
# Advanced Data Reduction Concepts

Ronald Pagani, Jr.
Open Technology Partners

Author:
SNIA - Data Protection & Capacity Optimization (DPCO) Committee

# SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- 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 Committee

- 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
  - Online DPCO Knowledge Base: www.snia.org/dpco/knowledge
  - Online Product Selection Guide: http://sniadataprotectionguide.org
- 2016 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 at** www.snia.org/education/tutorials

- **Introduction to Data Protection**
- **Trends in Data Protection**
- **Protecting Data in the Big Data World**
- **Privacy vs Data Protection – The impact of EU Legislation**
- **Encryption: What, Where & Why**

# Abstract

*Since arriving on the scene ~20 years ago, the adoption of data reduction has become widespread throughout the storage and data protection community. This tutorial assumes a basic understanding of data reduction techniques and covers topics that attendees will find helpful in understanding today's expanded use of this technology.*

Topics will include:

- Trends in data reduction design and usage
- Practical data reduction of primary storage
- Using data reduction techniques to reduce storage network traffic
- Pervasive data reduction across storage tiers

# Terminology from SNIA Dictionary

**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.

# Data Reduction Trends

- The value of data reduction technologies has not changed:
  - Satisfy ROI/TCO requirements
  - Manage data growth
  - Increase efficiency of storage and backup
  - Reduce overall cost of storage
  - Reduce network bandwidth requirements
  - Reduce operational costs including:
    - Infrastructure costs: space, power and cooling
      - Movement toward a greener data center
  - Reduce administrative costs
- Increasing integration with OSes, file systems and applications
  - e.g., Windows Server 2012 ReFS, ZFS, Cloud Gateways

# Data Reduction Techniques

- Compression
- Deduplication
  - File level (Single Instance Storage, aka “SIS”)
  - Block level (hash-based or delta block)
  - Content-aware or application-aware
  - Inline vs. post-process vs. hybrid
- Thin Provisioning

Note: Some techniques may be combined

# Deduplication and Compression

- Dedupe and compression are similar
  - Both are dependant on data patterns
    - Results can vary from little/no optimization to high percentage
  - Both consume system resources
  - Both can optimize required storage capacity or bandwidth utilization
- Dedupe and compression are different
  - Dedupe and compression can be complementary
  - But some knowledge about the data pattern is helpful
  - Some data is best optimized via dedupe
  - Some data is best optimized via compression
  - Some data can be optimized via dedupe **and** compression
- Sequence of optimization is important when encryption is used
  - Typically dedupe first, then compress, and encrypt last (reverse order at other end)

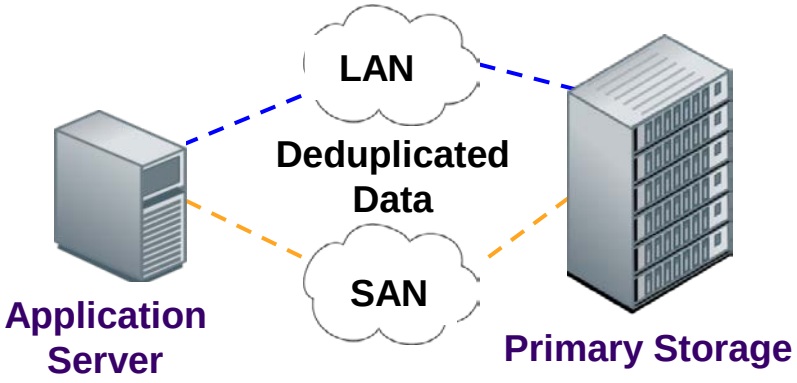
# Data Reduction Scope

The scope of data reduction is broadening:

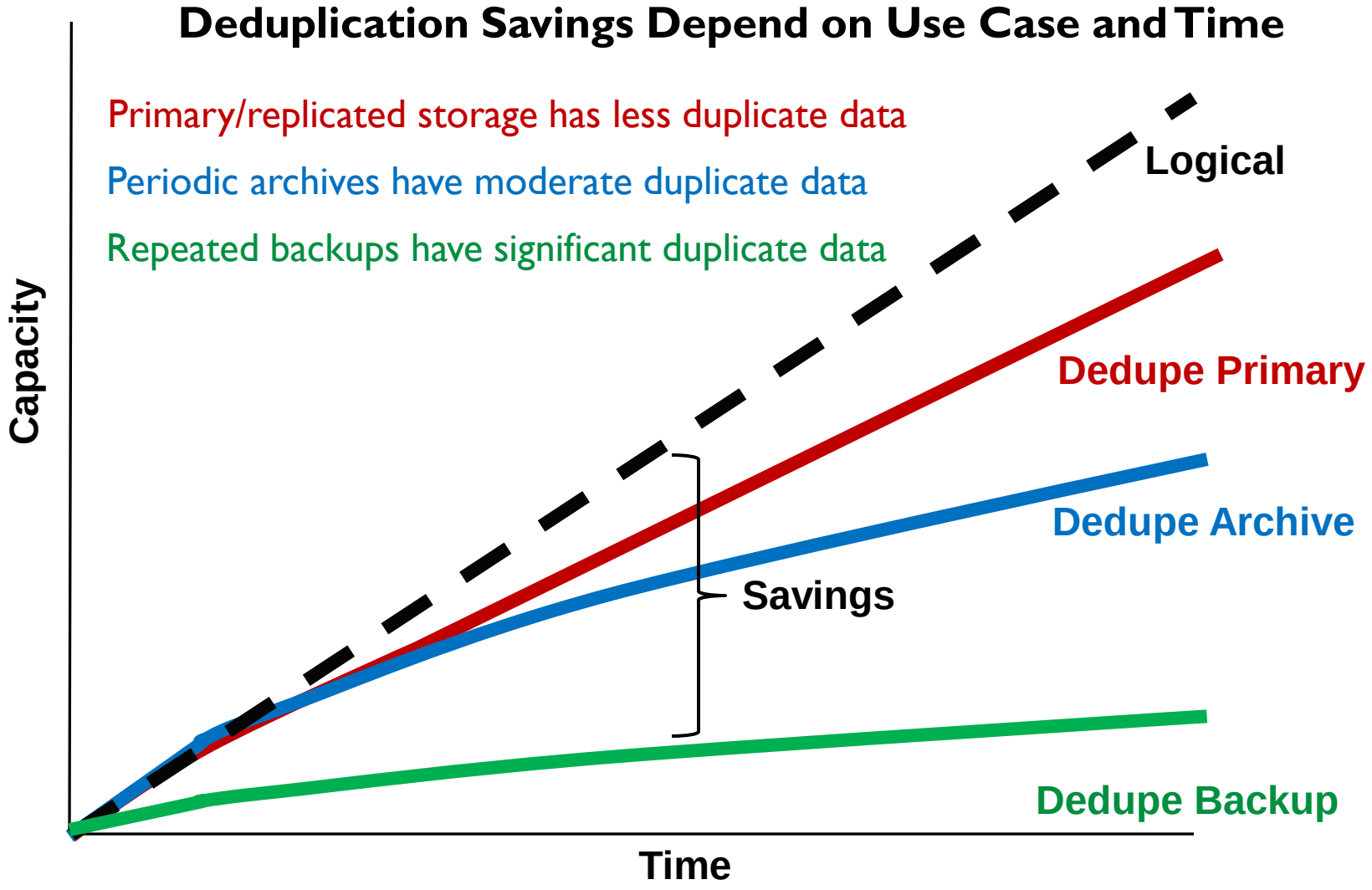
- **Primary Storage**
  - Reduced physical capacity for storage of active data
- **Data Protection**
  - Reduced capacity for backup with longer retention periods
- **Replication**
  - Reduced capacity for disaster recovery and business continuity
- **Archivals**
  - Reduced capacity for data retention and preservation
- **Movement / Migration of data, especially to/from Cloud**
  - Reduced bandwidth requirements for data-in-transit

# Data Reduction in Primary Storage

- Performance/Capacity Tradeoffs: a factor for compression or deduplication
  - Inline ingestion
  - Network-based
  - Post-processing
- Potentially high ROI with higher cost solid-state storage
- Deduplication works best with applications with high data redundancy
  - Virtual servers and desktops
  - Collaborative file “sharing”
  - Email (software SIS replacement)
- Compression varies by data type



# Deduplication Savings Expectation

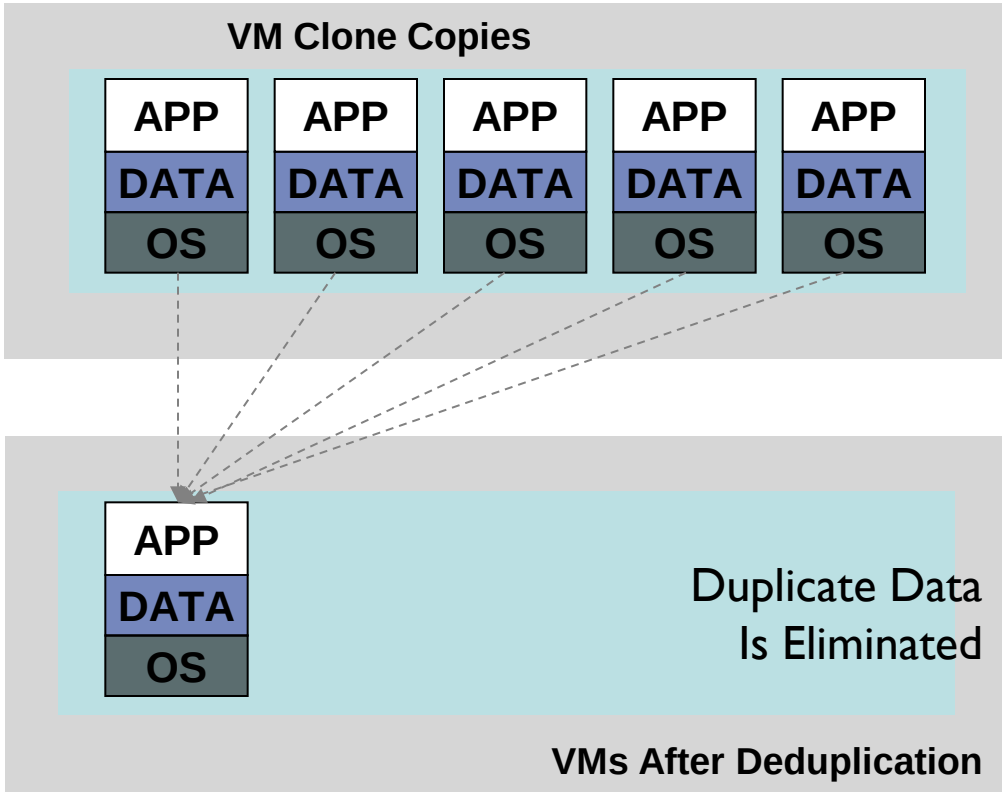


# Primary Storage with Cache

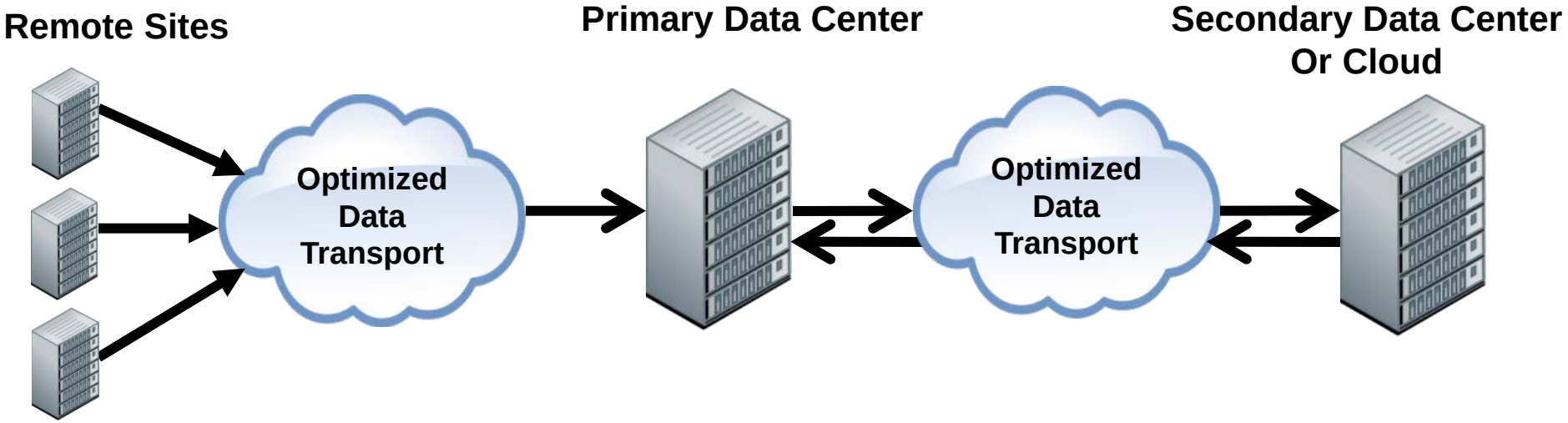
- Cache
  - Intelligent cache can be "dedupe-aware"
  - Hot data is cached with dedupe attributes
  - Reduces rotating media latencies
  - Example: Virtual Desktop "boot storms"

# Primary Storage Considerations

- Balance the tradeoff between cost savings and performance impact
- Some workloads lend themselves better to data reduction
  - Storage resource sharing across VMs
- Walk before you run
  - Use estimation tools
  - Perform POCs
  - Implement one workload at a time



# Data Reduction and Replication

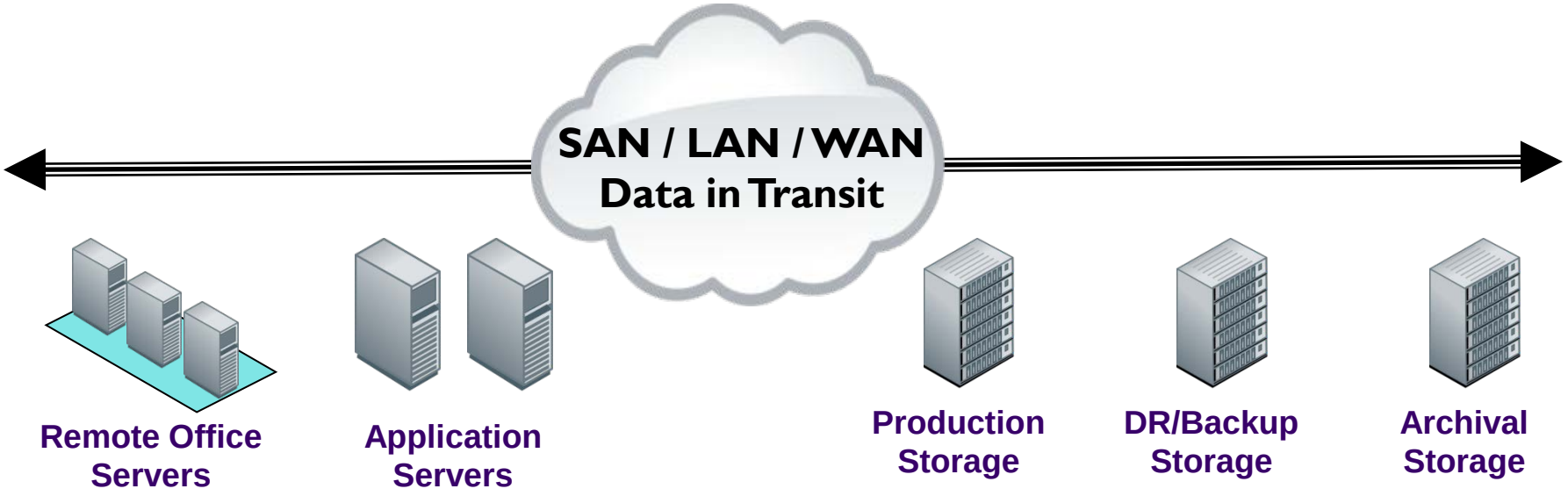



- Can be one-way, bi-directional, multi-hop, or cascade
- Optimized location(s) can be configured based on bandwidth constraints and data volume
- Data reduction makes replication more affordable
- Data reduction enables replication on constrained networks

# 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*
  - RPO & RTO are usually spelled out in the SLA(s)
- Is DR site planned as failover site?
  - If so, need to consider handling of data reduction re-hydration
- Does your organization plan to leverage cloud resources?
  - If so, need to consider availability of data for recovery (see SLAs)

# Data Reduction and Network Traffic




- Consider use of data reduction for any/all network transfers
- Increased SAN / LAN / WAN Efficiency
  - Compression and/or deduplication for data in flight
  - Transfer data references instead of data objects
  - Shorten data transfer times by sending less data

# Deduplication and Backups

## The Original Promise: (delivered!)

- 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
- Makes longer retention possible

## What's New?

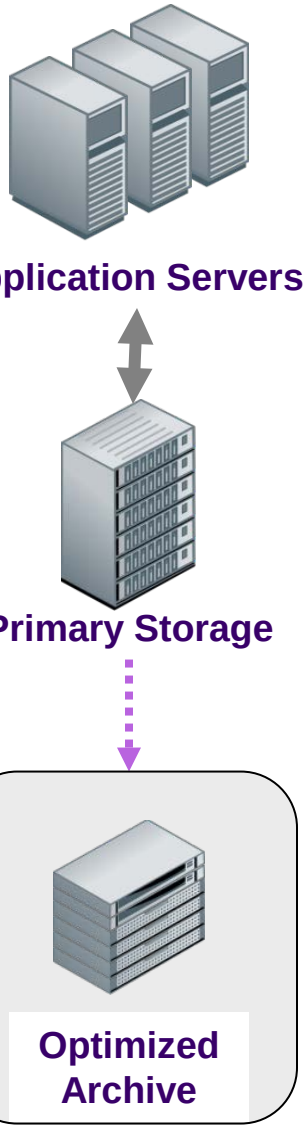
- Wide use as part of backup software
- Scalability of deduplication appliances
- Deduplication across appliances
- Cloud for backup, archive: throughput and metadata considerations
- Deduplication when using tape

# Backup Considerations

- Deduplication at appliance or in backup server (software)
- Inline or post-process?
- Source or target deduplication?
- Variable or fixed-length deduplication?
- File or sub-file deduplication?
- Compression with deduplication?

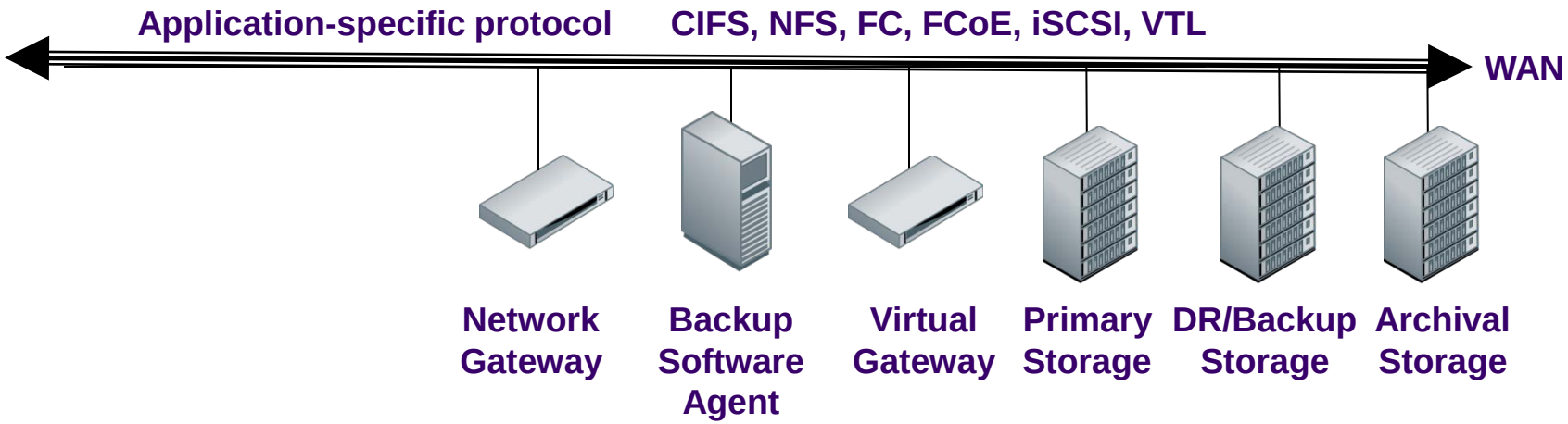
- Answers depend on the problem you are trying to solve

# Data Reduction and Archival




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

# Data Reduction Location Considerations




- Data Reduction Location can reduce:
  - Network Traffic
  - Physical Capacity
  - Backup Time
  - Recovery Time
  - Replication Time
  - Media Latency

# Summary

- The primary value of data reduction is in reducing costs and helping manage data growth
- The scope of data reduction is broadening
  - All storage tiers including primary and cloud
  - Various levels of granularity
  - Bandwidth reduction
  - Data subject to regularity rules
- Content-Aware and application-aware data reduction is becoming more prevalent
  - Potential for greater data reduction with more knowledge of specific data structures and data types

# Summary (Cont.)

- Is it Necessary to Optimize All Data?
  - Mission-critical applications
  - May have regulatory issues for some data
  - Some data types not conducive to data reduction
  - Replicate incremental changes only, without other optimization
- New use cases and new technologies bring new challenges
  - And new opportunities!

# Where to Get More Information

- Related tutorials
  - Advanced Data Reduction Concepts
  - Trends in Data Protection and Restoration Technologies
  - Managing Backup and Recovery in Today's Agile, Complex and Heterogeneous Data Centers
  - Protecting Data in the Big Data World
  - Retaining Information for 100 Years
- Visit the Data Protection and Capacity Optimization Committee (DPCO) website
  http://www.snia.org/forums/dpco/
- Best Practices for Data Protection white paper
  http://www.snia.org/forums/dpco/
- DPCO online Product Selection Guide
  http://sniadataprotectionguide.org/

# Attribution & Feedback

The SNIA Education Committee thanks the following individuals for their contributions to this Tutorial:

## Authorship History

**Original Author: DPCO Committee, 3/2011**

**Updates:**

**DPCO Committee, 9/2011**
**DPCO Committee, 3/2012**
**DPCO Committee, 8/2012**
**DPCO Committee, 2/2013**
**DPCO Committee, 8/2013**
**DPCO Committee, 3/2015**

## Additional Contributors

**David A. Chapa**
**Kevin Dudak**
**Mike Dutch**
**Michael Fishman**
**Larry Freeman**
**David G. Hill**
**Tom McNeal**
**Gene Nagle**
**Ronald Pagani**
**Thomas Rivera**
**Tom Sas**
**Gideon Senderov**
**Paul Talbut**
**SW Worth**

*Please send any questions or comments regarding this SNIA Tutorial to* ***tracktutorials@snia.org***