Education

Deduplication's Role in Disaster Recovery

Thomas Rivera, SEPATON

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, a group of more than 60 people representing 36 SNIA members
- The mission of the DPCO is to foster the growth and success of the market for data protection and capacity optimization technologies
- 2010 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

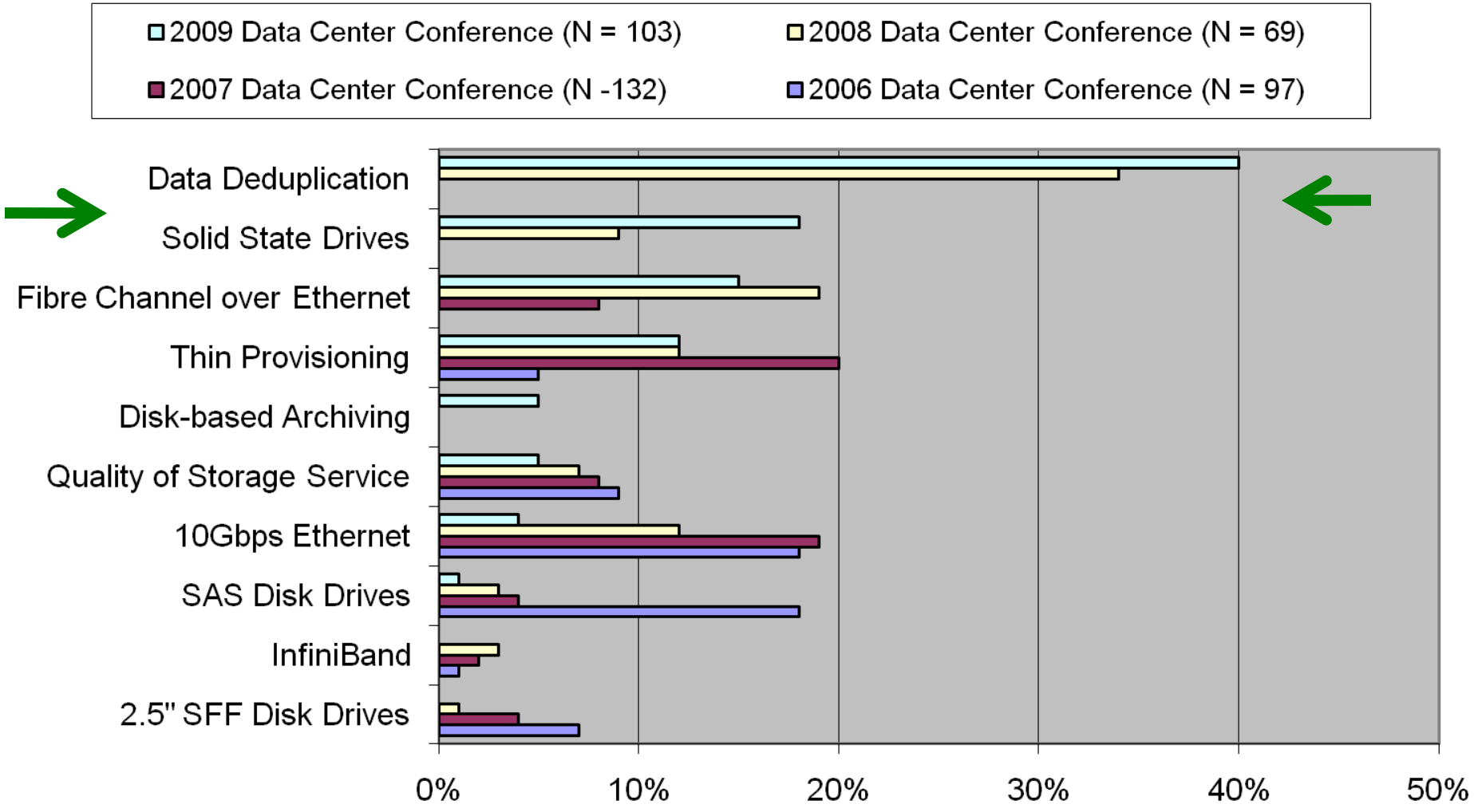
Abstract

Data deduplication can enhance Disaster Recovery (DR) because deduplication significantly reduces the amount of bandwidth required to replicate data.

This technical session will:

- Review Data Deduplication Concepts
- Cover the impact of Deduplication on WAN Replication
- Discuss Deduplication effects on meeting SLAs for DR

Which of the following technologies will most affect your storage infrastructure during the next three years?



Gartner
(2009)

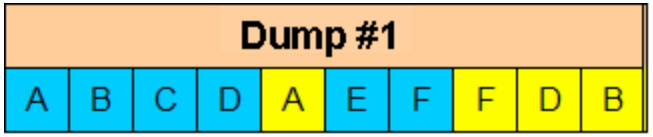
Capacity Optimization Terminology

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

- **Single Instance Storage** is form of data deduplication that operates at a granularity of an entire file or data object
- **Subfile Data Deduplication** is a form of data deduplication that operates at a finer granularity than an entire file or data object

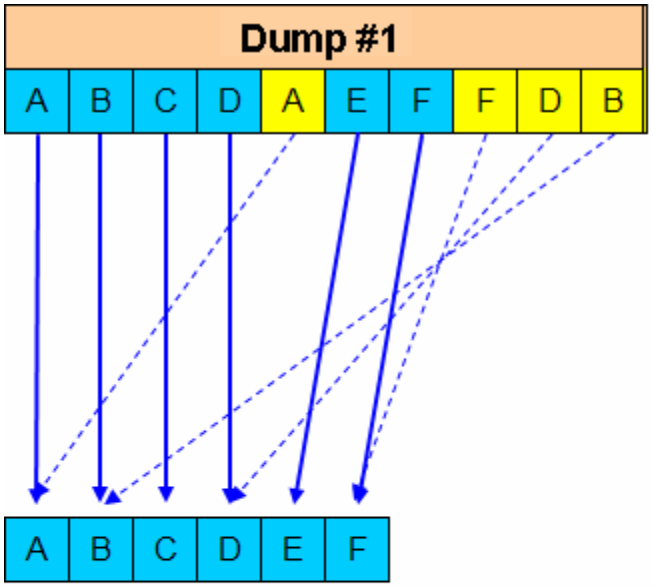
Compression is the encoding of data to reduce its storage requirement - compressed data can also be deduplicated

Data Deduplication Simplified



= new unique data
= repeat data

Data Deduplication Simplified

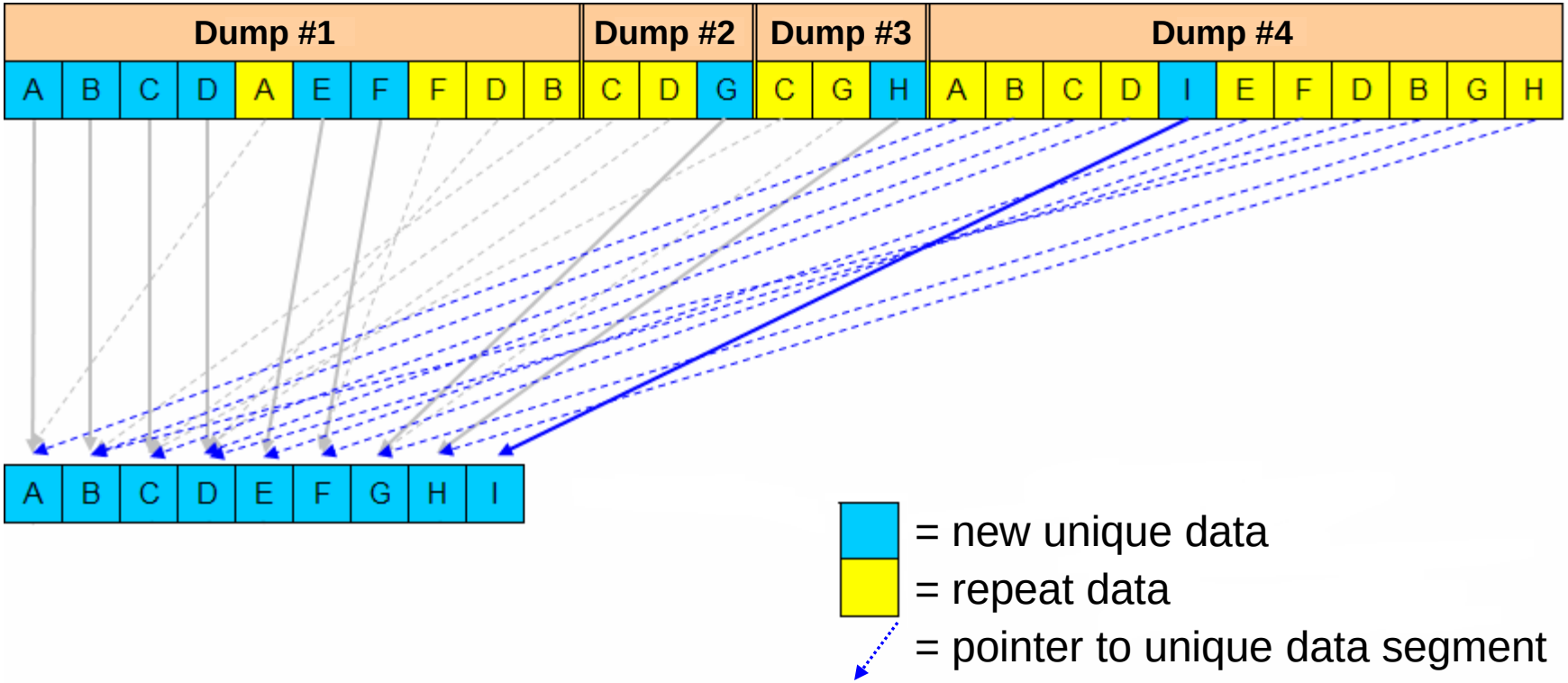


= new unique data

= repeat data

= pointer to unique data segment

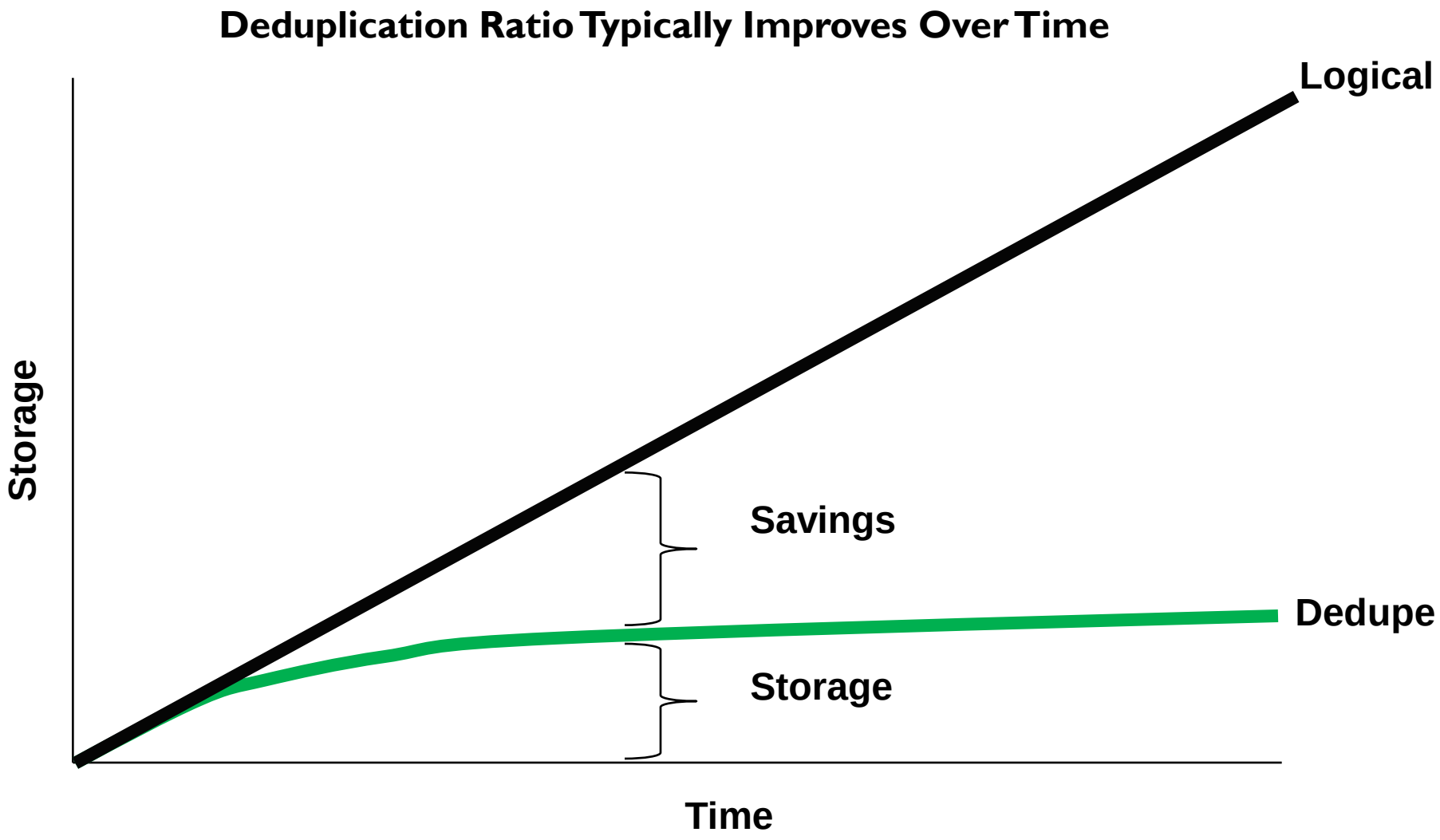
Data Deduplication Simplified



Check out SNIA Tutorial:

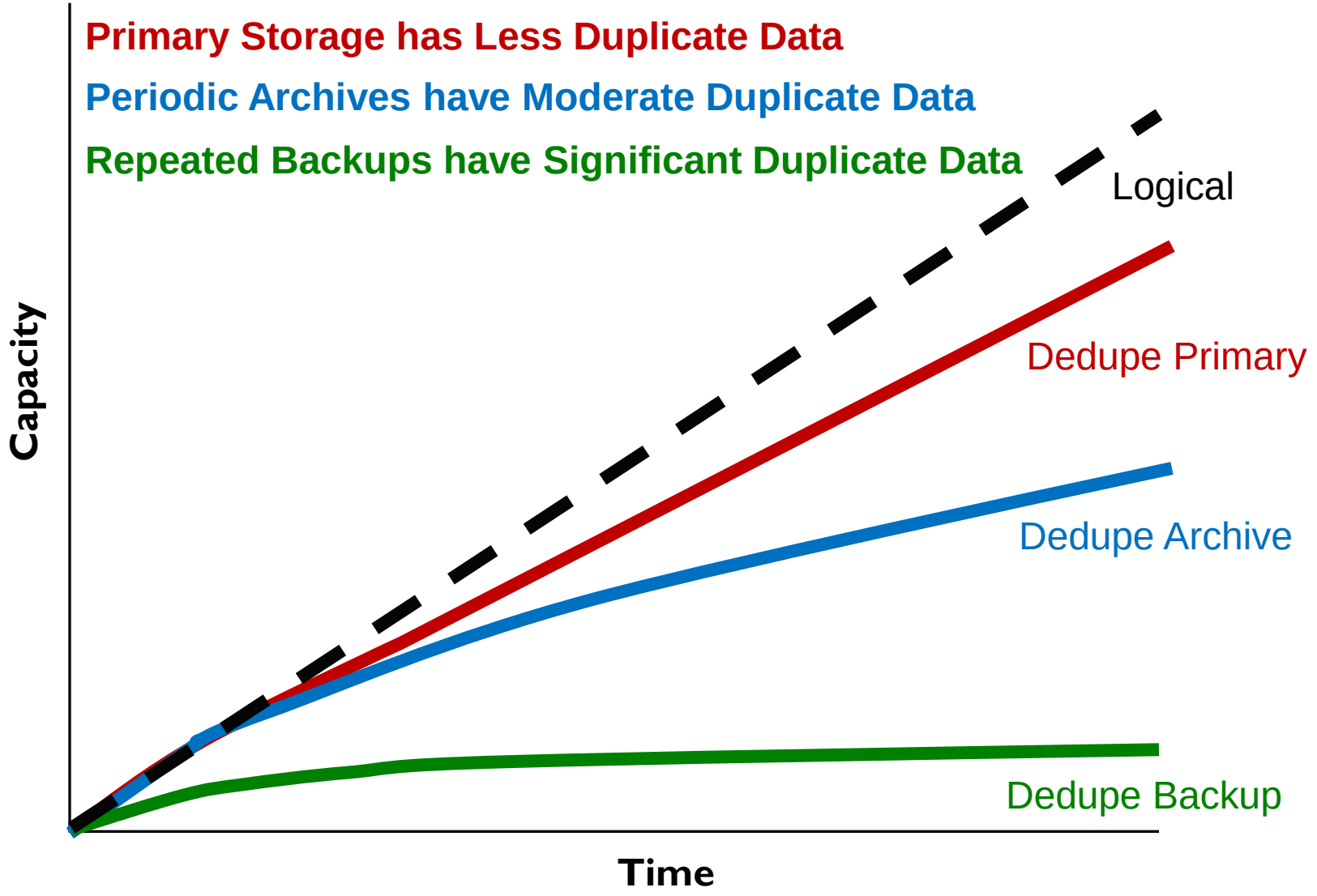
Understanding Data Deduplication

Deduplication Controls Growth

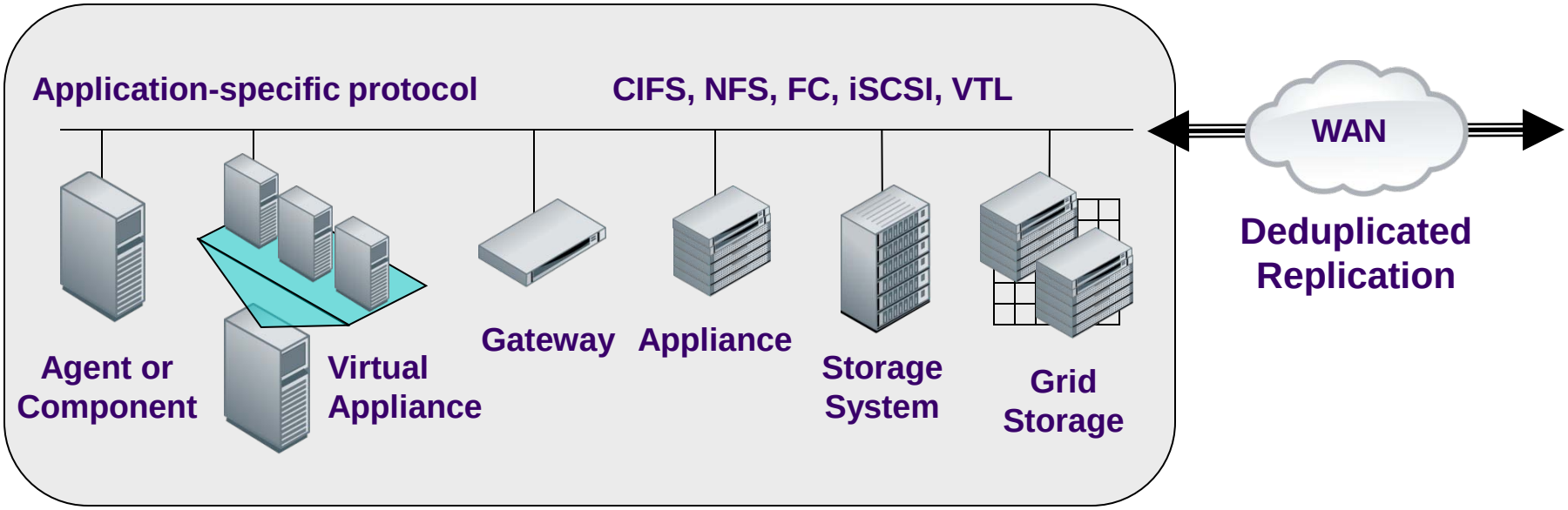


Deduplication Ratio: Depends on Use Case, Time

Education SNIA



Deduplication Implementations



- Vendors Provide Deduplication Solutions for nearly Every Point at which Data is Stored or Transmitted
- The Decision as to "Where" to Deduplicate is Determined by Which Problem you are Trying to Solve

Data Reduction Becomes Ubiquitous

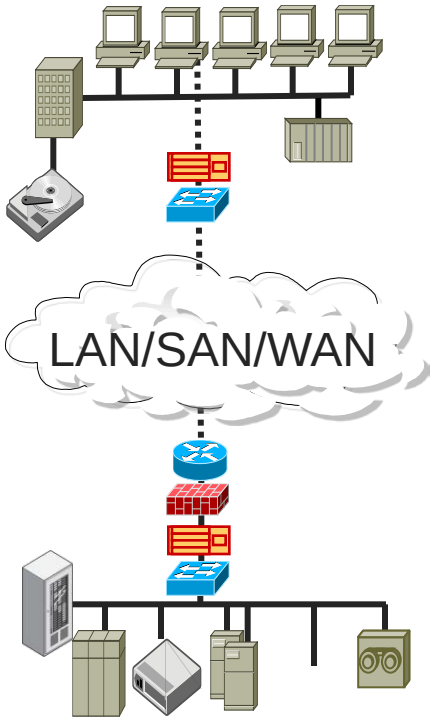
Education SNIA

Use Cases Expand:

- **Backup**
- **Archive**
- **Primary Data**

Techniques:

- **Compression**
- **Single-Instance Storage**
- **Subfile Data Deduplication**



"Deduplication will be widely available in 2012 for blocks & files, and deployable in application software, middleware, operating systems, appliances & storage arrays."

"By 2014, some form of primary data reduction, such as compression and/or deduplication, will be used for at least 20% of all enterprise workloads, up from the low single digits in 2009."

Gartner (2010)

Data Deduplication Benefits

- Data Deduplication can help organizations:
  - Help Satisfy ROI/TCO Requirements
  - Manage Data Growth Costs
  - Increase efficiency of Storage and Backup
  - Reduce overall Expenditure on Storage
  - Reduce Network Bandwidth
  - Reduce Operational Costs including:
    - Infrastructure costs required space, power and cooling
  - Reduce Administrative Costs

Definitions

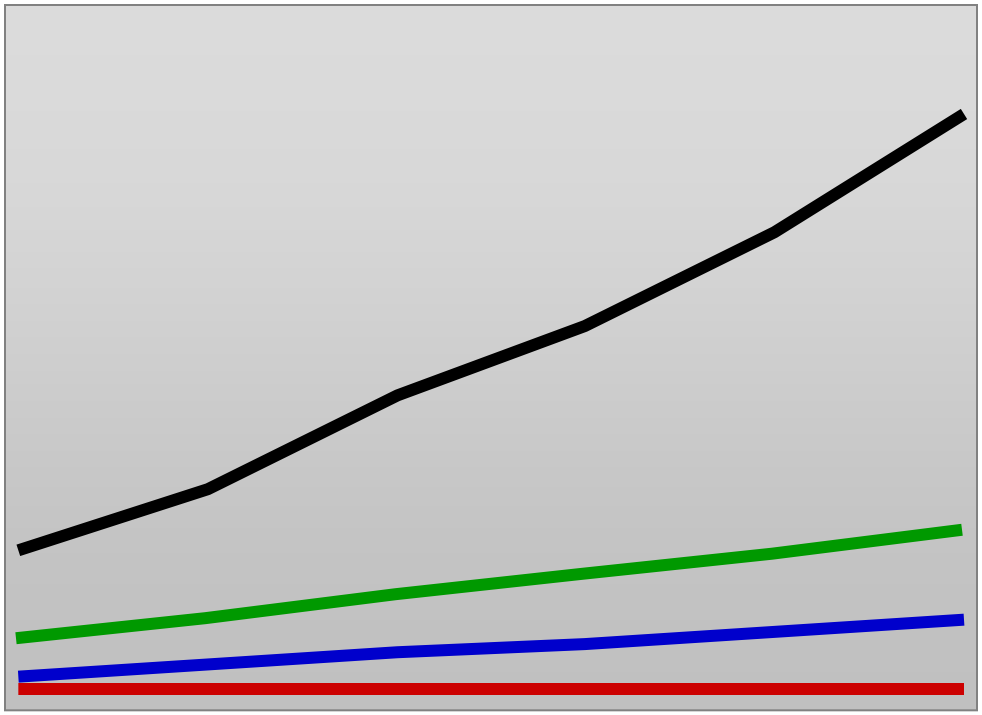
- Replication is (in this context):
  - The electronic transport of a collection of data between primary and secondary sites
  - There are multiple "Use Case" scenarios, which we will cover later
- Disaster Recovery is:
  - The recovery of data, access to data, and associated processing through a comprehensive process of setting up a redundant site (equipment & work space) with recovery of operational data to continue business operations after a loss of use of all or part of a data center

Challenges of Enterprise Scale DR

- Data Volumes Too Large for Timely Replication
- Bandwidth Constraints / Costs
- Exceeding Backup Windows
- Satisfying RPO/RTO Metrics
- Added Complexity

- These Challenges result in:
  - Not meeting SLAs (Backup & Recovery)
  - Added Complexity (Cost $$ for Admin, HW/SW, etc.)

Challenges of Enterprise Scale DR



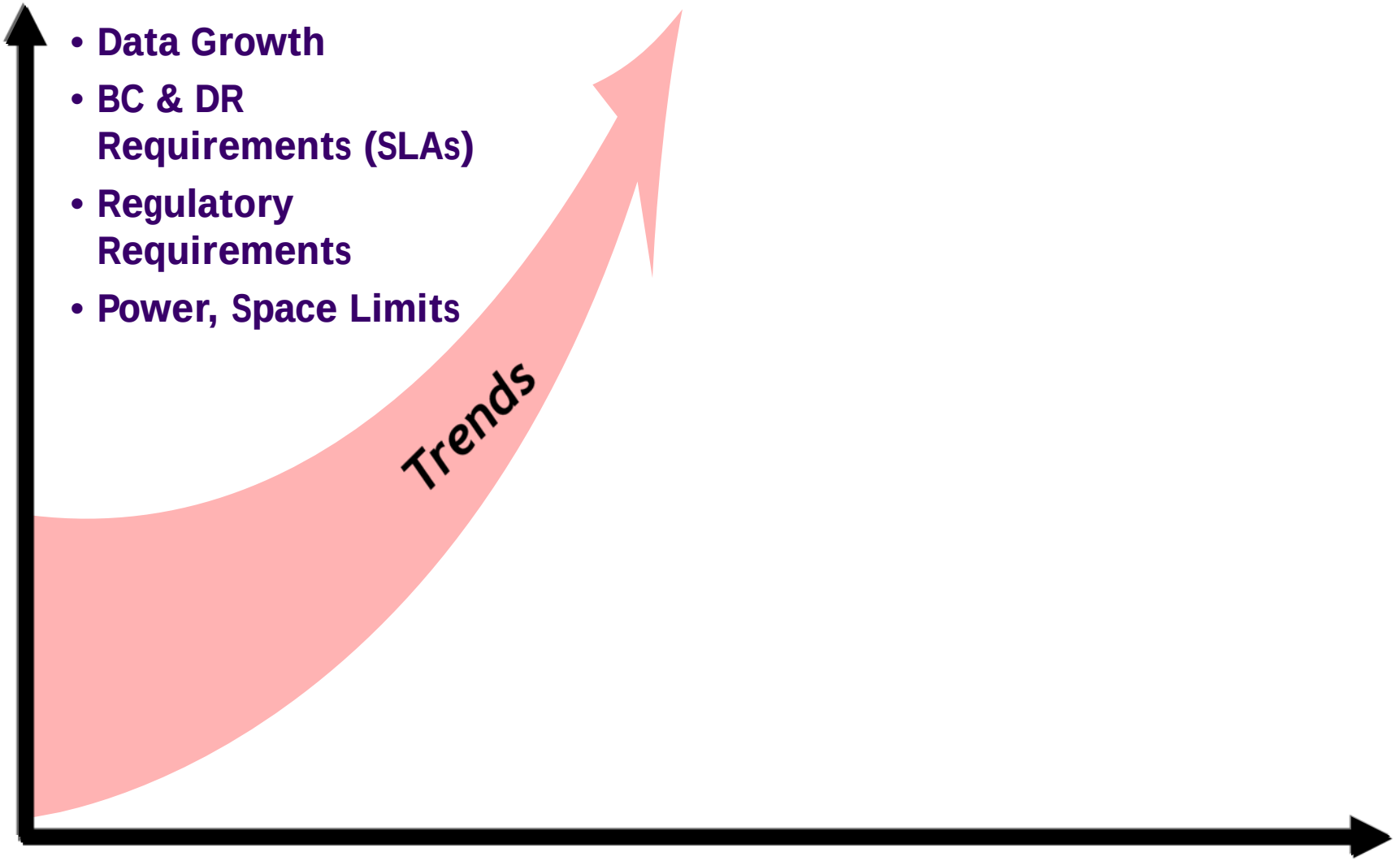
- IT Budgets
- Storage as a % of IT Budgets
- Data Growth
- Cost of Storage Mgmt as a % of Storage

IT Challenges

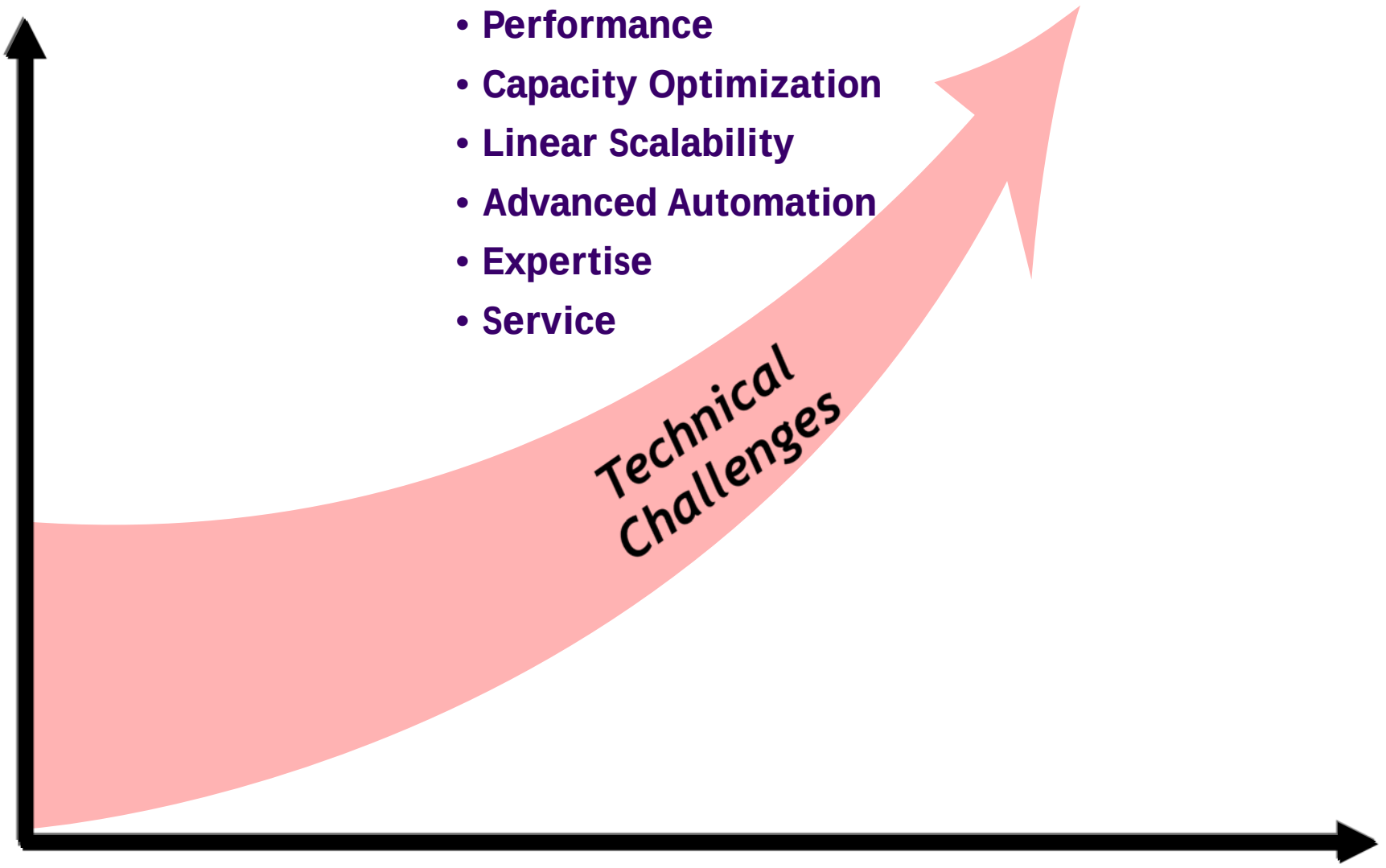
- Explosion of Online Data
- Infrastructure Complexity
- Inflexible Architectures
- Simplifying the Storage Infrastructure
- Antiquated Recovery Infrastructure
- Increase Staff Productivity
- Meeting SLAs within Restricted Budgets

Gartner (2009)

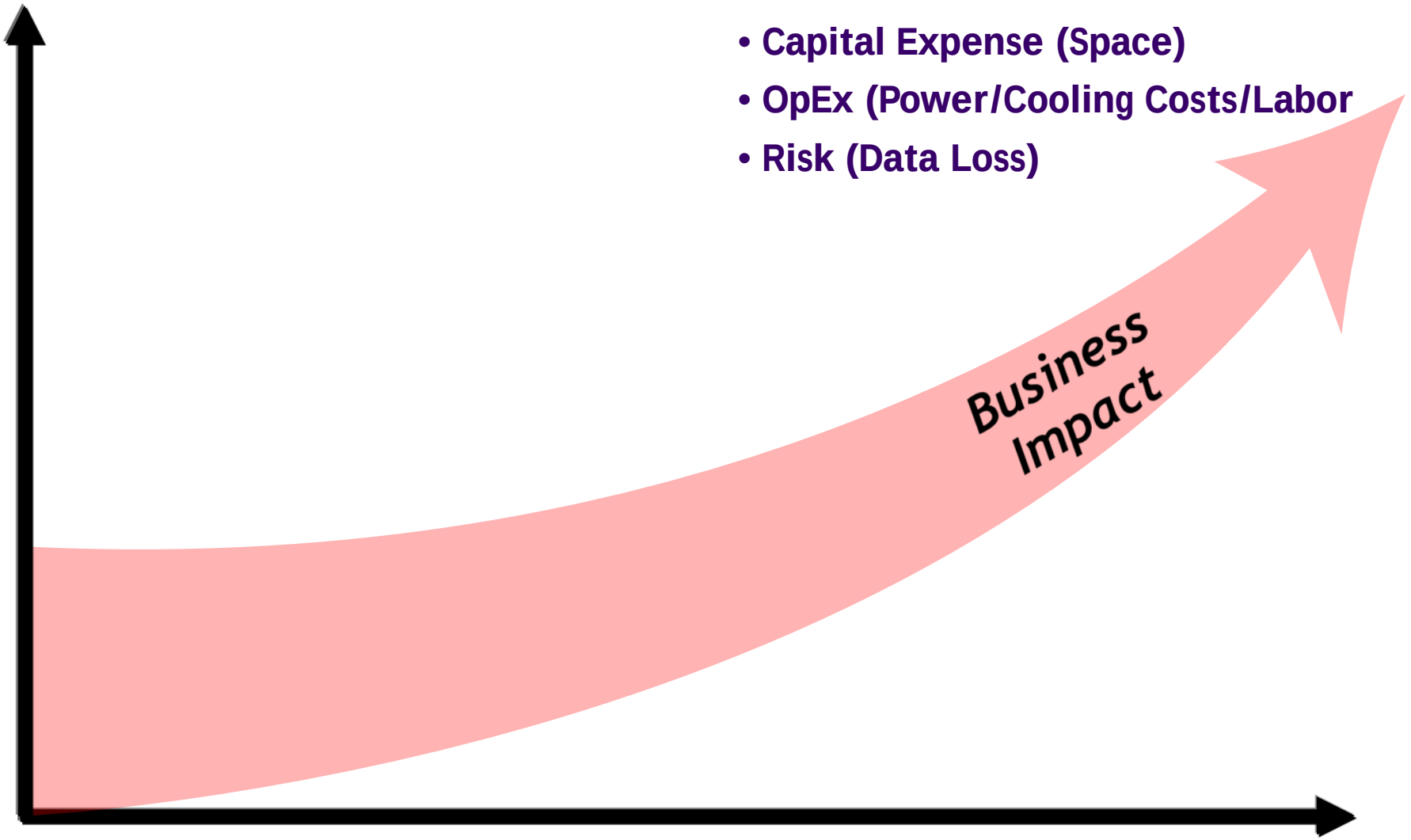
Increasing Cost & Risk: Trends



Increasing Cost & Risk: Technical Challenges



Increasing Cost & Risk: Business Impact



Meeting SLAs

Education SNIA

Business Issues & SLAs

Rapid Data Growth

- ✓50% CAGR
- ✓Increased Backup Costs

SLAs for BC/DR

- ✓Downtime Costs
- ✓RTO / RPO

Space/Power Limitations

- ✓Data Center Footprint
- ✓Power Costs

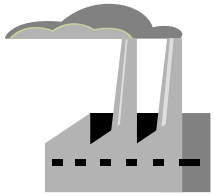
Regulatory Requirements

- ✓Online Retention
- ✓Added Complexity

Lower TCO & Better ROI

Education SNIA

Measurable TCO & ROI



Reduce Capital Expense

- ✓Lower Acquisition Cost
- ✓Scalability

Reduce Operating Expense

- ✓Non Disruptive
- ✓Less Labor: Automation
- ✓Less Power & Space

Avoid Costs

- ✓Frees IT Staff Time
- ✓More Data per FTE
- ✓No Human Error

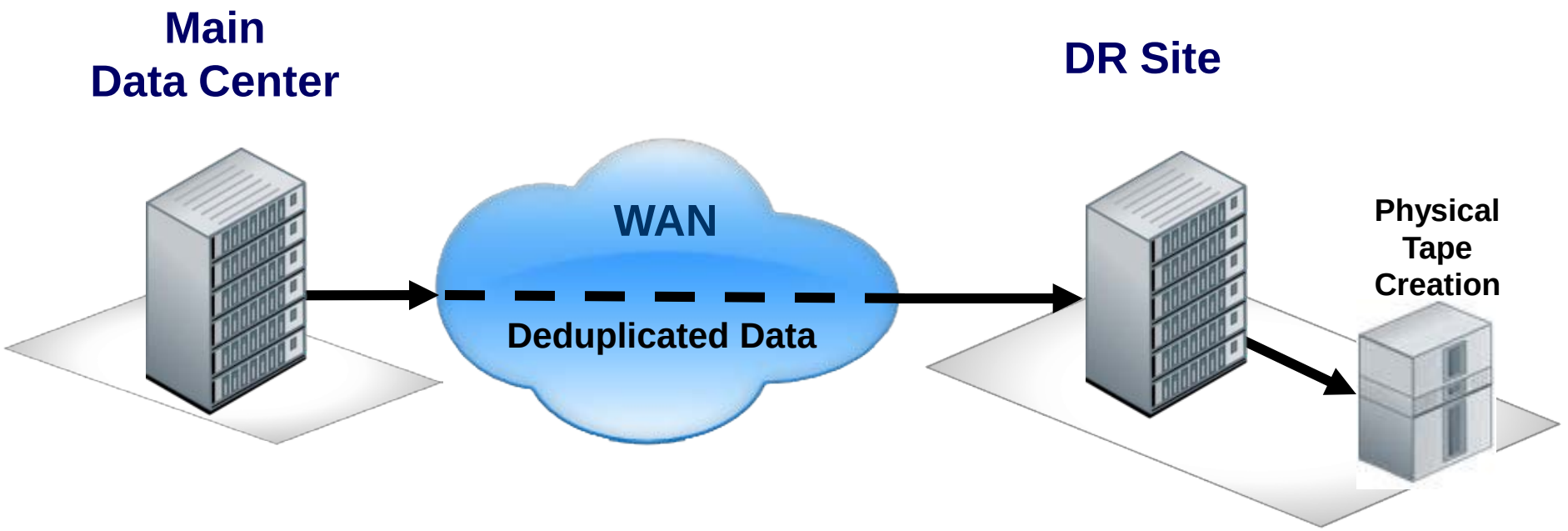
Backup: Deduplication for Data Movement

- Disaster Recovery
  - Replicate all Data after Deduplication for Bandwidth Efficiency
  - Meet Offsite Requirements without Physical Transport
- Bandwidth Optimization
  - Increasing WAN Efficiency
    - Transfer more information per pipe
  - Support Remote Office Protection
  - Enable Backup Centralization
  - Consolidate Physical Tape Creation

Dedupe in DR: Automation Benefits

Automation

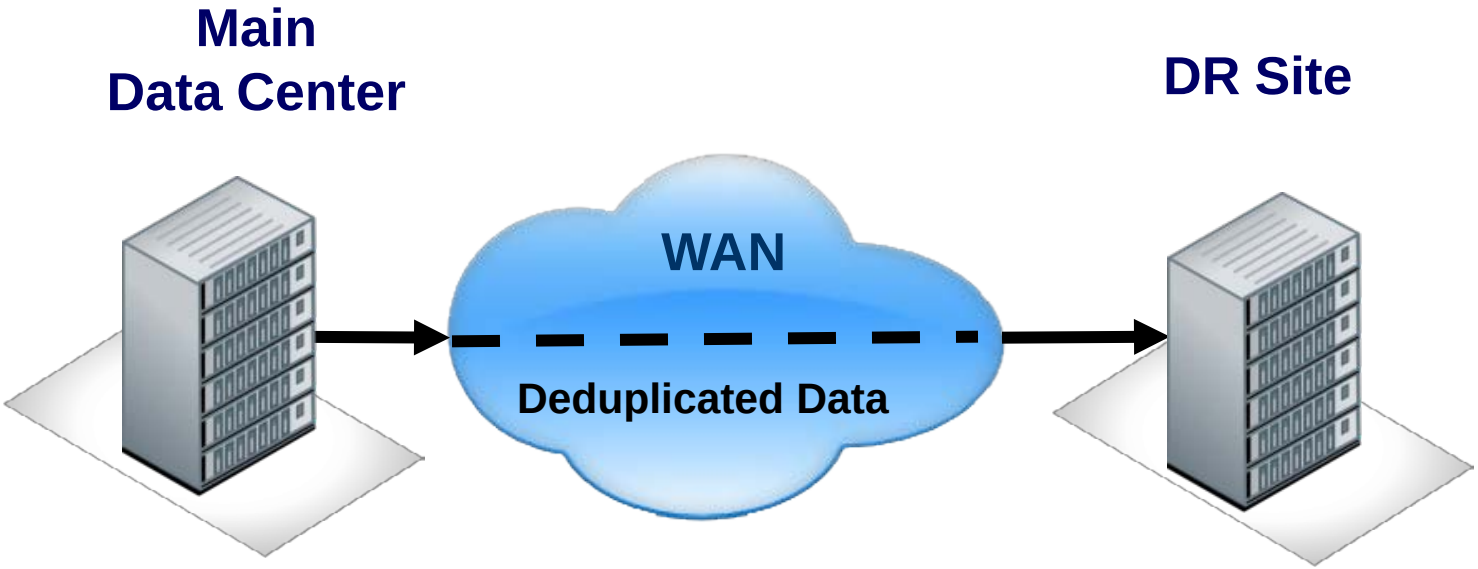
- Simplifies the Offsite Process
- Minimize risk of Data Loss/Data Theft
- Leverage Existing Bandwidth



Dedupe in DR: Network Efficiency

Network Efficiency

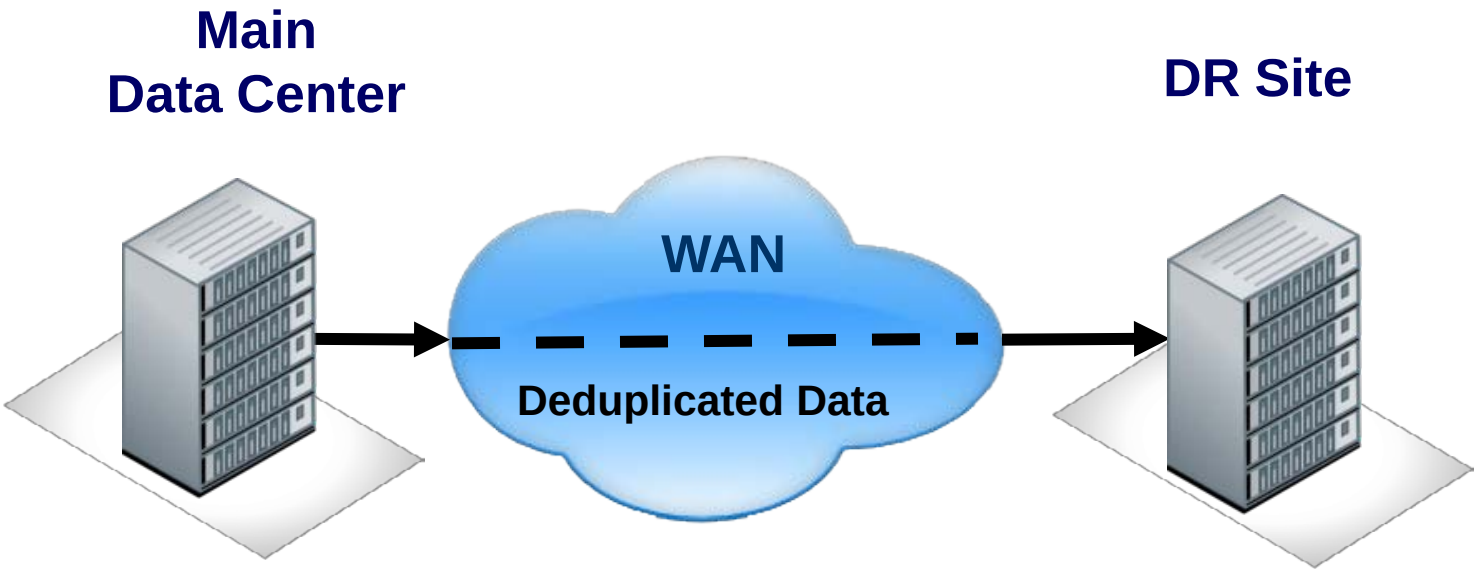
- Deduplication Dramatically Reduces Bandwidth Usage



Dedupe in DR: Risk Reduction

Risk Reduction

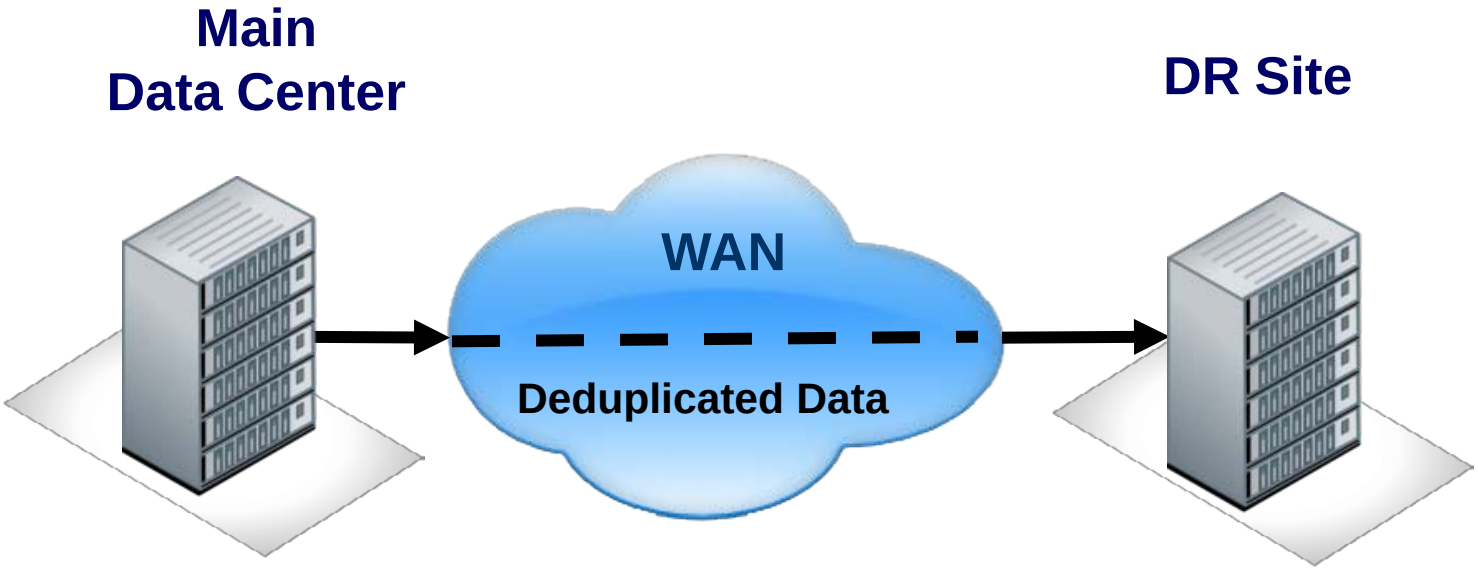
- Human Error reduced with automation
- Regulatory Compliance more easily achieved
- Improve Data Access Reliability



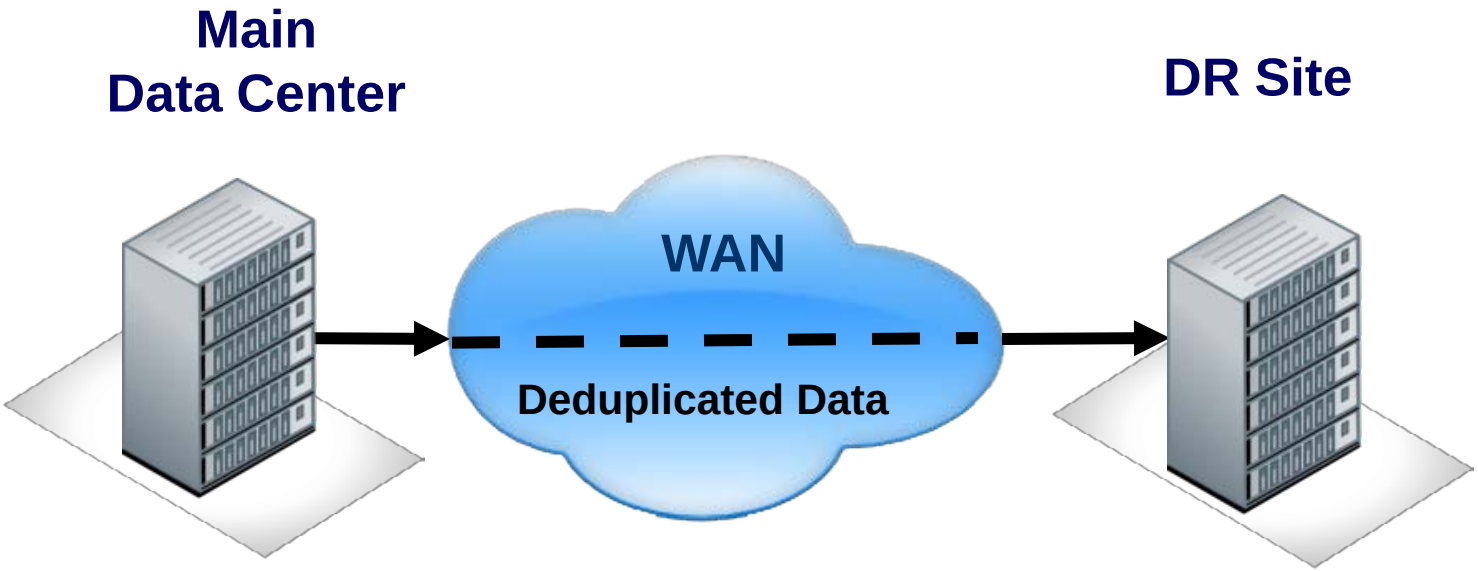
Dedupe in DR: Cost Savings

Cost Savings

- Reduced Manual Media Handling
- Reduce Tape Archival Services
- Minimize Data Loss

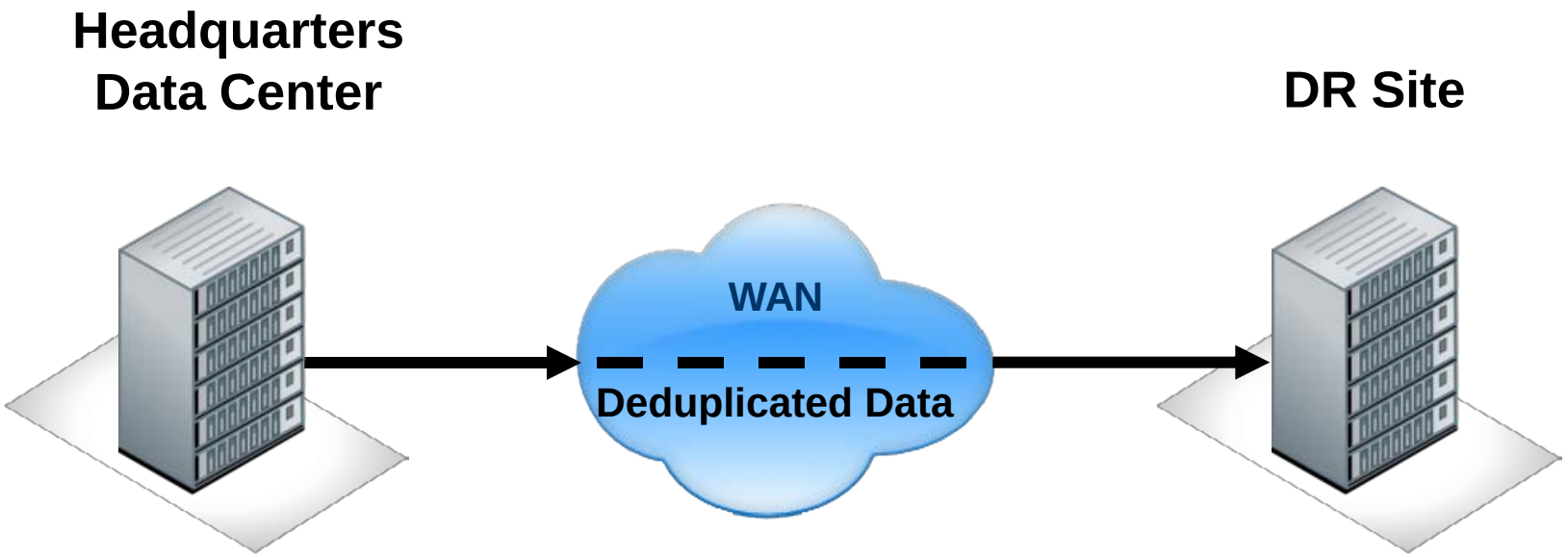


Dedupe in DR: Requirements



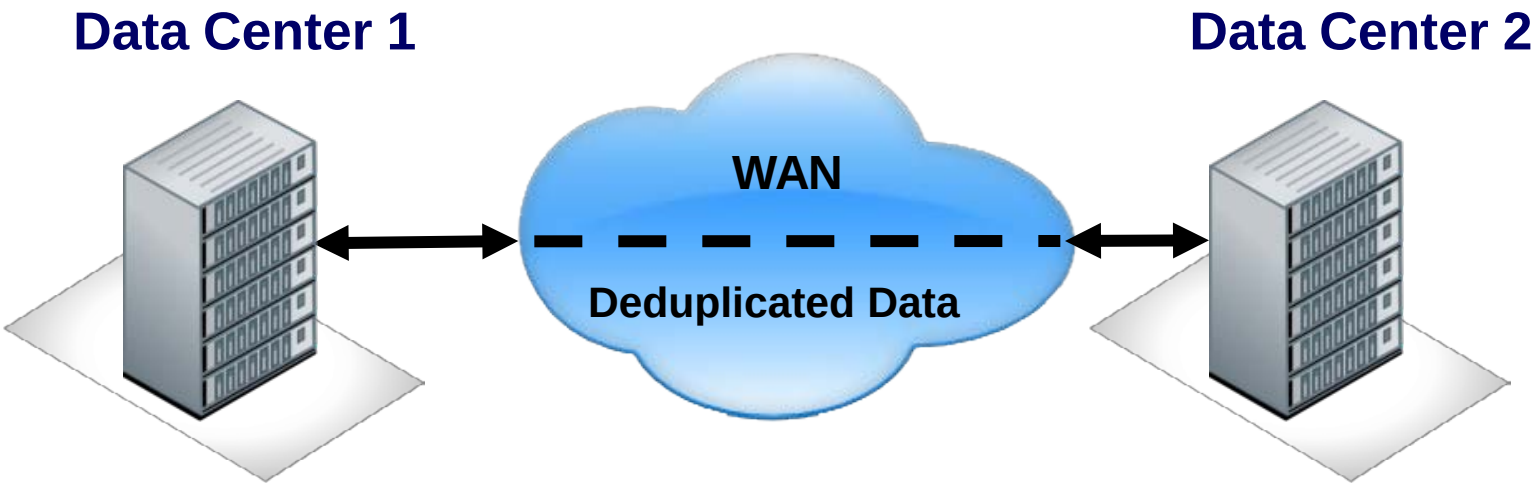
- Replicate Large Data Volumes
- Send only "Changed Data" over the Network
- Perform Fast Data Restores from Remote Site
- Provide Control over Replication/Restoration Process
- Provide Resiliency/High Availability

Use Model: HQ to DR Location



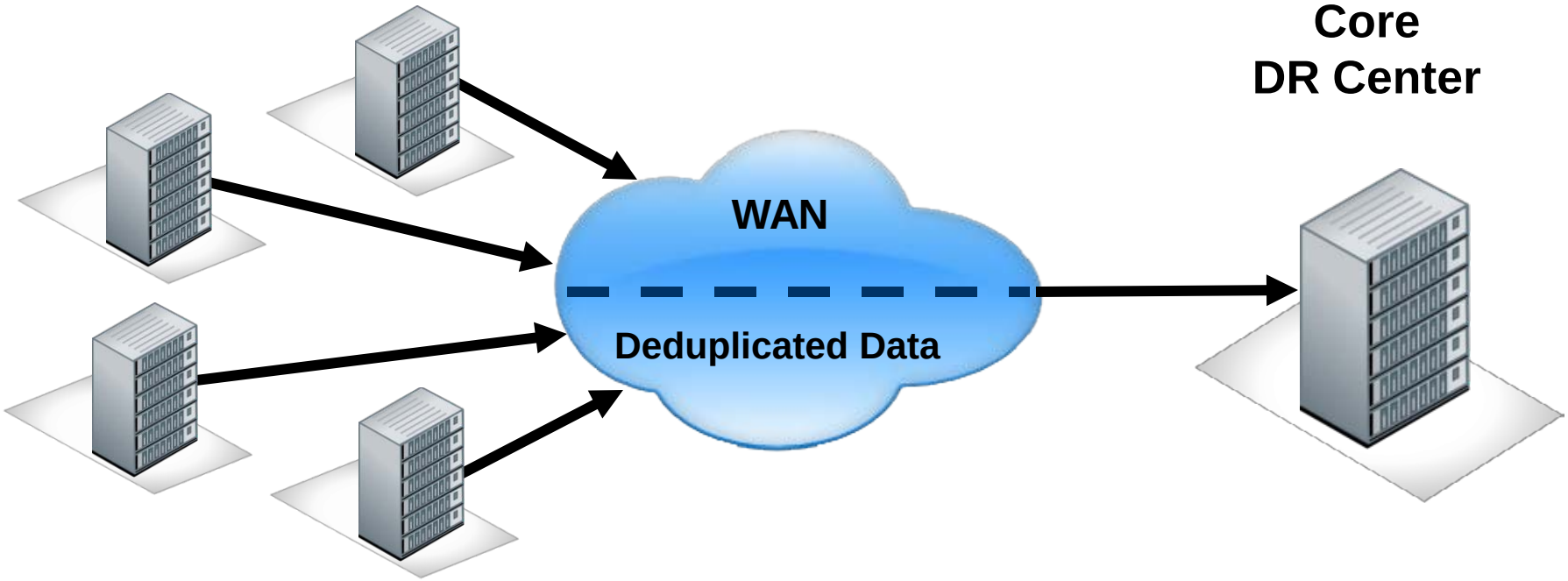
- Data is deduplicated and replicated to a DR site
- In event of data becoming unavailable at the headquarters data center:
  - Data can be restored to headquarters data center
  - Data can be used at the DR site

Use Model: Data Center to Data Center



- Data is deduplicated & replicated bi-directionally between two production data centers
  - Each Data Center acting as a “DR Site” for the other

Use Model: Edge to Core DR



- Data is deduplicated and replicated from multiple regional data centers to a main DR center
  - Core DR center acting as a "DR Site" for all production data centers

Deduplication: Potential Issues

- Be Aware of the Potential Pitfalls
  - May Decrease Data Ingestion Performance
  - Can Negatively Impact Restore Performance
  - May not Scale in Performance
  - May not Scale in Capacity
  - May not Offer Resiliency/HA Features
  - Encrypted Data Limits Deduplication
- Vendor Implementations Make Trade-offs between Deduplication Ratios and the Above
- Easy to Under-Estimate the Bandwidth Required
  - Changed Data Size ÷ Replication Window = Data Rate Needed

Dedupe in DR: What Really Matters?

- Focus on your Service Level Agreements (SLAs)
  - Needs to Meet Window for *Replication*
  - Needs to Meet Window for *Restore*
- Is it Necessary to Dedupe all Data?
  - May Have Regulatory Issues for Some Data
  - Some Data Types not Conducive to Deduplication
- Can the Dedupe Solution Scale to meet your Needs?
  - Needs to Scale in Capacity & Performance
  - Different Dedupe Approaches Yield Different Reduction Ratios
  - CapEx & OpEx Savings will be Higher

Review

- Using Dedupe in DR can Help Organizations:
  - Satisfy ROI/TCO Requirements
  - Manage Data Growth
  - Increase Efficiency of Storage and Backup
  - Reduce Overall Cost of Storage
  - Reduce Required Network Bandwidth
  - Reduce Operational Costs Including:
    - Infrastructure costs Required Space, Power and Cooling
    - Movement toward a Greener Data Center
  - Reduce Administrative Costs

Summary

- Multiple Elements to Consider when Evaluating Deduplication Technologies for DR Projects:

CPU Utilization and/or Power Consumption

Restore Performance of Deduped Data

WAN Efficiency of Deduped Data

Replication Scalability of Deduped Data

Resiliancy/HA of Deduplication Solution

Summary (cont.)

- There is no “Right” Solution for Everyone!
  - The Appropriate Solution will Vary by Environment and Requirements

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

Michael Alvarado
Matthew Brisse
Don Deel
Mike Dutch
Larry Freeman
Bernd Henning
Gene Nagle
Ronald Pagani
Thomas Rivera
Tom Sas
Gideon Senderov