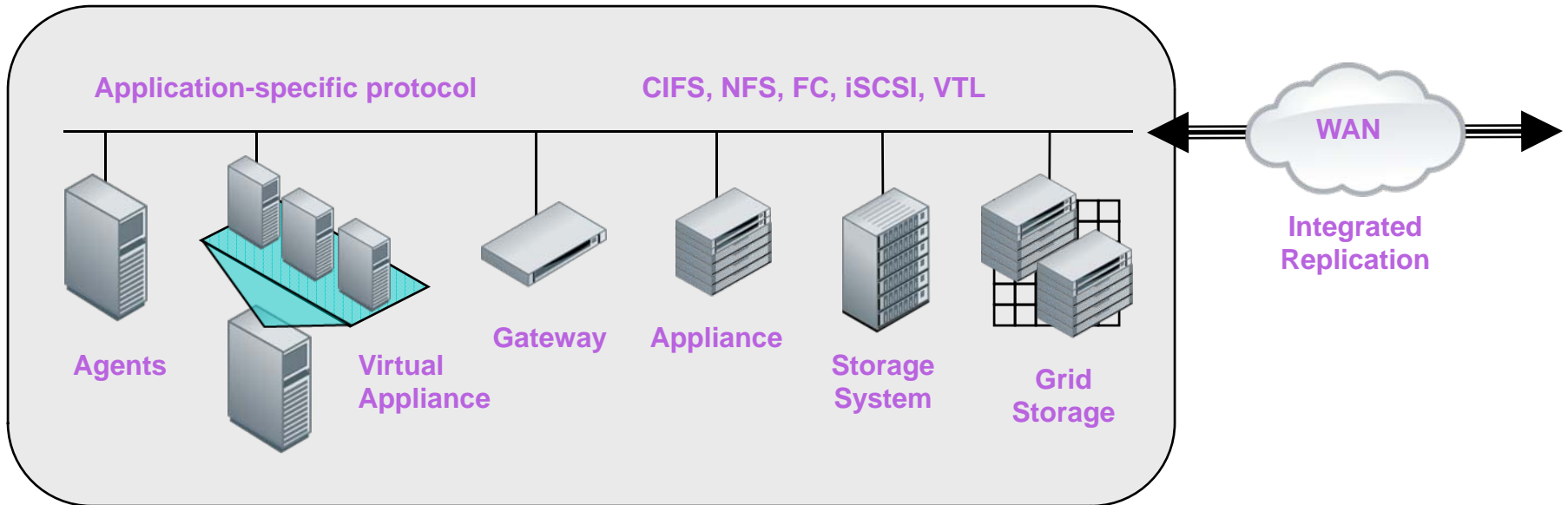
➤ Grid Storage

- ◆ 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 deduplication

- ◆ Identifies duplicate data at the source
- ◆ Transfers unique segments to a central repository
- ◆ Separate client and server components

➤ Target deduplication

- ◆ Identifies duplicate data at the target
- ◆ Stores unique segments
- ◆ Standalone system

➤ Considerations

- ◆ Neither approach inherently enables a greater or lesser space savings
- ◆ Scope of data deduplication may vary by implementation

➤ 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 – carbon footprint friendly
 - ◆ Reduce administrative costs

Scope of Data Deduplication

➤ Global Deduplication

- ◆ Data deduplication which stores only unique data across multiple deduplication systems

➤ Local Deduplication

- ◆ Data deduplication which removes duplicate data only within a single deduplication system

➤ Can span multiple applications

➤ Can span multiple datacenters

➤ Can span multiple solution components

Applications for Deduplication

➤ Backup and Recovery

- ◆ Backup to disk efficiently with long retention – recoverability
- ◆ Replication for offsite data movement

➤ Archiving

- ◆ Long-term retention for archive requirements

➤ 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
- Global Deduplication and Scope
- Deduplicated Data Resiliency
- Scalability
 - ◆ Capacity
 - ◆ Performance

Backup: Factors Impacting Space Savings

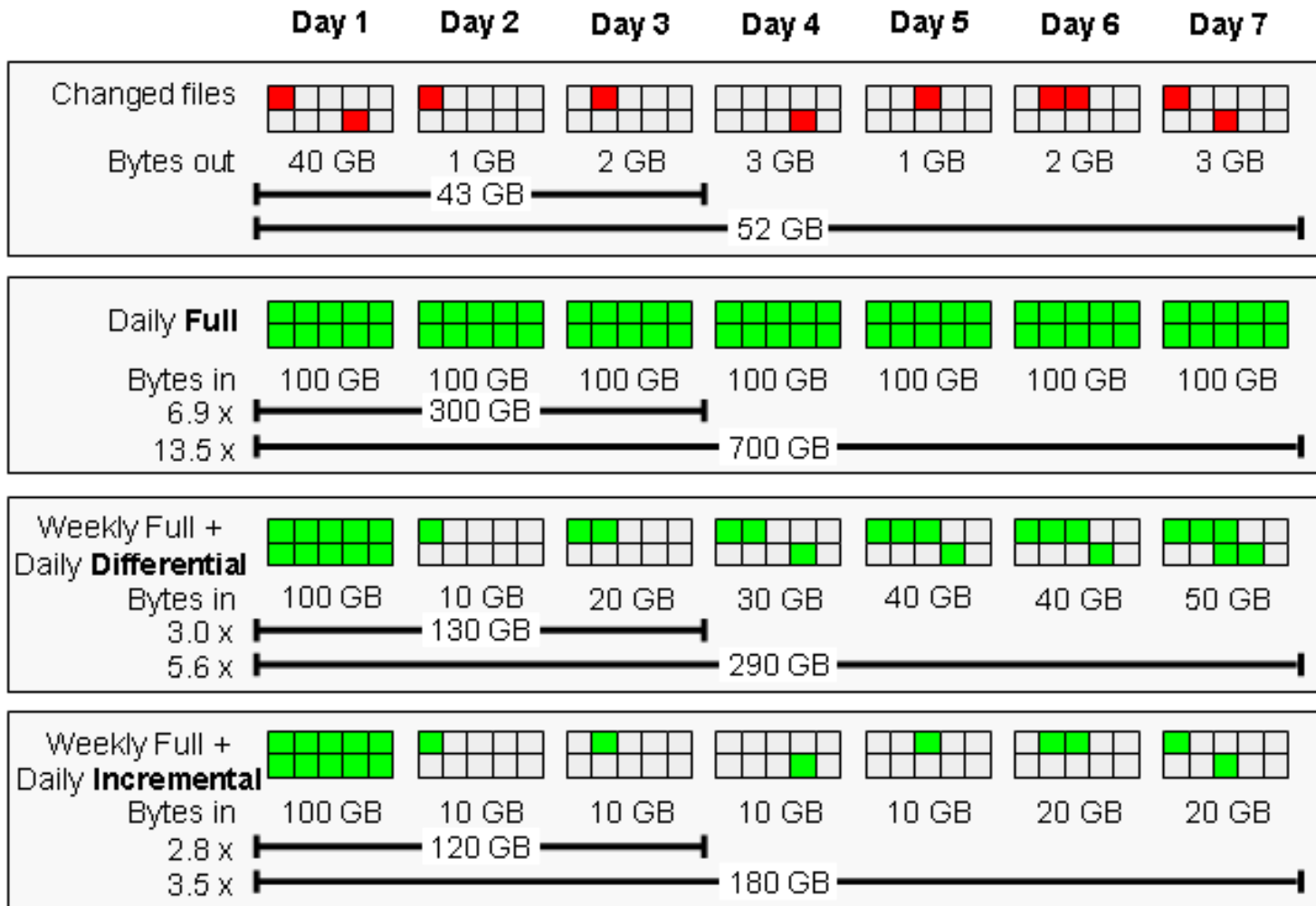
Factors associated with higher data deduplication ratios	Factors associated with lower data deduplication ratios
Data created by users	Data captured from mother nature
Low change rates	High change rates
Reference data and inactive data	Active data, encrypted data, compressed data
Applications with lower 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

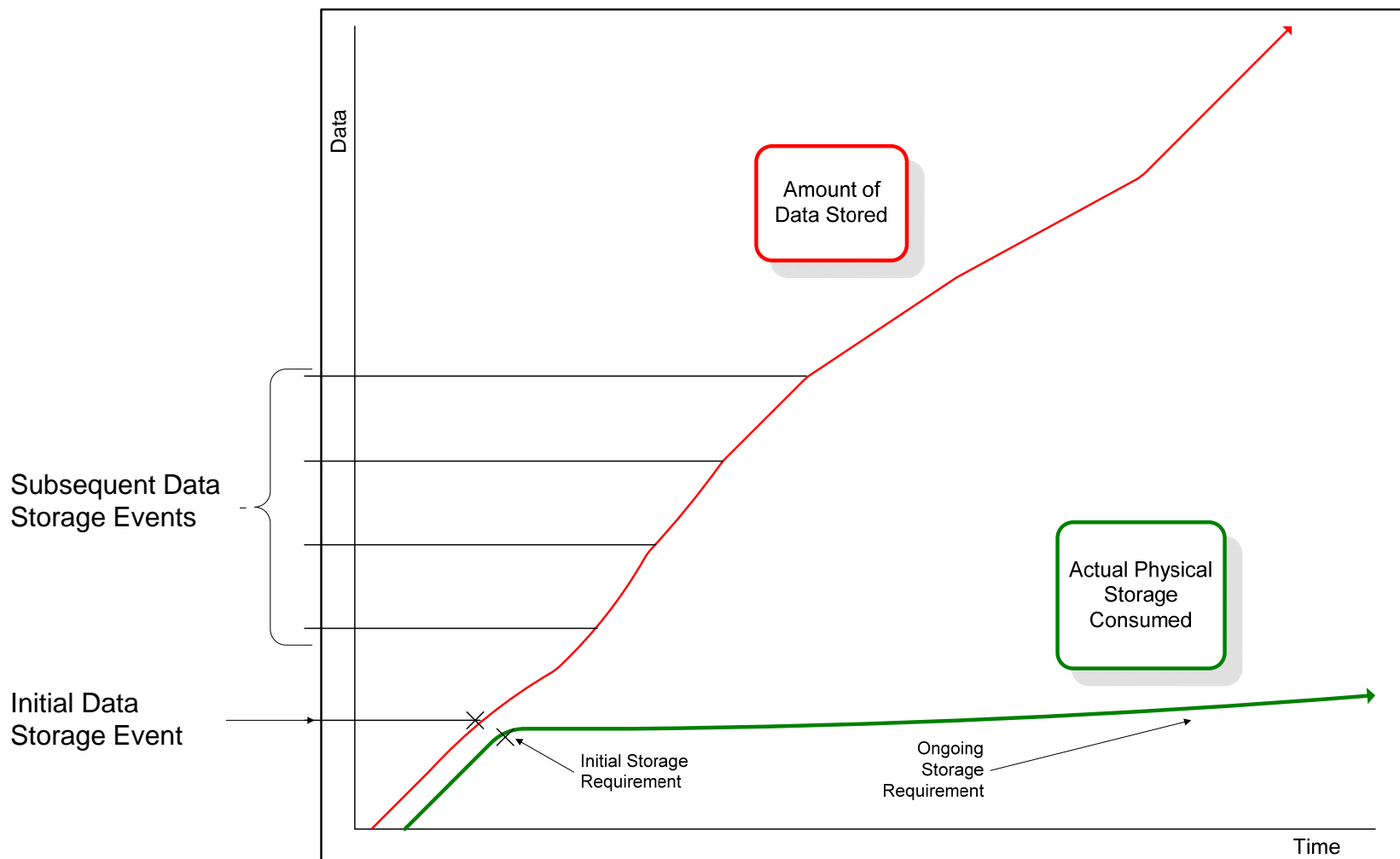


Understanding Data Deduplication Ratios

Backup: Influence of Backup Methodology



Backup: Capacity Savings Over Time



Backup: Deduplication for Data Movement

➤ Disaster Recovery

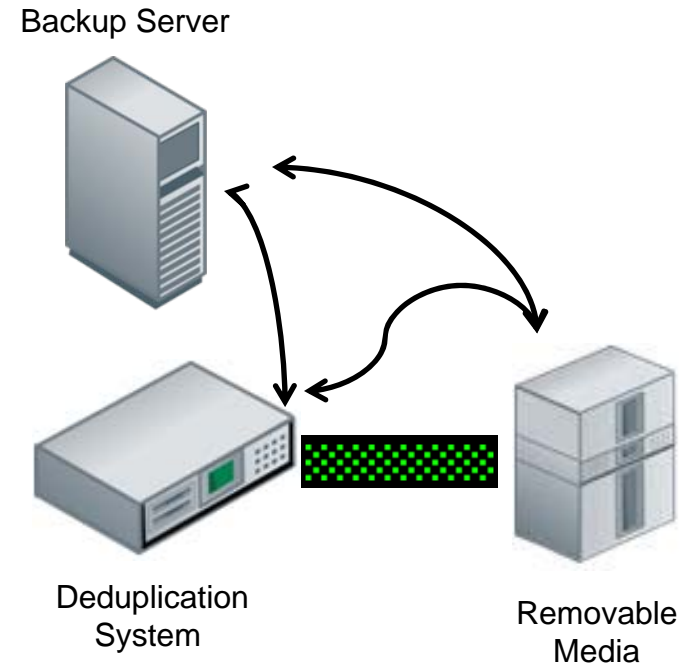
- ◆ Replicate all data after deduplication for bandwidth efficiency
- ◆ Replication of deduplicated data meeting offsite requirements
- ◆ Offsite archive

➤ Bandwidth Optimization

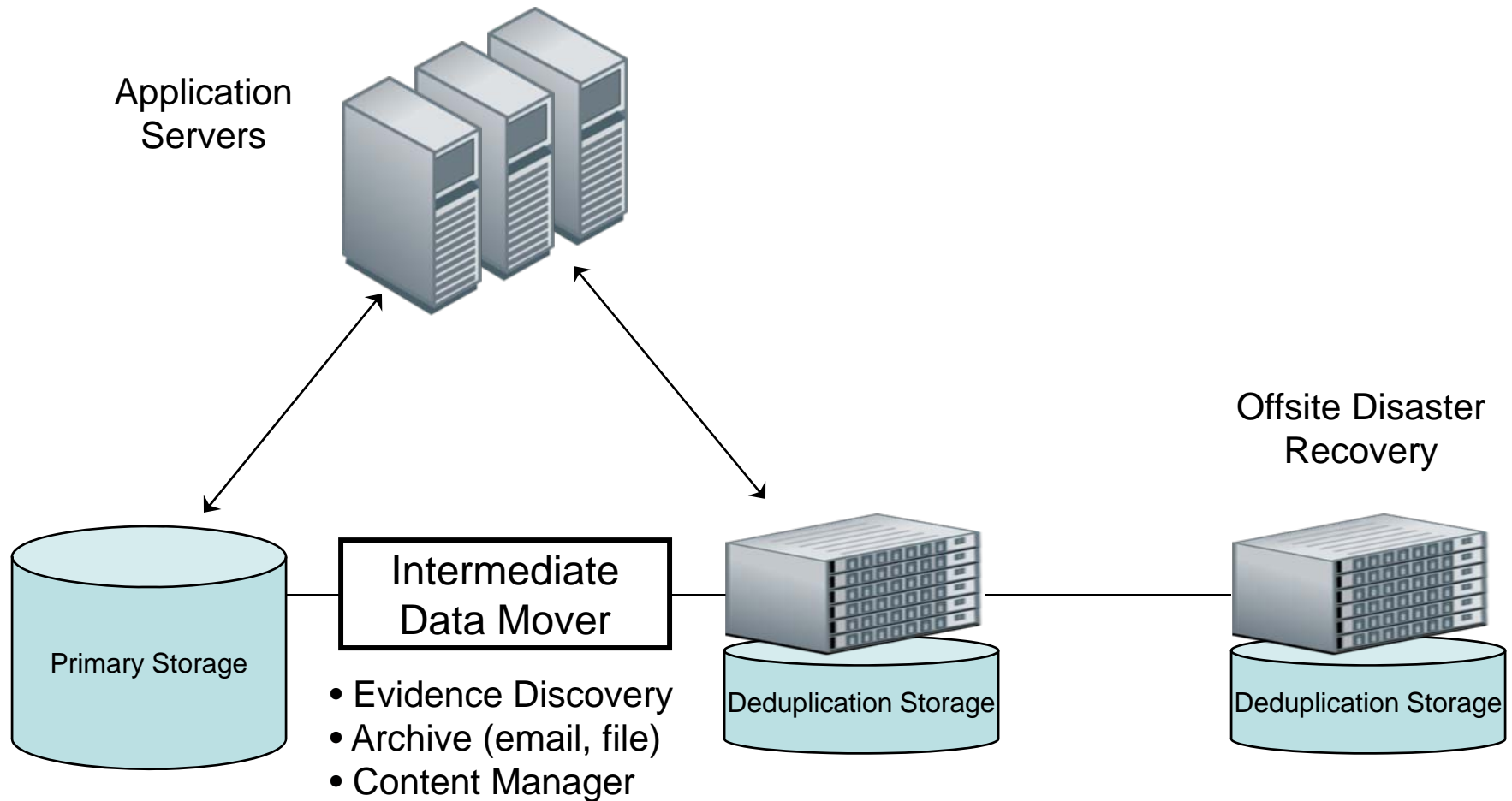
- ◆ Increasing WAN efficiency
 - More applications per pipe
- ◆ Branch Center Consolidation
- ◆ Backup Centralization

Backup: Removable Media Integration

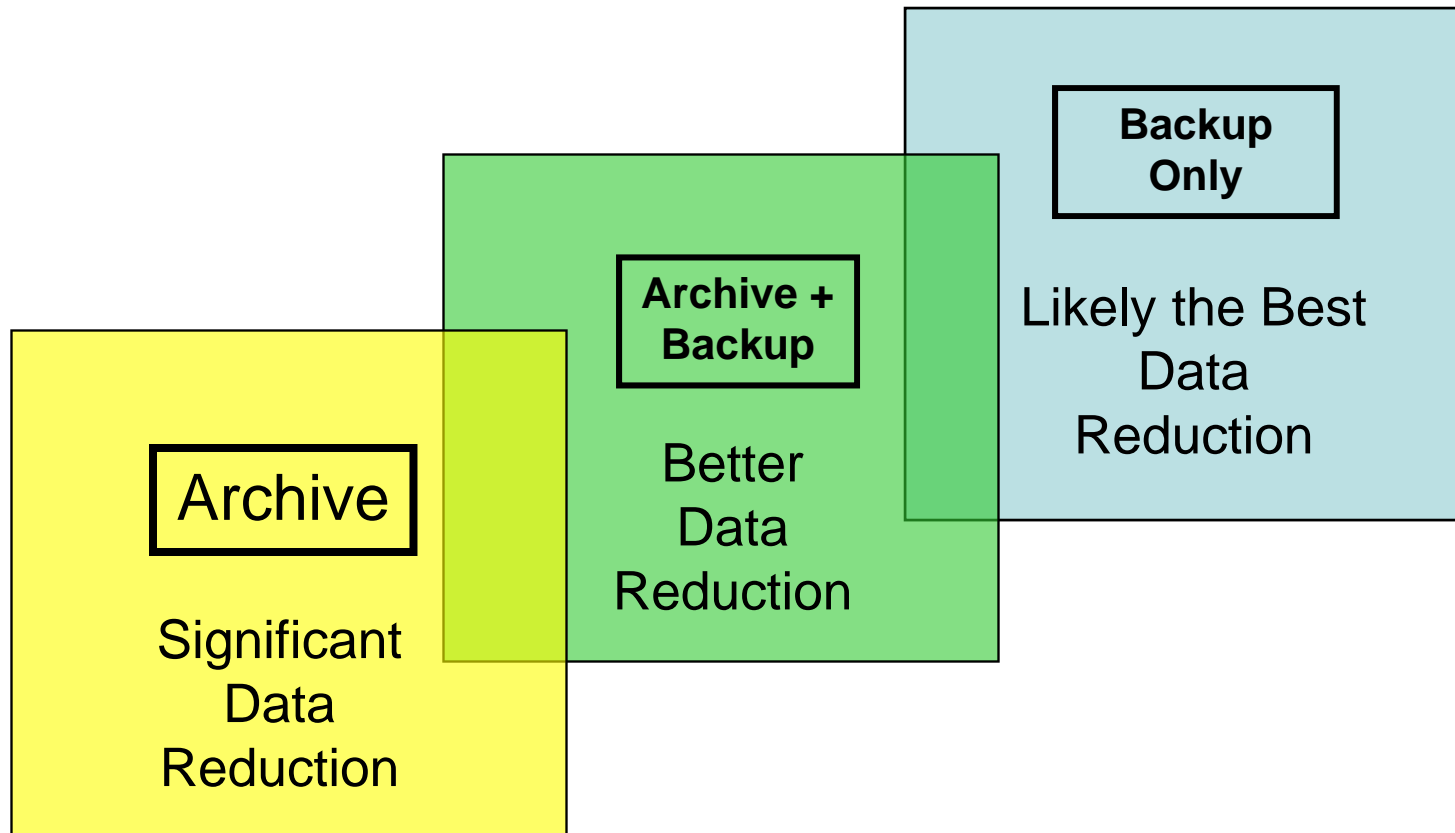
- Create long-term removable media storage for compliance and archive
- Different data path approaches
 - ◆ Path through backup server
 - ◆ Path direct from Deduplication to removable media storage



Archive: Deployment Example

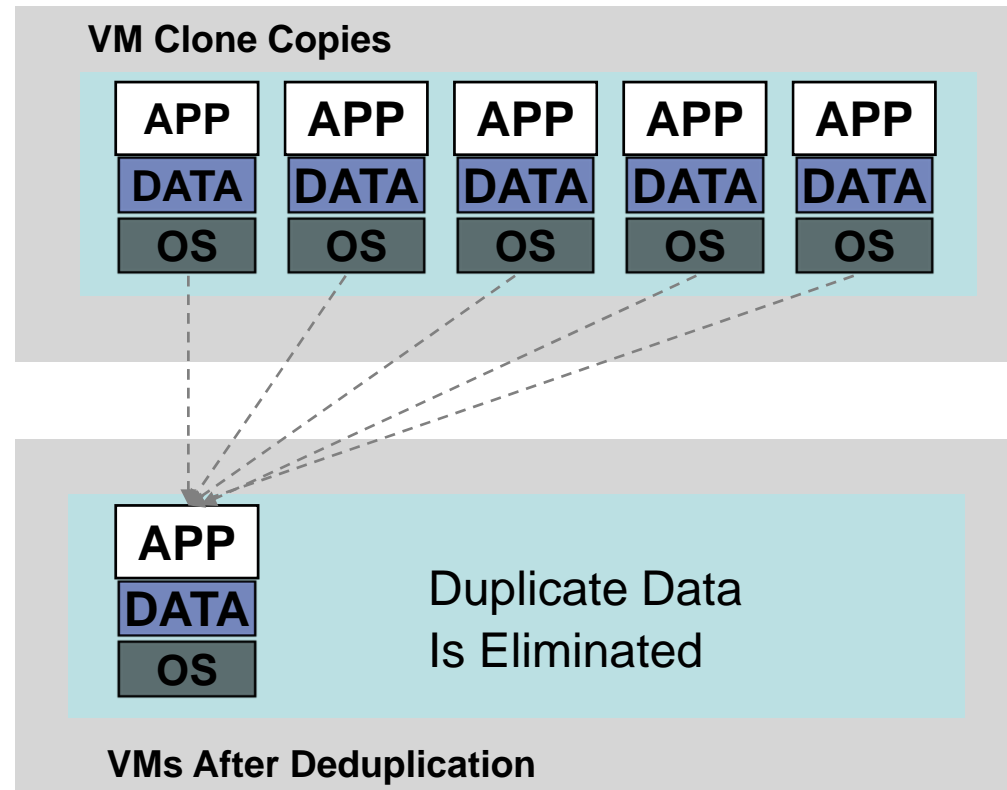


Archive & Backup Combined



Primary Storage: Deployment Example

- 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

Data Deduplication and Space Reduction SIG

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

**Jason Iehl
Shane Jackson
Toby Marek
Jeff Porter
Thomas Rivera
Tom Sas
Gideon Senderov**