
“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	<ul style="list-style-type: none">• Stale, orphaned, obsolete, non-business data
Incremental Forever	<ul style="list-style-type: none">• Only new or changed files are transmitted• Avoids regular full backups required by many backup solutions!• Conserves network bandwidth and server storage
Compression	<ul style="list-style-type: none">• 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)	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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 TotalStorage Productivity Center for Data Server: toasterVM1 -- tpcuser.Large Files 2 yrs > 300MB

File View Connection Preferences Window Help

Navigation Tree

- Administrative Services
 - Data Manager
 - My Reports
 - System Reports
 - tpcuser's Reports
 - Large Files 2 yrs > 300MB
 - Tier 1 File Sys Aged Access
 - Tier 1 Group Utilization
 - Tier 1 Total Utilization
 - Unix Orphan Summary
 - Windows Orphan Users
 - Batch Reports
 - Monitoring
 - Groups
 - Discovery
 - Pings
 - Probes
 - Scans
 - Profiles
 - Alerting
 - Computer Alerts
 - Storage Subsystem Alerts
 - Filesystem Alerts
 - Directory Alerts
 - Alert Log
 - Policy Management
 - Quotas
 - Network Appliance Quotas
 - Constraints
 - Filesystem Extension

Selection Network-wide

Largest Files: Network-wide

Number of Rows: 22

	Physical Size	Computer	Filesystem	Filename	Logical Size	Access Time
	2.00 GB	paris	D:/	SANMSRM.GHO	2.00 GB	Sep 16, 2003 2:11:51 PM
	650.94 MB	w2w-dist	C:/	SLES-8-ppc-Int-RC7-CD2.iso	650.94 MB	Dec 16, 2003 4:49:19 PM
	646.47 MB	w2w-dist	C:/	SLES-8-ppc-Int-RC7-CD1.iso	646.47 MB	Dec 16, 2003 5:49:54 PM
	646.13 MB	w2w-dist	C:/	shrike-i386-disc2.iso	646.12 MB	Oct 28, 2003 10:59:53 AM
	645.44 MB	w2w-dist	C:/	SLES-8-ppc-Int-RC7-CD3.iso	645.44 MB	Dec 18, 2003 8:01:10 AM
	638.13 MB	w2w-dist	C:/	shrike-i386-disc1.iso	638.12 MB	Dec 3, 2003 6:08:05 PM
	585.94 MB	adrian	/home/db2inst1/ftbsp	cpuutil	585.94 MB	Aug 14, 2003 4:34:53 PM
	580.40 MB	w2s-prod2	E:/	ITSANM_1.3.0.50_mgr_ISO.iso	580.40 MB	Dec 18, 2003 1:29:35 PM
	537.36 MB	w2s-prod2	E:/	vsprod3.zip	537.36 MB	Feb 17, 2003 1:46:01 AM
	485.03 MB	w2w-dist	C:/	shrike-i386-disc3.iso	485.03 MB	Oct 28, 2003 11:08:06 AM
	465.68 MB	w2s-prod2	E:/	ITSRM_1.3.0.38_ISO_1.iso	465.68 MB	Dec 8, 2003 10:21:35 AM
	447.31 MB	w2s-prod2	E:/	vsprod2.zip	447.30 MB	Feb 17, 2003 1:46:01 AM
	443.86 MB	w2w-gmp	C:/	archive.pst	443.86 MB	Sep 23, 2002 5:08:21 PM
	430.08 MB	paris	D:/	SANMS001.GHS	430.07 MB	Sep 16, 2003 2:13:01 PM
	400.00 MB	nts-wsm	E:/	tempcont.dat	400.00 MB	Feb 14, 2003 5:57:59 AM
	398.79 MB	orla	C:/	ITSANM_1.3.0.50_mgr_wwin.zip	398.79 MB	Dec 15, 2003 12:37:40 AM
	351.02 MB	adrian	/home/db2inst1/scripts	bpefftest.adrian.061103_142549	351.02 MB	Aug 6, 2003 3:49:22 PM
	328.75 MB	adrian	/TSM/data1	c4967.tar	328.75 MB	Aug 6, 2003 3:41:43 PM
	311.65 MB	nts-wsm	D:/	LCD7-0468-00.iso	311.65 MB	May 16, 2003 11:25:10 AM
	308.45 MB	w2w-tzo	C:/	outlook.pst	308.45 MB	Jun 6, 2003 1:45:48 AM
	308.31 MB	nts-wsm	E:/	StorageAlert.iso	308.31 MB	Feb 14, 2003 5:52:34 AM
	307.39 MB	nts-wsm	D:/	StorageAlert.iso	307.39 MB	May 16, 2003 11:23:54 AM

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 transparent 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

The image shows a Windows XP desktop environment. On the left, the 'tsm532_module1.avi Properties' dialog box is open, displaying the 'General' tab. The file name is 'tsm532_module1.avi', its type is 'AVI Video', and it opens with 'RealPlayer'. The location is 'F:\TSM532Differences\TSM 5.3.2 RECORDINGS'. The size is 9.62 MB (10,094,592 bytes) and the size on disk is 9.62 MB (10,096,640 bytes). The creation date is Friday, October 28, 2005, 10:31:30 AM, and the modification date is Tuesday, November 01, 2005, 12:26:52 PM. The accessed date is Today, November 15, 2005, 1:54:18 PM. The attributes are 'Read-only' and 'Hidden'. On the right, a file explorer window shows a list of files. A yellow callout box with the text 'Before migrating standard document icons Logical disk size = physical disk size' is overlaid on the file list. The callout box has two green arrows pointing to the 'Size' and 'Size on disk' fields in the properties dialog. The file explorer window shows a table with columns 'Size', 'Type', and 'Modified'. The file 'tsm532_module1.avi' is listed with a size of 9,208 KB and a type of 'AVI Video', modified on 11/1/2005 at 12:48 PM. The taskbar at the bottom shows the file size as 9.62 MB and the current directory as 'My Computer'.

**Before migrating
standard document icons
Logical disk size = physical
disk size**

Size	Type	Modified
9,208 KB	AVI Video	11/1/2005 12:48 PM

Example After

The image shows two overlapping windows from a Windows XP desktop. The foreground window is titled "tsm532_module1.avi Properties" and has the "General" tab selected. It displays the following information:

- File name: tsm532_module1.avi
- Type of file: AVI Video
- Opens with: [Change...]
- Location: F:\TSM532Differences\TSM 5.3.2 RECORDINGS
- Size: 9.62 MB (10,094,592 bytes) ←
- Size on disk: 4.00 KB (4,096 bytes) ←
- Created: Friday, October 28, 2005, 10:31:30 AM
- Modified: Tuesday, November 01, 2005, 12:26:52 PM
- Accessed: Today, November 15, 2005, 2:00:18 PM
- Attributes: Read-only Hidden [Advanced...]

The background window shows a folder view with a table of contents:

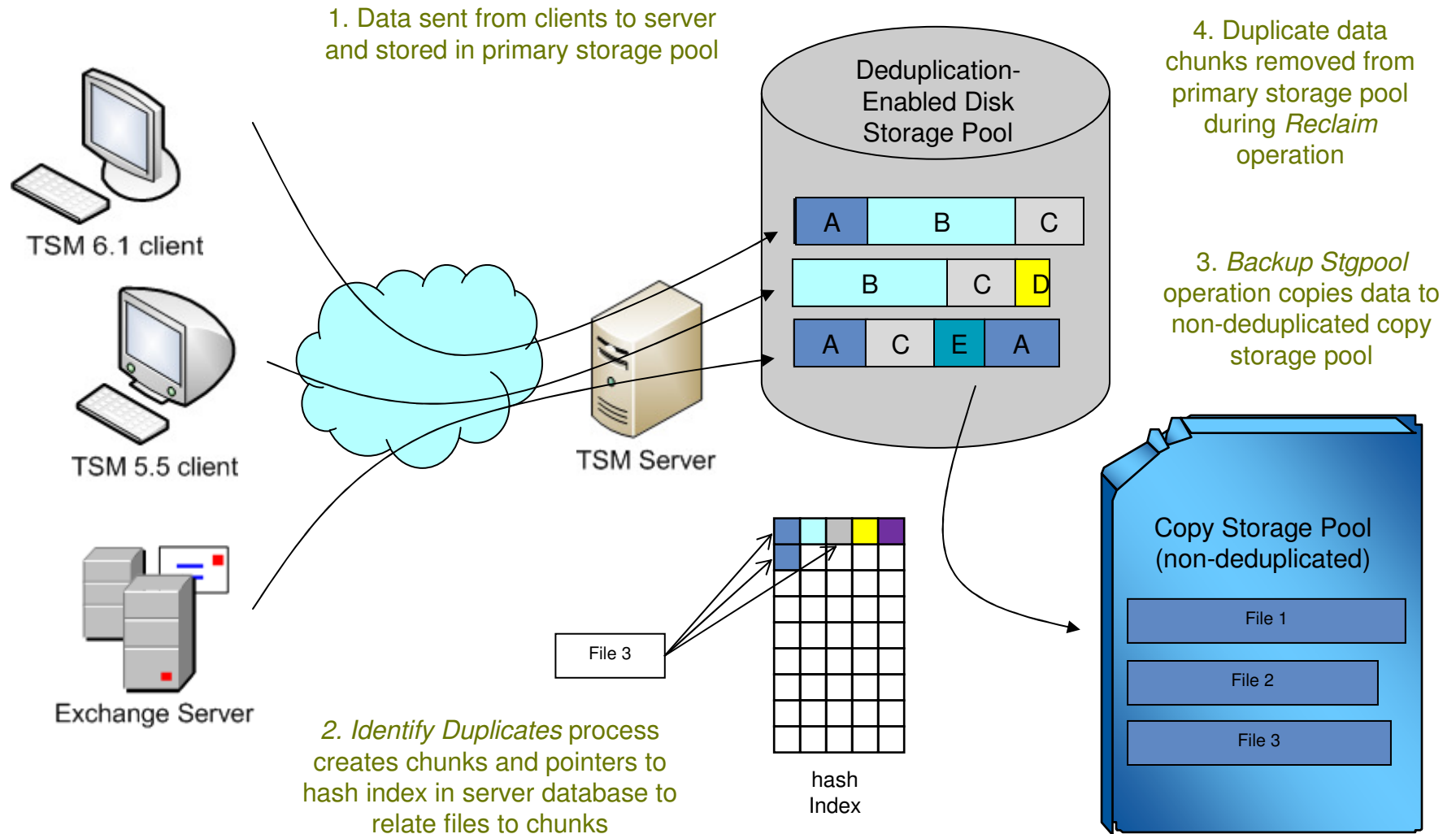
Size	Type	Modified
	File Folder	11/4/2005 2:46 PM

A yellow callout box with black text is overlaid on the right side of the image, containing the following text:

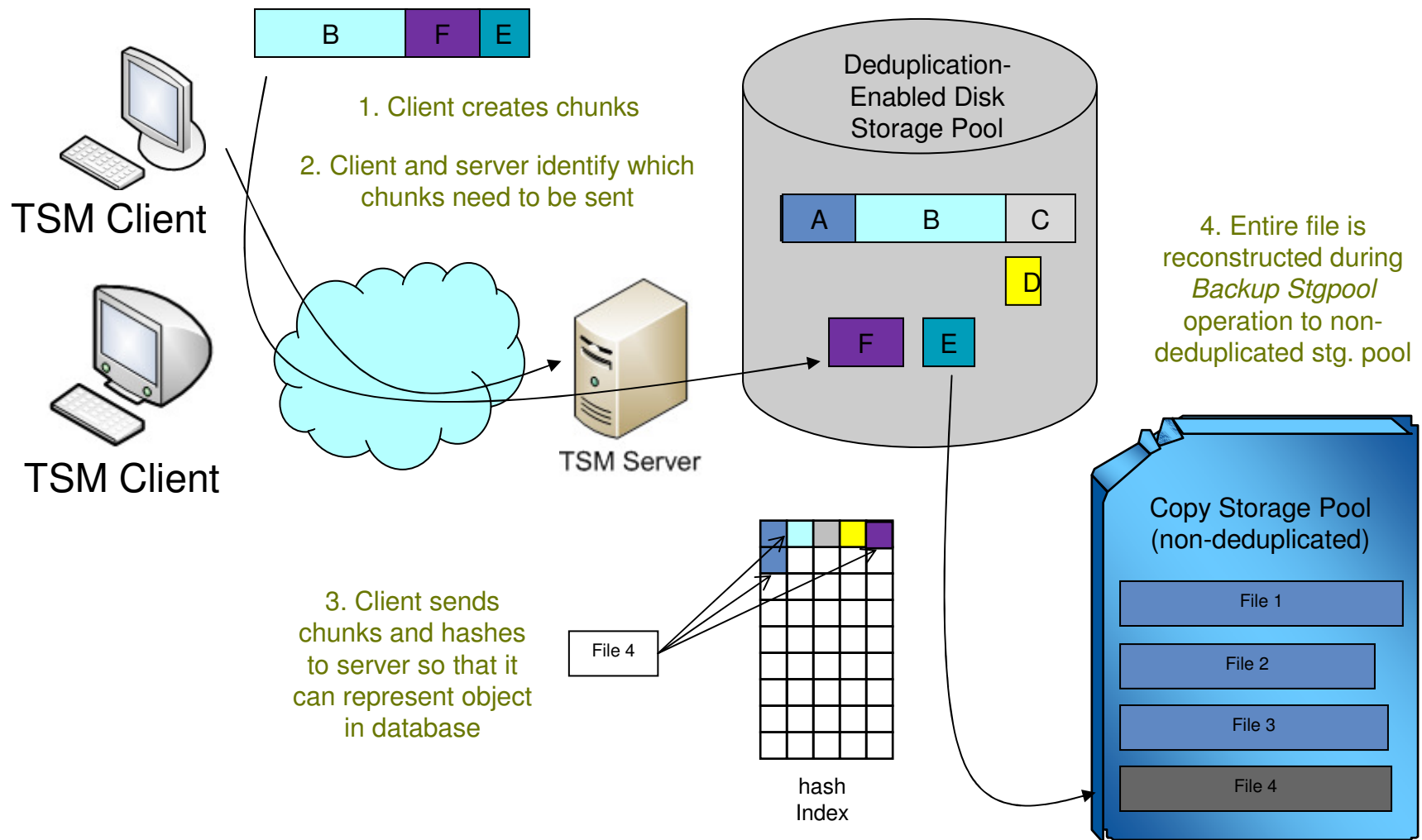
After migrating
The icon has been modified (with clock)
Logical disk size is unchanged but the physical disk size is reduced to size of one disk cluster

At the bottom of the background window, a status bar shows "9.62 MB" and "My Computer".

TSM Server Side Data Deduplication (TSM 6.1)



TSM Client Side Data Deduplication (TSM 6.2)

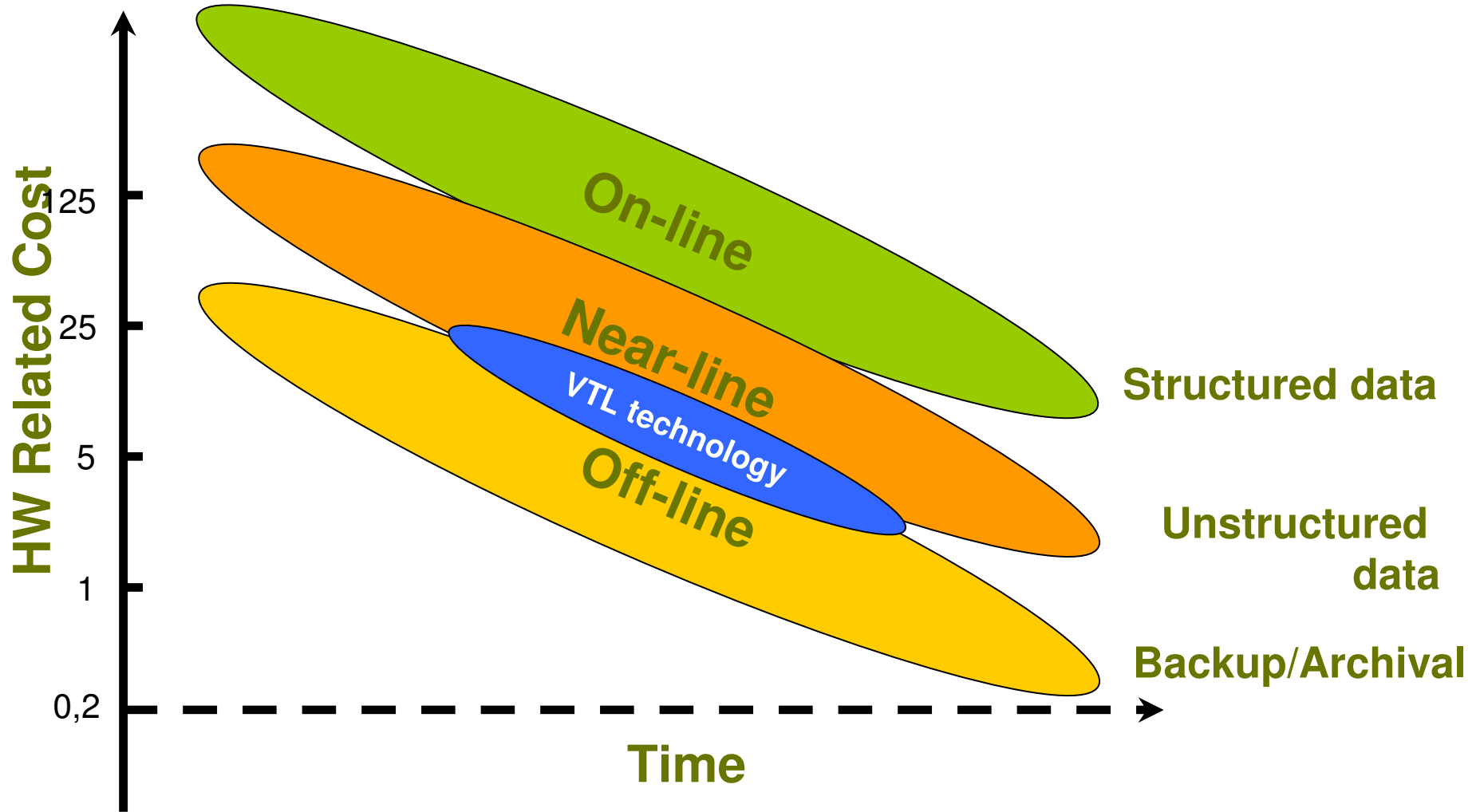


TS7650G ProtecTIER Overview



- 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

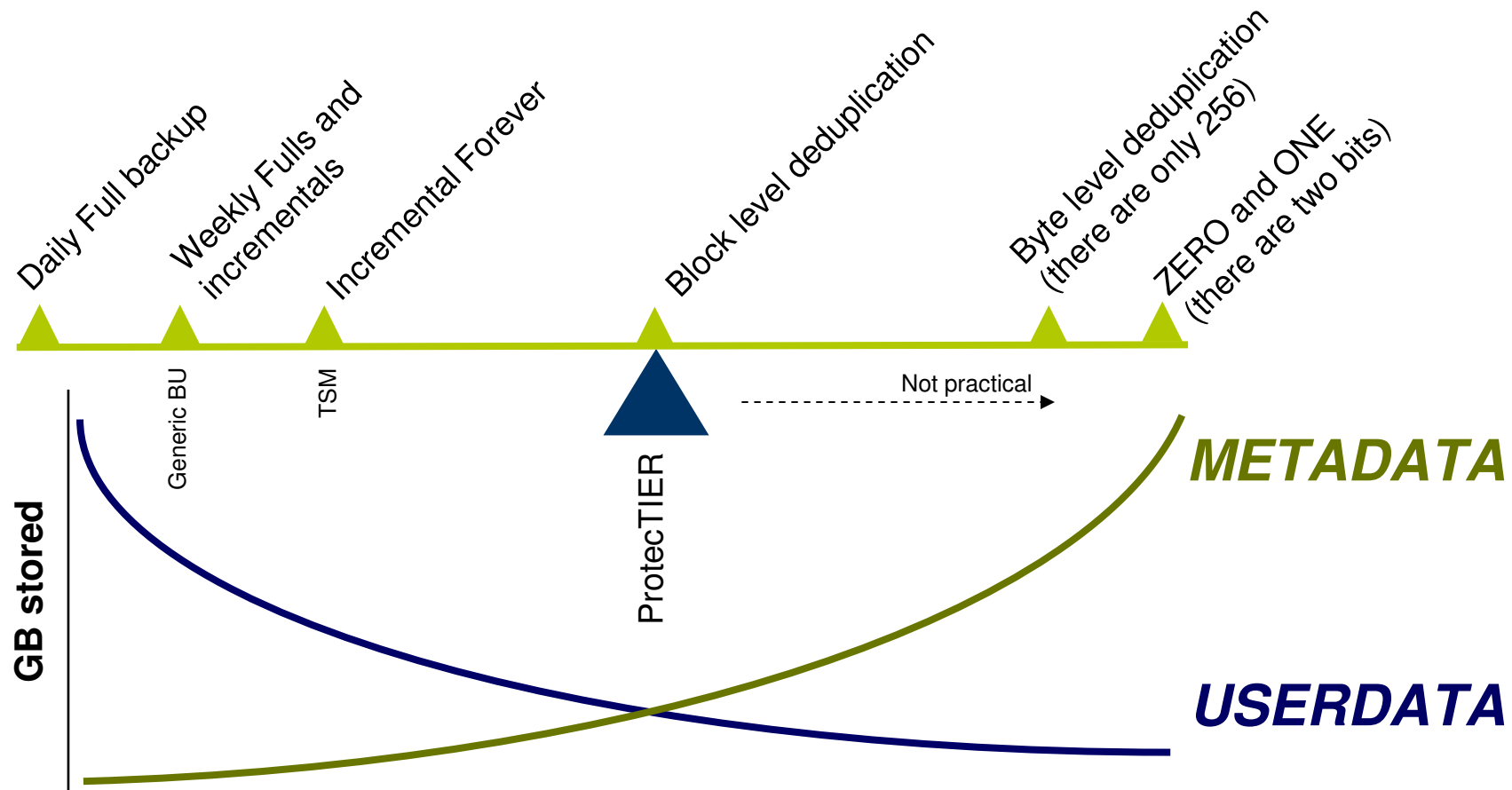
- Speed
- Flexibility
- Integration
- Replication

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

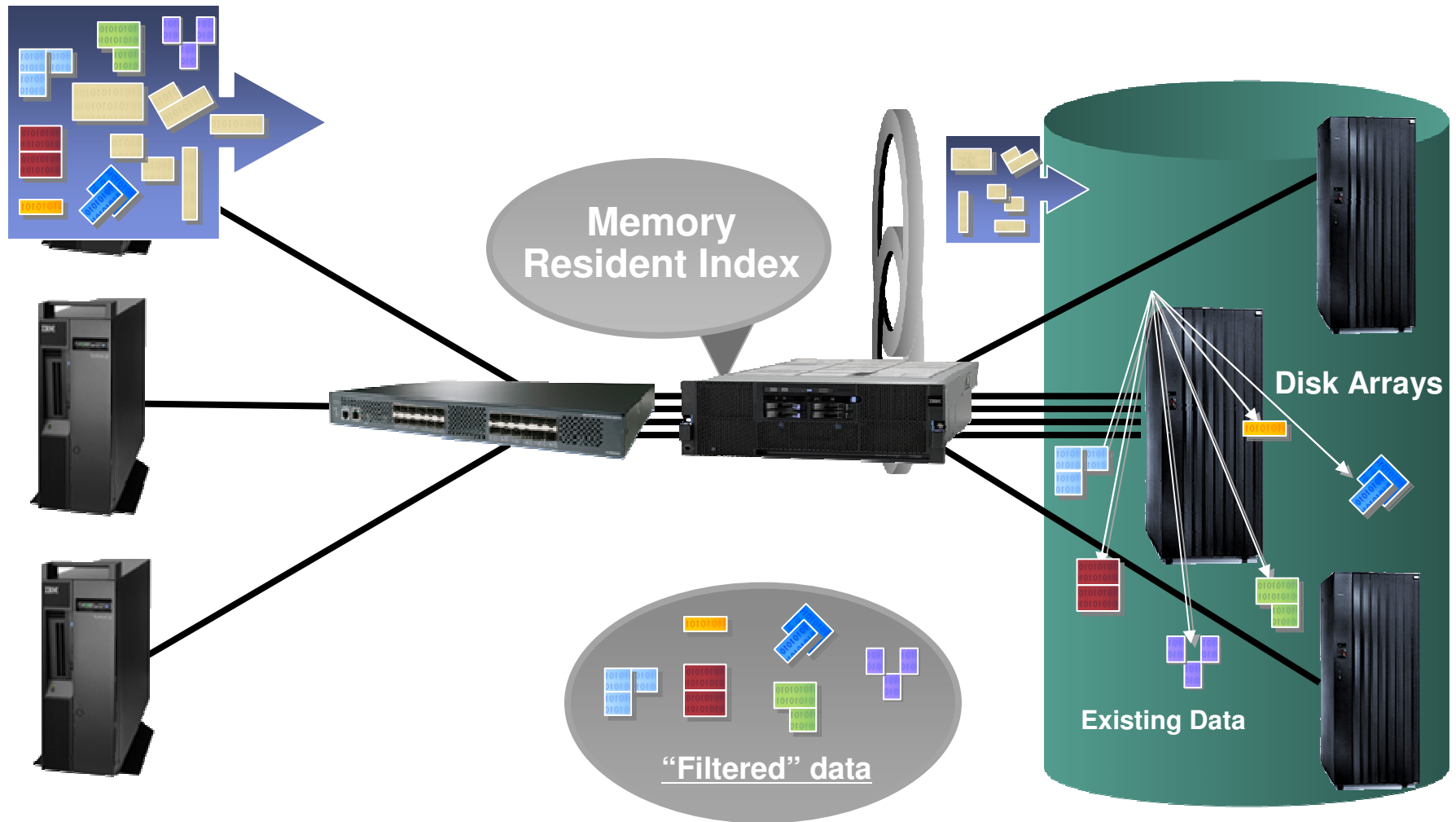
The Dedupe Balance



- 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

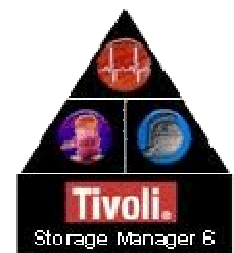
- Highest performance and capacity scaling are required!
- Up to 500 MB/sec (1GB/s with 2 node) deduplication rates are needed
- Deduplicated capacities up to 25 PB are required
- You desire deduplication be done inline during data ingest
- A VTL appliance model is desired
- Deduplicating across multiple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

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

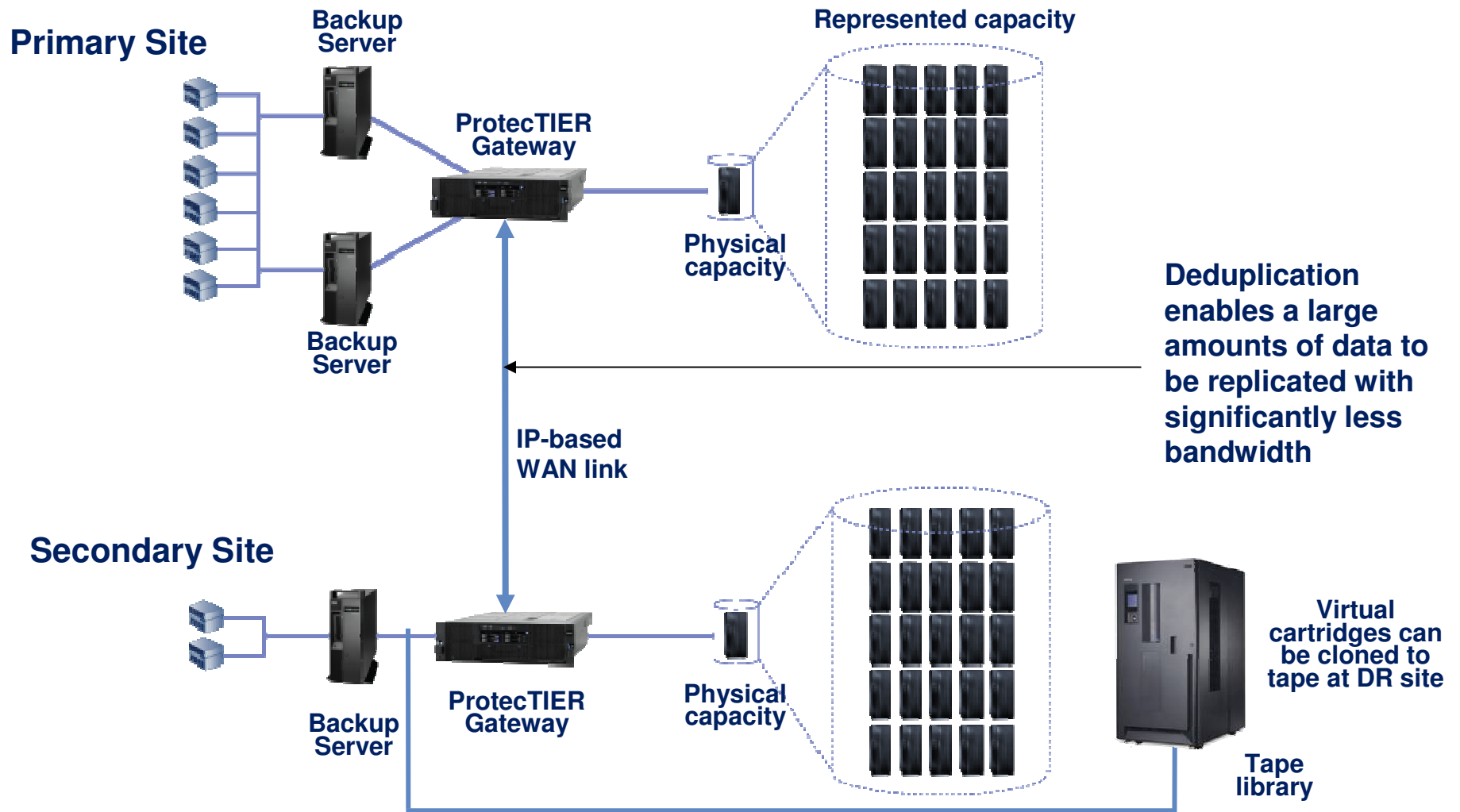


IBM
 ProtecTIER



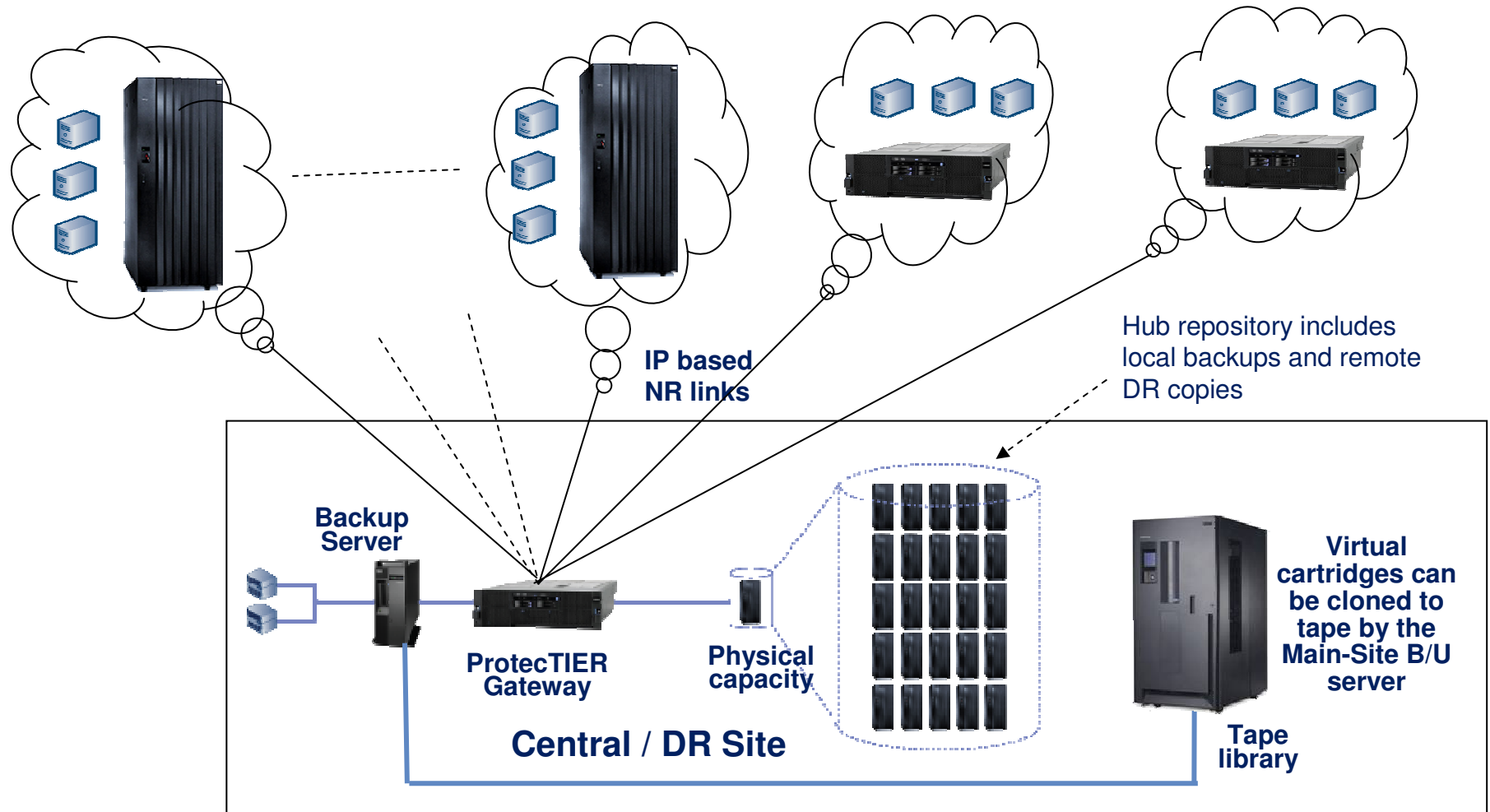
TSM

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



thank
you

- 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