Education

Introduction to Data Protection: Backup to Tape, Disk and Beyond

Jason Iehl, NetApp

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.

Abstract

- Introduction to Data Protection: Backup to Tape, Disk and Beyond
  - Extending the enterprise backup paradigm with disk-based technologies allow users to significantly shrink or eliminate the backup time window. This tutorial focuses on various methodologies that can deliver an efficient and cost effective disk-to-disk-to-tape (D2D2T) solution. This includes approaches to storage pooling inside of modern backup applications, using disk and file systems within these pools, as well as how and when to utilize virtual tape libraries (VTL) within these infrastructures

About the SNIA DMF

This tutorial has been developed, reviewed and approved by members of the Data Management Forum (DMF)

- The DMF is an industry resource to those responsible for the accessibility and integrity of their organization's information
- The DMF focuses on the technologies and trends related to Data Protection, ILM and Long-term digital information retention

DMF Workgroups:		
Data Protection Initiative (DPI)	**Information Lifecycle Management Initiative (ILMI)**	**Long-term Archive and Compliance Storage Initiative (LTACSI)**
Defining best practices for data protection and recovery technologies such as Backup, CDP, Data deduplication and VTL	Developing, educating and promoting ILM practices, implementation methods, and benefits	Addressing the challenges of retaining, securing, and preserving digital information for the long-term

Backup to Tape, Disk and Beyond

- Fundamental concepts in Data Protection
- Overview of Backup Mechanisms
- Backup Technologies
- Appendix

Data protection is about data availability

There are a wide variety of tools available to us to achieve data protection, including backup, restoration, replication and recovery.

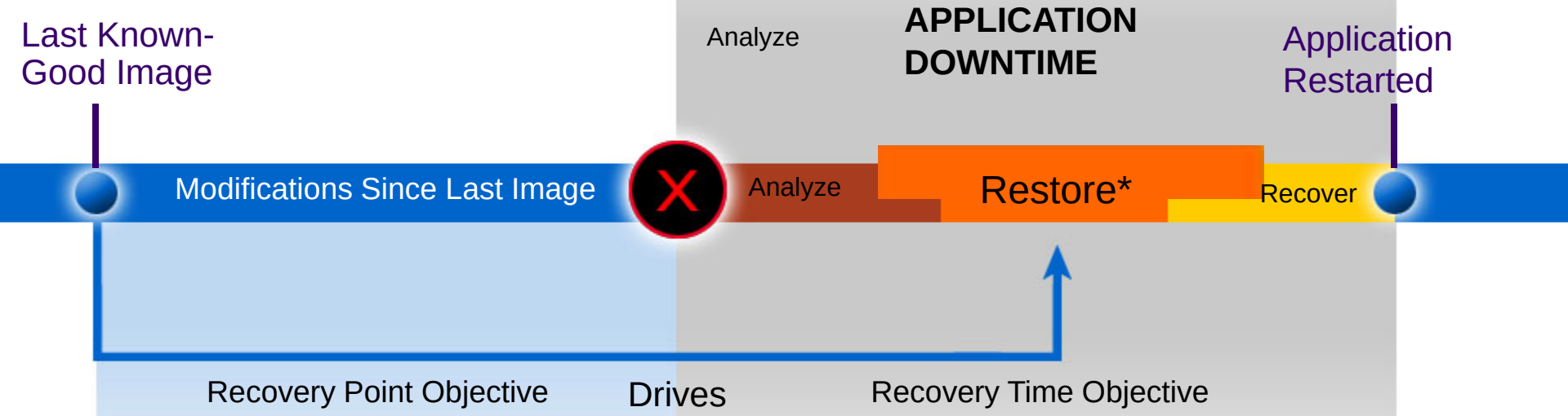
It is critical to keep focused on the actual goal -- availability of the data -- and to balance how we achieve this by using the right set of tools for the specific job.

Held in the balance are concepts like data importance or business criticality, budget, speed, and cost of downtime.

The Process of Recovery

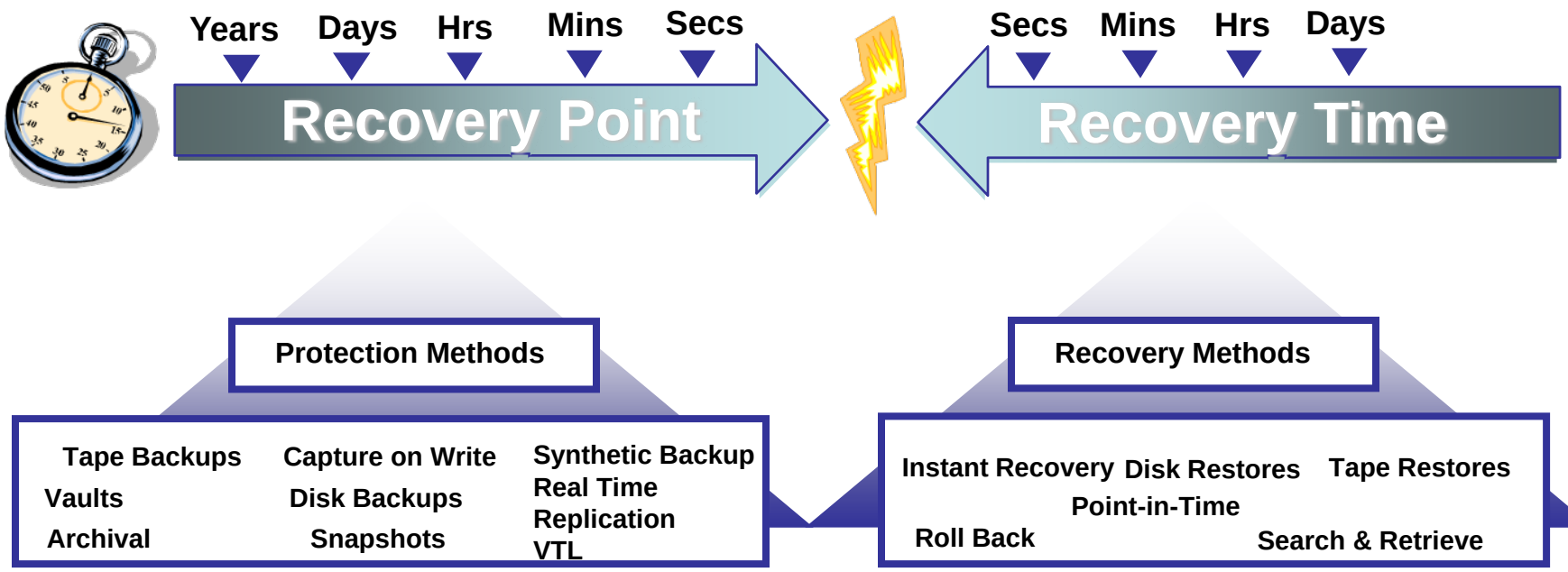
- **Detection**
  - Corruption or failure noted
- **Diagnosis / Decision**
  - What went wrong?
  - What recovery point should be used?
  - What method of recovery -- overall strategy for the recovery?
- **Restoration**
  - Moving the data
  - From tape to disk, or disk to disk, from the backup or archive (source), to the primary or production disks.
- **Recovery –** Almost done!
  - Application environment perform standard recovery and startup operations
  - Any additional steps
    - Log replays for a database
    - Journals replays for a file system
- **Test and Verify**

Traditional Recovery



*10TB = 4 hours from disk, 12.5 hours from tape

Protection Based on Recovery



Methodologies of Backup

- Cold
  - Offline image of all the data
  - As backup window shrinks and data size expands, cold backup becomes untenable.
  - Cheapest and simplest way to backup data
- Application Consistent
  - Application supports ability to take pieces of overall data set offline for a period of time to protect it - application knows how to recover from a collection of individual consistent pieces.
  - No downtime for backup window.
- Crash Consistent or Atomic
  - Data can be copied or frozen at the exact same moment across the entire dataset.
  - Application recovery from an atomic backup performs like a high availability failover.
  - No backup window.

Protection Design Trade-offs

- What's most important:
  - Backup Performance
    - Shorter backup window?
  - Recovery Time Objective (RTO)
    - Speed of recovery
    - How much does it cost to be down?
  - Recovery Point Objective (RPO)
    - Amount of data loss
    - How far back in time to recover data?
  - Move data offsite for DR or archive

- There are trade-offs everywhere
  - Newer technology alters but cannot remove trade-offs
    - Where is the bottleneck?
  - Need to identify the priority order, and establish SLA targets for each data
    - What is the cost of losing data?

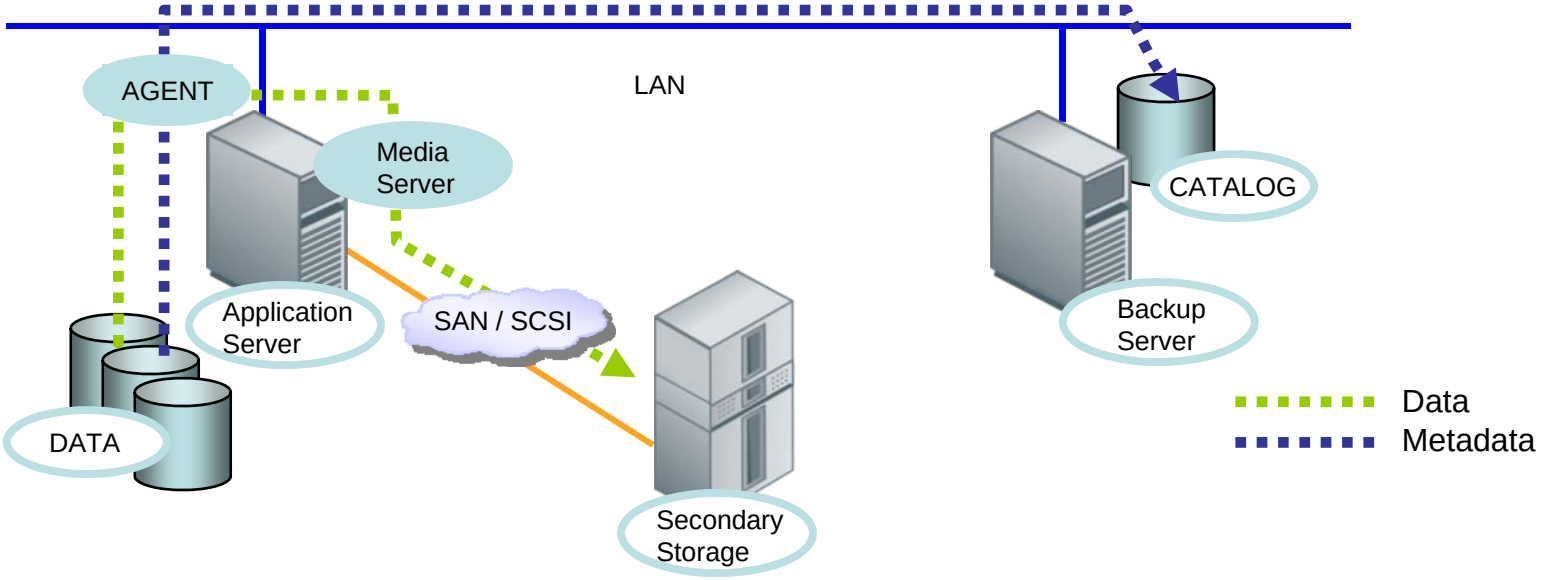
Backup to Tape, Disk and Beyond

- Fundamental concepts in Data Protection
- Overview of Backup Mechanisms
- Backup Technologies
- Appendix

Backup Topology Components

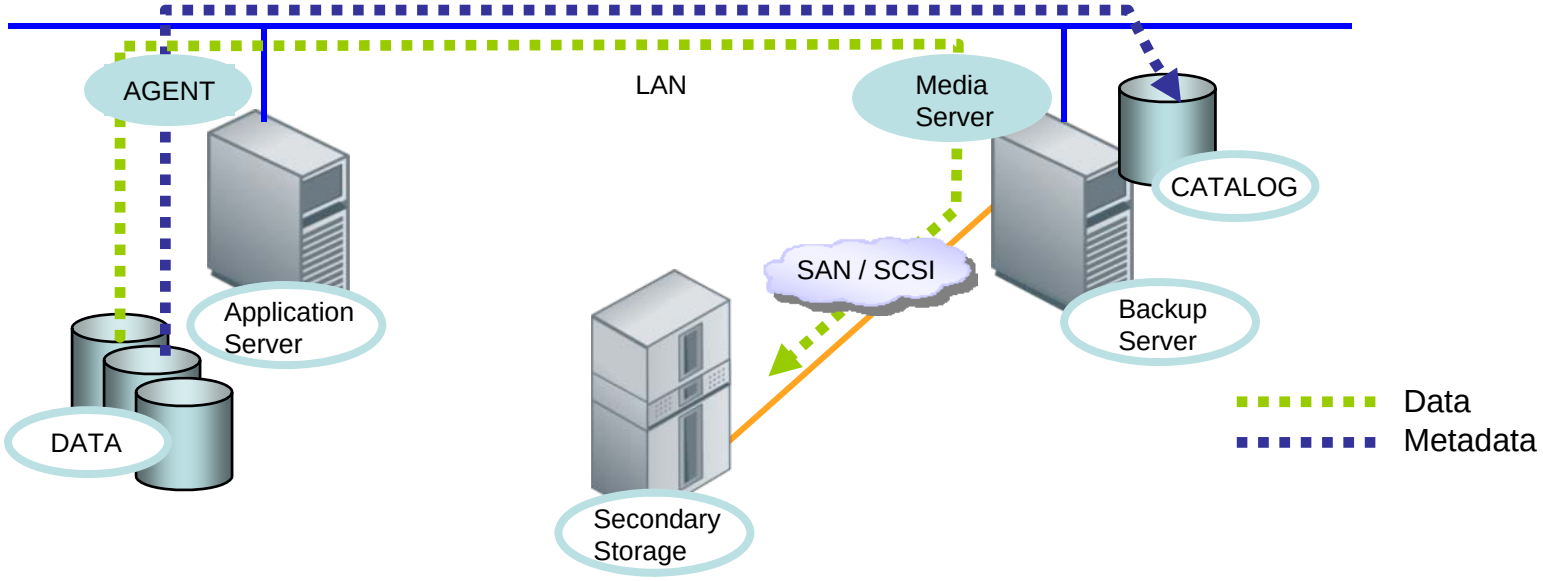
- Agent
  - Manages the collection of the data and Metadata according to the level requested by the backup server
- Storage Node or Media Server
  - Collects the data from the Agent
  - Read and writes to a secondary storage device
- Backup Server
  - Typically single point of administration
  - Owns the Metadata catalog
    - May offer DR for catalog data
- Application Server
  - Server that owns (produces) the data
  - Maybe structured or unstructured data
- Secondary Storage
  - Target for the backup data
  - Traditionally removable media with many moving to disk-based backup

Local Data Mover



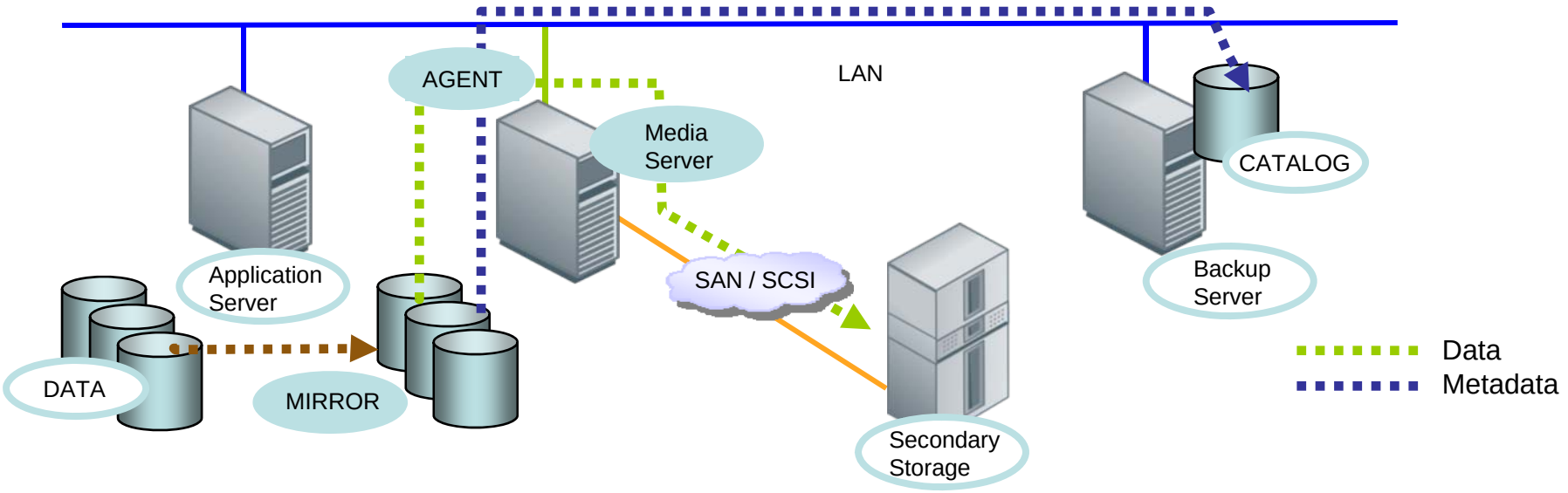
- Sometimes known as LAN-Free
- Application server reads and writes the data locally
  - Application server acts as a media server
  - Storage is accessible by the application server
  - Minimal LAN impact
    - Only Metadata transfers to the backup server
  - May impact bandwidth on application server when backup occurs

LAN Backup



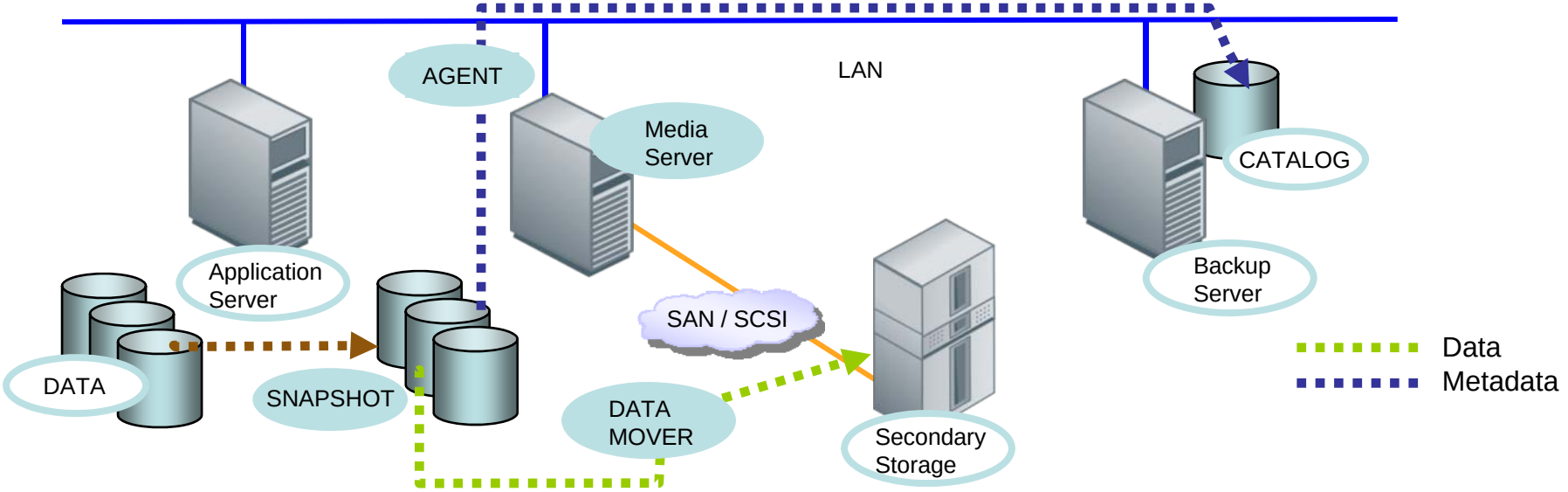
- Backup server receives data and Metadata from application server across the LAN
  - LAN is impacted by both backup and restore requests
  - Application server may be impacted by storage I/O
  - CIFS, NFS, iSCSI, NDMP, or vendor specific

(Application) Server-free Backup



- The application server allocates a snapshot/mirror of the primary storage volume to a media server that delivers the data over the LAN or SAN
  - Media server must understand the volume structure
    - Mirror: Application server impacted when creating the mirror
    - Snapshot: Application server impacted by volume access
  - Metadata go to the backup server

Server-free (Server-less) Backup



- Backup server delegates the data movement and I/O processing to a "Data-mover" enabled on a device within the environment
  - SCSI Extended Copy (XCOPY or "Third-Party Copy")
    - Metadata still sent to the backup server for catalog updates
    - Much less impact on the LAN
  - Network Data Management Protocol (NDMP)
    - NDMP is a general open network protocol for controlling the exchange of data between two parties

Traditional Backup Schedules

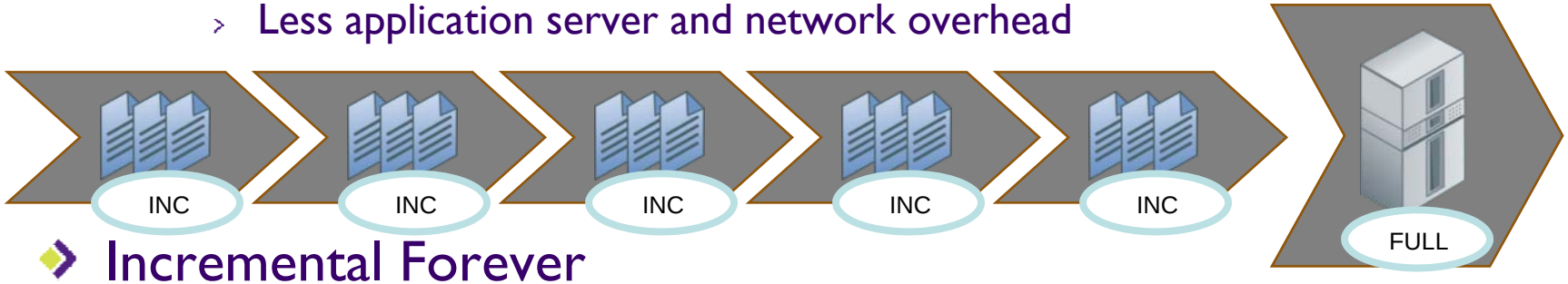
- Full Backup
  - Everything copied to backup (cold or hot backup)
    - Full view of the volume at that point in time
  - Restoration straight-forward as all data is available in one backup image
  - Huge resource consumption (server, network, tapes)
- Incremental Backup
  - Only the data that changed since last full or incremental
    - Change in the archive bit
  - Usually requires multiple increments and previous full backup to do full restore
  - Much less data is transferred
- Differential backup
  - All of the data that changed from the last full backup
  - Usually less data is transferred than a full
  - Usually less time to restore full dataset than incremental

What gets backed up and how

- File-level backups
  - Any change to a file will cause entire file to be backed up
  - Open files often require special handling SW
    - Open files may get passed over – measure the risks
  - PRO: File level backup simplifies both backup and recovery
  - CON: Small changes to large files result in large backups
- Block-level backups
  - Only the blocks that change in a file are saved
  - Requires additional client-side processing to discover change blocks versus entire file
  - PRO:
    - Reduce size of backup data thus improving network utilization
    - In some cases may speed backups
  - CON: Client-side impact may affect client performance
    - Increases backup and restore complexity

Synthetic Backup & Incremental Forever

- Synthetic Full Backups
  - Incremental backups are performed each day
    - Full backups are constructed from incrementals typically weekly or monthly
    - Less application server and network overhead



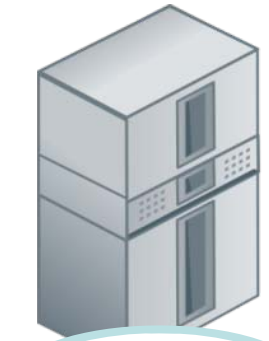
- Incremental Forever
  - Incremental backups are performed every day
  - Primary backups are often sent to disk-based targets
  - Collections of combined incrementals used for offsite copies
    - Usually consolidate images from clients or application and create tapes

Backup to Tape, Disk and Beyond

- Fundamental concepts in Data Protection
- Overview of Backup Mechanisms
- Backup Technologies
- Appendix

Introduction to Tape

- Sequential access technology
  - Versus random access
- Can be removed and stored on a shelf or offsite
  - Disaster Recovery
  - Encrypted, Archived for compliance?
  - Reduce power consumption
- Media replacement costs
  - Tape life, reusability
- Performance and Utilization
  - Can accept data at very high speeds, if you can push it
  - Streaming and multiplexing
- Typically Managed by backup and recovery software
  - Controls robotics (Inventory)
  - Media management



Tape Library

Tape is not Dead!

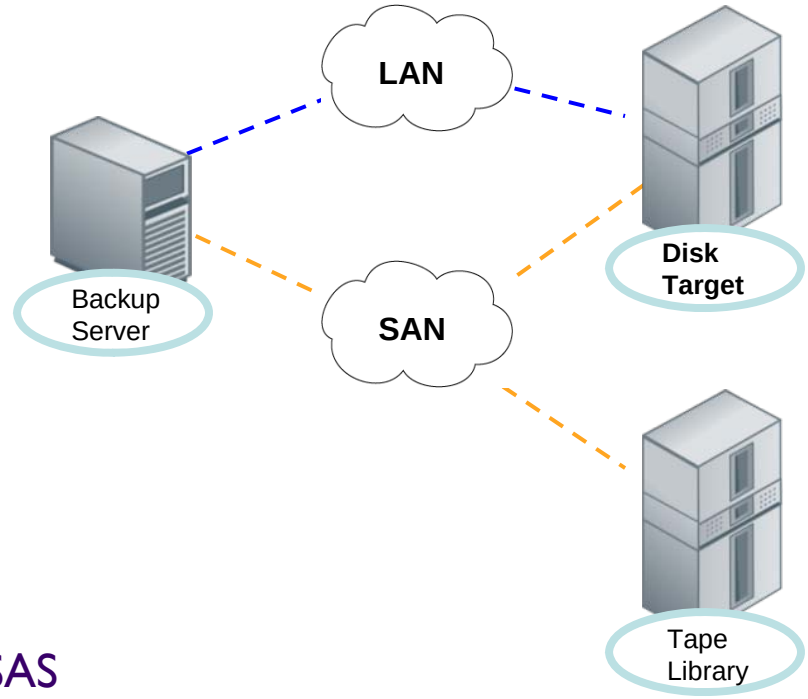
Introduction to B2D

What?

- Backup to Disk / Disk to Disk Backup
- Disk as a primary backup target

Why?

- Performance and reliability
  - Reduced backup window
  - Greatly improved restores
  - RAID protection
  - Eliminate mechanical interfaces
- Eliminate (tape) multiplexing
- Fewer shared devices



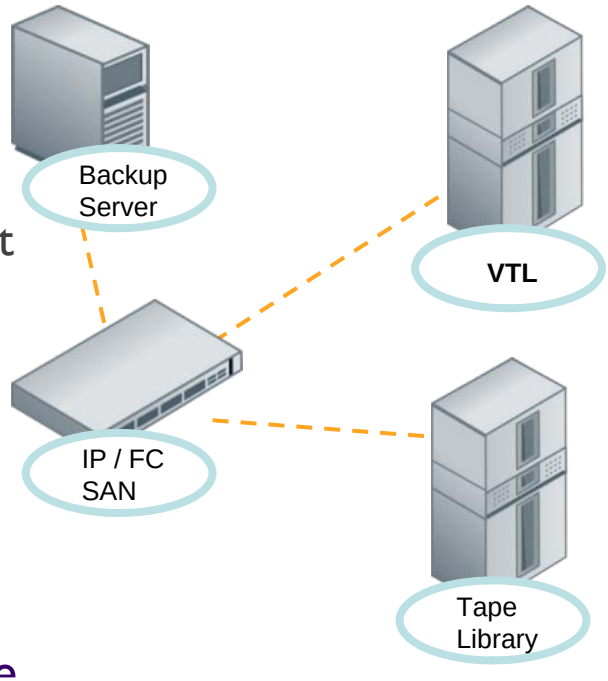
Considerations

- Fibre Channel Disks versus ATA versus SAS
  - I/O random access vs. MB/s sequential
- SAN, NAS or direct Attached
- May require updates to backup software or extra modules
- Consider a mix of Disk and Tape

Introduction to VTL

What:

- Open systems Virtual Tape Libraries
- Fits within existing backup environment
- Easy to deploy and integrate
  - Benefits of B2D by emulating existing tape format
  - Incremental changes to existing environment*
  - Leverage current processes and people
- Reduce / eliminate tape handling



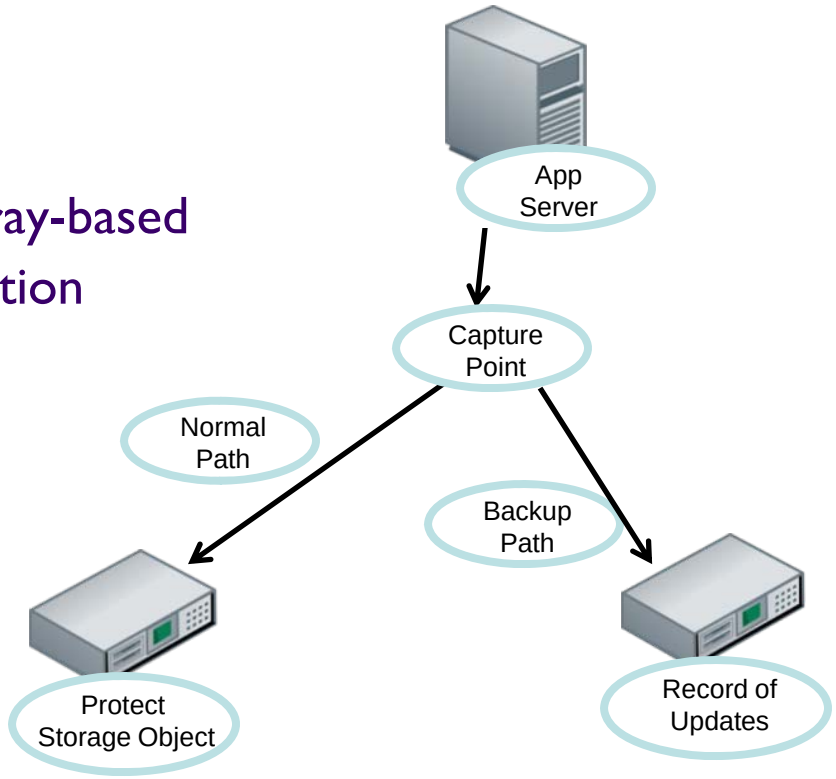
Why:

- Improved performance and reliability (see B2D)
- Reduced complexity versus straight B2D or tape
- Unlimited tape drives reduce device sharing, improve backup times *
- Enables technologies such as remote replication, data deduplication

Introduction to CDP

What:

- Continuous Data Protection
- Capture every change as it occurs
- May be host-based, SAN-based, array-based
- Protected copy in a secondary location
- Recover to any point in time



How:

- Block-based
- File-based
- Application-based

Why:

Implementations of true CDP today are delivering zero data loss, zero backup window and simple recovery. CDP customers can protect all data at all times and recover directly to any point in time.

Introduction to Data Deduplication

What?

- The process of examining a data-set or I/O stream at the sub-file level and storing and/or sending only unique data
- Client-side SW, Target-side HW or both

Why?

- Reduction in cost per terabyte stored
- Significant reduction in storage footprint
- Less network bandwidth required

Check out SNIA Tutorial:

Understanding Data Deduplication

Considerations

- Greater amount of data stored in less physical space
- Suitable for backup, archive and (maybe) primary storage
- Enables lower cost replication for offsite copies
- Store more data for longer periods
- Beware 1000:1 dedupe claims – Know your data and use case
- Multiple performance trade-offs

Next Steps in Data Protection

- Choose the appropriate level of protection
  - Assess risk versus cost versus complexity
  - Include your "customers" in your decisions
- Match RPO, RTO goals with technology
  - Consider resources required to support your decisions
  - Consider centralized versus distributed solutions
- Performance is ALWAYS a consideration
  - Assess your system today for strengths and weaknesses
  - A new box or new SW may NOT be the answer
- When in doubt, call in the experts

Where to Get More Information

- Related tutorials
  - Trends in Data Protection and Restoration Technologies
  - Deduplication – Methods of Achieving Data Efficiency
  - In the Face of Litigation: Best Practices for Retention, Discovery, and Deletion
  - A Crash Course in Wide Area Data Replication
- Visit the Data Management Forum website
  - http://www.snia-dmf.org
- Data Protection Buyers Guides available
  - Chapters on Continuous Data Protection, Deduplication, and Virtual Tape Libraries

Q&A / Feedback

Please send any comments on this tutorial to SNIA at: trackdatamgmt@snia.org

The DMF would like to thank the following individuals for their contributions to the development of this tutorial:

SNIA Data Protection Initiative
SNIA Data Management Forum

Michael Fishman	**Nancy Clay**	**Philippe Reynier**
Mike Rowan	**SW Worth**	**Jason Iehl**

SNIA Data Management Forum

It's easy to get involved with the DMF !

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

www.snia.org/dmf

Thank you for your feedback

Education SNIA

Questions and Answers

APPENDIX

Backup versus Mirroring

- **Backup**
  - Protecting data by making copies or allowing copies to be generated from saved data
  - Examples: snapshots, split mirrors, VTL, tape, CDP
  - When?
    - **Multiple Recovery Points needed**
    - **Recovery from data corruption**
    - **Archival and indexing**
- **Mirroring/Replication**
  - Protecting data by moving the data, usually as it changes, to a remote copy. Synchronous or Asynchronous mirroring
  - When?
    - ***Disaster Recovery Time Objective (DR/RTO centric usually)***
    - ***Data Migration***
    - ***Content Distribution***

Introduction to Snapshots

A disk based "instant copy" that captures the original data at a specific point in time. Snapshots can be read-only or read-write.

"A fully usable copy of a defined collection of data that contains an image of the data as it appeared at the **point in time** at which the copy was initiated. A snapshot may be either a ***duplicate*** or a ***replicate*** of the data it represents.

www.snia.org/dictionary"

Snapshot of Networked Storage

- **Terminology:** Snapshot, Checkpoint, Point-in-Time, Stable Image = Any technology that presents a consistent point-in-time view of changing data. *Many implementations exist.*
- **Why?** Allows for complete backup or restore, with application downtime measured in minutes (or less)
- Most vendors: Image only = (entire Volume)
- Backup/Restore of individual files is possible
  - If conventional backup is done from snapshot
  - Or, if file-map is stored with Image backup

Snapshot Comparison

	Full Copy Snapshot	Differential Copy Snapshot
Upsides	• No cost during "snapshot" process • Can be used for DR - independent copy	• Less storage consumption - typically 10-20% ◦ Depends on churn • Typically can take advantage of cheaper disk

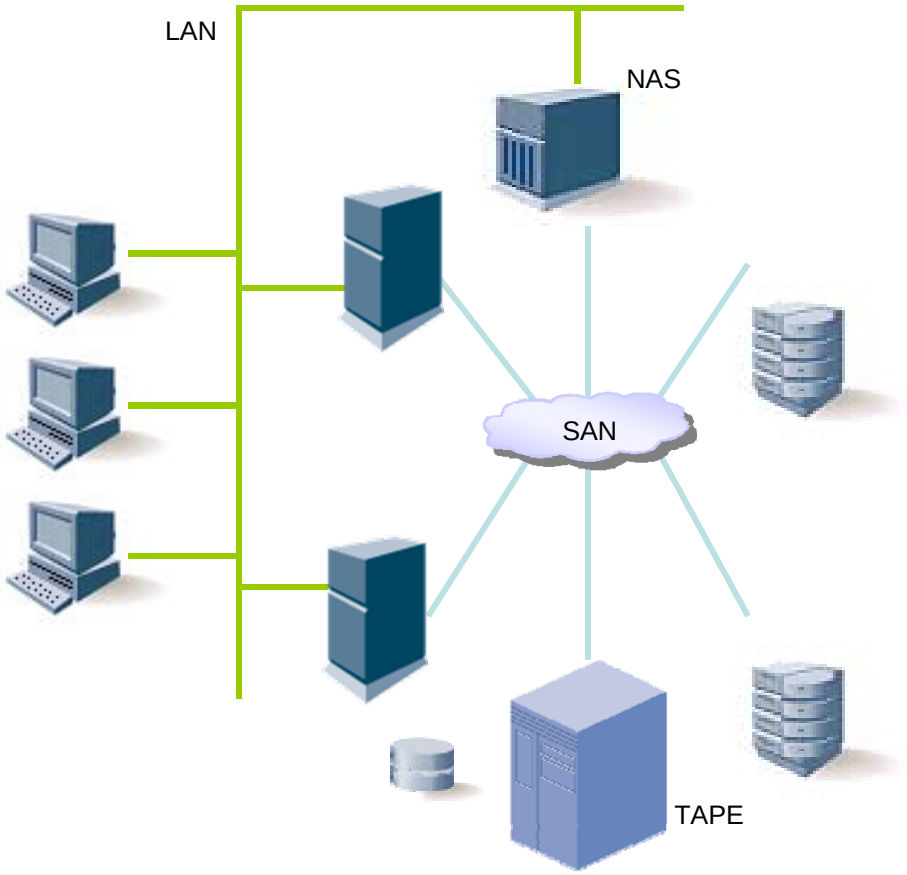
Downsides	• Massive storage cost
	  ◦ 1x of storage per RPO
	  ◦ Like disk - expensive
	• Often in the same disk frame
	  ◦ Loss of DR component
	• Consider re-sync time in schedules
		• Performance impacts while snapshot exists
		  ◦ Multiple implementations to optimize performance impact
		  ◦ Most vendors don't offer multiple implementations - pick at onset
		• Leverages main copy - not DR capable

Applications	• Disaster Recovery
	• Near zero backup window
	  ◦ 24x7 operations
	• Faster restore
	  ◦ Can do no-copy restore
	  ◦ Most run-books require copy
	• Can help with data repurposing
		• Backup source
		• Near zero backup window
		  ◦ 24x7 operations
		• Fast restore
		  ◦ copy based by definition
		• Can help with data repurposing
		  ◦ Beware performance impact

Performance Considerations

- Performance
  - Caching on storage node for multiplexing and stream management
    - Cache to Virtual Tape or Disk prior to tape
  - Synthetic full backup – offload backup engine
    - faster backup = Incrementals
    - faster restore = restore from full session
- Smart recovery functions
  - Most recent image
  - Consistent view = true image
  - Minimize downtime
    - time to diagnose
    - time to restore
  - Lost time + downtime = total loss time

Backup Scaling



Constant change and **heterogeneity** in technologies

- Operating Systems
- Disk Storage Appliances
- Network Architectures / Topologies
- Tape Storage Devices

Challenge

- **Protect** mission-critical data
- **Timely** backup and restore
- **Administration** overhead
- **Optimize** storage resources

Flexibility

- Heterogeneous Environment
  - Multiple platforms (HW / OS)
  - Multiple tape drives & libraires
  - Multiple applications
  - NAS and SAN
  - Snapshot facilites
- Advanced Tape Management
  - Tape mirroring
  - Off-site storage
  - Multiplexing
- Advanced Library Management
  - Sharing, partitioning
  - Port handling
- Security
  - Authentication & Encryption
  - DMZ / Firewall support

Manageability

- Centralized Administration
  - Web GUI & smart interface
  - Backup strategies
  - Scheduling
  - Onsite and offsite mmedia management
- Centralized Supervision
  - Real-time monitoring
  - Alarms
  - Event log
  - SNMP compliant, integration with
  - Frameworks
  - Bill back