Education

ADVANCED DEDUPLICATION CONCEPTS

Larry Freeman, NetApp Inc
Tom Pearce, Four-Colour IT Solutions

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 DPCO

- 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
- 2011 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

Check out these SNIA Tutorials:

- **Understanding Data Deduplication**
- **Deduplication's Role in Disaster Recovery**

Abstract

Since arriving on the scene 10 years ago, the adoption of data deduplication has become widespread throughout the storage and data protection community. This tutorial assumes a basic understanding of deduplication and covers topics that attendees will find helpful in understanding today's expanded use of this technology.

Topics will include:

- Trends in vendor deduplication design
- Practical deduplication of primary storage
- Using deduplication to reduce storage network traffic
- Pervasive deduplication across storage tiers
- Deduplication implications with storage array cache and SSD's
- Integration of Data Compression and deduplication

Terminology

Capacity Optimization Methods [Storage System]

Methods which reduce the consumption of space required to store a data set, such as compression, data deduplication, thin provisioning, and delta snapshots

Data Deduplication [Storage System]

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.

Compression [General]

The process of encoding data to reduce its size. Lossy compression (i.e., compression using a technique in which a portion of the original information is lost) is acceptable for some forms of data (e.g., digital images) in some applications, but for most IT applications, lossless compression (i.e., compression using a technique that preserves the entire content of the original data, and from which the original data can be reconstructed exactly) is required.

Vendor Deduplication Trends

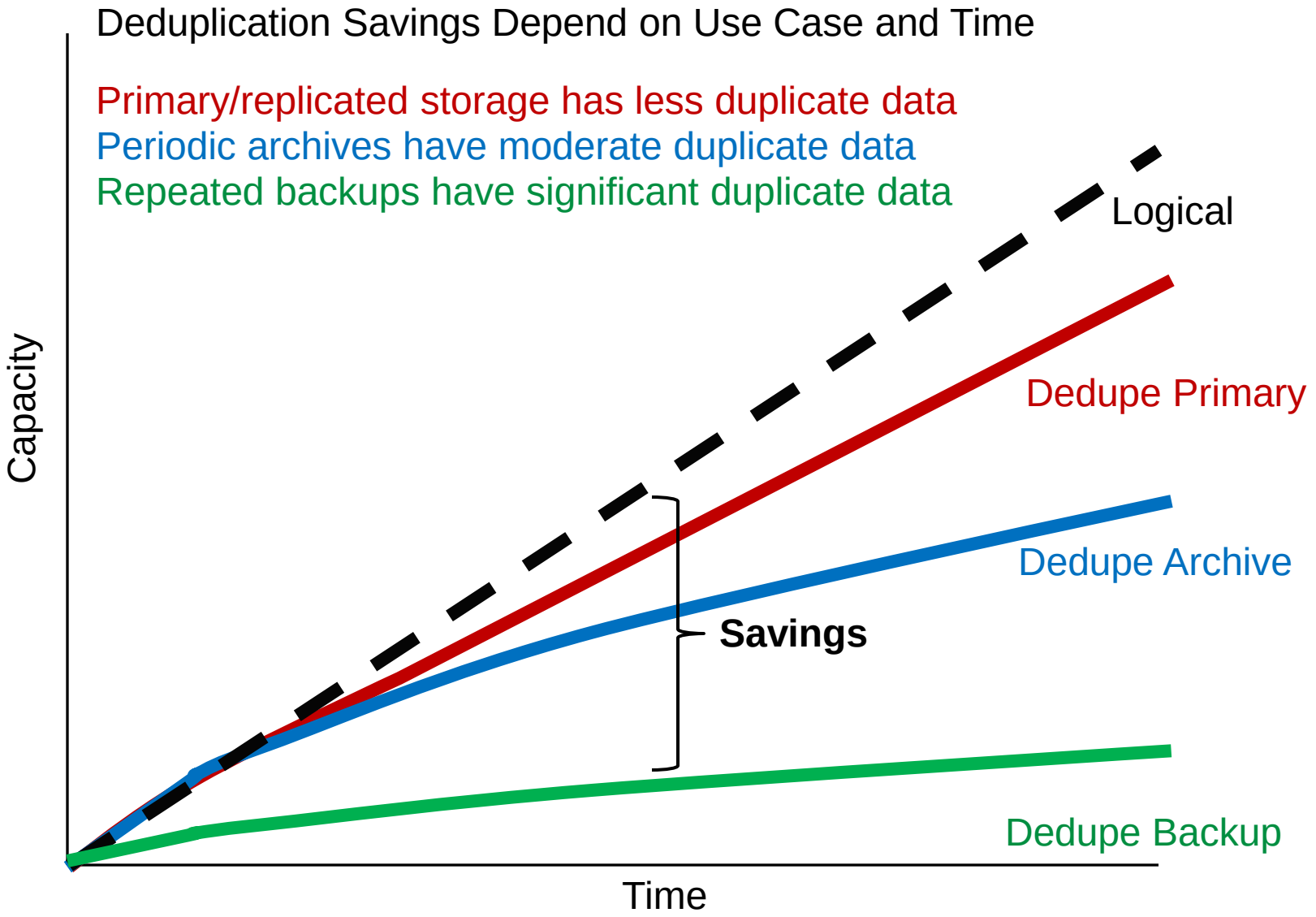
- Original value and justification has not changed:
  - Satisfy ROI/TCO requirements
  - Manage data growth
  - Increase efficiency of storage and backup
  - Reduce overall cost of storage
  - Reduce network bandwidth
  - Reduce operational costs including:
    - Infrastructure costs required space, power and cooling
    - Movement toward a greener data center
  - Reduce administrative costs

Deduplication Scope

The scope of deduplication is broadening:

- **Primary Storage**
  - Reduced capacity requirement for storage of active data
- **Replication**
  - Reduced capacity requirement for DR and business continuity
- **Data Protection**
  - Reduced capacity requirement for backup and/or longer retention periods
- **Archivals**
  - Reduced capacity requirement for data retention and preservation
- **Movement/Migration of data**
  - Reduced bandwidth requirements for data-in-transit

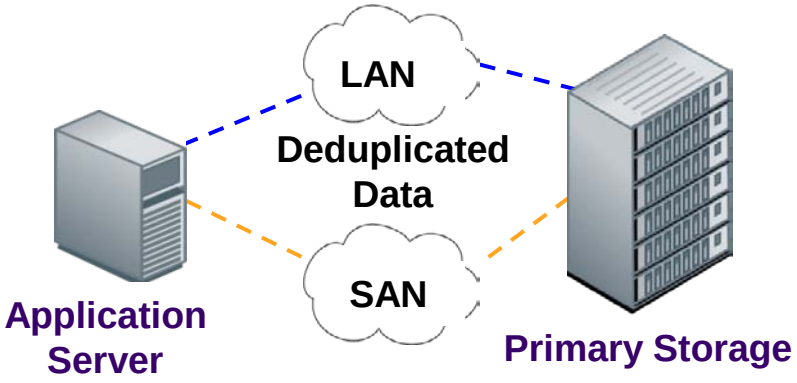
Deduplication Savings Expectation



Deduplication and Primary Storage

Effective for specific workloads

- Acceptable performance
  - Post-processing
  - Inline ingestion
- Primary storage applications with high data redundancy
  - Virtual servers and desktops
  - Collaborative file "sharing"
  - Email (Software SIS replacement)



Primary Storage Array Cache/SSD

Array Cache

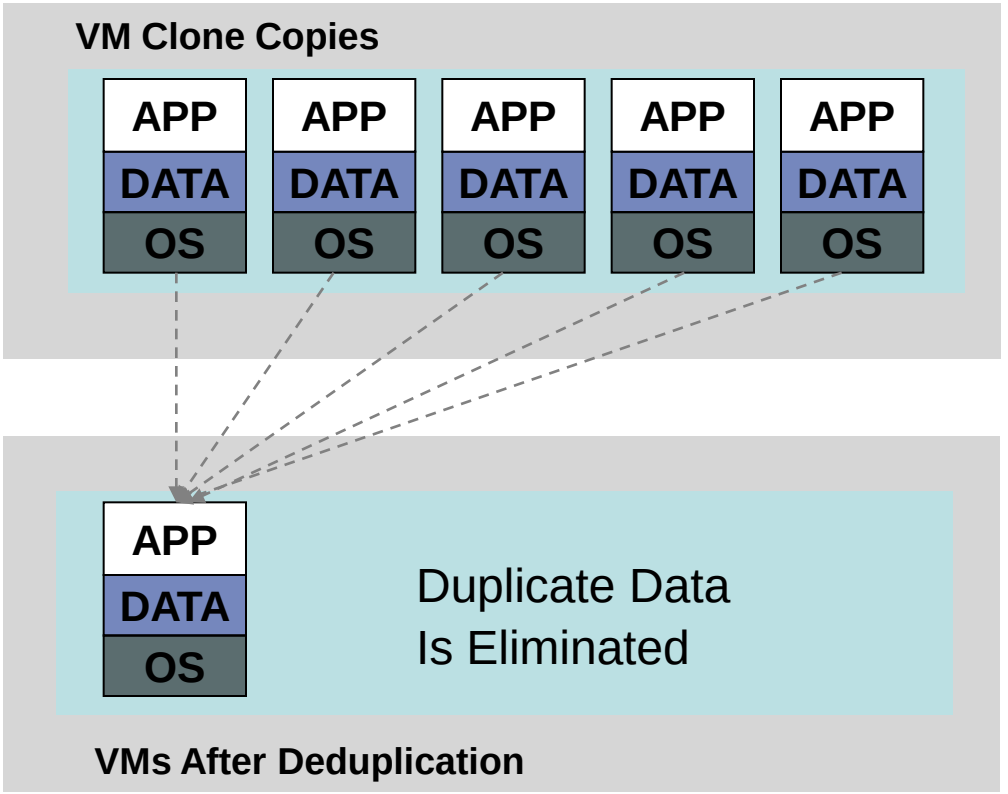
- Intelligent cache can be "dedupe-aware"
- Hot data is cached with dedupe attributes
- Reduces rotating media latencies
- Virtual Desktop "boot storms" is one example

Solid State Drives

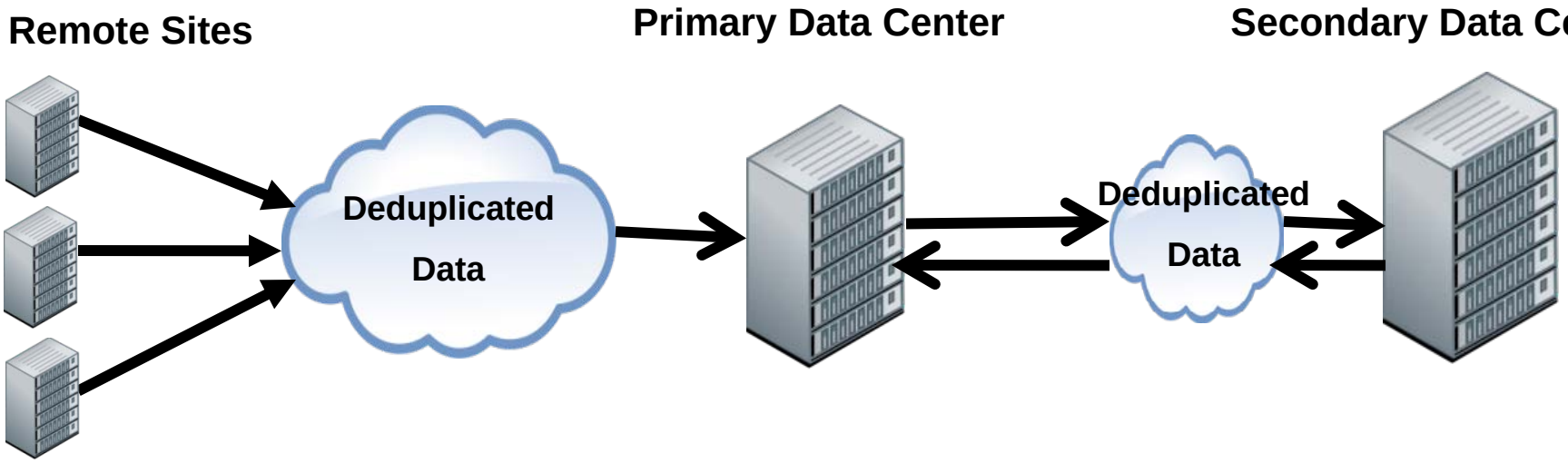
- Deduplication helps offset the higher cost/GB of SSD's
- High performance applications with highly redundant data will benefit

Primary Storage Considerations

- Balance the tradeoff between cost savings and performance impact
- Walk before you run
  - Use estimation tools
  - Perform POCs
  - Implement one workload at a time



Deduplication and Replication



- Can be one-way, two-way, multi-node, or multi-hop
- Deduplication points can be determined by Users
- Bandwidth cost savings / smaller pipes
- For very small pipes, dedupe can make replication possible

Replication Considerations

- Focus on your Service Level Agreements (SLAs) first
  - Needs to meet window for *Replication*
  - Needs to meet SLA for System *Recovery or Data Restore*
- Is it Necessary to Dedupe All Data?
  - Mission-critical applications
  - May have regulatory issues for some data
  - Some data types not conducive to deduplication

Deduplication and Backups

The Original Promise

- Faster data recovery from disk
- Reduction in D2D cost per terabyte stored
- Reduction in D2D backup storage footprint
- Less network bandwidth required for D2D backups

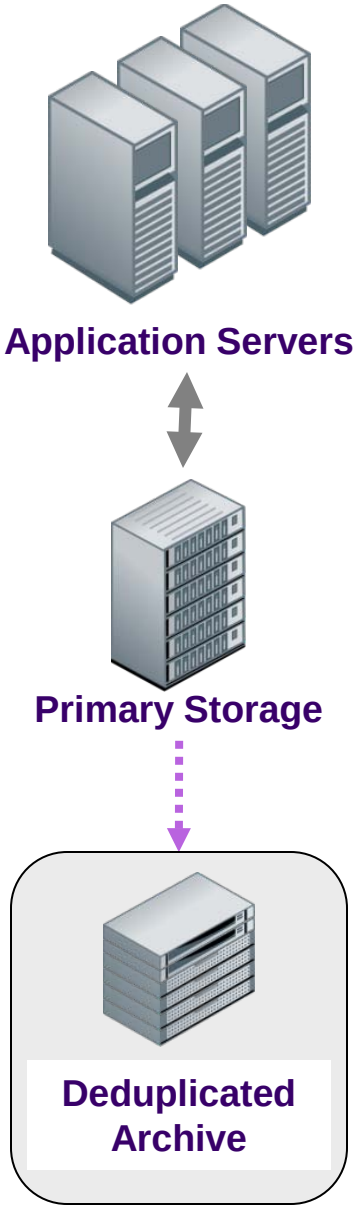
What's New?

- Deduplication embedded in backup software
- Scalability of deduplication appliances
- Deduplication across appliances
- Deduplication to tape

Backup Considerations

- Hardware or software deduplication?
  - Workload burden placement e.g. a hardware appliance or one or more server/storage devices in software.
- Source or target deduplication?
  - Locates the work-load. Dedupe at source conserves bandwidth.
- Variable or fixed-length deduplication?
  - Describes the segmentation of data for evaluation and influences the ability to detect data changes.
- File or sub-file deduplication?
  - Sub-file is more granular, has potentially further reductions but must be consistent with compliance requirements.
- Answers depend on the problem you are trying to solve

Deduplication and Archival

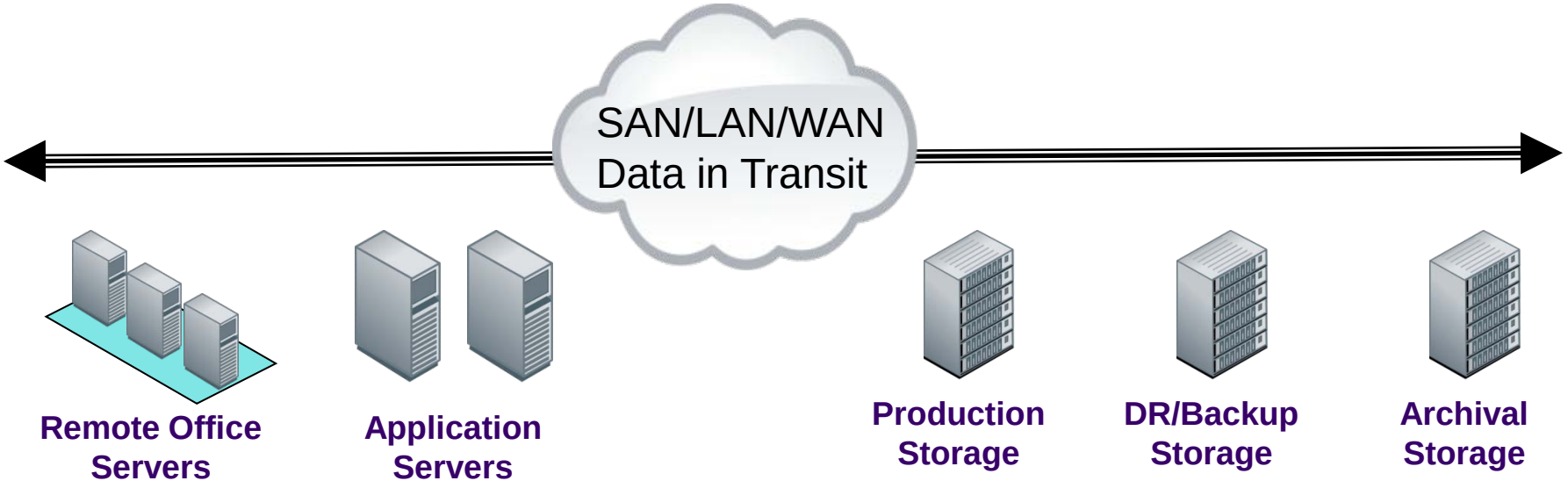


- Deduplication can reduce the cost of online archive repositories
- These repositories are often required for regulatory compliance
- No standard exists today for "approved" use of data reduction techniques with regulatory data
- Vendors should provide assurances that the ability to retrieve data in its original form is not impaired

Deduplication and Compression

- Dedupe and compression are similar
  - Both are dependant on data patterns
  - Both consume system resources
  - Both can optimize required storage capacity
- Dedupe and compression are different
  - Some data can only be optimized via dedupe
  - Some data can only be optimized via compression
  - Some data can be optimized via dedupe **and** compression
  - Some data cannot be optimized at all
- Dedupe and compression are complementary
  - But some knowledge about the data pattern is required

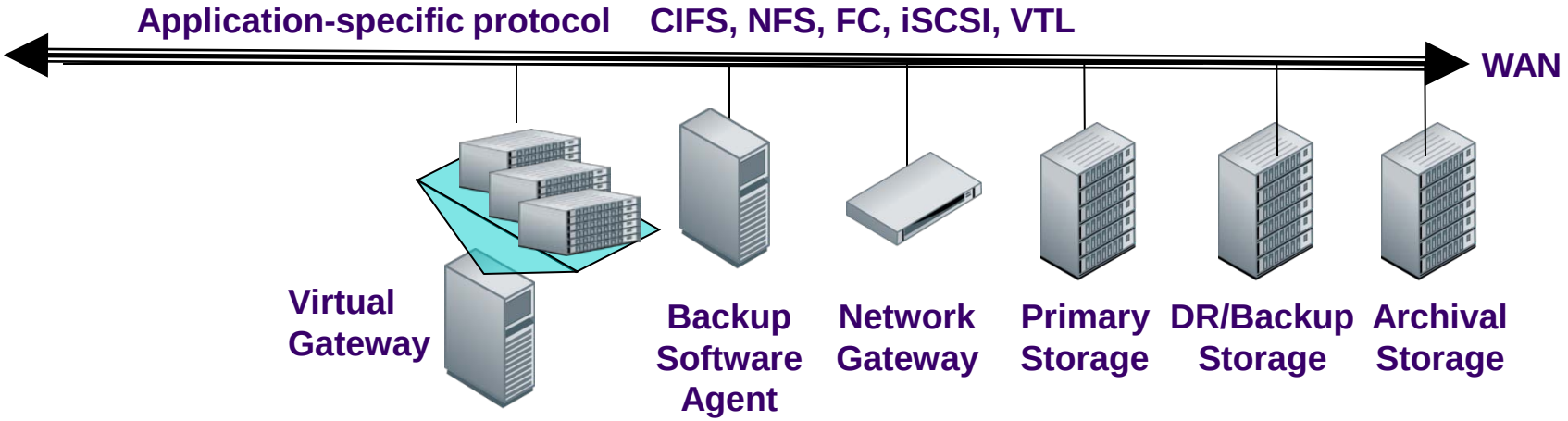
Deduplication and Network Traffic



Increased SAN/LAN/WAN Efficiency

- Transfer data references instead of data objects
- Shorten data transfer times by sending less data
- Increasing WAN efficiency
  - More applications per pipe

Consideration Matrix



Reduce Network Traffic		✖	✖		✖	
Reduce Physical Capacity	✖	✖		✖	✖	✖
Reduce Backup Time		✖				
Reduce Recovery Time		✖			✖	
Reduce Replication Time			✖	✖		
Reduce Media Latency	✖			✖		

Summary

- The original value of deduplication has not changed
- The scope of deduplication is broadening
  - All storage tiers
  - Bandwidth reduction
  - Regulatory data
- New use cases and new technologies bring new challenges
  - And new opportunities

Refer to the Hands-On Lab

Hands-On LAB COMPUTERWORLD SNIA SNW

Check out the Hands-On Lab

Data Deduplication

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 Protection and Capacity Optimization (DPCO) Committee:

Mike Dutch
Larry Freeman
Gene Nagle
Tom Pearce
Thomas Rivera
Tom Sas

SNIA Data Management Forum

It's easy to get involved with the DPCO !

- Find a passion
- Join a committee
- Gain knowledge & influence
- Make a difference

www.snia.org