SEPATON

Benefits of Virtual Tape Libraries and Data De-duplication Technologies with TSM

Stephen Firmes
ITSM Solutions Architect

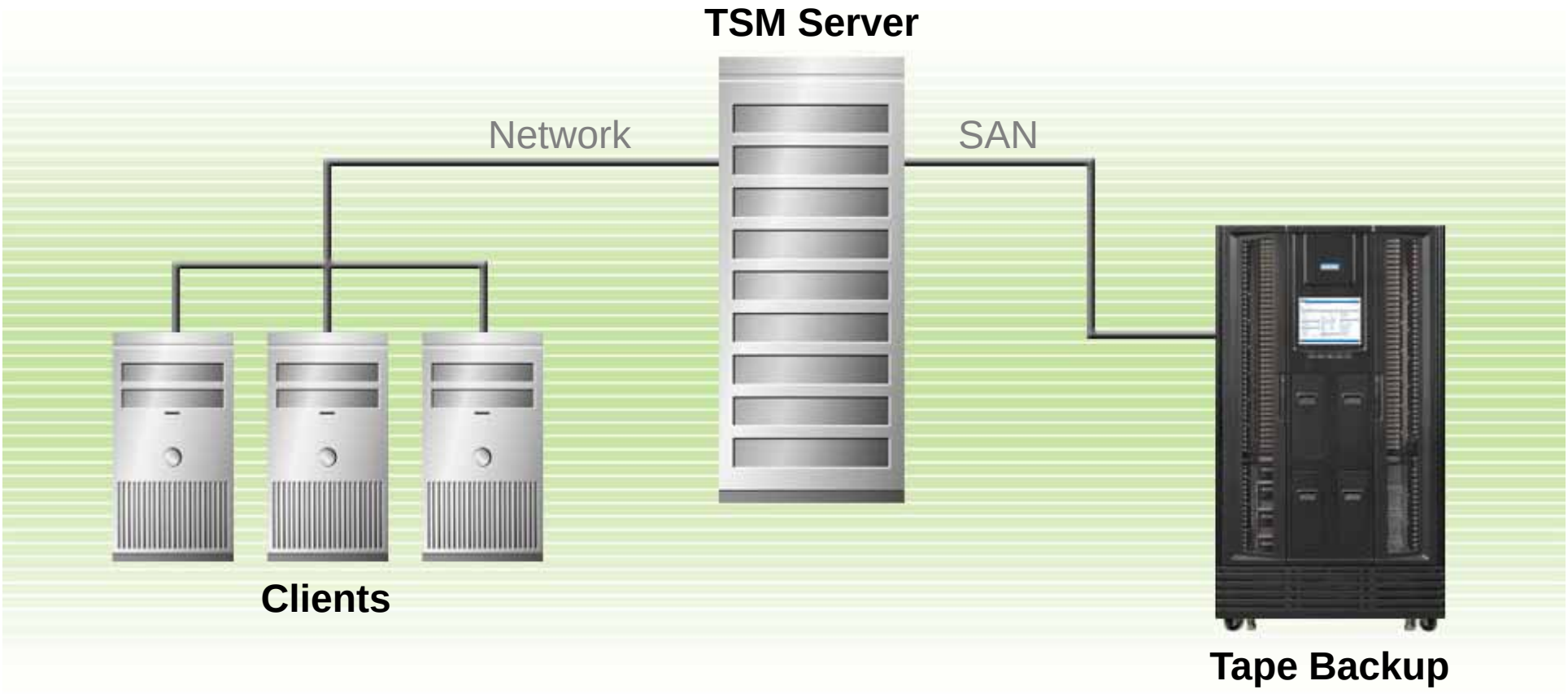
Topics to be Discussed

- **Data Protection Overview**
- **The Virtual Tape Solution**
- **VTL in a TSM Environment**
- **De-duplication 101 & Beyond**
- **Q & A**

SEPATON

Data Protection Overview

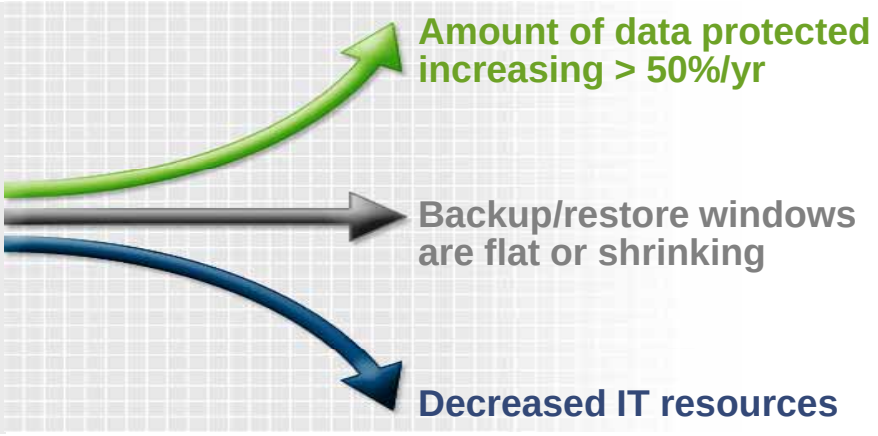
Data Protection Today



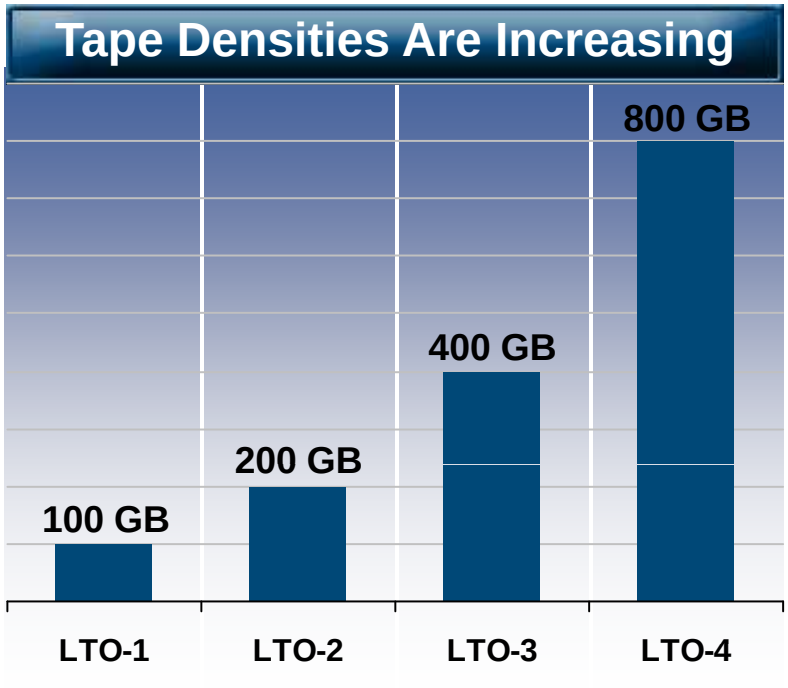
Today's Unique Data Protection Challenges

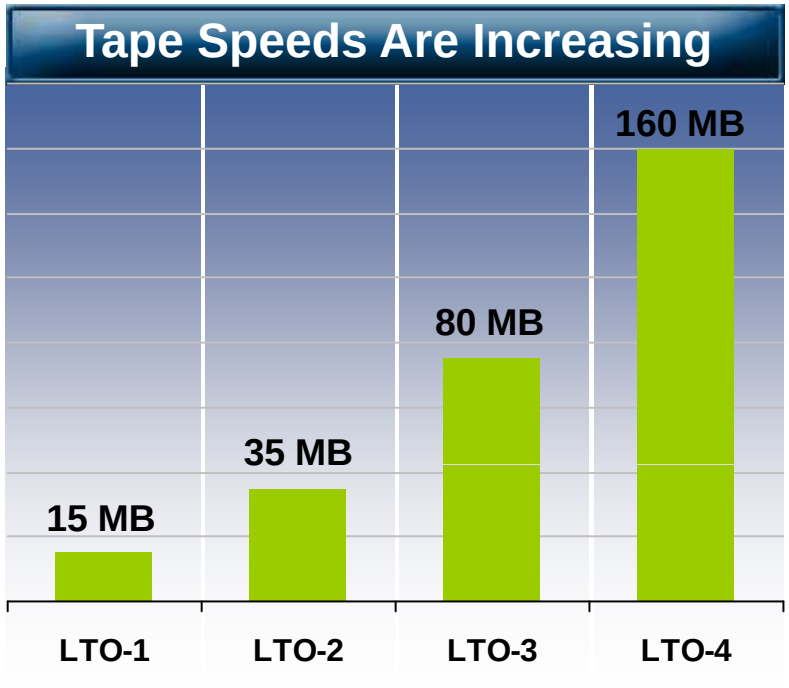
- **IT Cost Containment**
  - Tape environments not designed with today's amount of data in mind
  - Space, power, cooling, operational costs are skyrocketing
- **Compliance**
  - New business SLAs, regulatory requirements drive IT
  - Requires online data retention for fast access, search and recovery
- **Complexity**
  - Tape is increasingly sensitive to highly-tuned backup environments
  - Error prone, multiple points of risk, and delicate
  - Tapes drain over-committed IT resources

Disk based data protection is faster, more reliable, less expensive, and easier to use than tape



Tape Technology Is Evolving

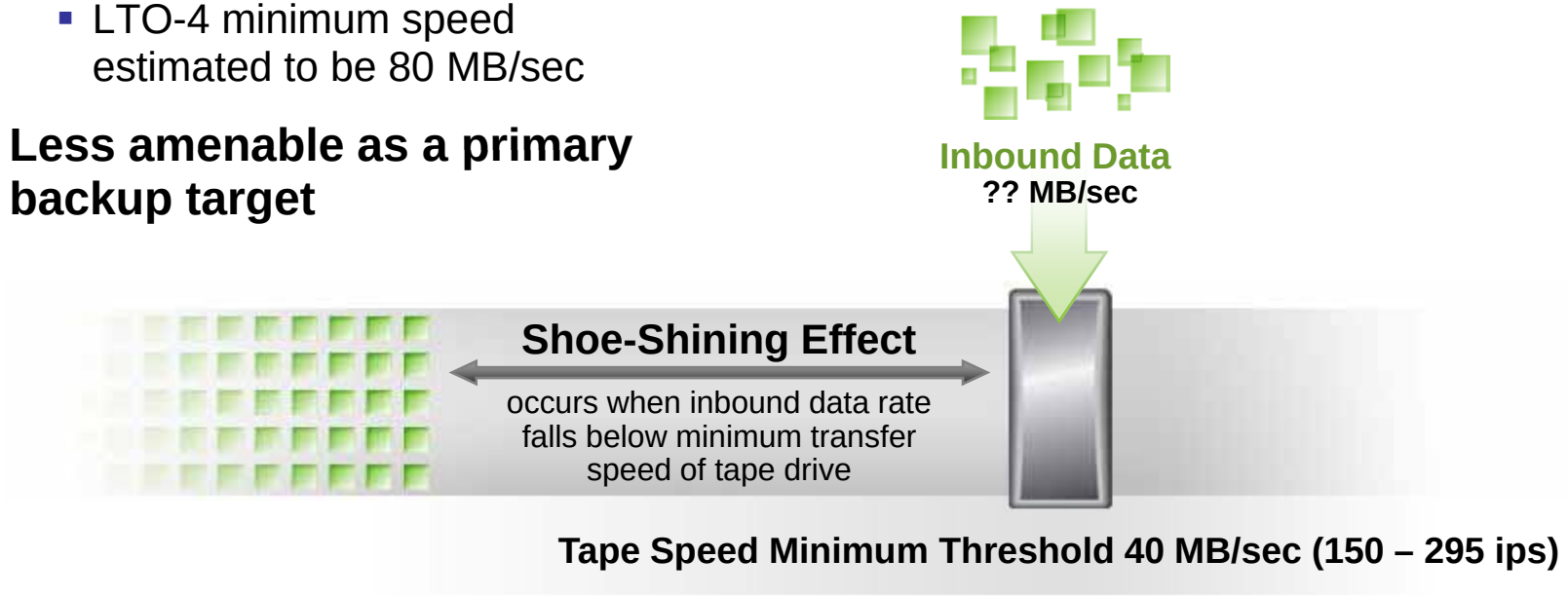




So What's Wrong With Tape?

Evolving Technology Makes Using Tape More Difficult

- **Increasingly sensitive to highly-tuned backup environments**
- **Minimum threshold increases with tape speed to prevent shoe-shining**
  - LTO-3 minimum speed is 40 MB/sec
  - LTO-4 minimum speed estimated to be 80 MB/sec
- **Less amenable as a primary backup target**



Performance vs. Tape

Performance Specification	Physical Tape	S2100-ES2
Tape load and thread to ready:	**16 seconds**	<1 seconds
Average file access time (first file):	**62 seconds**	<1 seconds
Average rewind time:	**91 seconds**	<1 seconds
Unload time:	**16 seconds**	<1 seconds
Theoretical data transfer rate: (uncompressed)	**80 MB/sec (Single stream)**	150 MB/sec **(Single and multi-stream)**
Real world transfer rate: (uncompressed and includes factors list above)	**28 MB/sec (Single stream)**	150 MB/sec **(Single and multi-stream)**

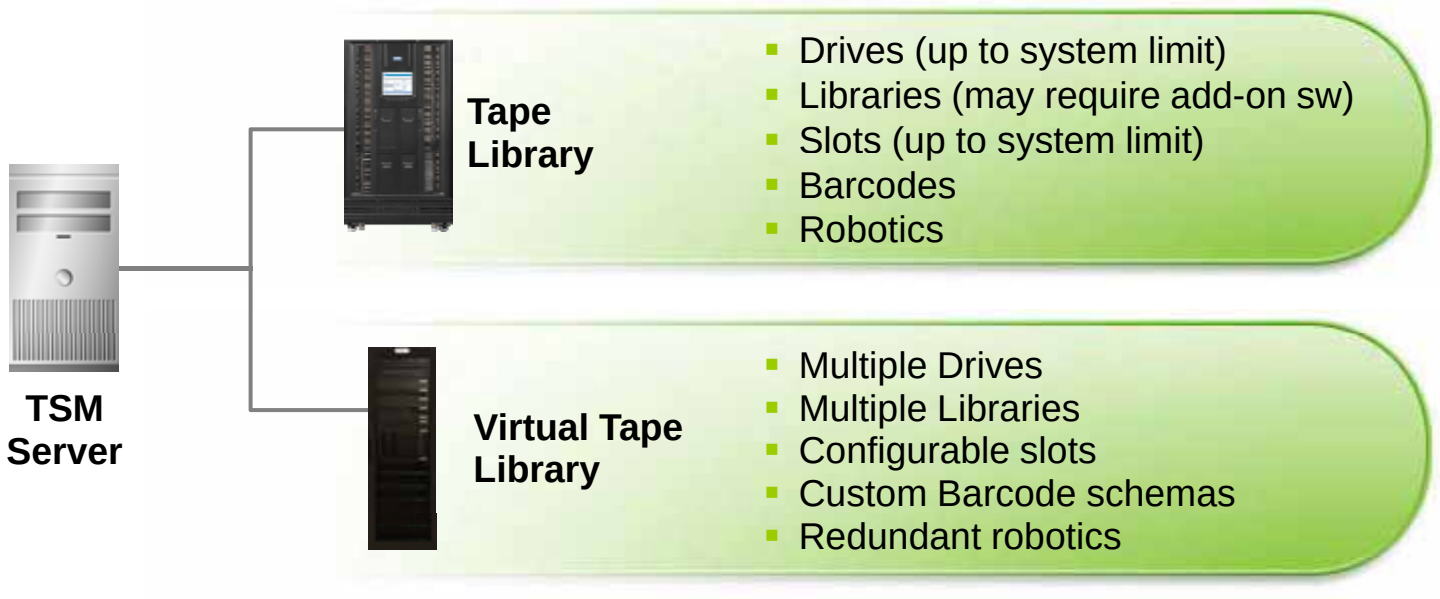
Cost vs. Tape

	Physical Tape	Virtual Tape
Throughput	**Tape Drive & Interface 28 MB/Sec**	**Node & Port 150 MB/Sec 64 Streams – 64 Virtual drives**
Costs	**$12K per drive**	$14K per port
Performance	**100 GB/Hr.**	540 GB/Hr.
Time to recover a 200GB file	**2 Hrs.**	24 Minutes
Units required to backup 3.5 TB in 12 hours	**5 @ $12K Per**	1 @ $14K Per
Units required to backup 15 TB in 12 hours	**27 @ $12K = $324K**	5 @ $14K =$72K

SEPATON

The VTL Solution

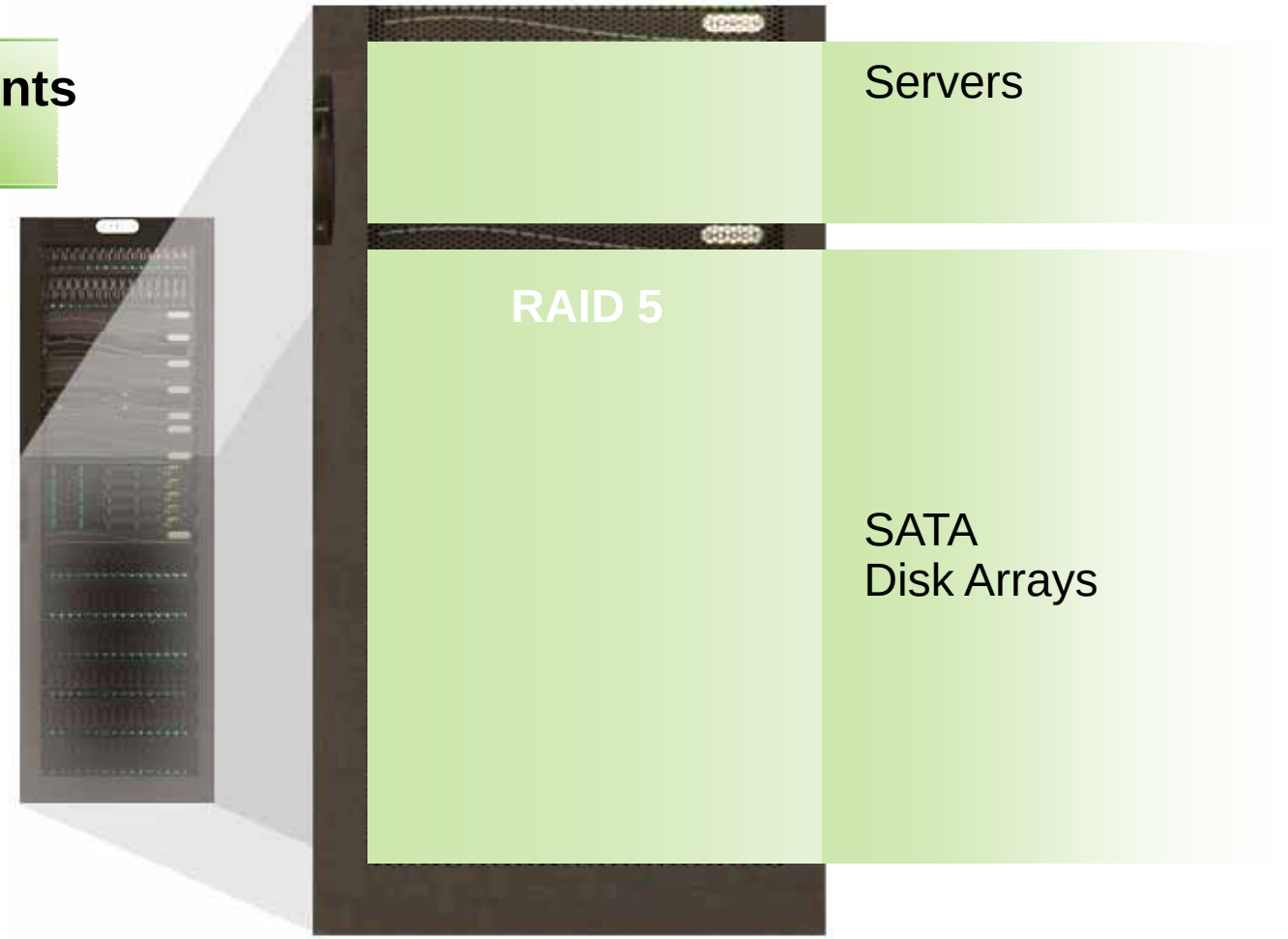
Virtual Tape Overview



- **A VTL appears exactly like a physical tape library except that it offers much greater configuration flexibility**
  - Configurable drive/library/slot/barcode parameters
- **Physical tape can still used for long-term offsite archival**

VTL Conceptual Architecture

Components



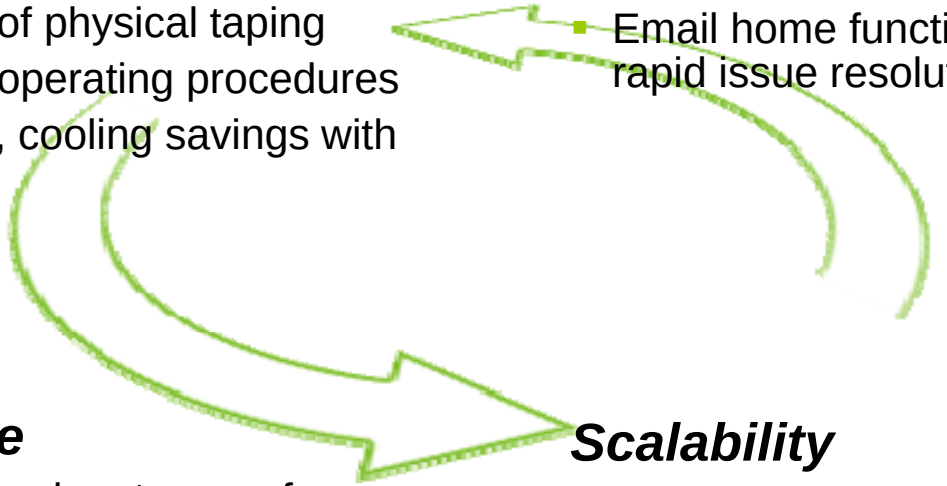
Key Benefits of a VTL

Cost Savings

- Less manpower required to manage backups
- Reduced use of physical taping
- No change in operating procedures
- Space, power, cooling savings with de-duplication

Reliability

- Redundant architecture enables consistently successful backups
- Email home functionality ensures rapid issue resolution



Performance

- Fast backup and restore performance
- Faster than tape

Scalability

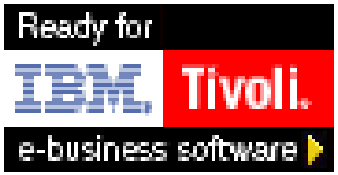
- Negligible increase in management cost as system grows
- Expand storage non-disruptively

SEPATON

VTL in a TSM Environment

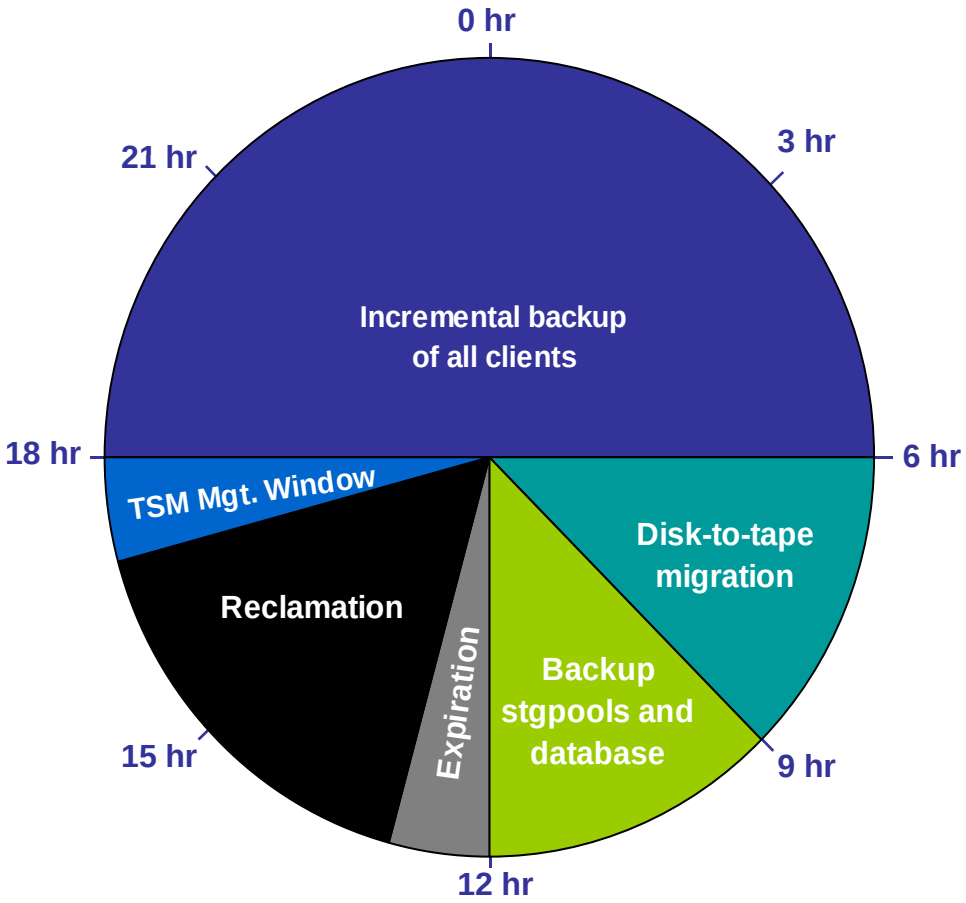
Accepted Technology in TSM Environments

- Supported compatibility with TSM
  - IBM Total Storage Proven
  - IBM Certified Tivoli Ready
  - TSM Certified
  - IBM Business Partner



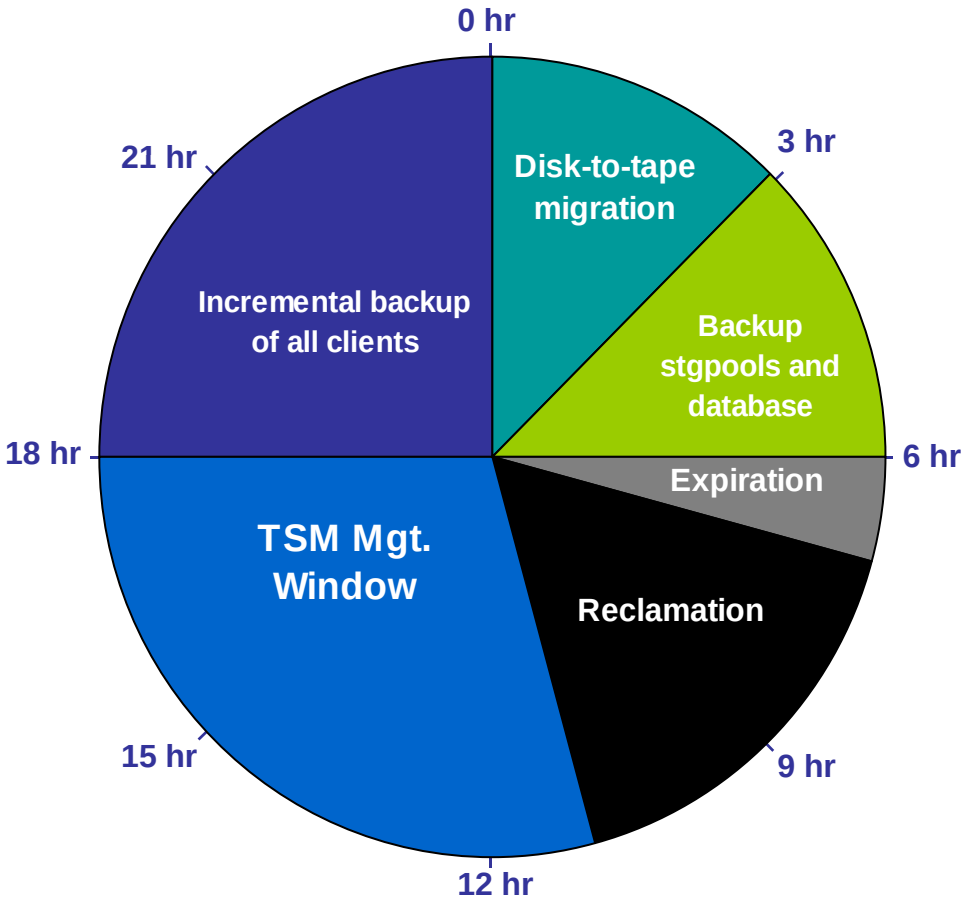
Backup and Restore

- **Backup window can be improved**
  - LAN-Free
  - Skip Disk Pool
- **Restore times will be reduced**
  - Now you are restoring at disk speeds from “tape”
  - Multi-threaded restores
  - Elimination of load and seek times



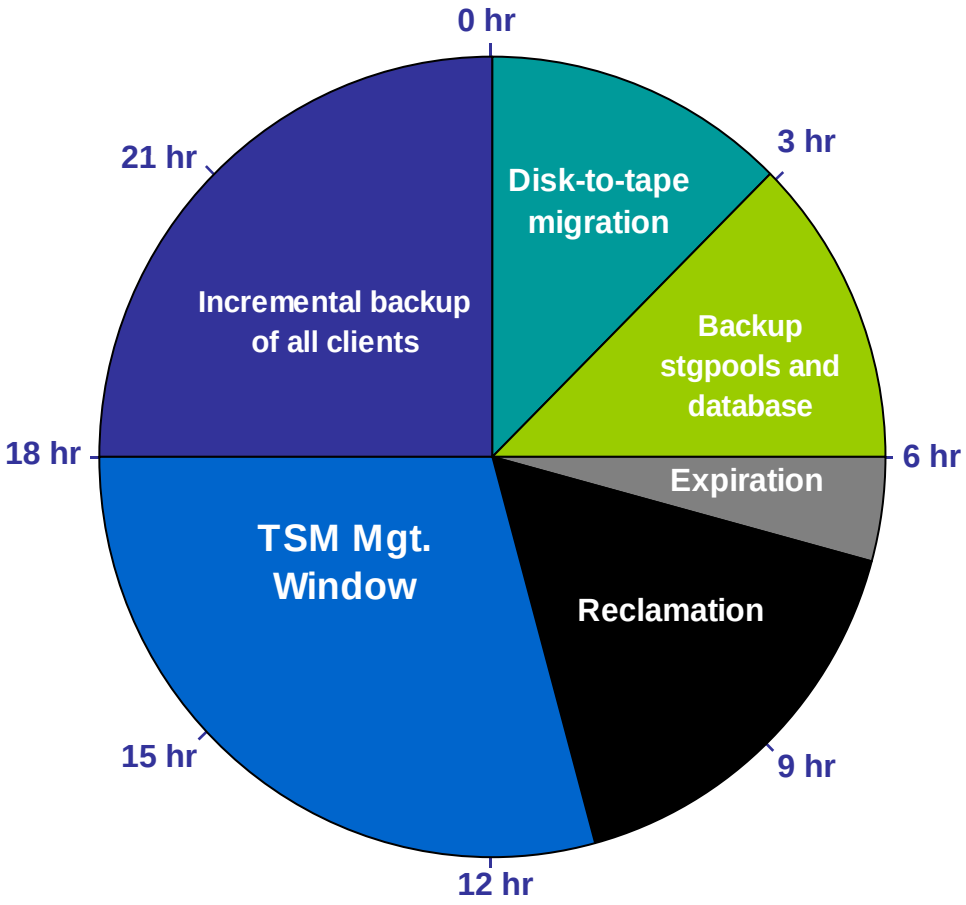
Backup and Restore

- **Backup window can be improved**
  - LAN-Free
  - Skip Disk Pool
- **Restore times will be reduced**
  - Now you are restoring at disk speeds from “tape”
  - Multi-threaded restores
  - Elimination of load and seek times



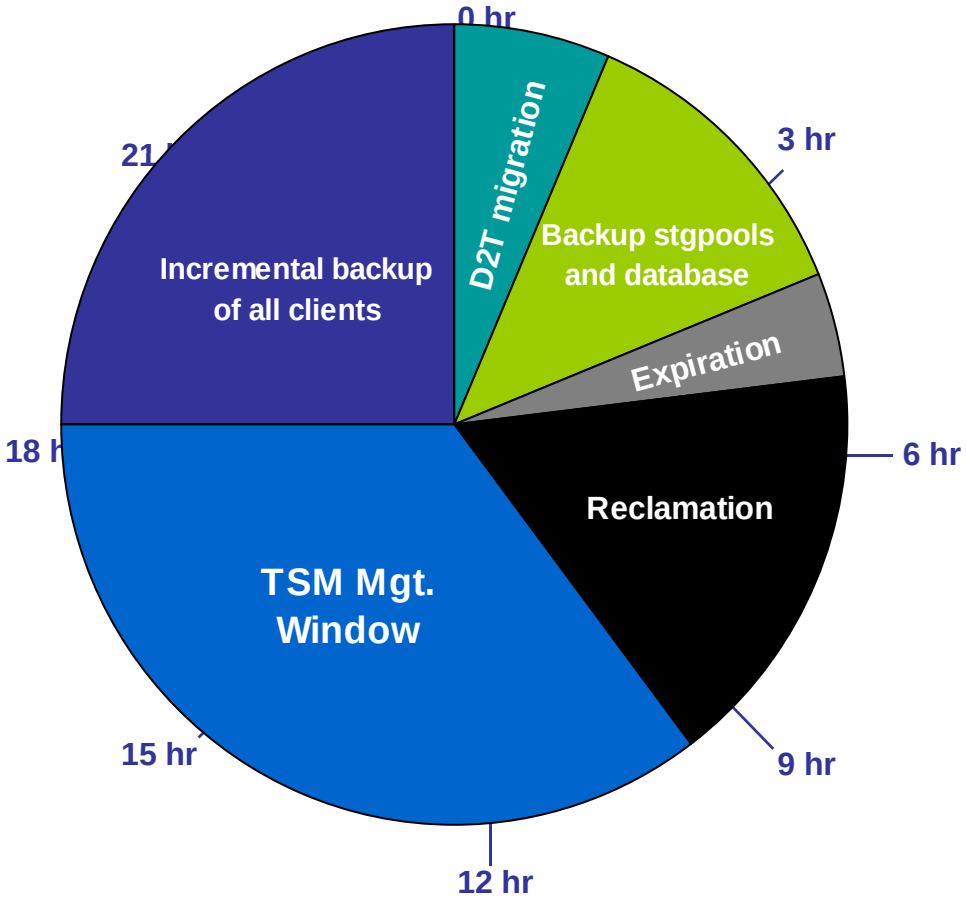
Storage Pool Migration

- **Decrease migration time from disk to tape**
- **Disk to Disk operation**



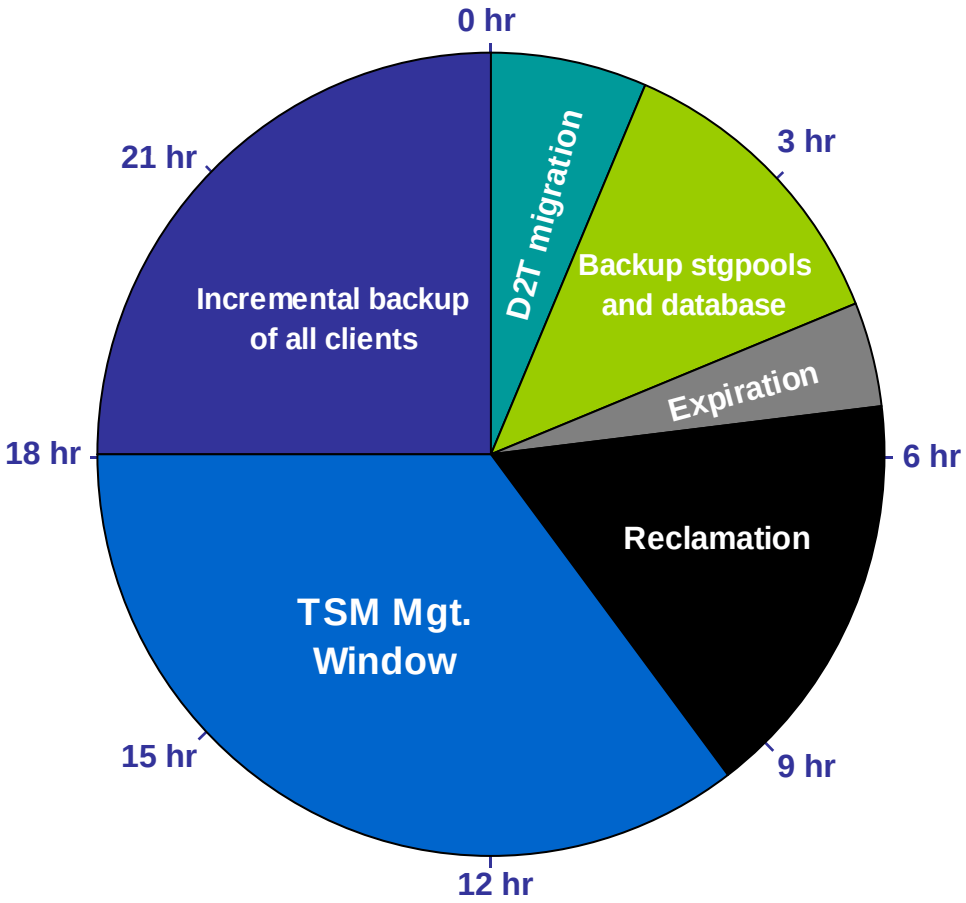
Storage Pool Migration

- **Decrease migration time from disk to tape**
- **Disk to Disk operation**



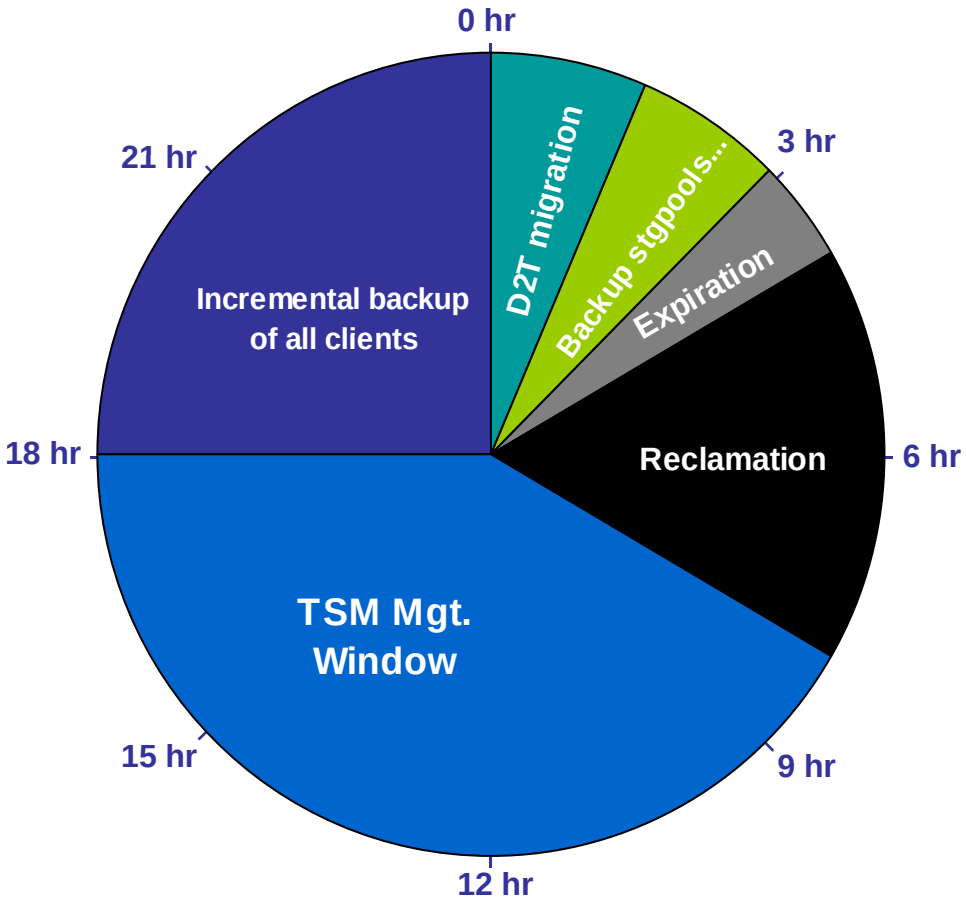
Backup Storage Pool and DBB

- **Creation of Copy Pool tapes faster**
  - Disk to tape
- **Onsite backup of the ITSM database extremely fast**



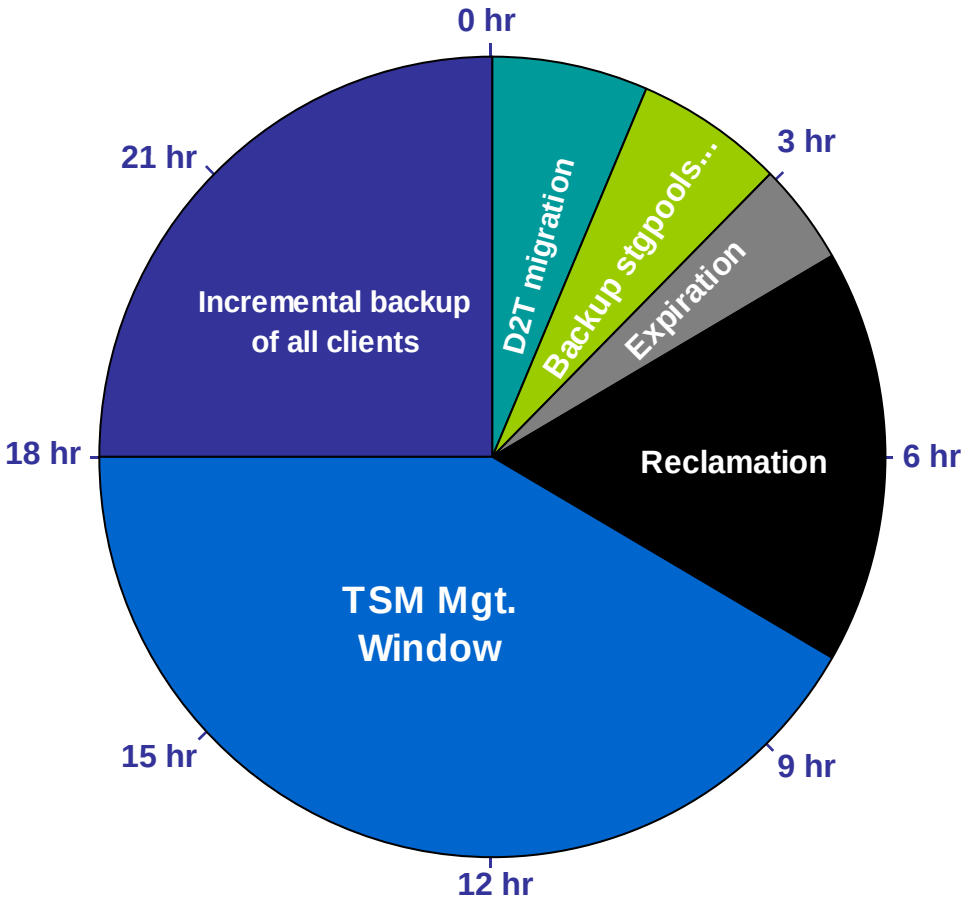
Backup Storage Pool and DBB

- **Creation of Copy Pool tapes faster**
  - Disk to tape
- **Onsite backup of the ITSM database extremely fast**



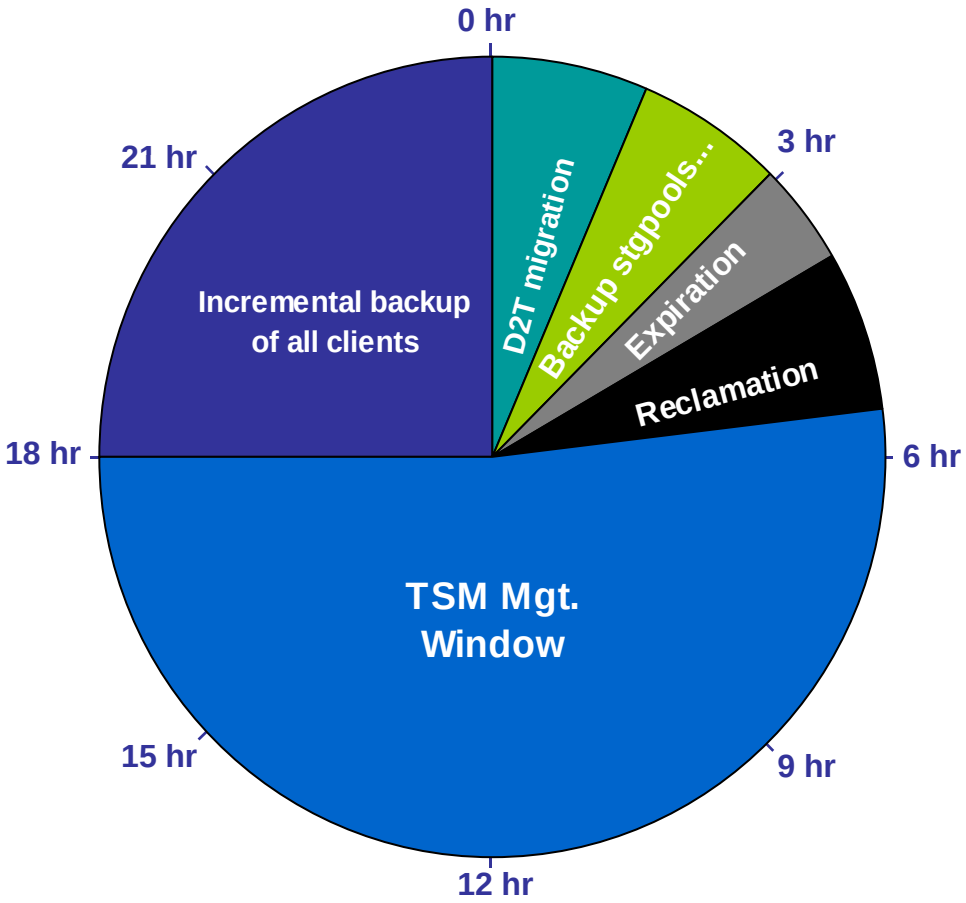
Reclamation

- **Mount/Dismount times eliminated**
- **Primary Pool reclamation extremely fast**
- **Reduce Copy Pool tape creation time**

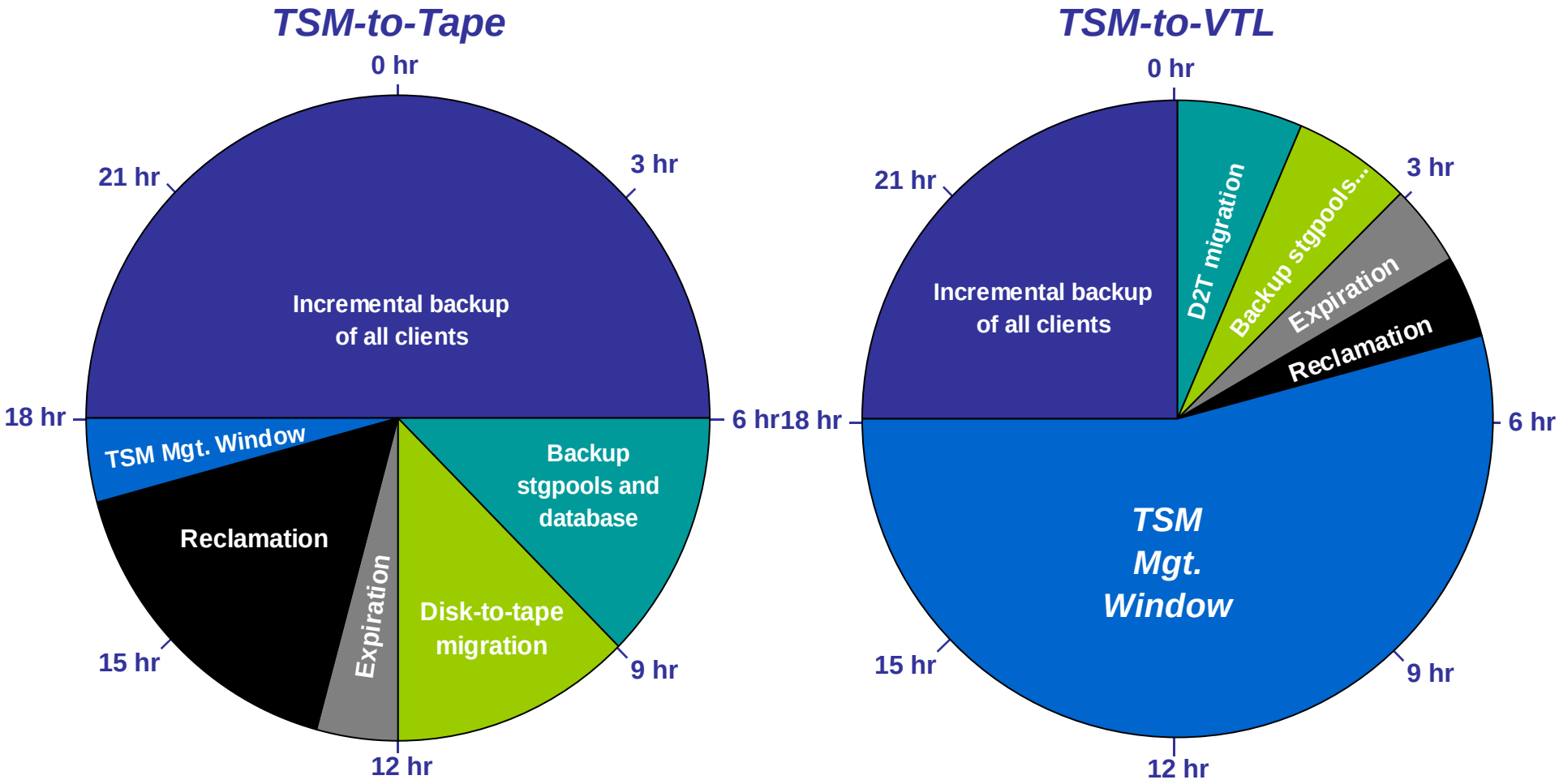


Reclamation

- **Mount/Dismount times eliminated**
- **Primary Pool reclamation extremely fast**
- **Reduce Copy Pool tape creation time**



Comparing Data Protection Models



Hardware Savings

- **Eliminate onsite physical library**
  - Multiple virtual libraries can exist within one VTL system
    - Logical partitioning
    - Departmental libraries
    - Chargeback
- **Minimize the size of onsite physical tape library**
  - High priority clients
  - LAN-Free clients
- **Reduce the size of disk storage pool(s)**
  - No need for caching
  - Introduce nightly migration?

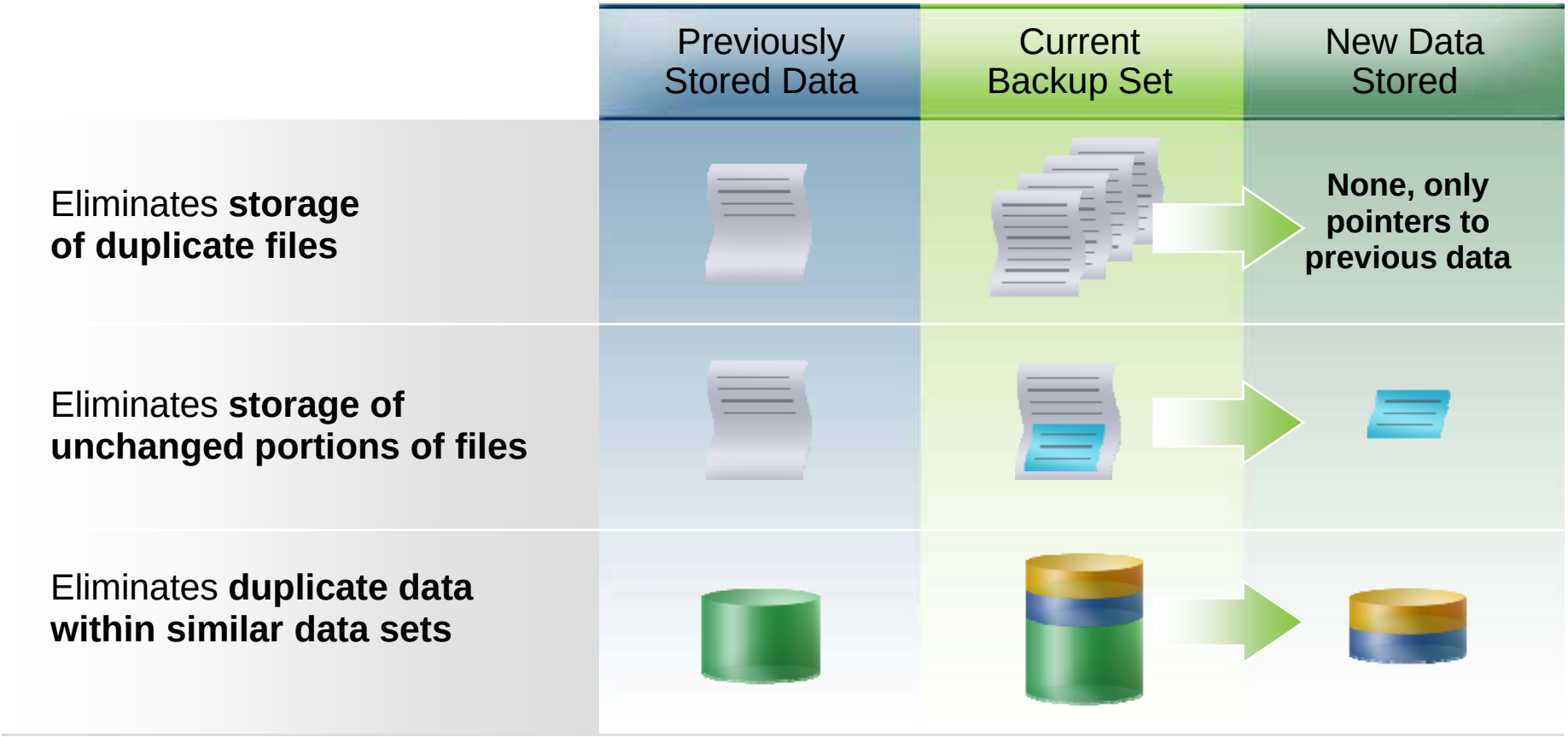
SEPATON

De-duplication 101 & Beyond

Delta Technology

- **Different technique versus commonplace Lempel-Ziv algorithm**
- **Requires two or more copies of the same or similar data object** (i.e. 2 versions of this PPT)
- **Achieves a “de-duplication effect” by eliminating portions of redundant data with pointers to existing data somewhere else**

Delta Technology Examples



In-band vs. Post-process

- **Determines when data de-duplication occurs**
- **In-band de-dupe happens during data backup**
  - Can affect how quickly data backups occur
- **Post-process de-dupe happens once data has been stored to disk**
  - Can affect how quickly de-duplication occurs

Method 1: Hash-based Redundancy Elimination

- **Data is broken into blocks**
- **A “hash” is assigned to each block**
  - The hash is typically a security checksum, such as MD-5 or SHA-1
- **The computed hash is compared to an index of existing hashes**
  - If the hash already exists, the data is assumed redundant and only a pointer is written
  - If the hash doesn’t exist, the data is assumed unique and written to disk

Hash-based De-Duplication Pros & Cons

- **Pros**
  - By only comparing the hash information, disk I/O can be minimized
  - Simple to adapt this algorithm to solve “over-the-wire” de-duplication of data
- **Cons**
  - Hash “collisions” can cause data corruption
  - Scalability of the de-dupe solution is typically constrained by the hash table memory footprint
  - Speed of computing hashes limits performance

Method 2: Pattern-matching Redundancy Elimination

- **Data is broken into blocks**
- **The block data is pattern-matched against other similar blocks**
- **Blocks that "probably" match are explicitly compared**
  - A data block is read from disk and compares with the "probable match" in memory
  - If the block matches, only a pointer is written
  - If the block doesn't match, it is written to disk

Pattern Matching Pros & Cons

- **Pros**
  - Guarantees data integrity
  - Memory footprint requirements can be minimized
- **Cons**
  - Can increase the amount of disk I/O activity during de-duplication
  - High latency of matching and confirmation steps can impede high throughput
  - Scalability still requires replicating memory tables

Method 3: Sliding Window Redundancy Elimination

- **Two similar data sets are associated**
  - i.e. backup sets, files, or databases
- **Marker data is collected from both data sets**
- **One object (“version object”) is scanned versus the other (“base object”)**
- **Matching byte streams are identified an eliminated as redundant data**
  - The “leftover” data is considered unique and stored to disk

Sliding Window Pros & Cons

- **Pros**
  - Byte stream-based de-duplication generally yields higher de-dupe ratios than block-based
  - Easier to parallelize processing without clustering
  - Less sensitive to structure of the data or changes
- **Cons**
  - Can increase the amount of disk I/O activity during de-duplication
  - Harder to adapt to replication solutions
  - Working with large objects requires special consideration

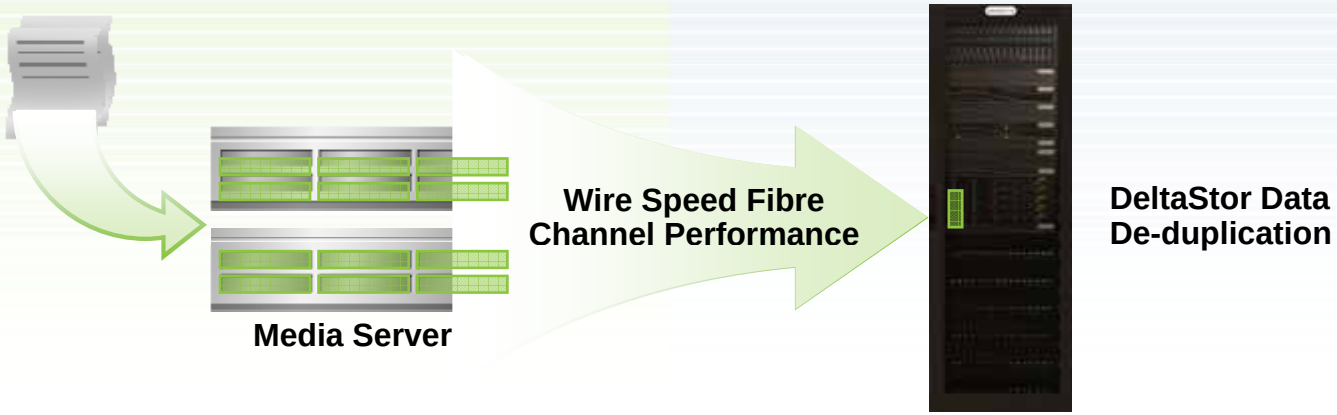
DeltaStor™ De-duplication

Unparalleled De-duplication

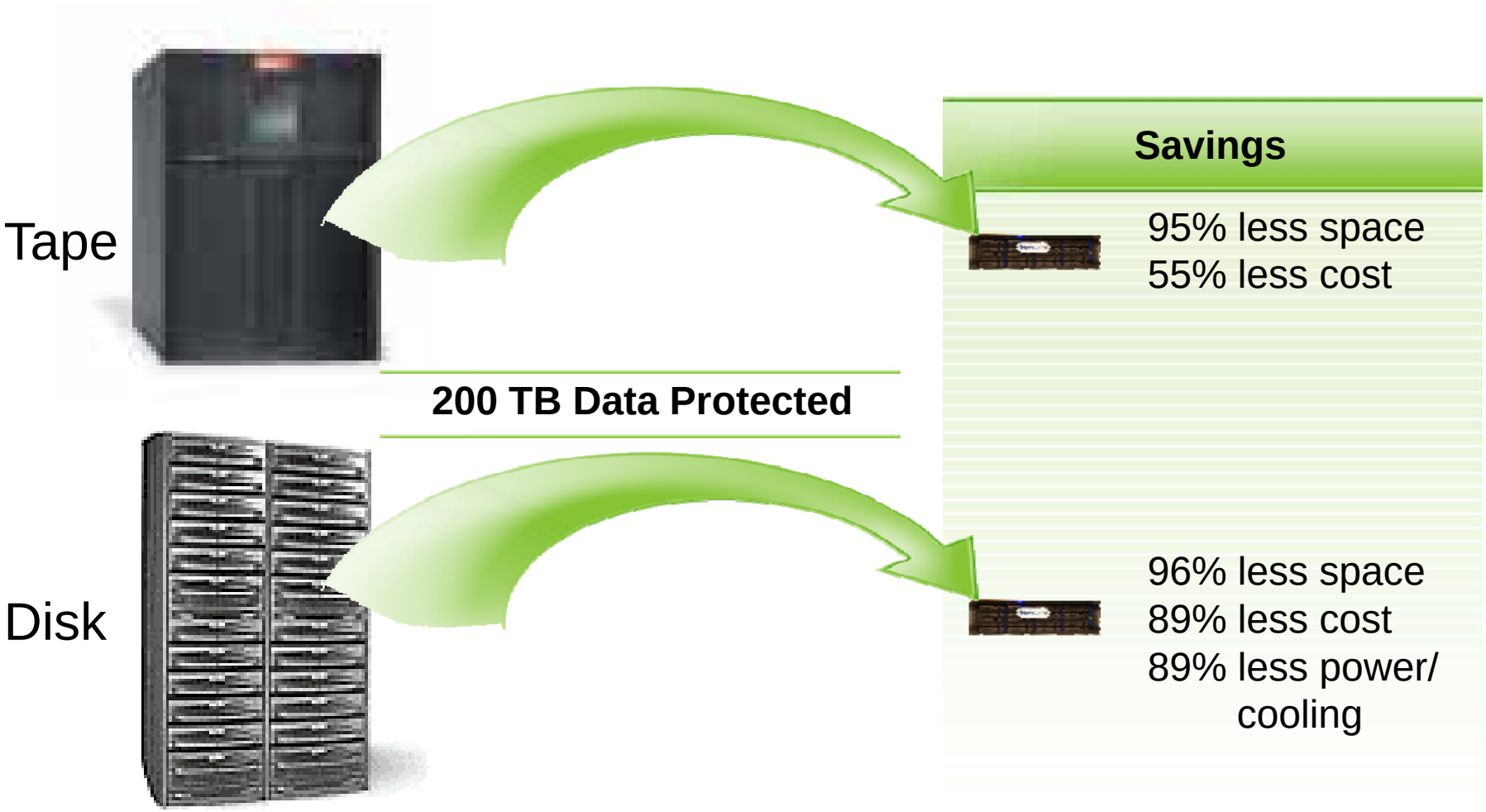
- De-duplication ratios reached immediately due to object-based comparisons
- Unparalleled de-duplication ratios
- Configurable de-duplication based on policies, clients, etc.

Forward Differencing for Fast Restores

- Eliminates need to rebuild files before restore
- Newest delivered data is used as reference
- Older duplicate data is replaced with pointers to the newer data



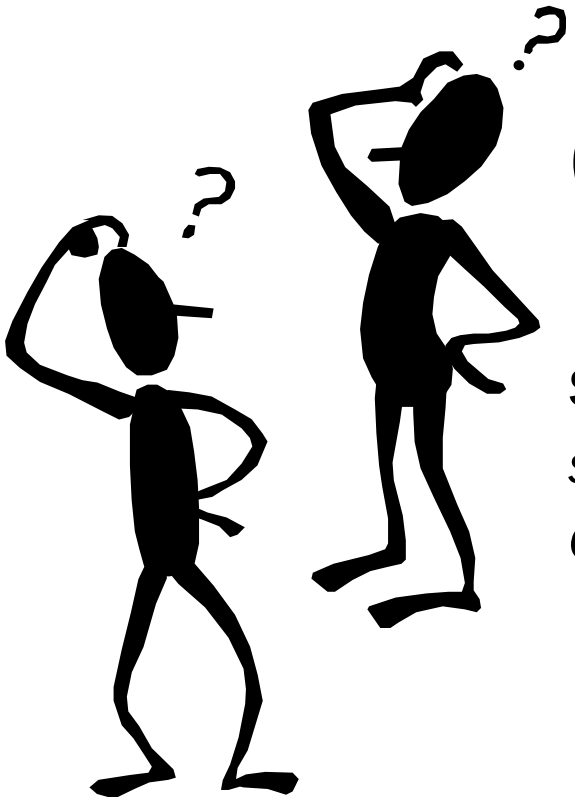
Economics of De-Duplication



Key Questions To Ask

- **How are de-duplication results measured?**
  - Full backups?
  - Incremental backups?
  - “Average user” results?
- **How long does it take to achieve these results?**
- **Is there a backup or restore performance impact (compared to running without de-dupe)?**
- **How is data integrity managed?**

SEPATON



Q & A

Stephen Firmes, ITSM Solutions Architect

sfirmes@sepaton.com

Office: 508-490-7921