



Education

# Trends in Data Protection and Restoration Technologies

Mike Fishman  
Education Chair: Data Management Forum

# SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
  - ◆ Any slide or slides used must be reproduced without modification
  - ◆ The SNIA must be acknowledged as 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 nor should be construed as legal advice or opinion. If you need legal advice or legal opinion please contact an attorney.
- The information presented herein represents the author's personal opinion and current understanding of the 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.**

Many disk technologies, both old and new, are being used to augment tried and true backup and data protection methodologies to deliver better information and application restoration performance. These technologies work in parallel with the existing backup paradigm,

This session will discuss many of these technologies in detail. Important considerations of data protection include performance, scale, regulatory compliance, recovery objectives and cost. Technologies include contemporary backup, disk based backups, snapshots, continuous data protection and capacity optimized storage.

Detail of these technologies interoperate will be provided as well as best practices recommendations for deployment in today's heterogeneous data centers.

- ◆ Understand legacy and contemporary storage technologies that provide advanced data protection
- ◆ Compare and contrast advanced data protection alternatives
- ◆ Gain insights into emerging DP technologies.

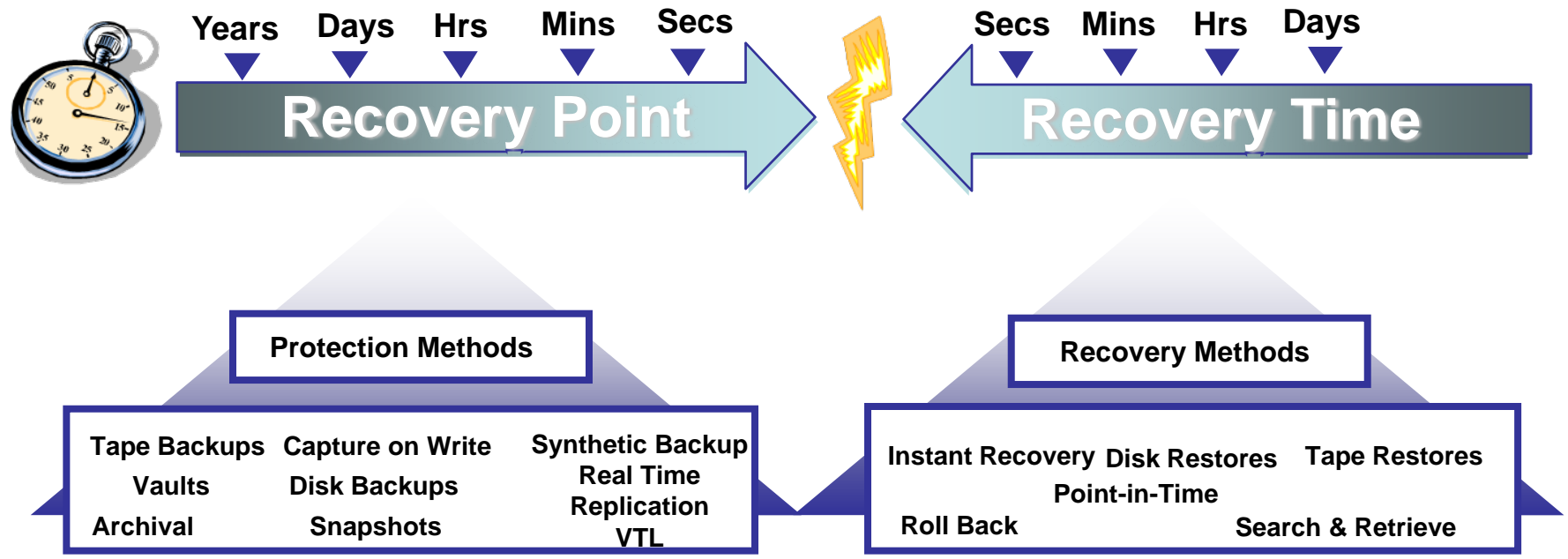
# About the SNIA DMF

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

<b>DMF Workgroups:</b>		
<b>Data Protection Initiative (DPI)</b>	<b>Information Lifecycle Management Initiative (ILMI)</b>	<b>Long-term Archive and Compliance Storage Initiative (LTACSI)</b>
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

# Protection Based on Recovery



## Enabling Technologies



When an application is running during the “copy” process, various techniques are available to ensure data consistency

Much like the “open files” issue when backing up a file system that is in use, applications (like databases, messaging systems, etc) allow for different approaches to capturing a holistic picture of the applications data during a copy process (such as a snapshot, a mirror-split, or CDP protection).

It is important to understand the consistency semantics of your application so that your data protection copies are recoverable.

# To Quiesce or Not?

- **Cold Snapshot**
  - ◆ Less complex, but backup window is downtime
  
- **Application Consistent Snapshot**
  - ◆ Application intervention
  - ◆ Application dependent
  - ◆ “Hot Backup” or “Online Backup”
  
- **Atomic or Crash Consistent Snapshot**
  - ◆ Ability to take snapshot for entire dataset at exactly same moment
  - ◆ Can be done in multiple ways
  - ◆ Recovery domain same as high availability systems

# Data Protection and Data Management

## Data Protection

- Disk-Assisted and Disk-based protection methods
- Array and storage network based data protection
- Object based Archival
- Tape based data protection
- Backup to Virtual Tape
- Backup to Disk

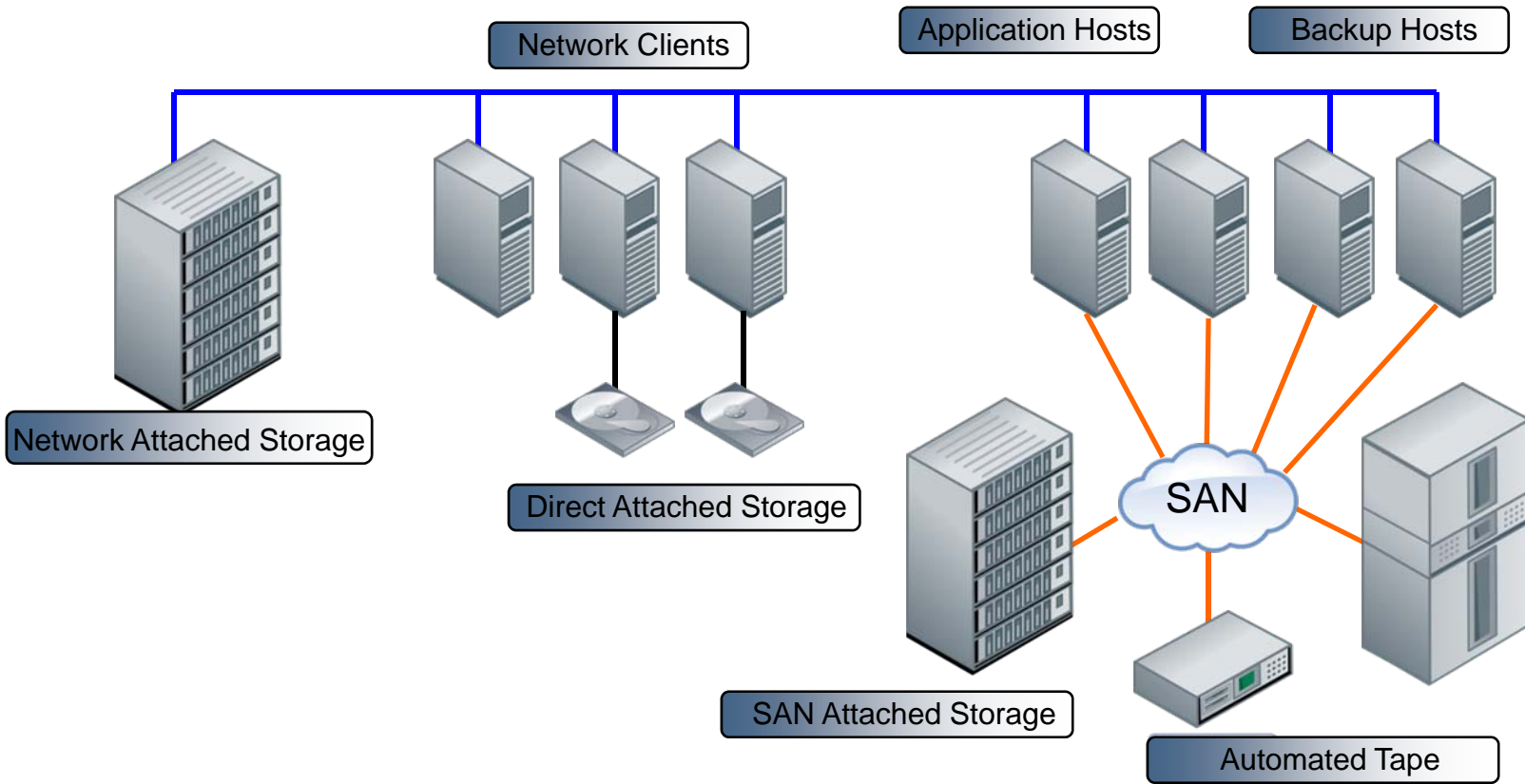
## Data Management

- Information classification
- Information valuation (\$\$\$)
- Information lifecycle management

## Tiered Storage

- Primary
- Secondary
- Archive
- Backup

# Backup Networking 101



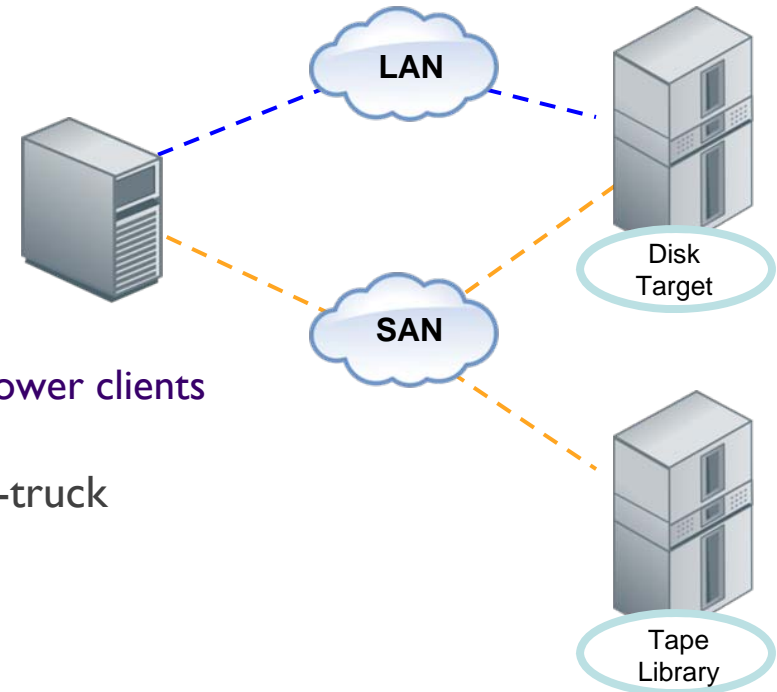
# Tape Based Backup, High Speed SANs

- Tape drives run faster than most backup jobs – Is this good?
  - ◆ Matching backup speed is more important than exceeding it
  - ◆ Avoid shoe-shining
- Slower hosts can tie up an expensive drive
  - ◆ It's a shame to waste a drive on these hosts.
- Slower tapes can tie up expensive (important) servers.
  - ◆ It's a shame to let the tape drive throttle backup servers
  - ◆ Slow backup can impact production servers as well
- Replacing your tapes may not solve your backup challenges
  - ◆ A well designed backup architecture is the best answer
- If backup target speed is your issue:
  - ◆ Consider multiplexing – Good for backup, not-so-good for restore
  - ◆ Consider alternates such as virtual tape, B2D or use LAN backup.
- Security, security, security.....

# Backup to Disk (B2D)

## ➤ What and Why?

- ◆ Backup target is a LUN or a share
  - › Easy to implement
- ◆ Any type of disk or connection
  - › FC, SATA, SAS, DAS, NAS, etc
- ◆ Disk-based Snapshots
  - › Reduce impact on production host
  - › Improve tape streaming when backing up slower clients
- ◆ Single/Volume Restores are faster than tape
- ◆ Remote and local backups on-line versus on-truck



## ➤ What to watch out for

- ◆ Bottlenecks may NOT be the backup target
- ◆ Disks can be faster than tape, but when?
- ◆ More difficult and possibly slower than VTL (generally)
- ◆ Current versions of enterprise backup applications support
  - › May charge more for advanced B2D functionality
- ◆ Backup window issues may still exist - not “instant”

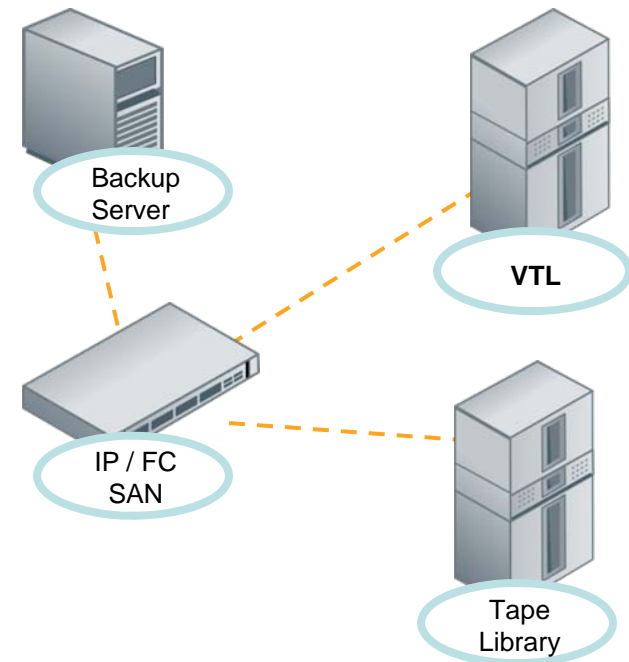
# Virtual Tape Library (VTL)

## What:

- Looks like Tape, Acts like tape
- Fits within existing backup environment
- Easy to deploy and integrate
- Takes advantage of current processes
- Reduces tape media handling

## Why:

- Improved speed and reliability
  - Exceeds speed of high end tape
  - No fast-forward, no rewind,
  - High performance without multiplexing
  - Enables faster access to data, faster restores
  - No detached tape leaders or mechanical failures
- Watch out:
  - Integration with physical tape
  - Consider total aggregate speed as well as speed per-drive
  - Backup savesets are highly redundant - Everyone needs deduplication.....

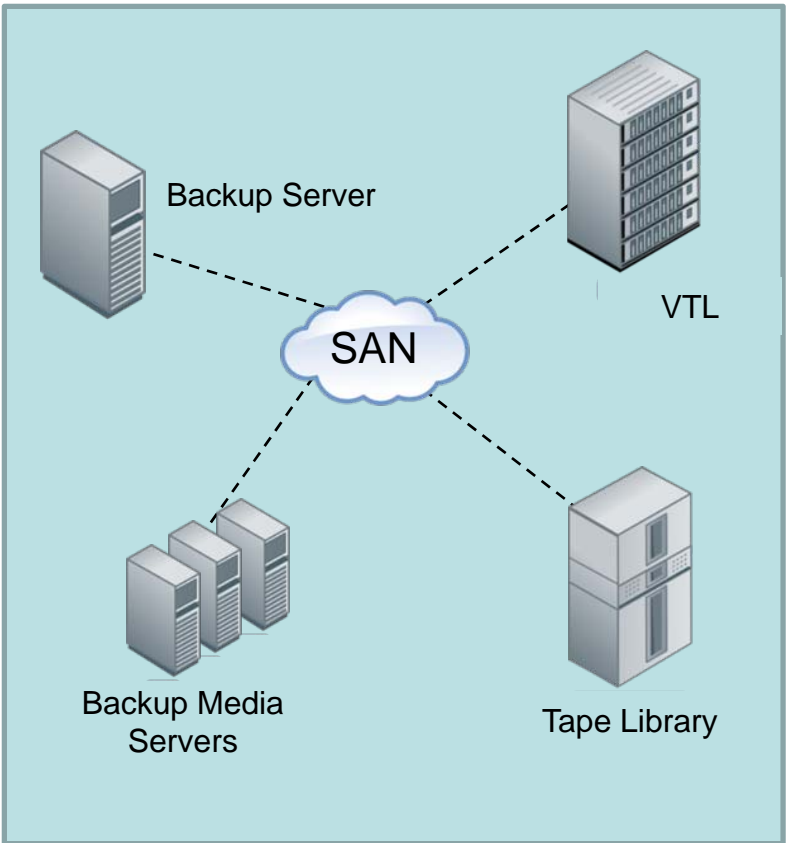


# The VTL Difference

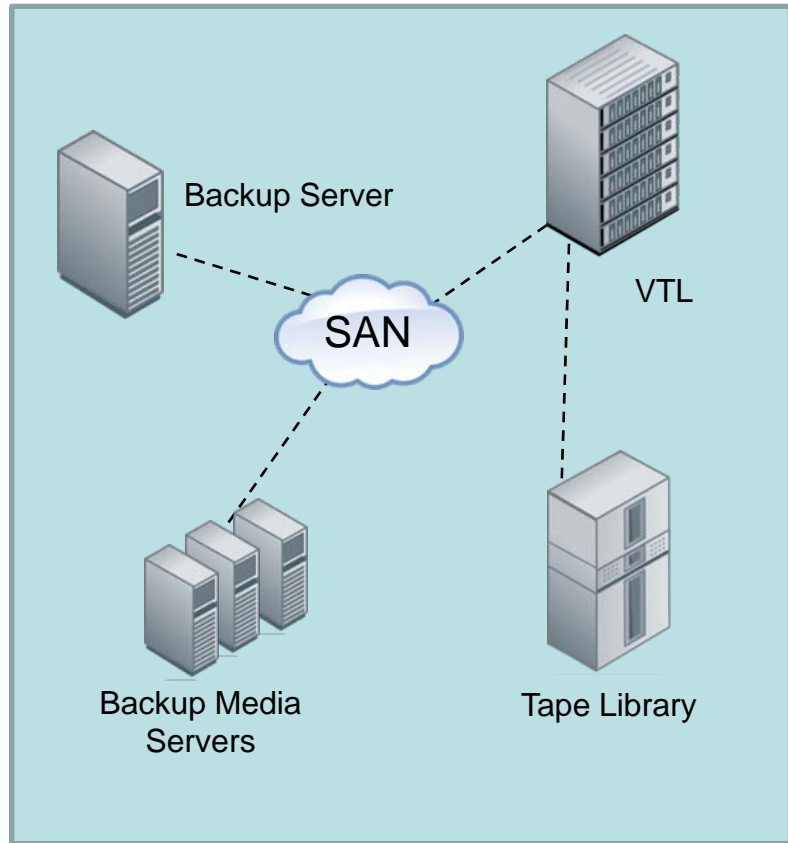
- Easy to manage in traditional backup software environment:
  - ◆ Works like normal tape library
    - › Fits into existing backup and restore processes
  - ◆ Viewed as open systems cartridges, robot, tape drives, and in some cases even a mail slot
  - ◆ Standard tape copy, cloning, or vaulting functions apply for off-site copies
    - › Used to replicate data to physical tape for long term retention
  
- Cost effective solution
  - ◆ Leverages lower cost disk, SAS, SATA
  - ◆ Deduplication enables higher density and network bandwidth reduction
  - ◆ Can extend the life of current physical tape investment
    - › Used as a front-end to the backup process
    - › Tape may still used for longer term retention

# VTL and Physical Tape Deployments

VTL as peer



VTL controls tape



# VTL Best Practices

- Use as primary backup target to reduce backup window
  - ◆ Leverage existing tape as backup target where appropriate
- Add storage to enable additional recovery time objectives
- Follow physical tape configuration and sharing rules
  - ◆ Match virtual drives per connection
  - ◆ Don't mix tape and disk on same ports
  - ◆ Use the right OS driver
- Tape redeployment
  - ◆ Eject process, controlled by the backup software
  - ◆ Cloning, vaulting, tape copies
  - ◆ Backend tape creation
- Offsite requirements
  - ◆ Bandwidth, connectivity, time to complete tape copies

# What are 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](http://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

	<b>Full Copy Snapshot</b>	<b>Differential Copy Snapshot</b>
<b>Upsides</b>	<ul style="list-style-type: none"> <li>◆ No cost during “snapshot” process</li> <li>◆ Can be used for DR - independent copy</li> </ul>	<ul style="list-style-type: none"> <li>◆ Less storage consumption - typically 10-20%                             <ul style="list-style-type: none"> <li>◦ Depends on churn</li> </ul> </li> <li>◆ Typically can take advantage of cheaper disk</li> </ul>
<b>Downsides</b>	<ul style="list-style-type: none"> <li>◆ Massive storage cost                             <ul style="list-style-type: none"> <li>◦ 1x of storage per RPO</li> <li>◦ Like disk - expensive</li> </ul> </li> <li>◆ Often in the same disk chassis                             <ul style="list-style-type: none"> <li>◦ Loss of DR component</li> </ul> </li> <li>◆ Consider re-sync time in schedules</li> </ul>	<ul style="list-style-type: none"> <li>◆ Performance may be impacted while snapshot exists                             <ul style="list-style-type: none"> <li>◦ Multiple implementations to optimize performance impact</li> <li>◦ Most vendors don't offer multiple implementations - pick at onset</li> </ul> </li> <li>◆ Leverages main copy - not DR capable</li> </ul>
<b>Applications</b>	<ul style="list-style-type: none"> <li>◆ Disaster Recovery</li> <li>◆ Near zero backup window                             <ul style="list-style-type: none"> <li>◦ 24x7 operations</li> </ul> </li> <li>◆ Faster restore                             <ul style="list-style-type: none"> <li>◦ Can do no-copy restore</li> </ul> </li> <li>◆ Can help with data repurposing</li> </ul>	<ul style="list-style-type: none"> <li>◆ Backup source</li> <li>◆ Near zero backup window                             <ul style="list-style-type: none"> <li>◦ 24x7 operations</li> </ul> </li> <li>◆ Fast restore                             <ul style="list-style-type: none"> <li>◦ copy based by definition</li> </ul> </li> <li>◆ Can help with data repurposing                             <ul style="list-style-type: none"> <li>◦ Beware performance impact</li> </ul> </li> </ul>

# Snapshot Considerations

- Snaps of production storage may impact production
  - ◆ Depending on use case
- Snap recovery tools may not be as mature
- Retention policy impact
  - ◆ Number of copies retained
  - ◆ Recovery granularity
  - ◆ Meeting off-site protection via distance replication
- On-array versus off-array alternatives
- Cost trade-offs and information classification
- ILM – Are you snapping old, un-used data

# Data Deduplication

## What?

- 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

## Why?

- Many data sets contain a high degree of redundant data
  - Example Repetitive Full Backups
- Reduction in cost per terabyte stored
- Enables storage of greater amounts of data
- Significant reduction in storage footprint
- Reduce power and cooling costs

## Considerations

- More data stored on fewer disks
- Enables low cost replication
  - ◆ Offsite copies
  - ◆ WAN Optimization
- Reduce backup, archive and primary storage
- Benefit from D2D backup
- Possible increase in workload
  - ◆ IO and Performance



**Check out SNIA Tutorial:  
Deduplication – Methods of  
Achieving Data Efficiency**

# Deduplication: What to Consider

- Factors that will impact your results:
  - ◆ Different applications or data types
  - ◆ Bandwidth and latency
  - ◆ Policies and Methodologies
  - ◆ Data Protection Overhead
  - ◆ Compression and Encryption
- Global Deduplication and Scope
- Deduplicated Data Resiliency
- Scalability
  - ◆ Capacity
  - ◆ Performance

# Backup: Factors Impacting Space Savings

<b>Factors associated with higher data deduplication ratios</b>	<b>Factors associated with lower data deduplication ratios</b>
<b>Data created by users</b>	<b>Data captured from mother nature</b>
<b>Low change rates</b>	<b>High change rates</b>
<b>Reference data and inactive data</b>	<b>Active data, encrypted data, compressed data</b>
<b>Applications with lower data transfer rates</b>	<b>Applications with higher data transfer rates</b>
<b>Use of full backups</b>	<b>Use of incremental backups</b>
<b>Longer retention of deduplicated data</b>	<b>Shorter retention of deduplicated data</b>
<b>Wider scope of data deduplication</b>	<b>Narrower scope of data deduplication</b>
<b>Continuous business process improvement</b>	<b>Business as usual operational procedures</b>
<b>Smaller segment size</b>	<b>Larger segment size</b>
<b>Variable-length segment size</b>	<b>Fixed-length segment size</b>
<b>Format awareness</b>	<b>No format awareness</b>
<b>Temporal data deduplication</b>	<b>Spatial data deduplication</b>

# Continuous Data Protection

## What:

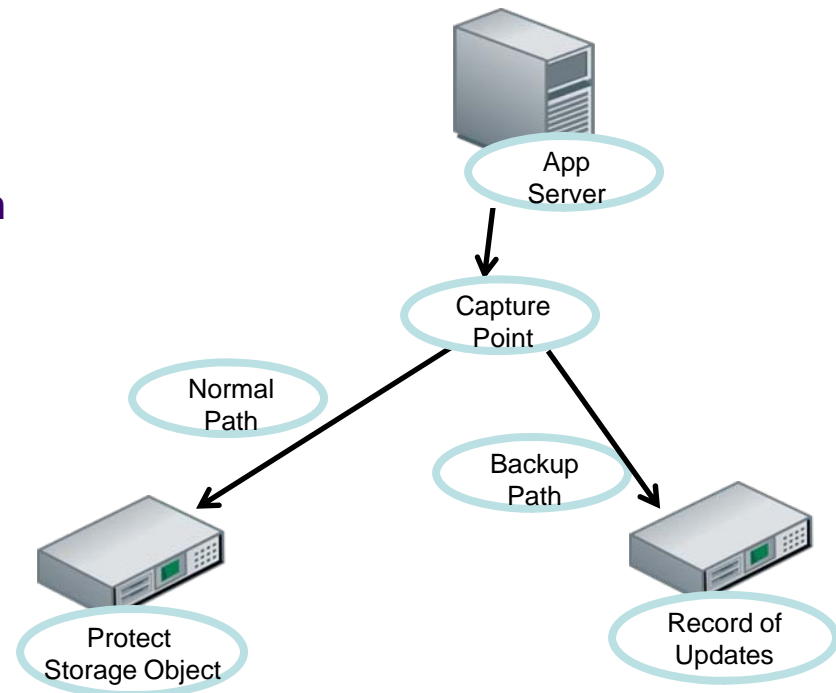
- Capture every change as it occurs
- Protected copy in a secondary location
- Recover to any point in time

## How:

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

## Why:

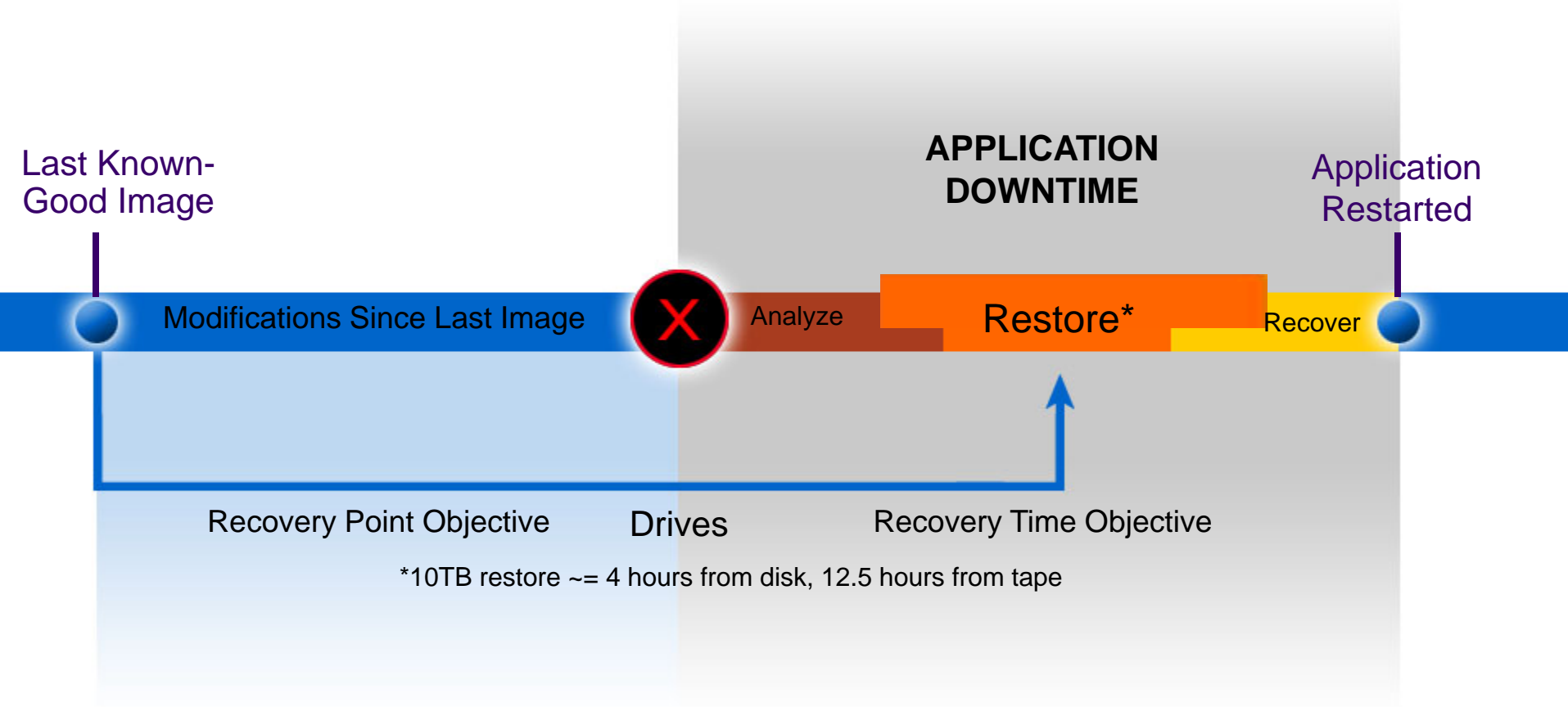
- Zero data loss, Zero backup window
- Simple recovery.
- CDP protects data at all times
- Granular recovery - directly to any point in time.



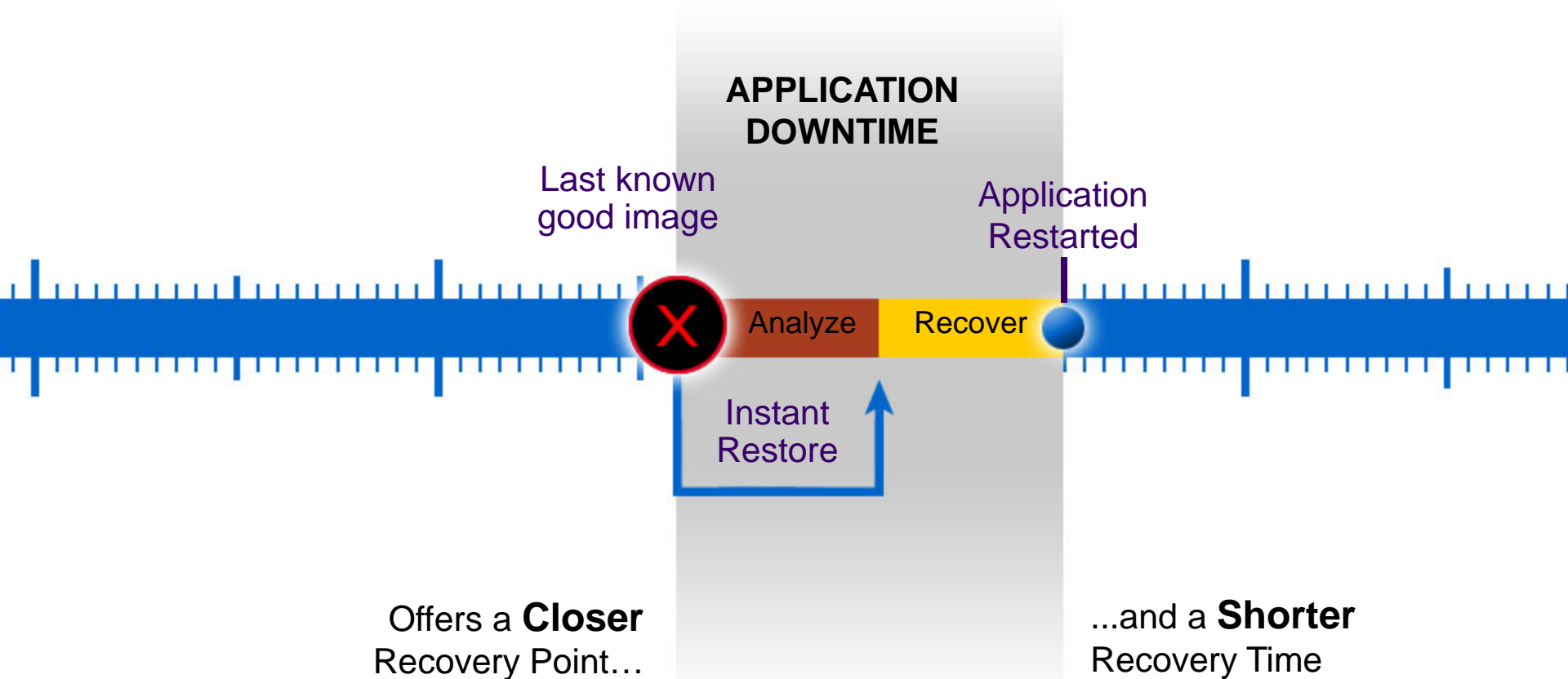
# The CDP (Continuous) Difference

- Replication is not CDP:
  - ◆ Maintains only a current copy of the data (Synchronous)
  - ◆ May be combined with some snapshot capabilities
  
- Snapshots are not CDP:
  - ◆ Snapshots are scheduled events (Asynchronous)
    - › Data loss possible if crash or corruption happens between snaps
    - › Snapshots frequently to same system as primary
    - › Lack continuous index with embedded knowledge of relationship of data to files, folders, application and server
  
- Scheduled events are not CDP:
  - ◆ Scheduled backup processes
  - ◆ Log collection for database style applications, rolling transactions forwards or backwards

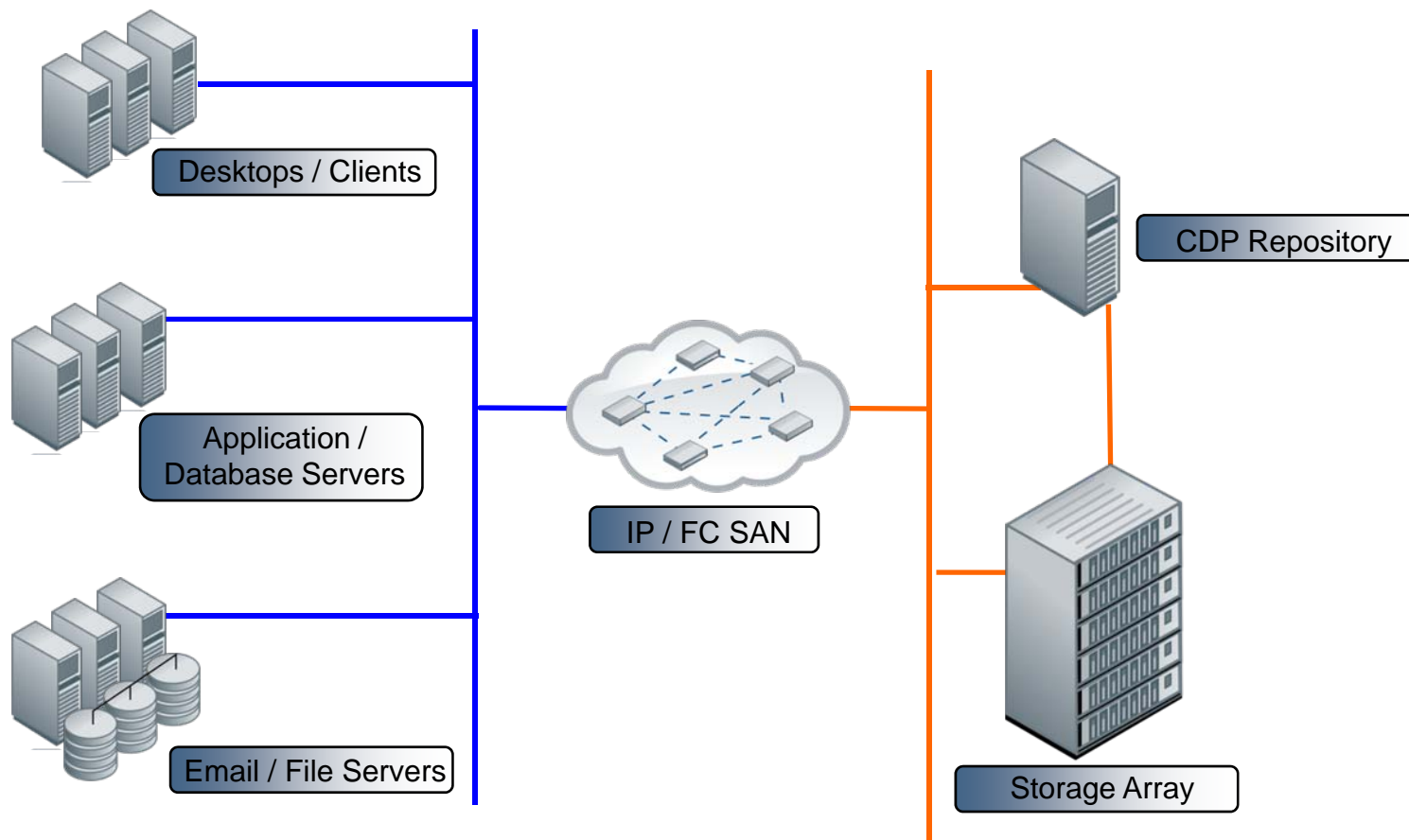
# Traditional Recovery



# Recovery with CDP



# CDP Deployment



# CDP Considerations

## ➤ Performance

- ◆ Can CDP offering keep up with highest change rates

## ➤ Scale

- ◆ Large changes in data can result in very large change logs

## ➤ Ease of use

- ◆ Changes in paradigm imply changes in process
- ◆ Educating an established team

## ➤ Product maturity

- ◆ Technology is rather new

## ➤ Impacts on production

- ◆ Data protection should never impact production

# Data protection summary

- Data growth requires us to plan for tomorrow
  - ◆ Investigate data and information management technology
- Information value determines data protection levels
  - ◆ Stop protecting employee home movies, last years news
  - ◆ Not all data assets are created equal
- Architecture
  - ◆ Understand your networks, hosts, applications
  - ◆ PLAN ahead – Avoid reactionary thinking
- Do your homework
  - ◆ SNIA and DMF offer seminars, classes, workshops.....

# SNIA Resources

## ➤ Related tutorials

- ◆ Disk and Tape Backup Mechanisms
- ◆ Disk Based Restoration Technology
- ◆ Deduplication
- ◆ Information Classification
- ◆ Information Lifecycle Management
- ◆ Archival

## ➤ Visit the Data Management Forum website at

<http://www.snia-dmf.org>

- ◆ Data Protection Buyers Guides available
  - Chapters on Continuous Data Protection, Deduplication, and Virtual Tape Libraries

Please send any comments on this tutorial to SNIA at:  
[trackdatamgmt@snia.org](mailto: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”  
SNIA Tech Council  
Nancy Clay  
Rob Peglar

Mike Fishman  
Jason lehl  
Mike Rowan  
SW Worth



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](http://www.snia.org/dmf)

# Thank you for your feedback

## Questions and Answers

# Copy On Write Snapshots

- Primary disks remain current
- Whenever a Write operation arrives, it is held:
  - ◆ First the current contents of the write-destination are read in
  - ◆ The old-contents from the primary disk is saved off somewhere and indexed
  - ◆ The new write is now allowed to pass through
- Read path of current disks remains optimized
- Write path of current disks is potentially impacted
- Read/Write path of “snapshot” disks impacted

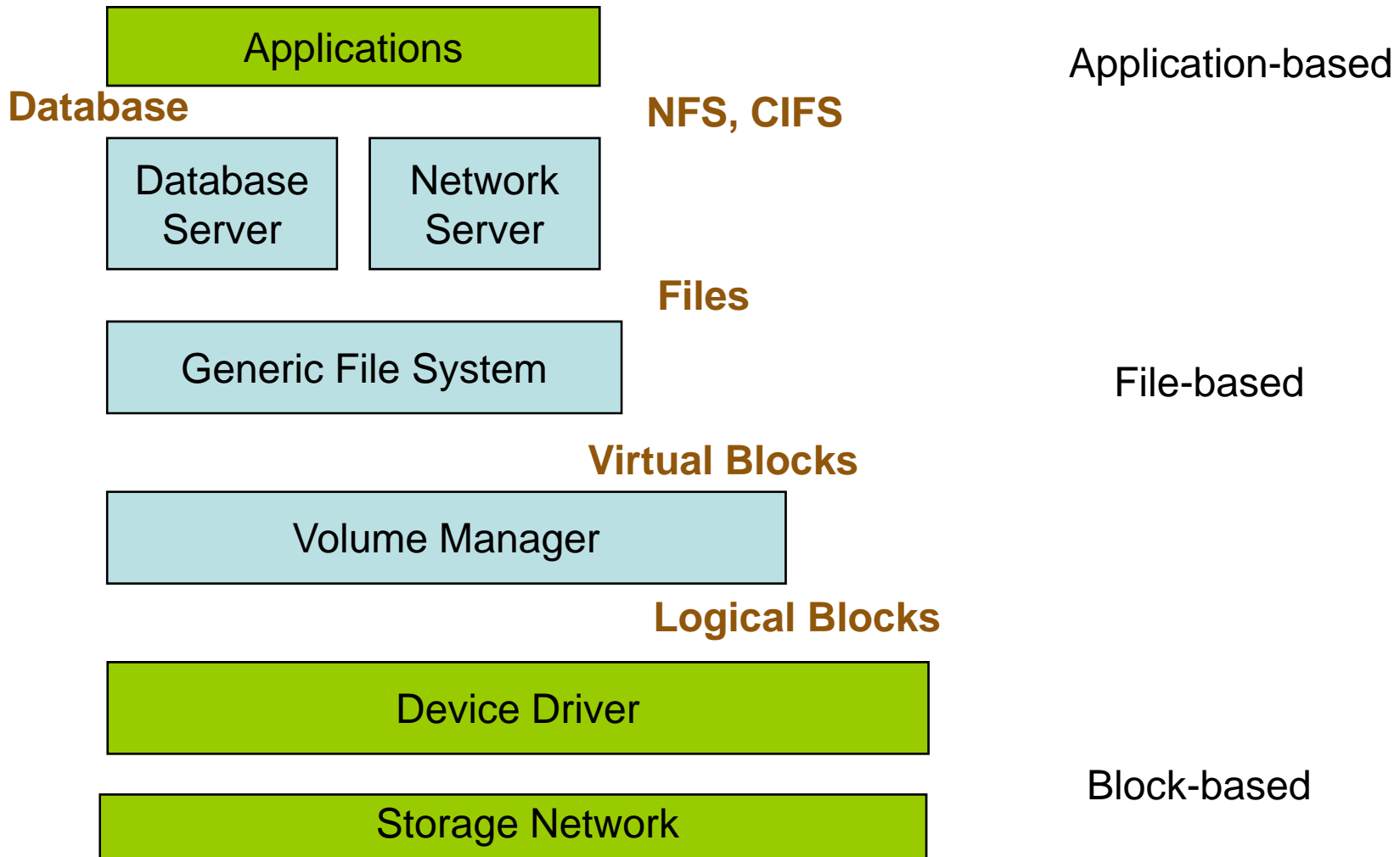
# Redirect on Write Snapshots

- Primary disks are frozen
- New write operations to primary disks are stored in a journal (and indexed)
  - ◆ To read current copy, the journal is checked first
  - ◆ To read the snapshot copy, the primary disks are used
  - ◆ When snapshot is “dissolved”, write journal must be applied to primary disks to “catch up”
- Read path of snapshot is optimized
- Write path of current disks is optimized (no copy)
- Read path of current disks is potentially impacted

# Write Anywhere

- All disk blocks are virtualized
  - ◆ Current disk is represented by a map to real blocks -- not directly mapped
  - ◆ Disk storage is larger than maps present
  - ◆ New writes, instead of “overwriting” blocks, are directed to free blocks
  - ◆ Maps are kept for “now” and potentially for multiple “snapshots”
  - ◆ Reference counts are kept for blocks “in-use”
- Performance doesn't generally change primary/snapshot
- Performance can be impacted by fragmentation depending implementation

# CDP Implementation Models



# Remote Replication

## ➤ What and Why?

- ◆ Extension of full snapshots, delta snapshots, DP images or CDP (depending on implementation/vendor)
- ◆ File or block replication across a communications pipe synchronously or asynchronously (remote)
- ◆ Combines DR with corruption protection
- ◆ Extend tier I data protection to second and thirds data tier's
- ◆ Can be used to consolidate tape creation to central location

## ➤ What to watch out for

- ◆ Speed of light/ bandwidth issues
- ◆ Site-to-site recovery can be awkward -- local corruption issues better dealt with locally

# Solving the problem

