



Education

Deduplication's Role in Disaster Recovery

Thomas Rivera, SEPATON

- 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.**

- This tutorial has been developed, reviewed and approved by members of the Data Protection and Capacity Optimization (DPCO) Committee
- 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

Data deduplication can be applied to the replication of data for disaster recovery (DR) projects, since 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

SNIA definitions:

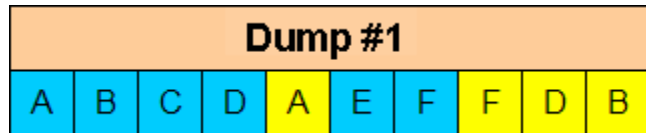
Data Deduplication is the replacement of multiple copies of data - at variable levels of granularity - with references to a shared copy in order to save storage space and/or bandwidth



Single Instance Storage is a 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

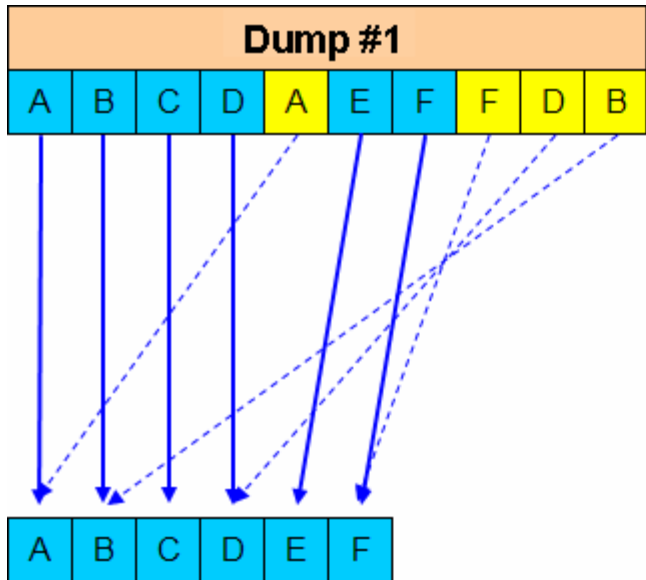
Compression is the encoding of data to reduce its storage requirement - compressed data can also be deduplicated




Data Deduplication Simplified



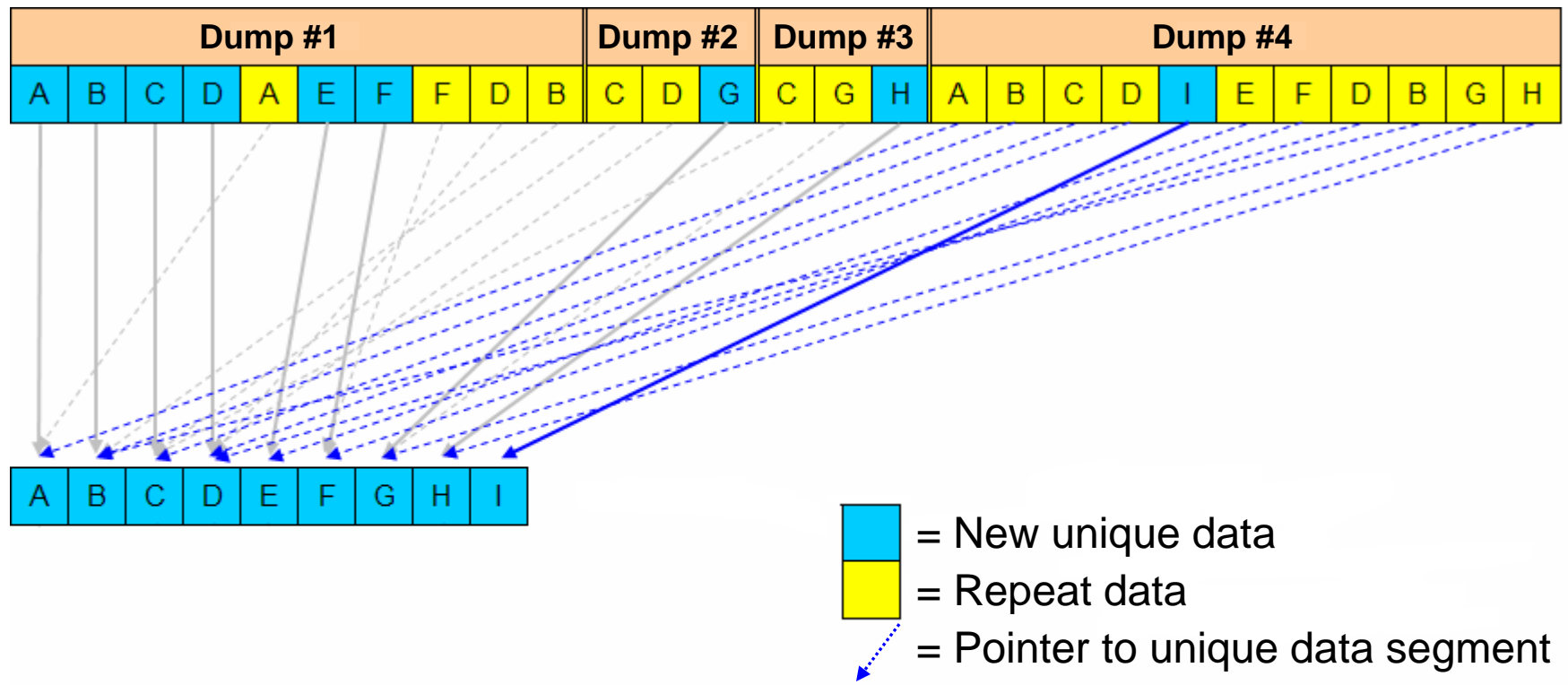
 = New unique data
 = Repeat data

Data Deduplication Simplified



-  = New unique data
-  = Repeat data
-  = Pointer to unique data segment

Data Deduplication Simplified



**Check out SNIA Tutorial:
 Understanding Data Deduplication**

➤ Replication is (in this context):

- ◆ The transport 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

Data Reduction Becomes Ubiquitous

Use Cases Expand:

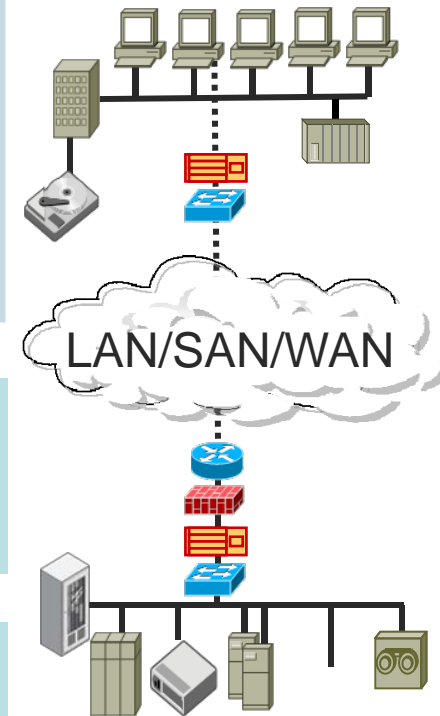
- Backup
- Archive
- Primary data

Techniques:

- Compression
- Single-instance store
- Deduplication

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

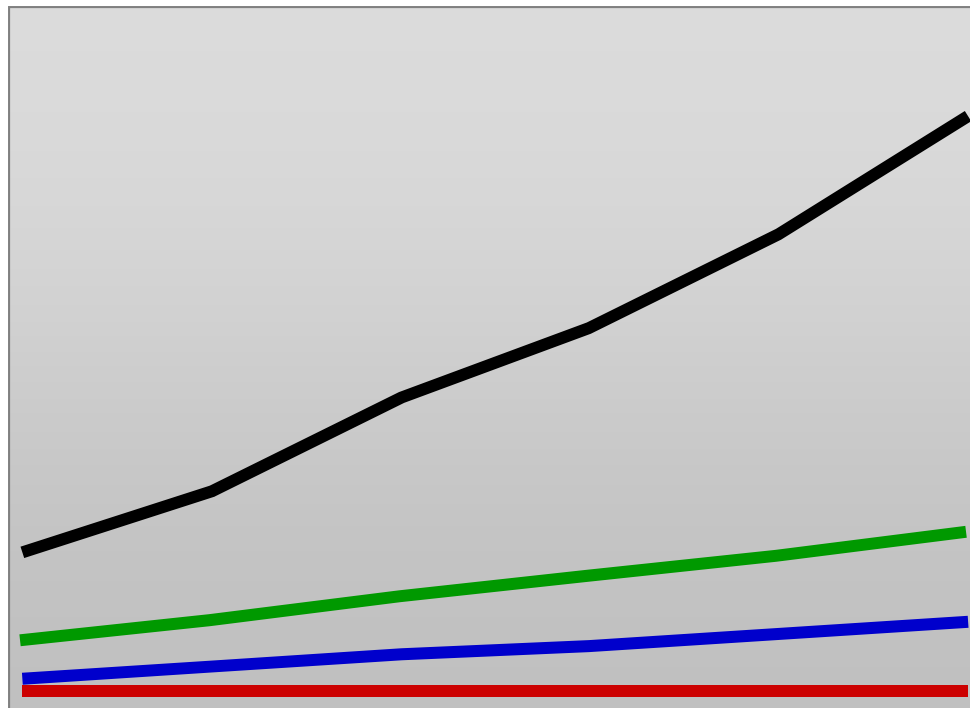
“By 2014, some form of primary data reduction, such as compression and/or deduplication, will be used for at least 20% of all enterprise workloads, up from the low single digits in 2009.”



- Data deduplication can help organizations:
 - ◆ 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 requiring space, power and cooling
 - ◆ Reduce administrative costs

- 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.)



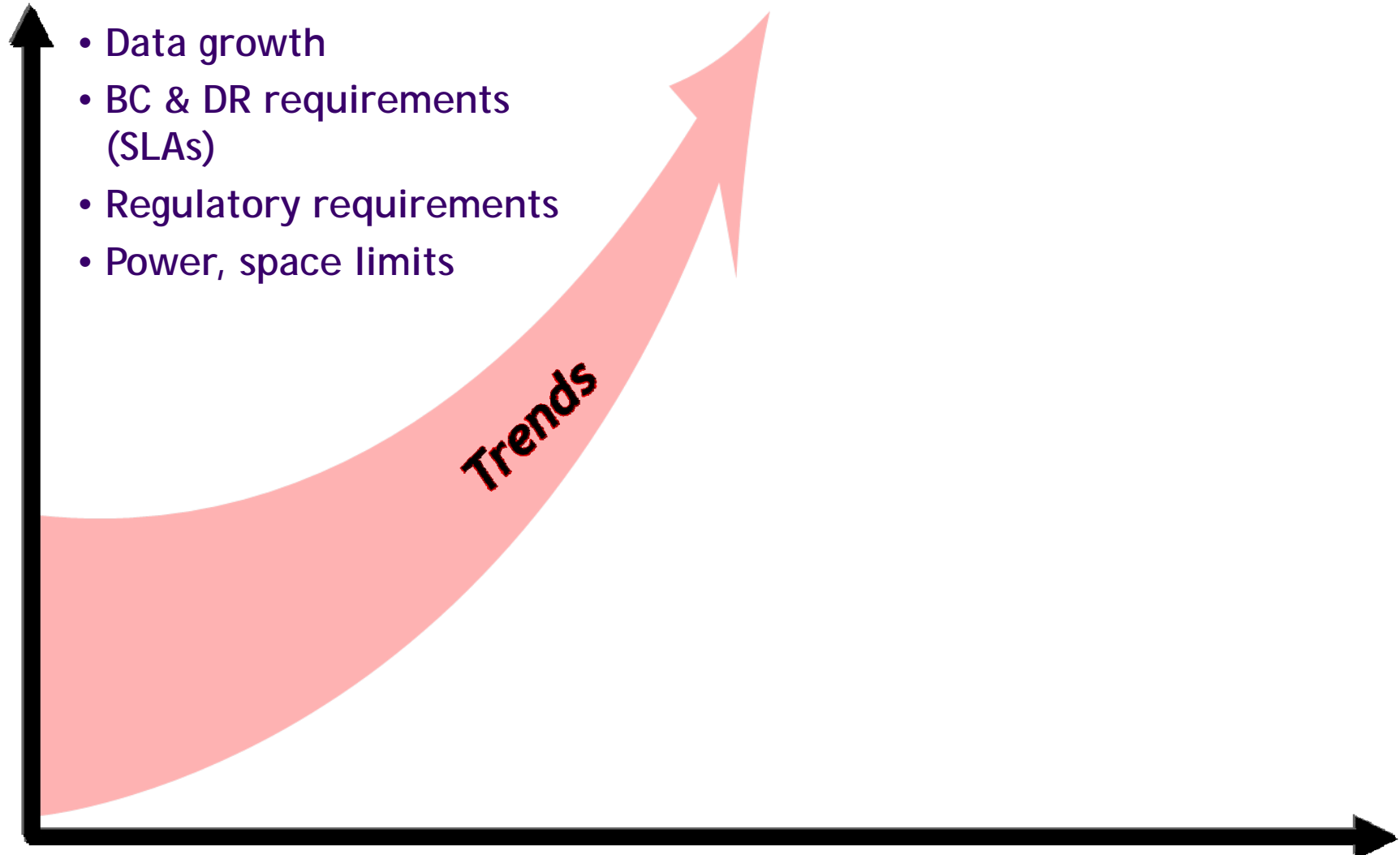
— IT Budgets
— Storage as a % of IT Budgets

— 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

Increasing Cost & Risk: Trends



Increasing Cost & Risk: Technical Challenges

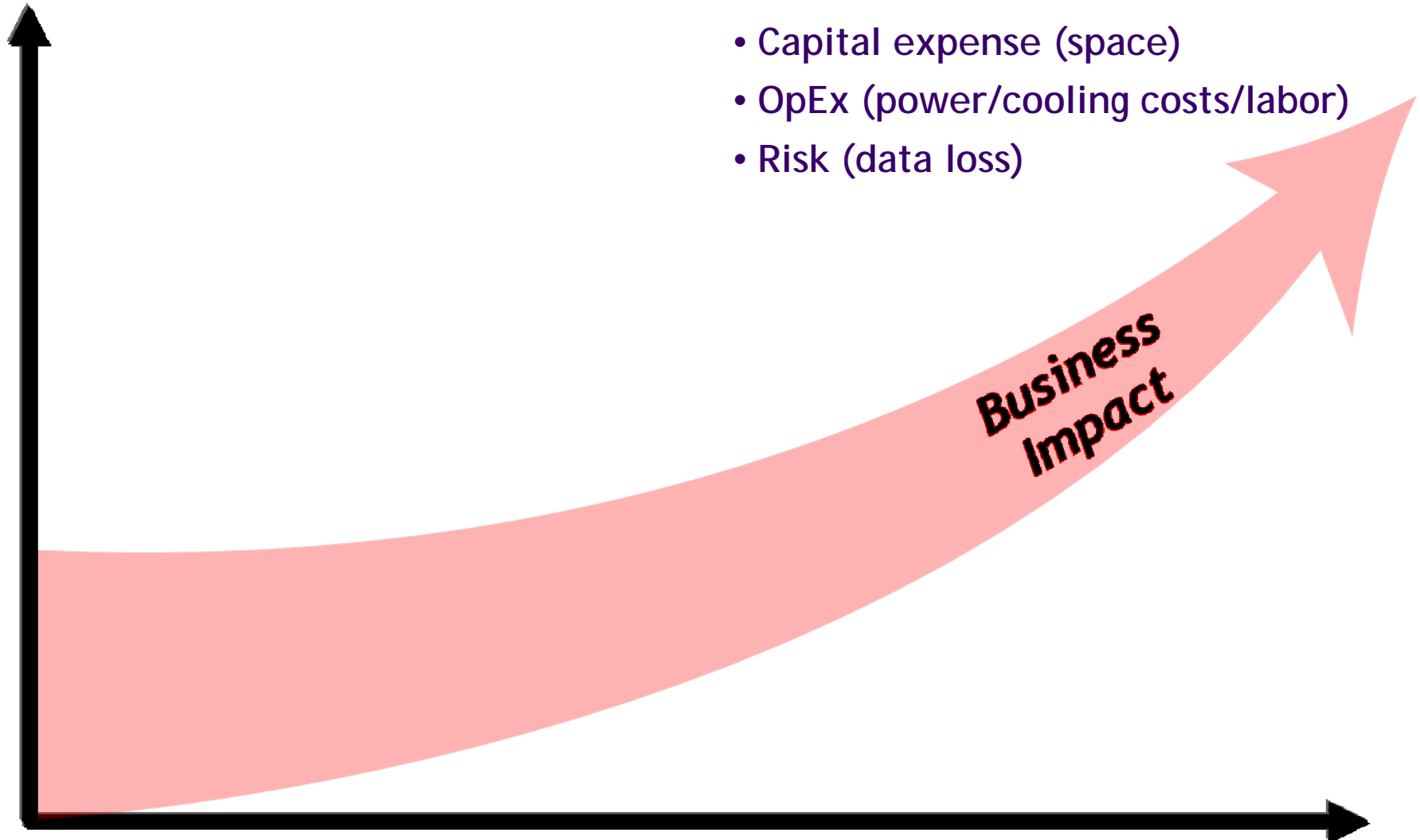
- Performance
- Capacity optimization
- Linear scalability
- Advanced automation
- Expertise
- Service



**Technical
Challenges**

Increasing Cost & Risk: Business Impact

- Capital expense (space)
- OpEx (power/cooling costs/labor)
- Risk (data loss)



Business Issues & SLAs



Rapid Data Growth

- ✓ High CAGR
- ✓ Increased backup costs



SLAs for BC/DR

- ✓ Downtime costs
- ✓ RTO / RPO



Space/Power Limitations

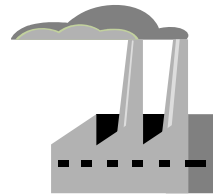
- ✓ Data center footprint
- ✓ Power costs



Regulatory Requirements

- ✓ Online retention
- ✓ Added complexity

**Measurable
TCO & ROI**



Reduce Capital Expense

- ✓ Lower acquisition cost
- ✓ Scalability



Reduce Operating Expense

- ✓ Non disruptive
- ✓ Less labor: Automation
- ✓ Less power & space

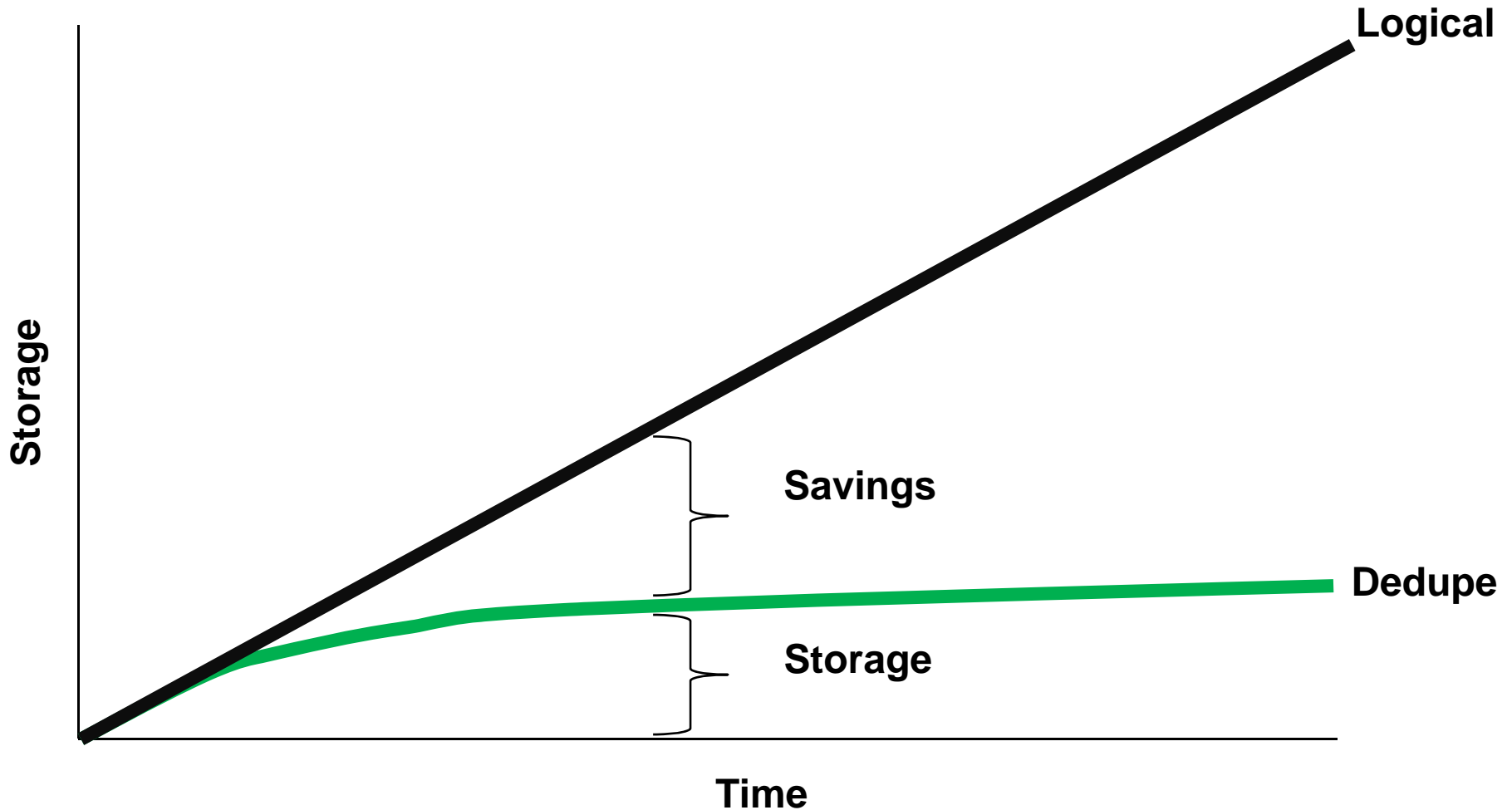


Avoid Costs

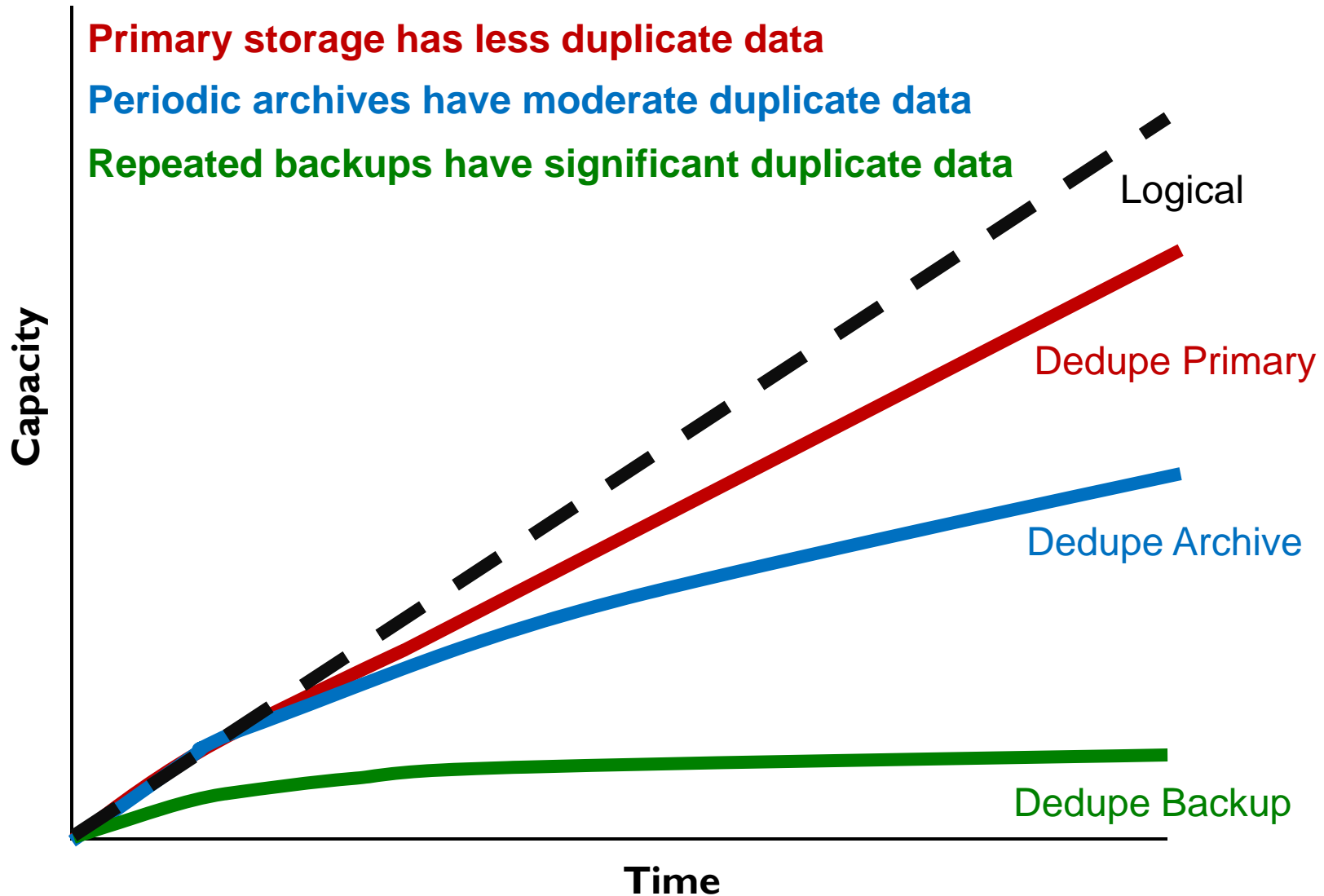
- ✓ Frees IT staff time
- ✓ More data per FTE
- ✓ No human error

Deduplication Controls Growth

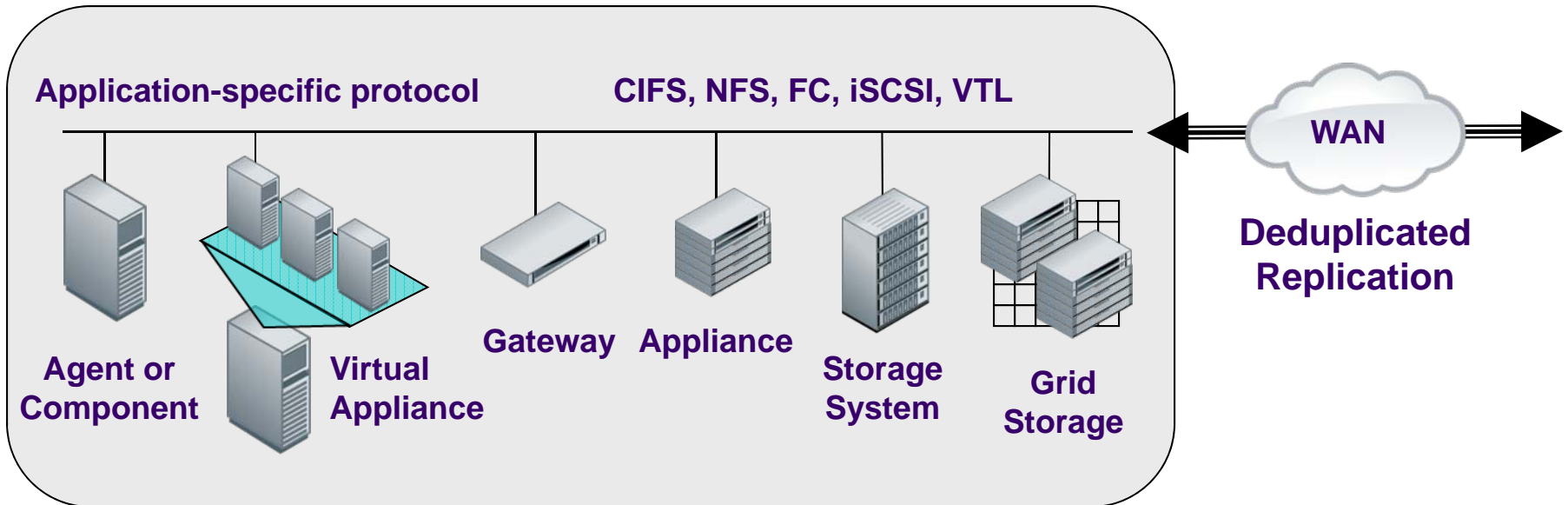
Deduplication ratio typically improves over time



Deduplication Ratio: Depends on Use Case, Time



Deduplication Implementations



- Many options to choose from
- Decision will be influenced by the project goals:
 - ◆ SLAs for data backup and recovery, regulations, etc.

- Focus on your service level agreements (SLAs)
 - ◆ Needs to meet allotted time for *replication*
 - ◆ Needs to meet allotted time for *restore*

- Is it necessary to dedupe all data?
 - ◆ Regulated data may require unique rules
 - ◆ Not all data deduplicates effectively

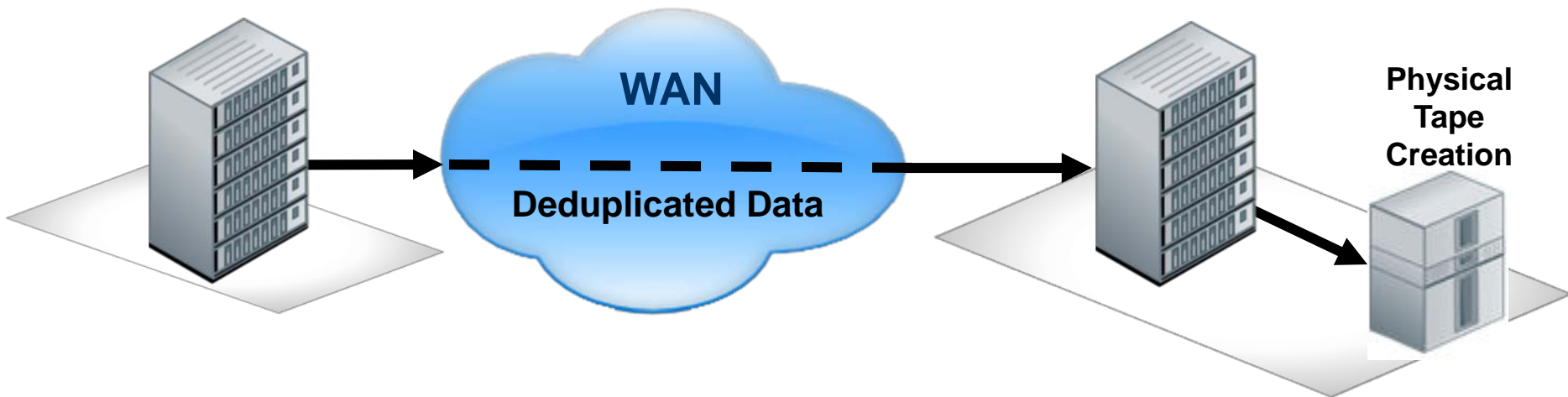
- Can the dedupe solution scale to meet your needs?
 - ◆ Needs to scale in capacity & performance
 - ◆ Different dedupe approaches yield different reduction ratios
 - ◆ CapEx & OpEx savings can be higher (one system vs. multiple)

Automation

- Simplifies the offsite process
- Minimize risk of data loss/data theft
- Leverage existing bandwidth

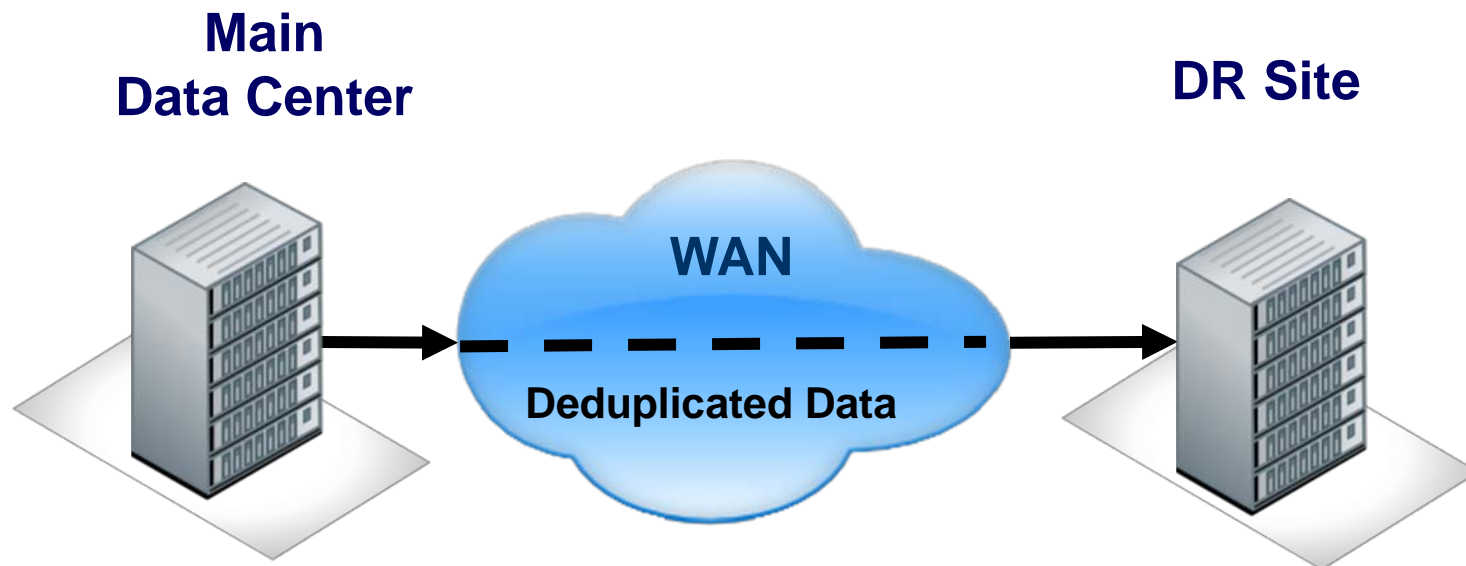
**Main
Data Center**

DR Site



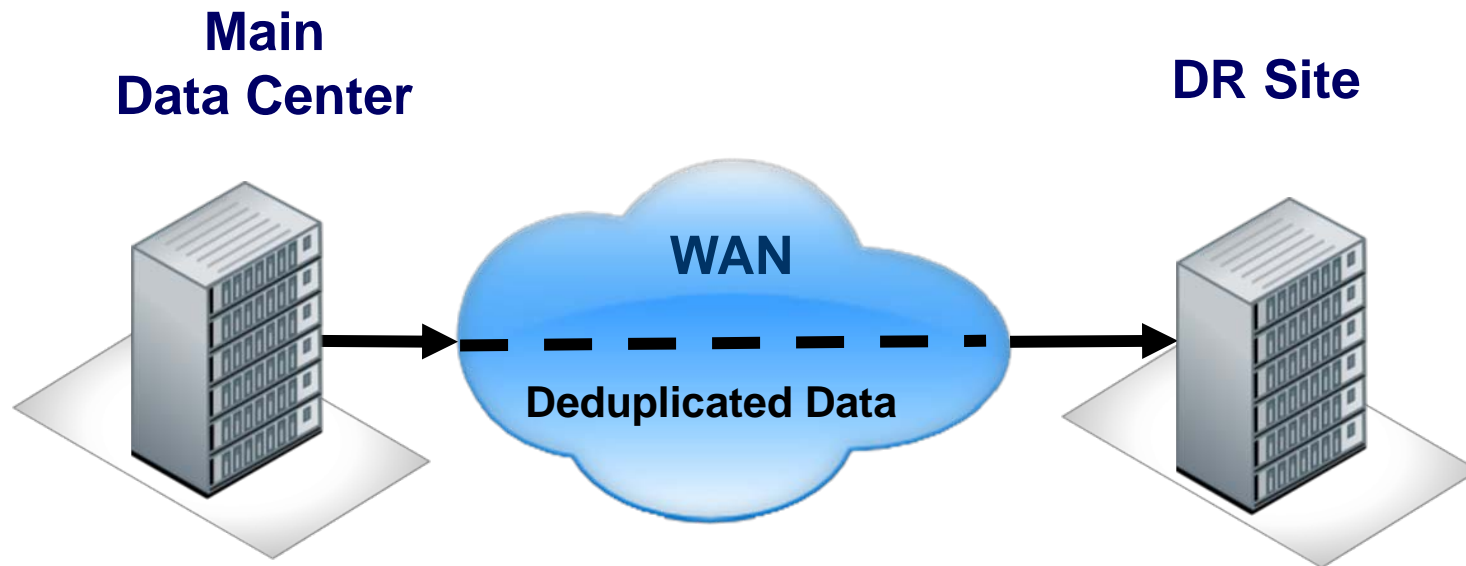
Network Efficiency

- Deduplication dramatically reduces bandwidth usage



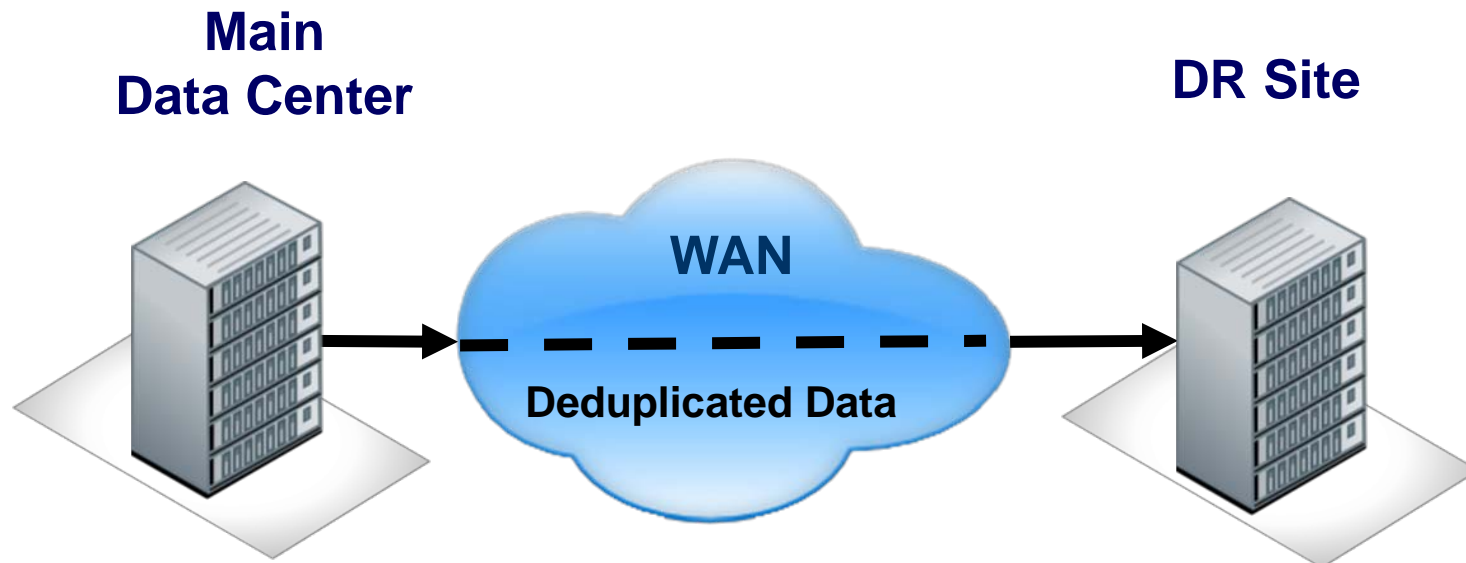
Risk Reduction

- Human error
- Regulatory noncompliance
- Improve data access reliability

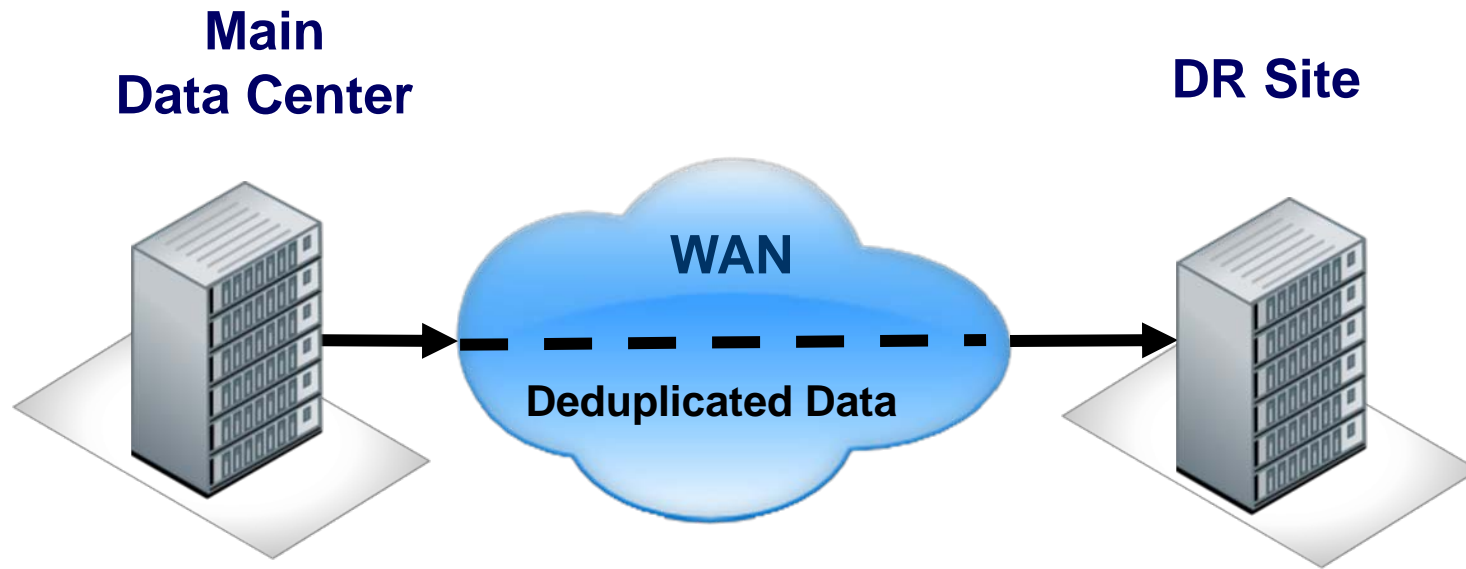


Cost Savings

- Reduced manual media handling
- Reduce tape archival services
- Minimize data loss



Dedupe in DR: Requirements

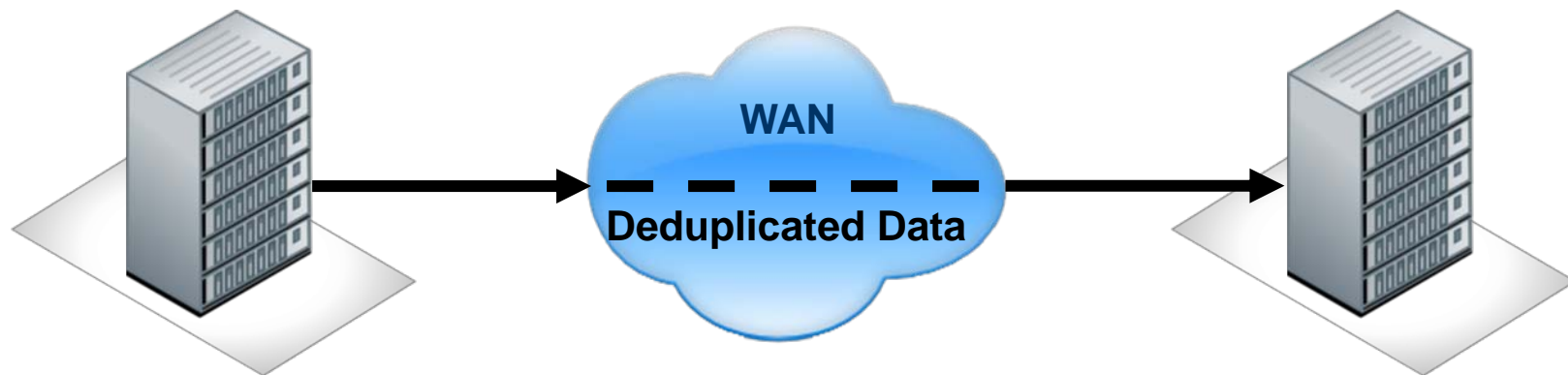


- Replicate large data volumes
- Send only “changed data” over the network
- Perform fast data restores from remote site
- Provide control over replication/restoration process
- Provide resiliency / high availability

Use Model: HQ to DR Location

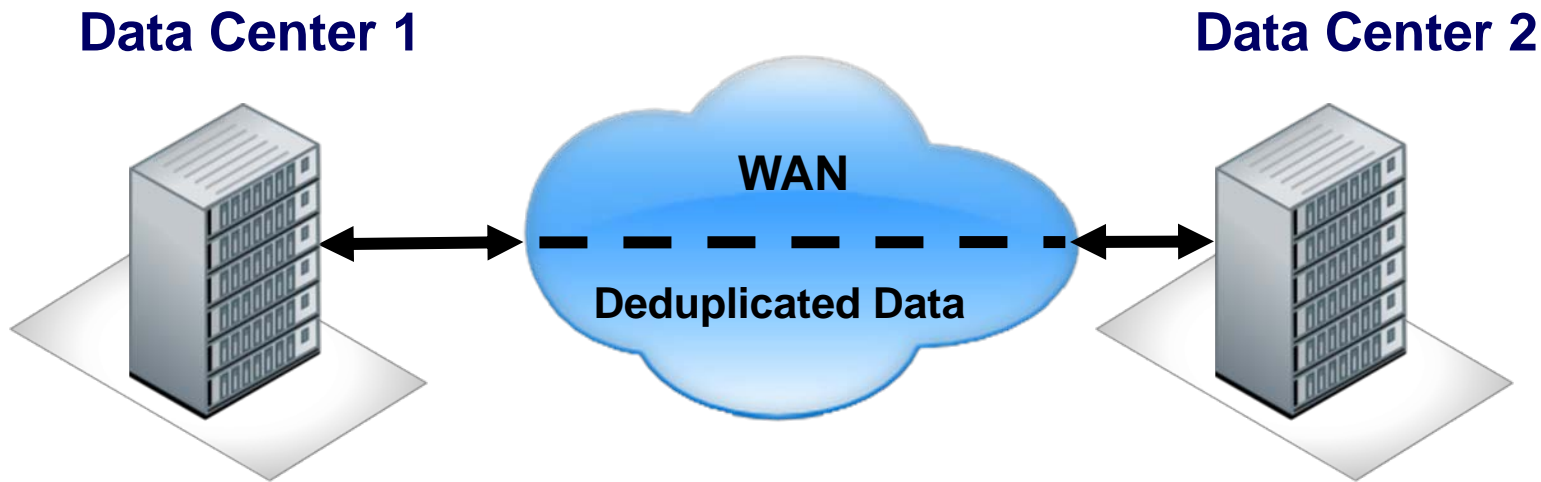
**Headquarters
Data Center**

DR Site



- **Data is deduplicated & replicated to a DR site**
 - ◆ Recover operations in the event of data becoming unavailable at main data center

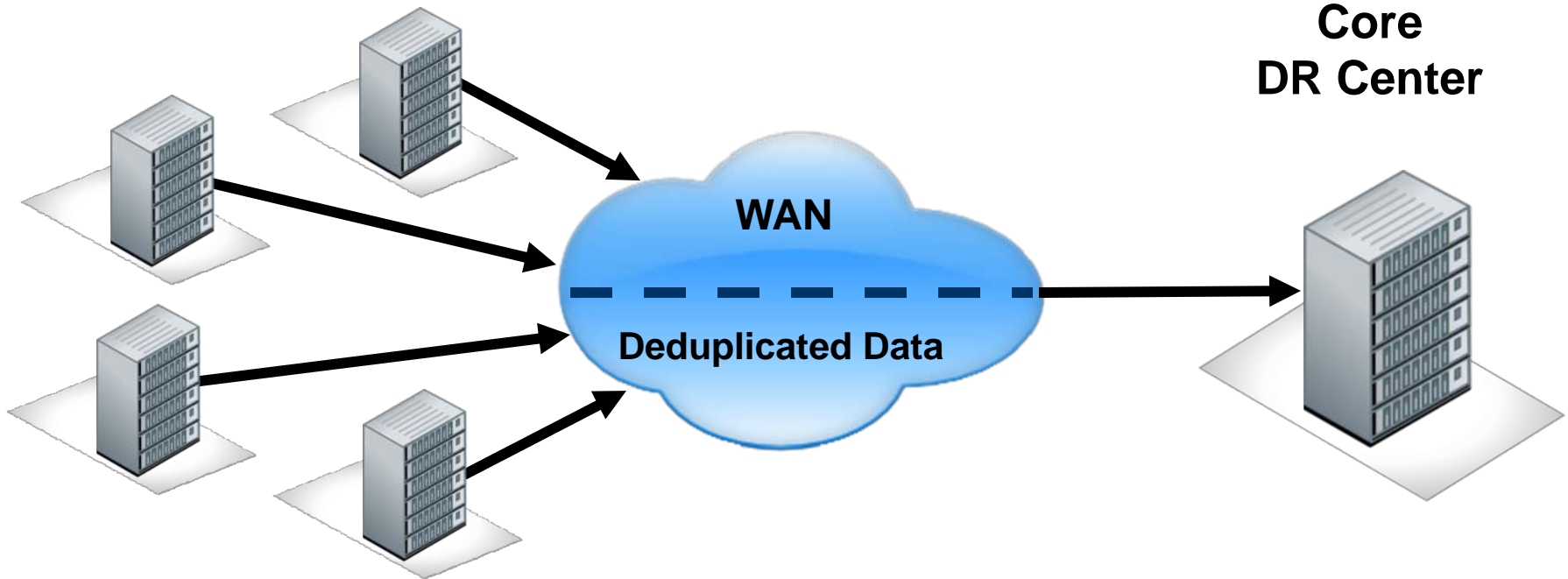
Use Model: Data Center to Data Center



- 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

Regional Data Centers



- Data is deduped & replicated from multiple regional data centers to a main DR center
 - ◆ Core DR center acting as a “DR site” for all production data centers

- Be aware of the challenges
 - ◆ 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
- Choices exist that trade between strengths and weaknesses
- Easy to under-estimate the bandwidth required
 - ◆ $\text{Changed data size} \div \text{replication window} = \text{data rate needed}$

➤ Using deduplication 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 requiring space, power and cooling
 - › Movement toward a greener data center
- ◆ Reduce administrative costs

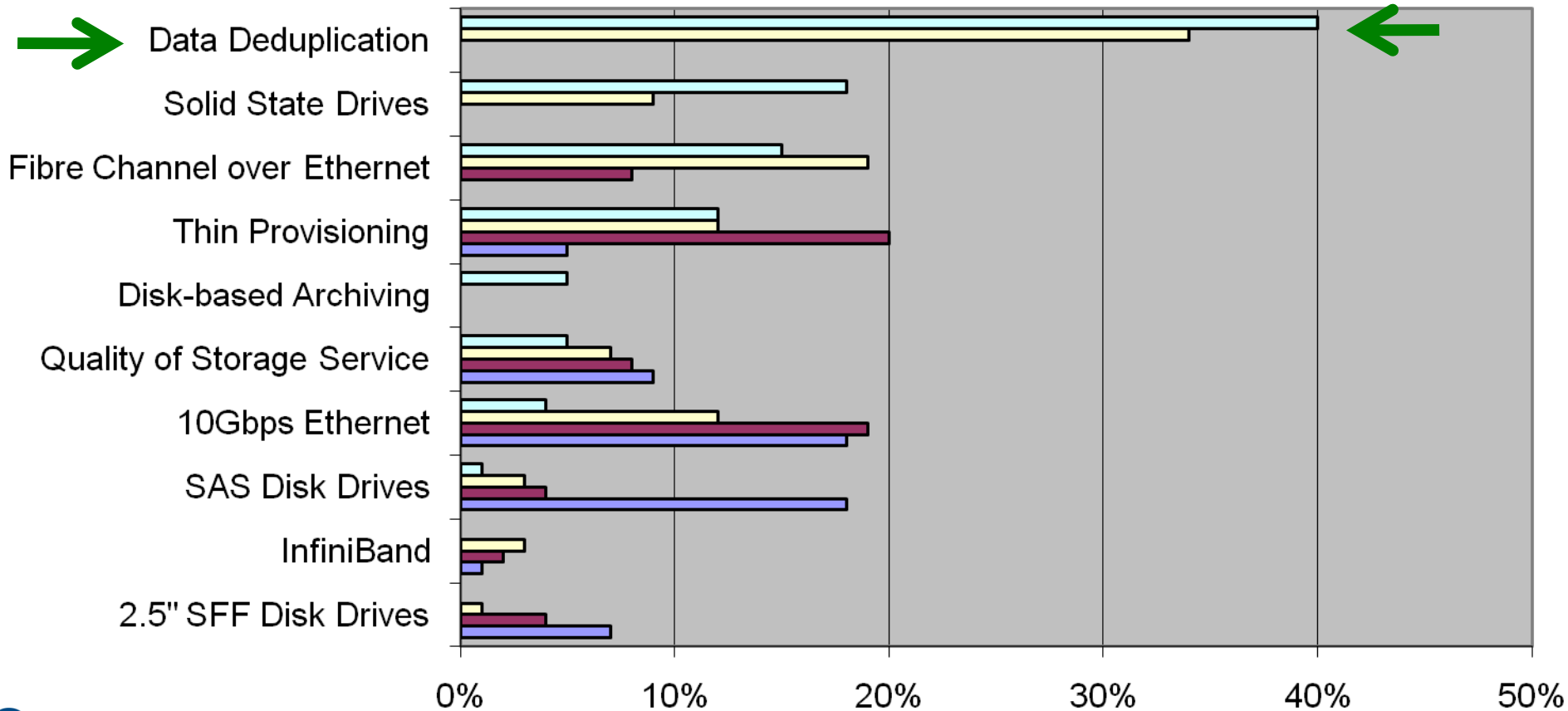
Which of the following technologies will most affect your storage infrastructure during the next three years?

2009 Data Center Conference (N = 103)

2008 Data Center Conference (N = 69)

2007 Data Center Conference (N = 132)

2006 Data Center Conference (N = 97)



- 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**

**Resiliency/HA
of
Deduplication
Solution**

- There is no “right” solution for everyone!
 - ◆ The appropriate solution will vary by environment and requirements
 - ◆ Determine service Level objectives (RTO/RPO) *first* - before selecting and implementing technology
 - ◆ Work with trusted advisors to assess your environment and recommend appropriate solutions

- 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

**Matthew Brisse
David Chapa
Don Deel
Mike Dutch
Larry Freeman
David Hill
Bernd Henning**

**Judy Leach
Gene Nagle
Richard Reitmeyer
Thomas Rivera
Tom Sas
Gideon Senderov**



It's easy
to get
involved
with
the DPCO !

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

www.snia.org/dpco