“Storage Deduplication” using IBM System Storage and Tivoli Software

Rob Turk

15 September 2010

The Data Juggernaut

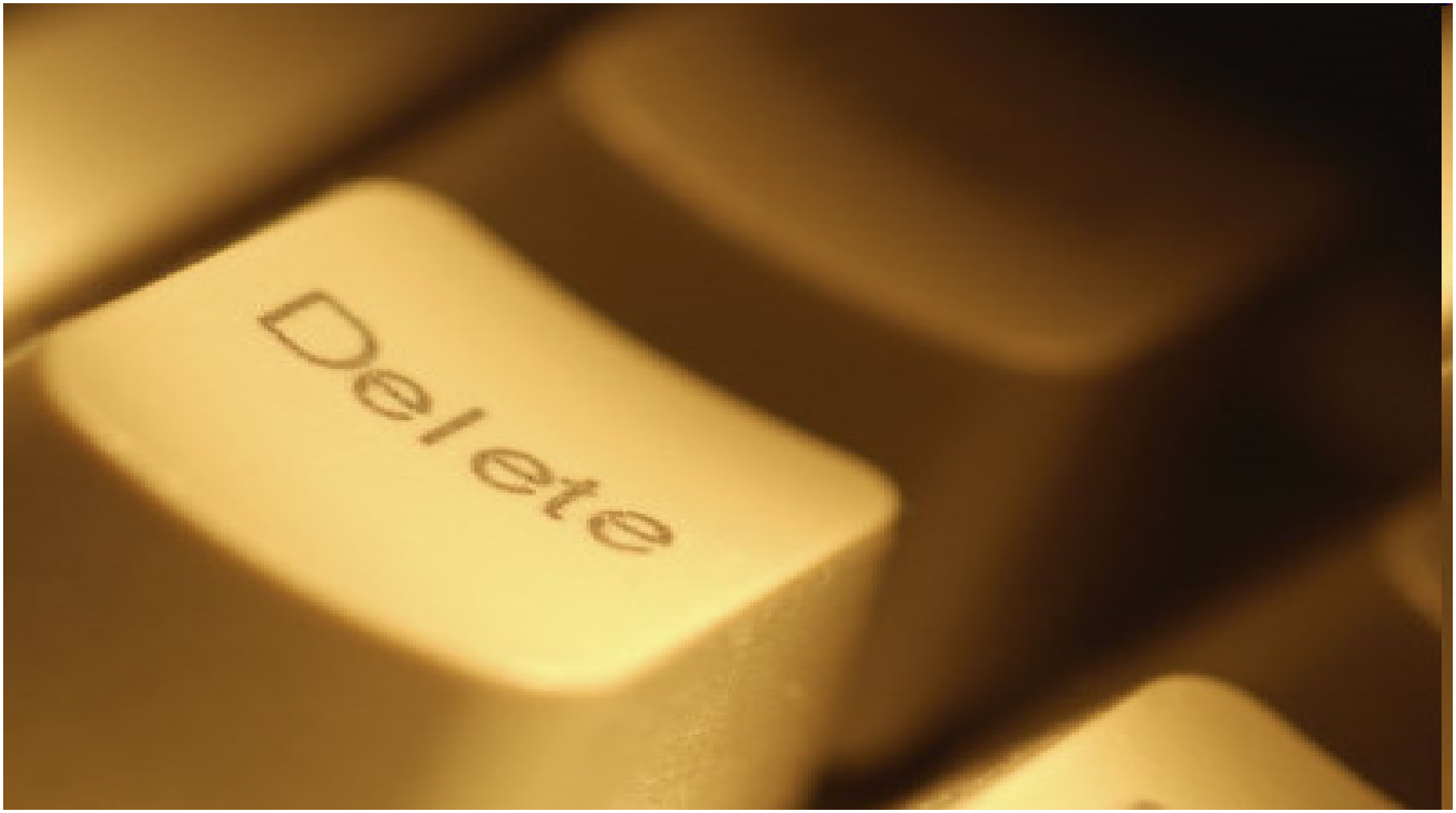
- **The amount of digital information continues to grow**
  - IT needs to manage more storage with shrinking budgets ...
- **More data has to be kept, longer**
  - Losing data is unacceptable …
  - Backup per definition multiplies the amount of data that is kept ...

We Need to do More with Less, and we need to do it smarter

What is data reduction

- Data reduction is the **removal of old, stale or orphaned data** from active disk and **removing unnecessary data** to save time and space thus improving the ROI and TCO or an existing investment

The Greenest Button on the Keyboard



There are a Variety of Data Reduction Methods

Eliminate Data	• Stale, orphaned, obsolete, non-business data
Incremental Forever	• Only new or changed files are transmitted • Avoids regular full backups required by many backup solutions! • Conserves network bandwidth and server storage
Compression	• Encoding of data to reduce size • Typically localized, such as to a file, directory tree or storage volume • Also done within storage devices themselves
Single instance store (SIS)	• A form of compression, usually applied to a large collection of files in a shared data store • Only one instance of a file is retained in the data store • Duplicate instances of the file reference the stored instance • Also known as redundant file elimination
Data deduplication	• A form of compression, usually applied to a large collection of files in a shared data store • In contrast to SIS, deduplication often refers to elimination of redundant subfiles (also known as chunks, blocks, or extents) • Only one instance is stored for each common chunk • Duplicate instances of the chunk reference the stored instance

Discovering the Benefits of Data Reduction

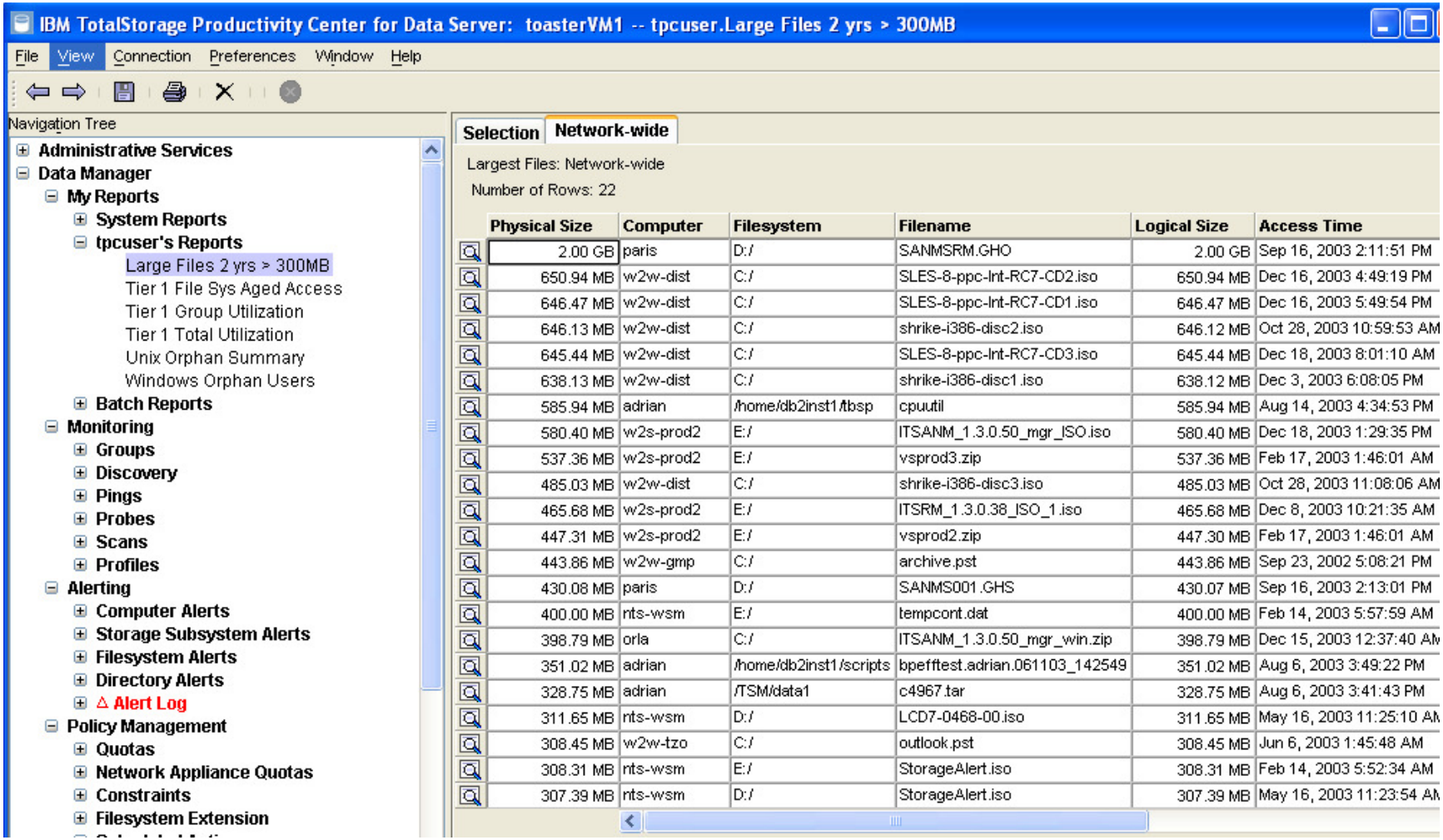
TotalStorage Productivity Center for data can:

- Find all data in your environment by age, owner etc
- Sort this data according to policy
  - ie: all data not accessed over 6 months
- Can find all duplicate files
- Can take action against that data

Uses:

- Find all data with no owner (archive)
- Find all data no longer accessed (Archive or HSM)
- Find all trash data (temp, core, dump files) (delete)

Screenshot of TPC Analysis



IBM's Space Management

What it does

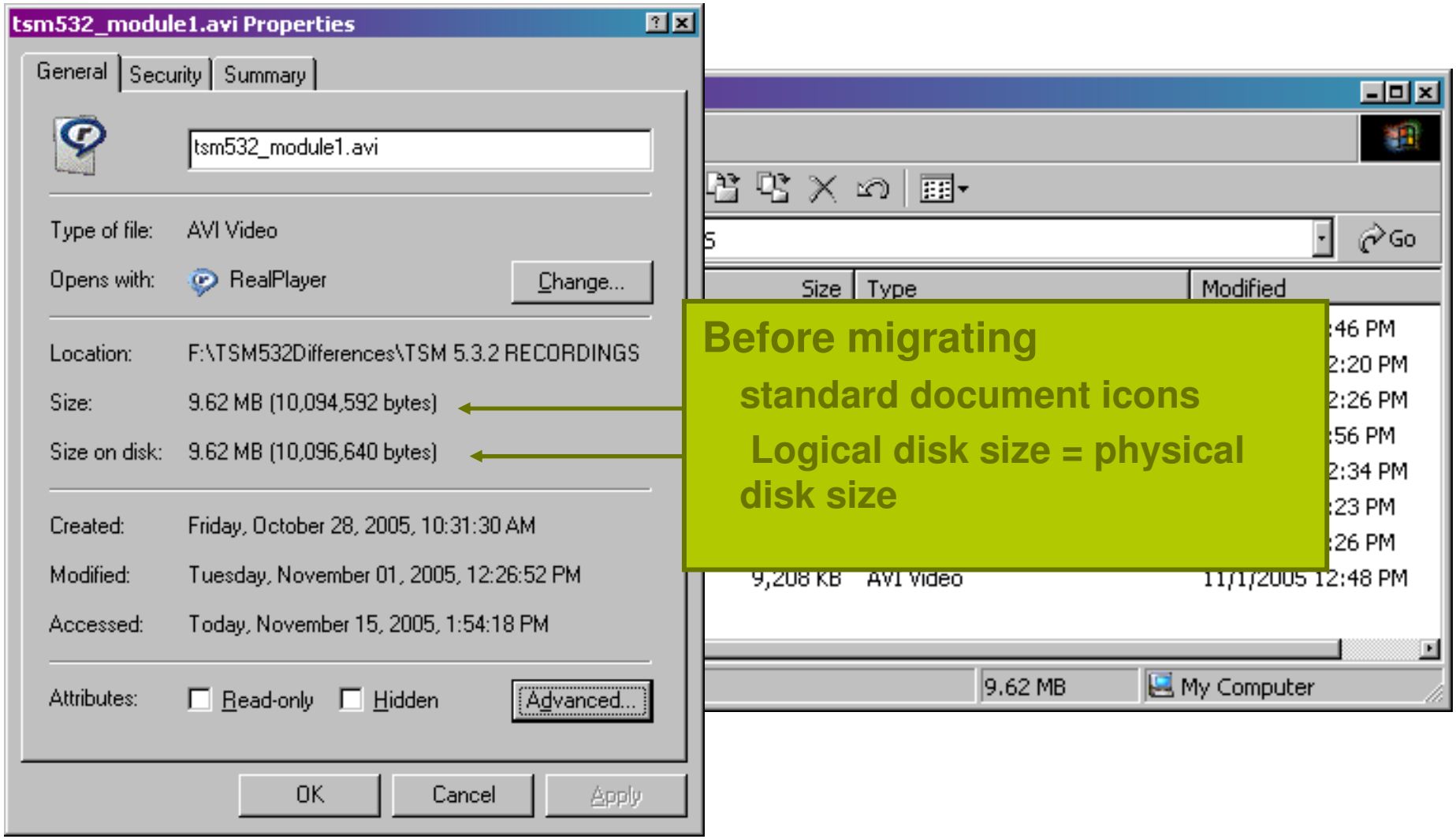
- Automatically scans production machine
- Moves these files to a better suited storage pool
- Replaces the real files with a shortcut
- Is transparant for users and applications

Why is this important

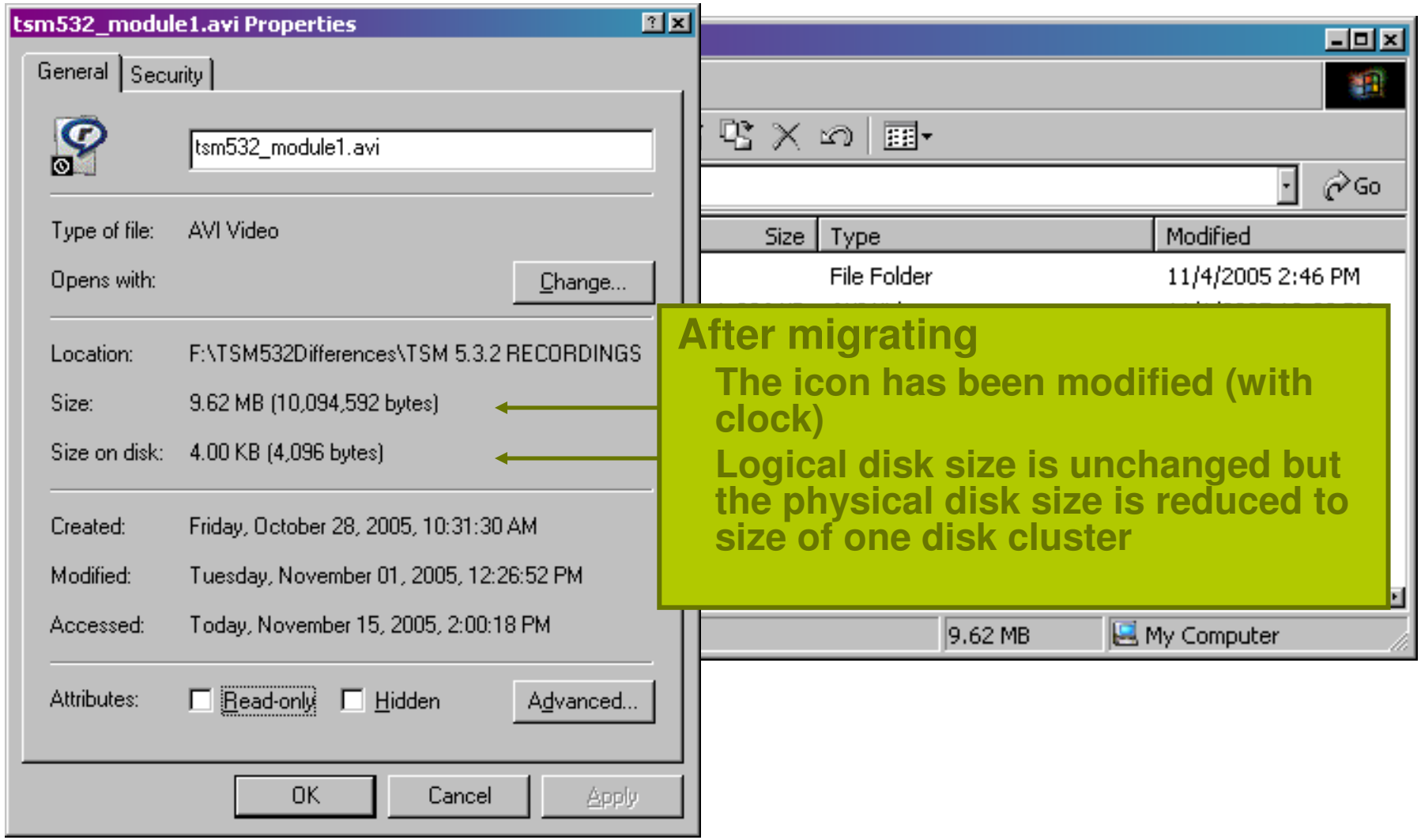
- Allows the retention of data for longer periods of time
- Transparent to the users
- Hardware purchase avoidance
- Better ROI and TCO of existing infrastructure
- Applications and files systems are faster
- Backup and recovery faster

Tivoli Storage Manager for HSM Windows
Tivoli Storage Manager for Space Management (UNIX)

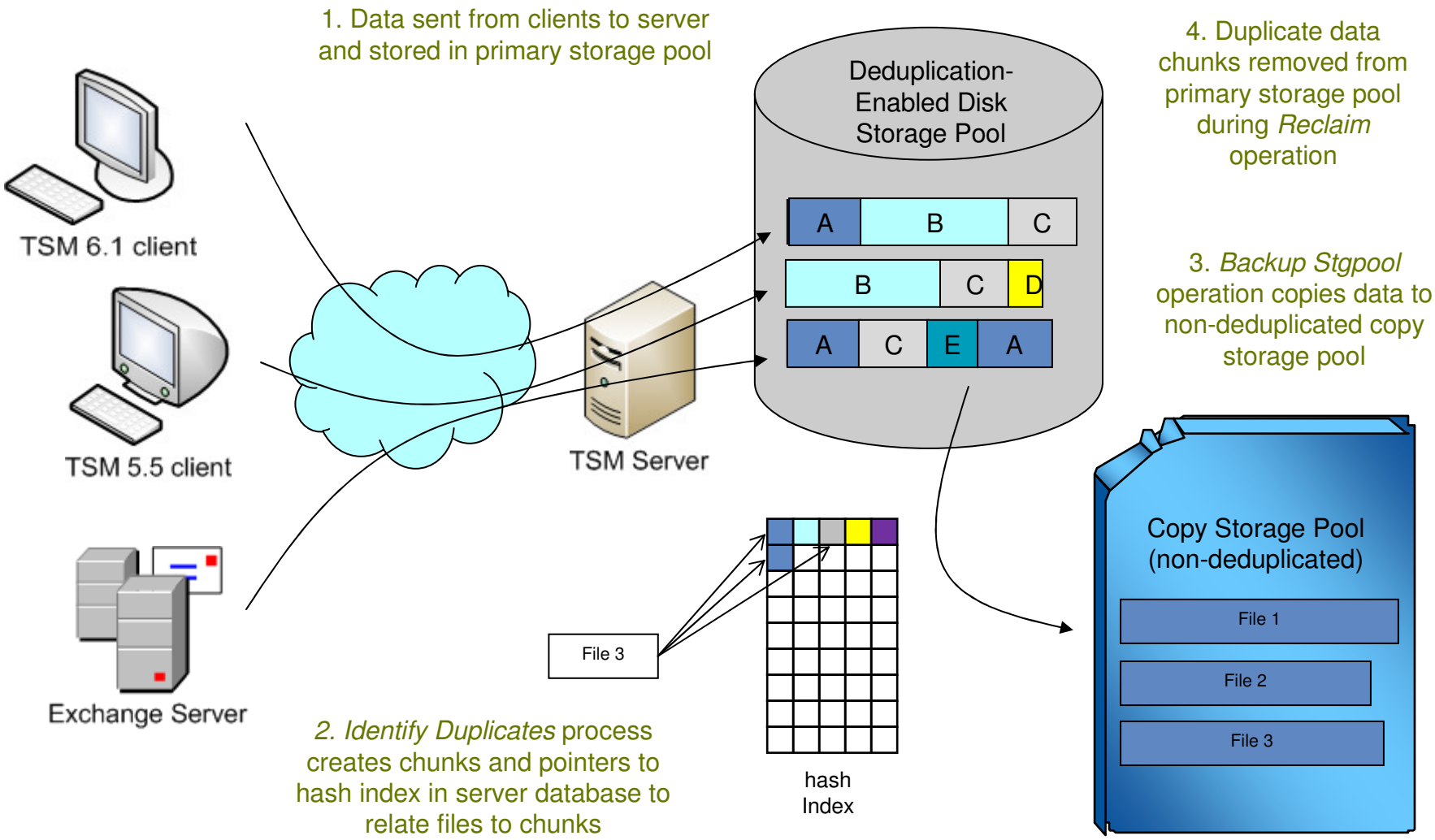
Example Before



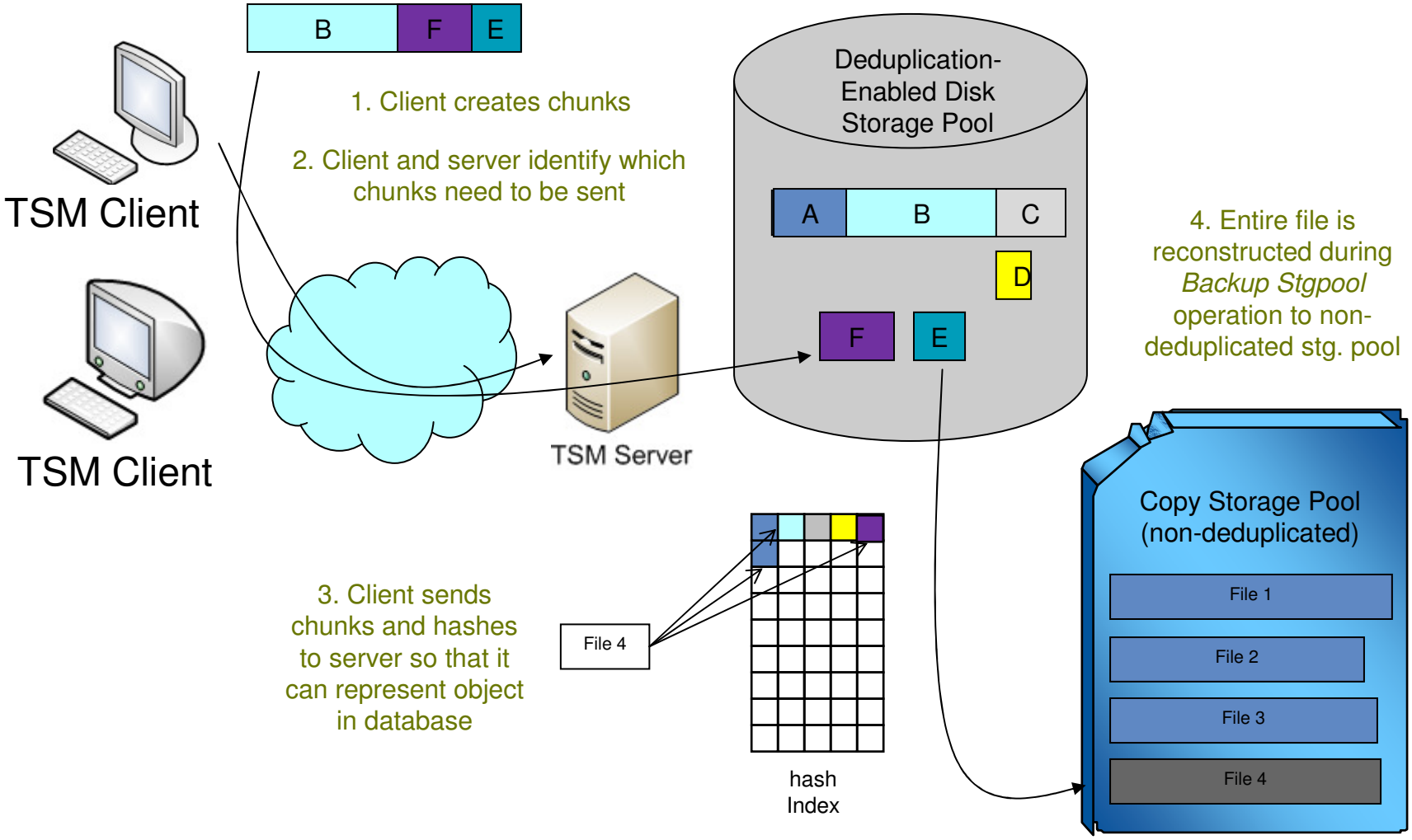
Example After



TSM Server Side Data Deduplication (TSM 6.1)



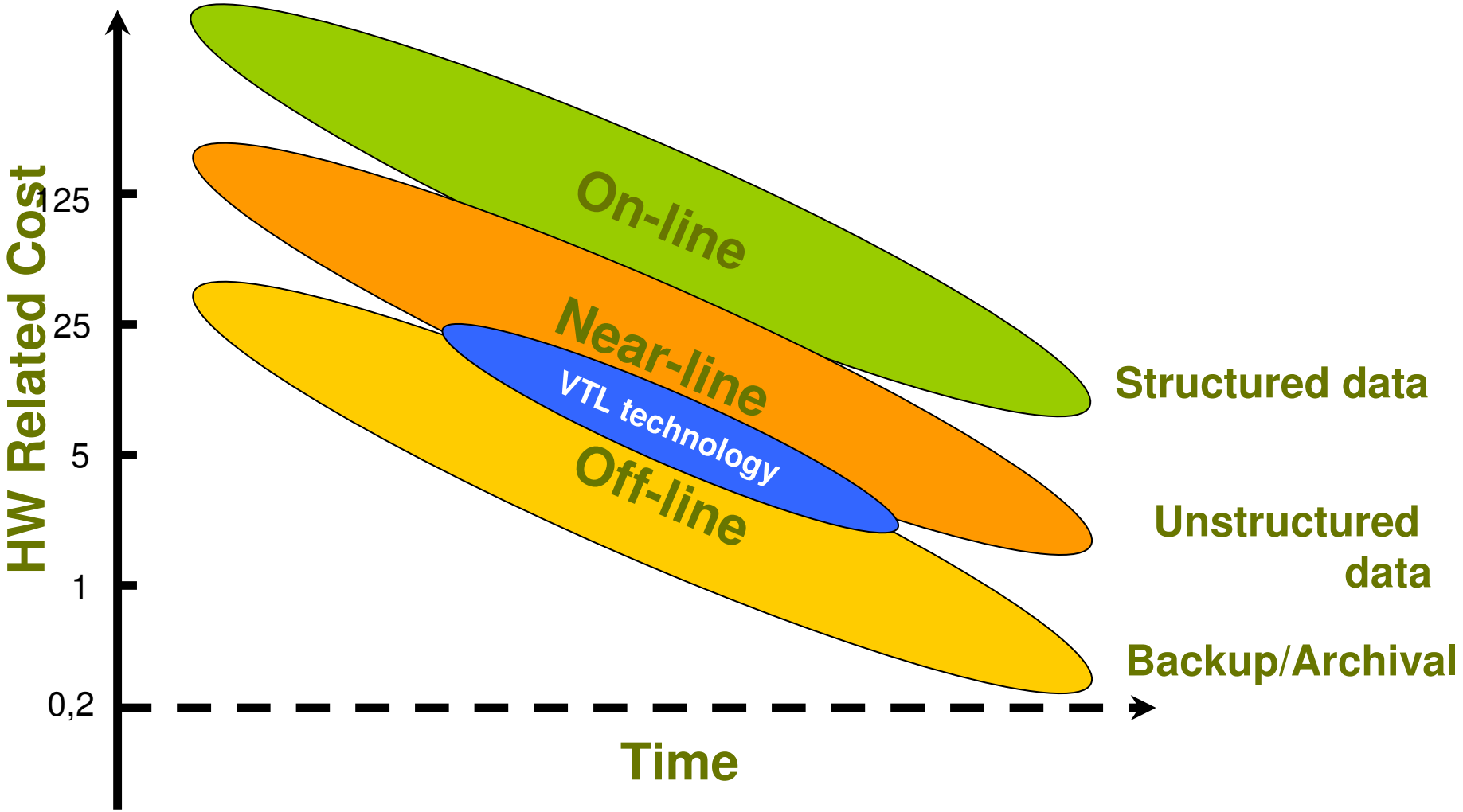
TSM Client Side Data Deduplication (TSM 6.2)



TS7650G ProtecTIER Overview

- ProtecTIER software resides on TS7650G Deduplication Gateway
- Emulates a tape library unit, including drives, cartridges and robotics
- Uses FC-attached disk array as the backup medium

Storage Classes



IBM offers Storage Systems for all storage classes

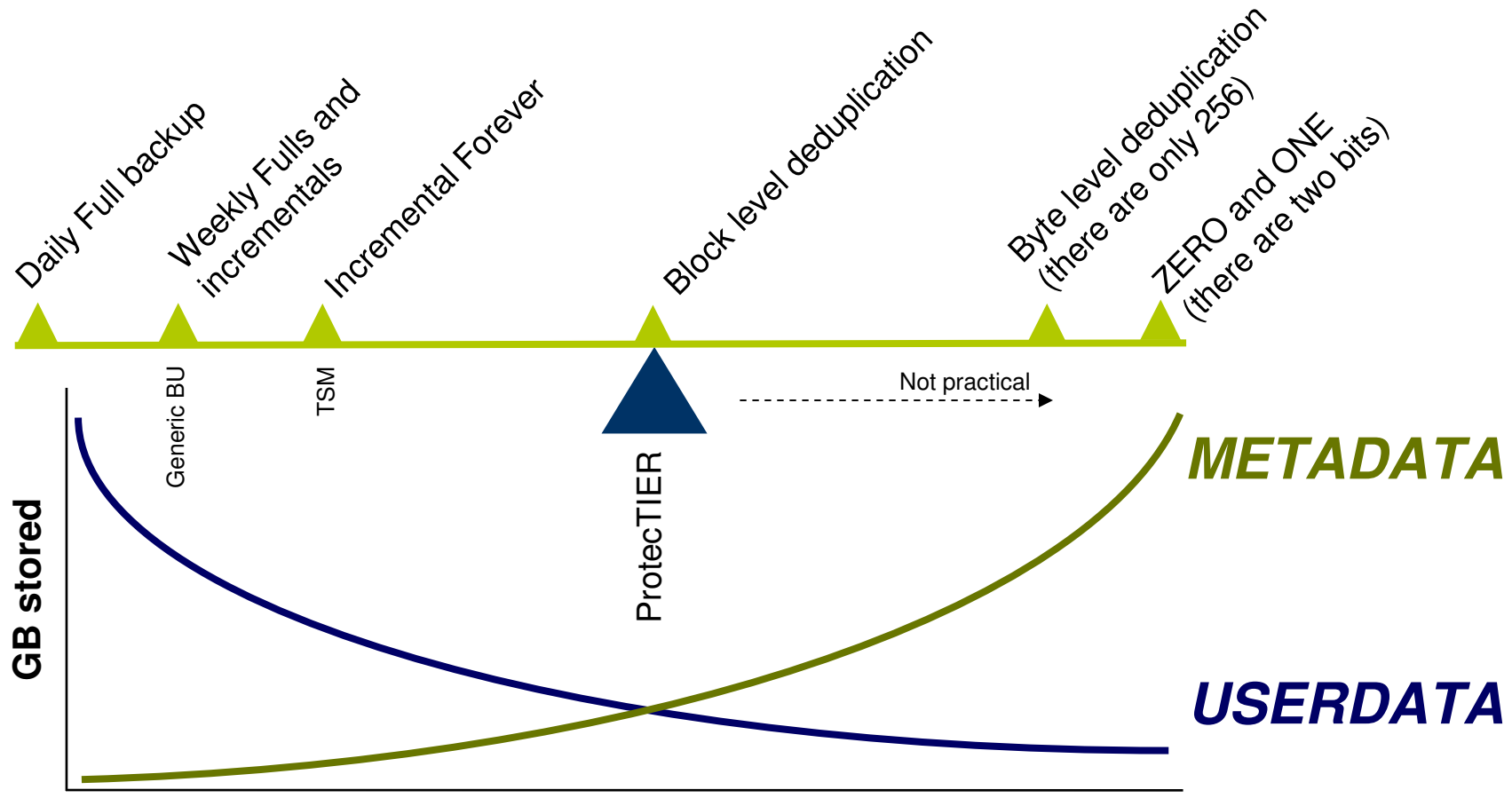
VTL Benefits

- Speed
- Flexibility
- Integration
- Replication

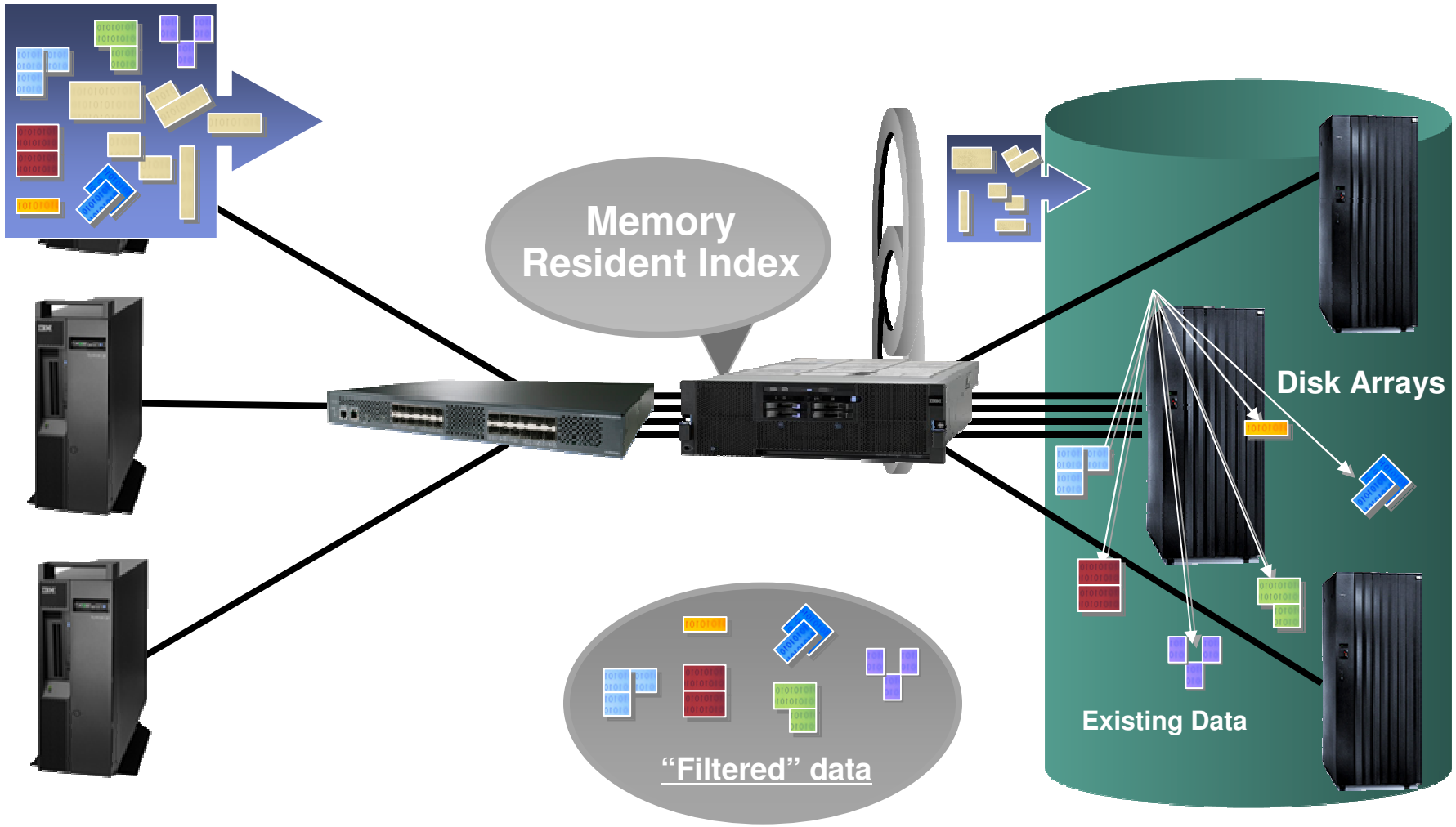
But.. Not at the price point of traditional tape.
Deduplication will help!

The Dedupe Balance

- Deduplication requires extra information -> Metadata
- More dedupe means more Metadata



Inside ProtecTIER TS7650G



ProtecTIER versus TSM 6 Built-in Deduplication

Both Solutions Offer the Benefits of Deduplication

- Greatly reduced storage capacity requirements
- Lower operational costs, energy usage and TCO
- Faster recoveries with more data on disk

Use ProtecTIER When

- Highest performance and capacity scaling are required!
- Up to 500 MB/sec (1GB/s with 2 node) deduplication rates are needed
- Deduplicated capacities up to 25 PB are required
- You desire deduplication be done inline during data ingest
- A VTL appliance model is desired
- Deduplicating across multiple TSM (or other backup) servers

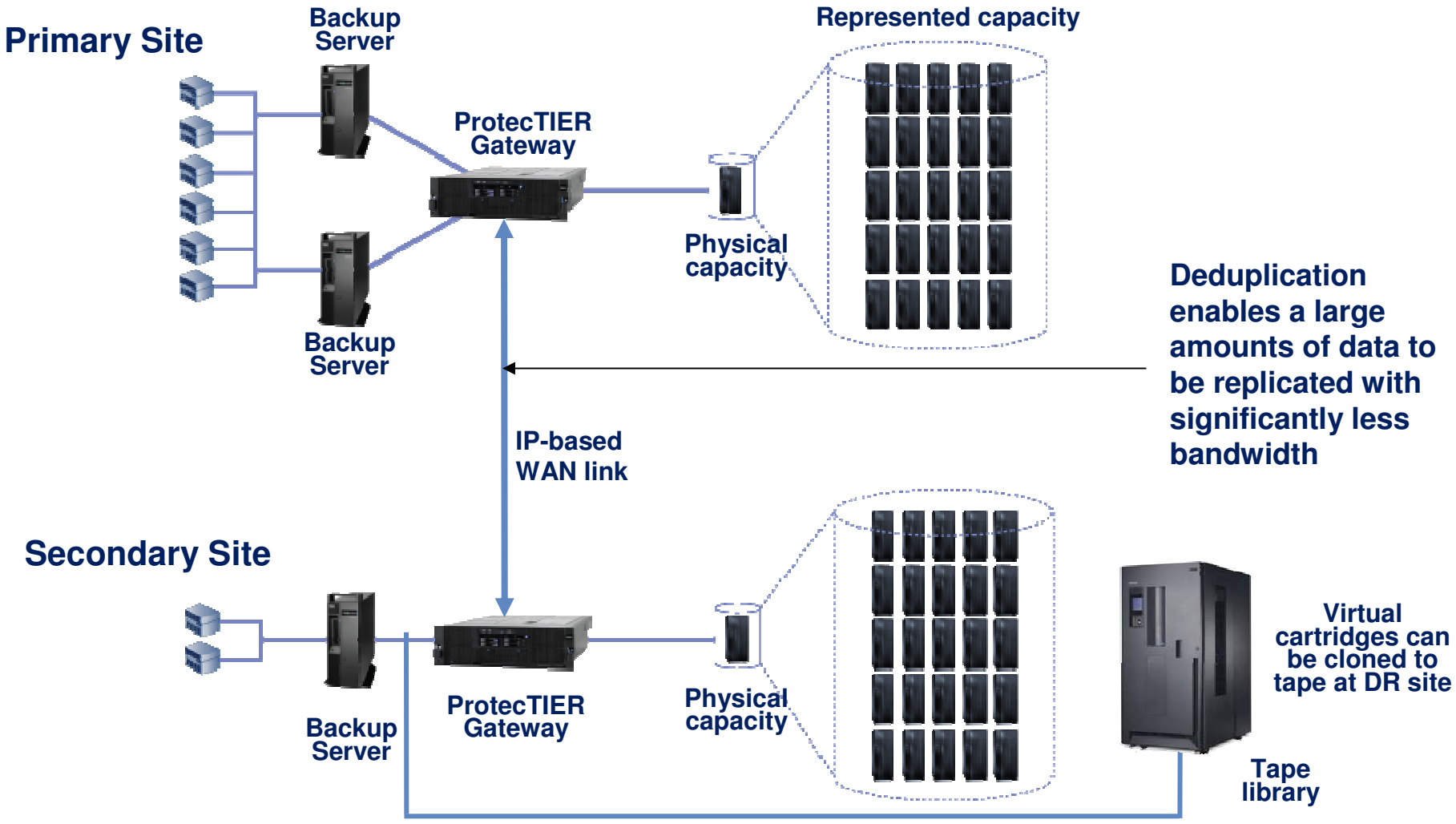
IBM ProtecTIER

Use TSM 6 Built-in Deduplication When

- Sufficient TSM server resources can be made available and you desire deduplication operations be completely integrated within TSM
- The benefits of deduplication are desired without separate hardware or software dependencies or licenses (ships with TSM Extended Edition)
- You desire end to end data lifecycle management with minimized data store

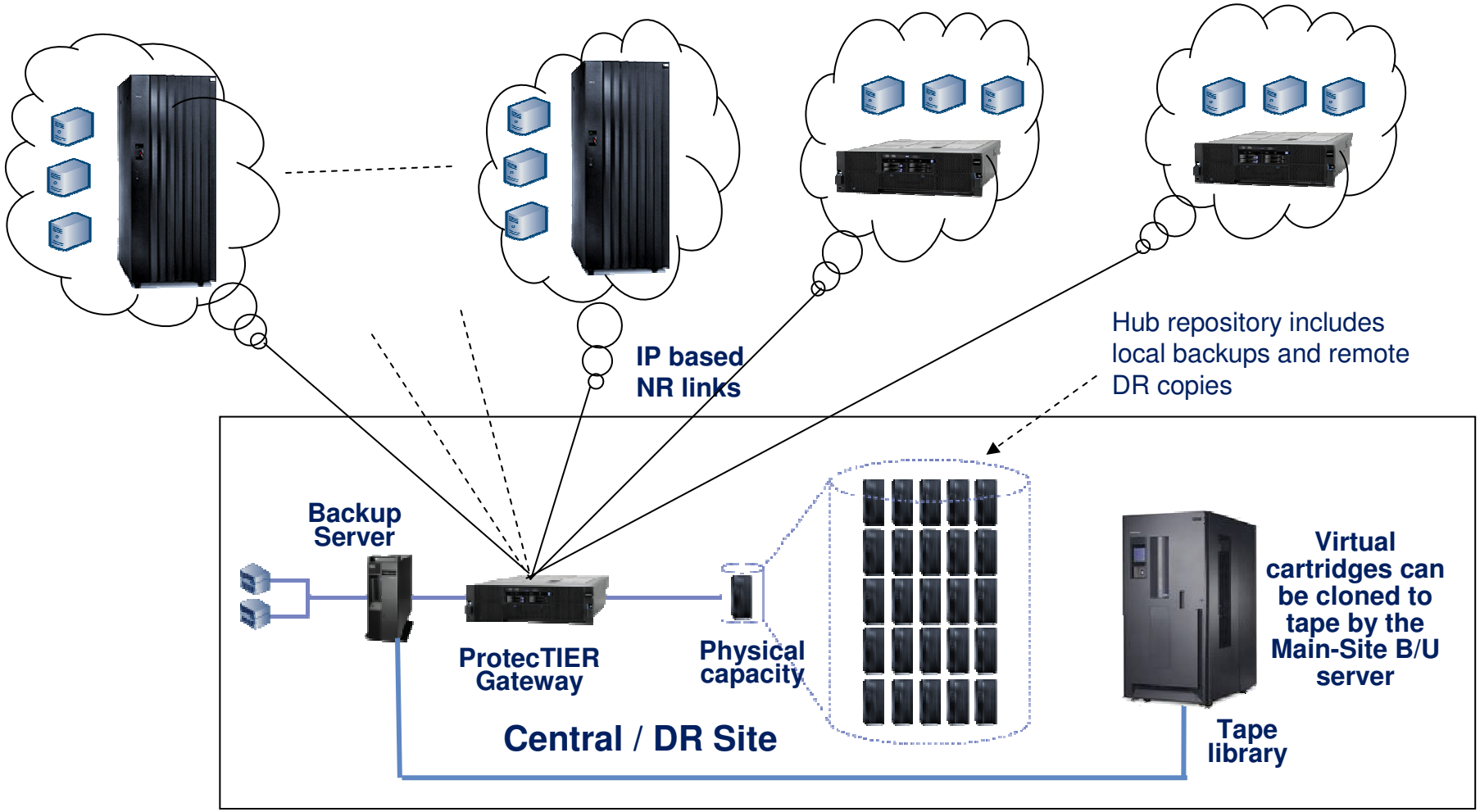
TSM

ProtecTIER Native IP Replication



Many-to-1 Native Replication DR Solution

Up to 12 Branch Offices (spokes): Gateways and/or Appliances
1 target (hub): Appliance, Gateway, single or two-node cluster



Questions

- We welcome your questions during the closing drink following on this lecture
- For more information you can reach Rob Turk via email: rob.turk1@nl.ibm.com