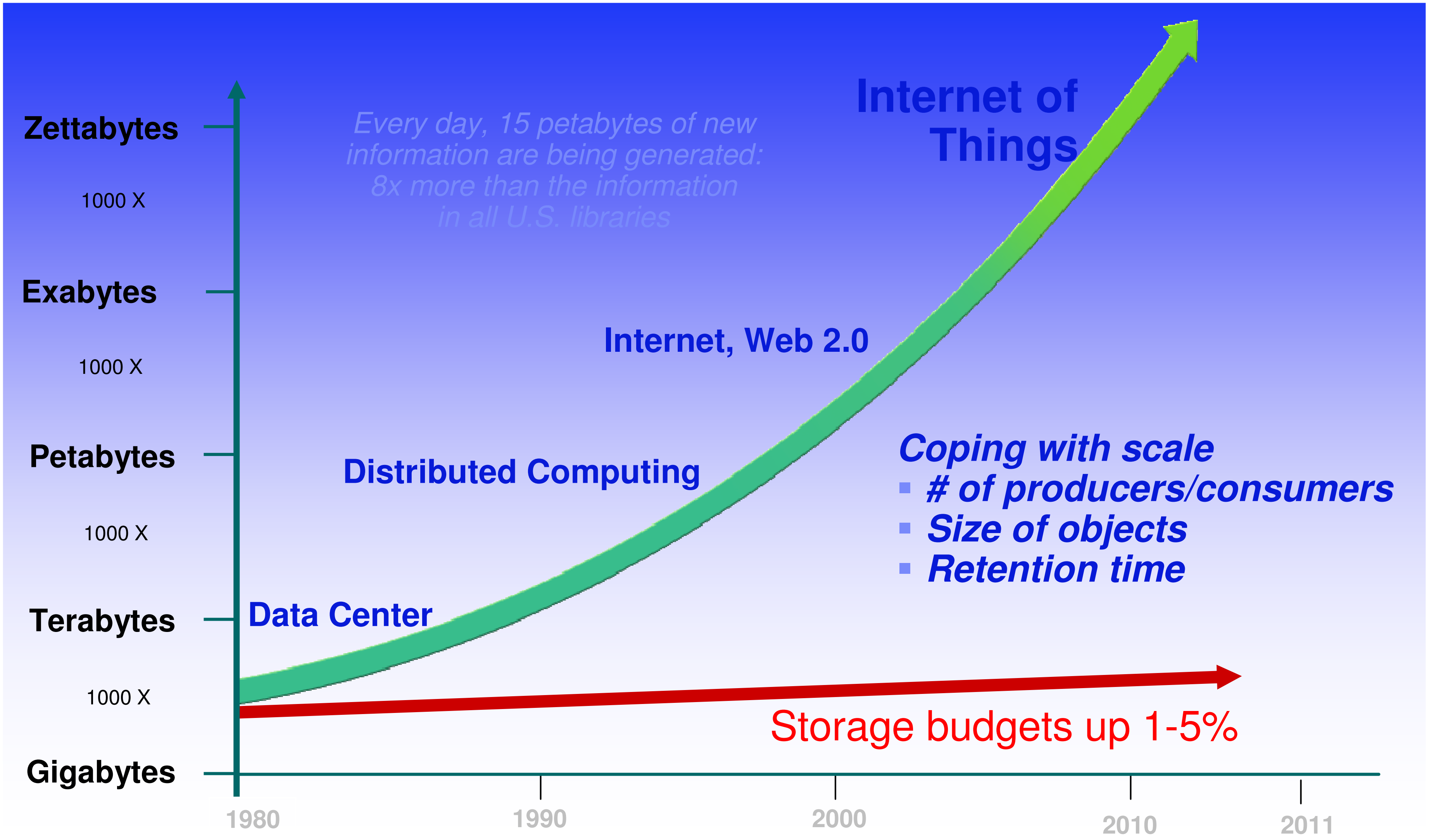
Introducing the IBM GPFS Storage Server

Jim Roche| IBM

Agenda

- Quick Industry Overview in Technical Computing Storage Devices
- Introducing the IBM System x GPFS Storage Server
- Extending GPFS with LTFS
- Summary

Storage Requirements Devouring Resources

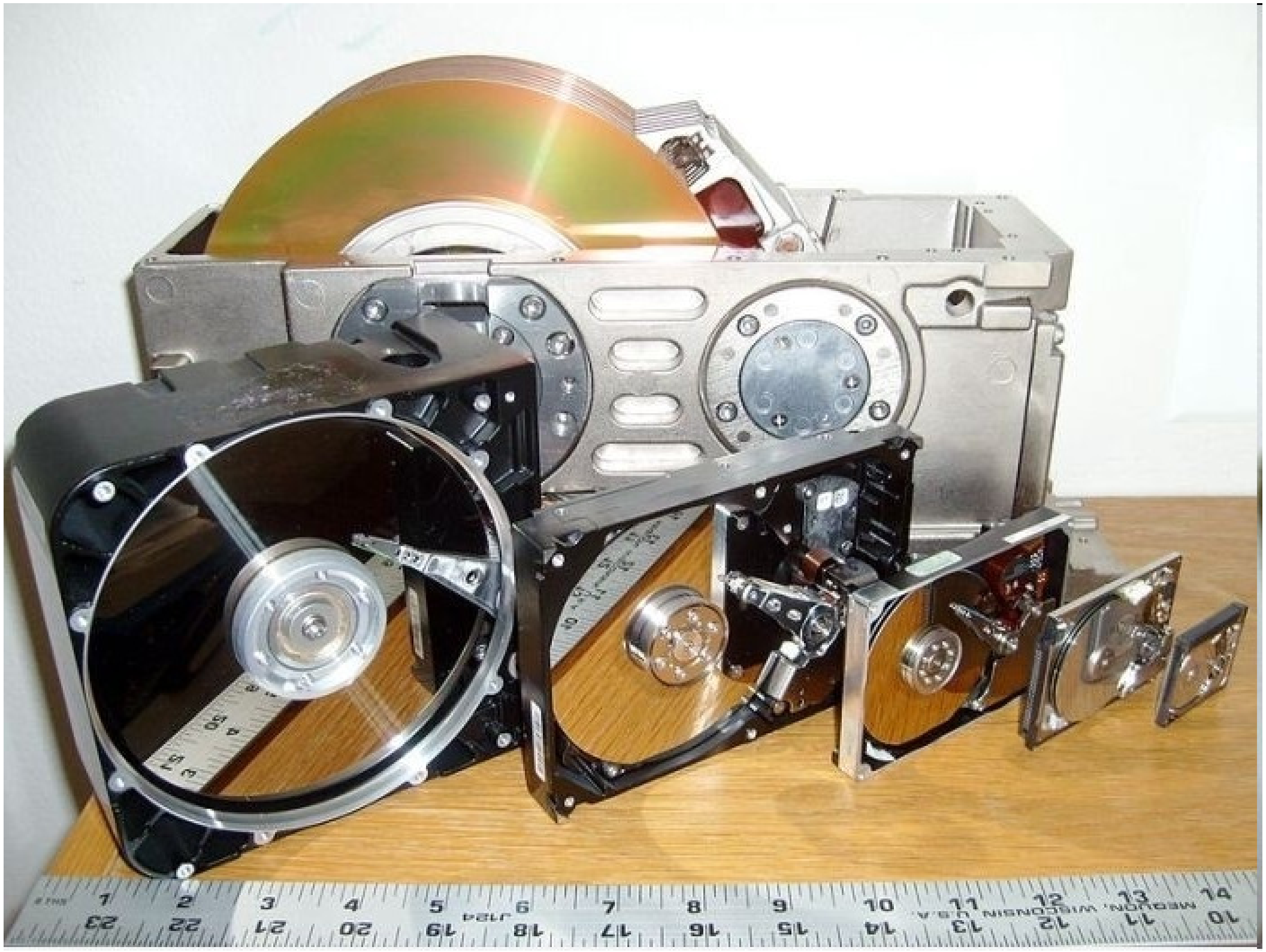


Every day, 15 petabytes of new information are being generated: 8x more than the information in all U.S. libraries

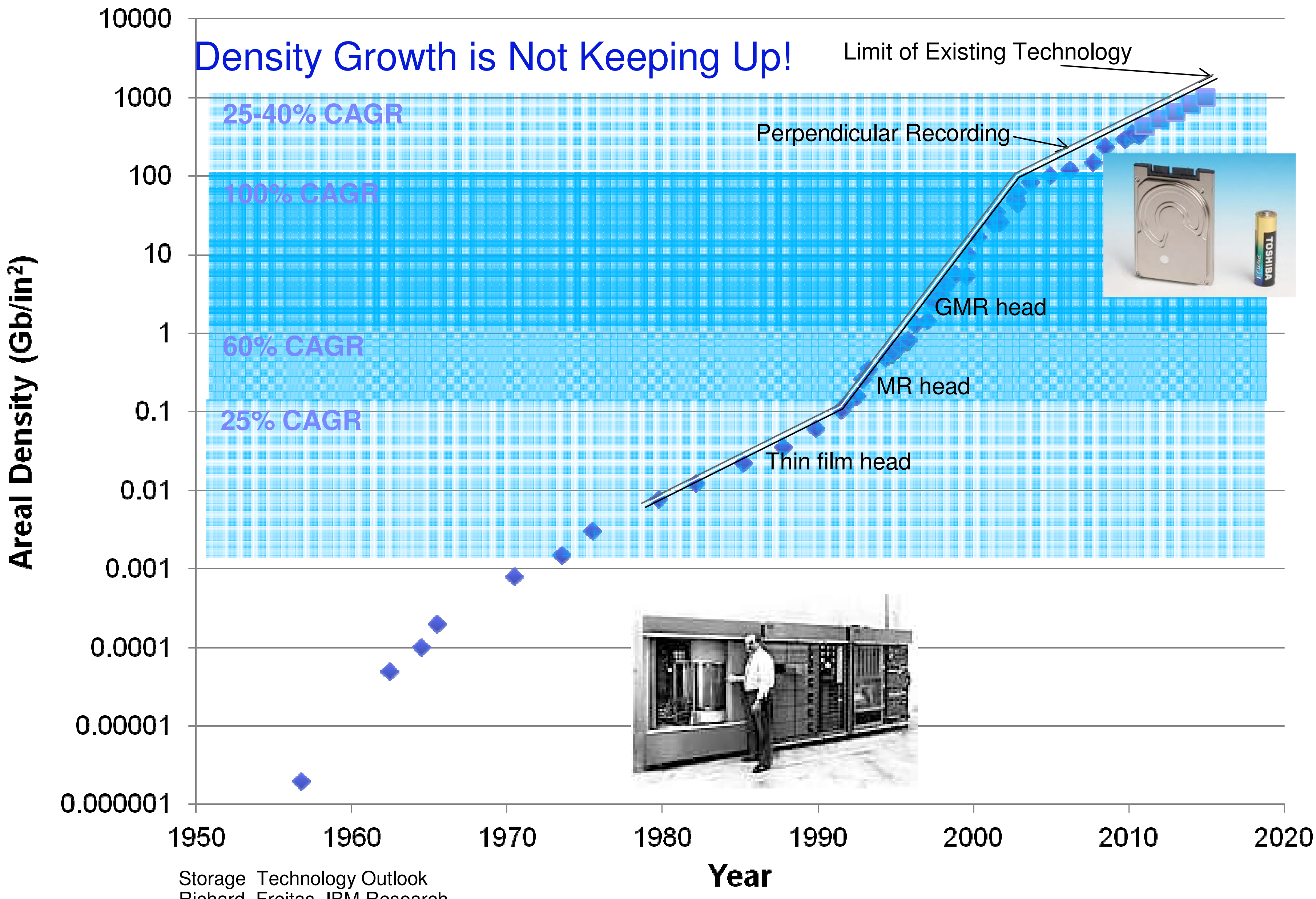
Coping with scale

- ***# of producers/consumers***
- ***Size of objects***
- ***Retention time***

Disk Drive Sizes over the Years

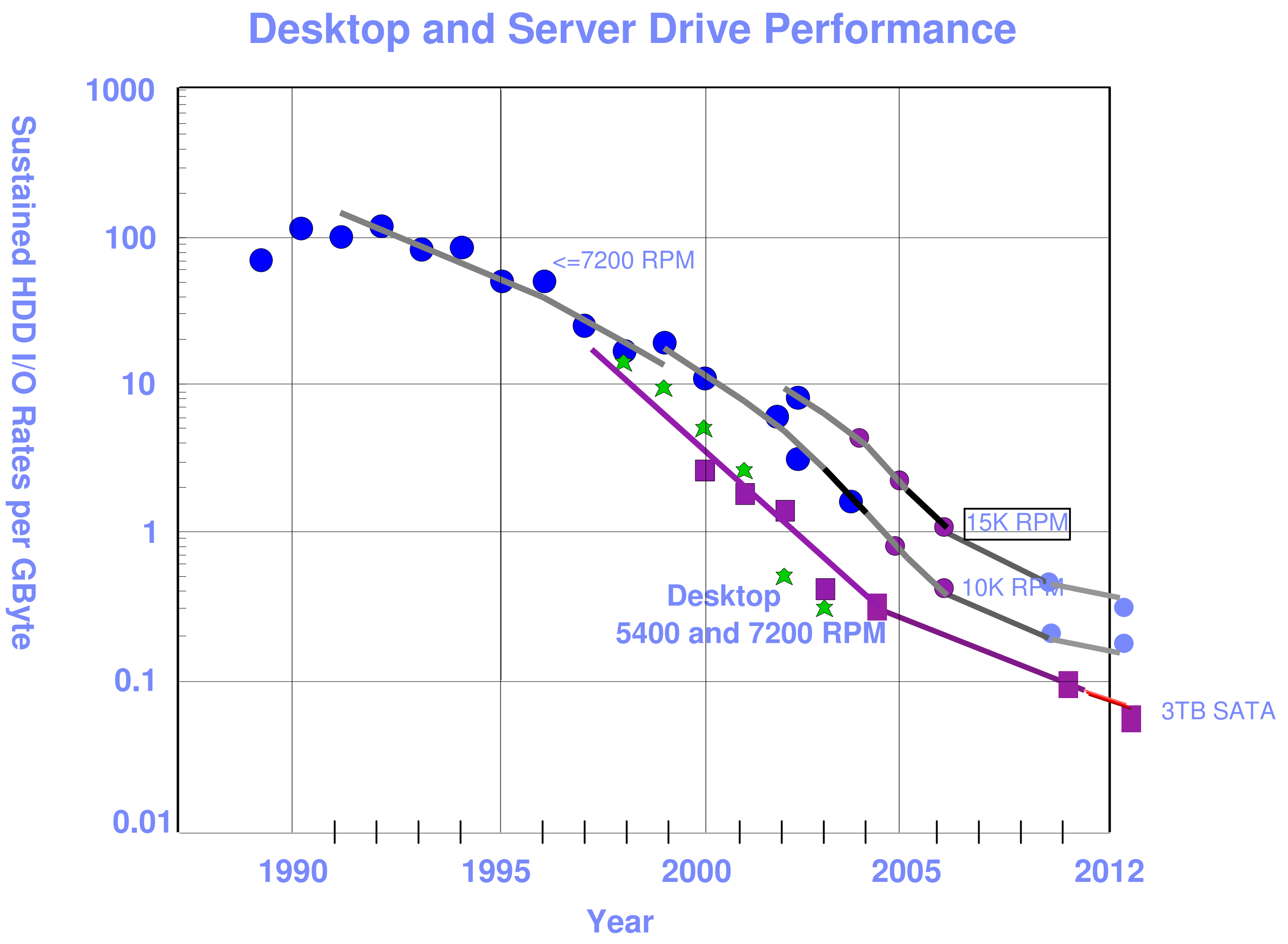


Density Growth is Not Keeping Up!



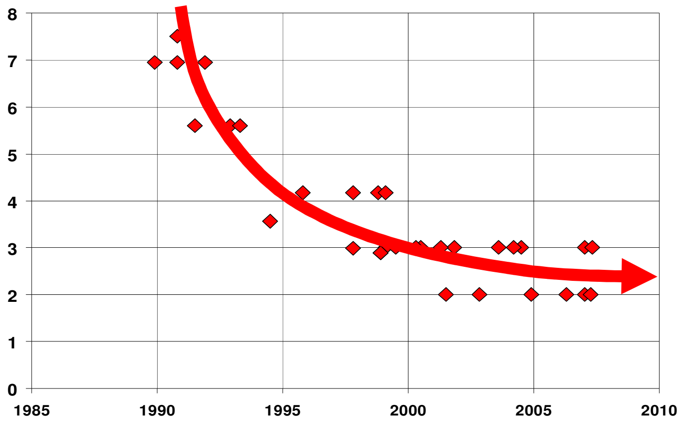
Storage Technology Outlook
Richard Freitas, IBM Research

Disk Performance Falling Behind

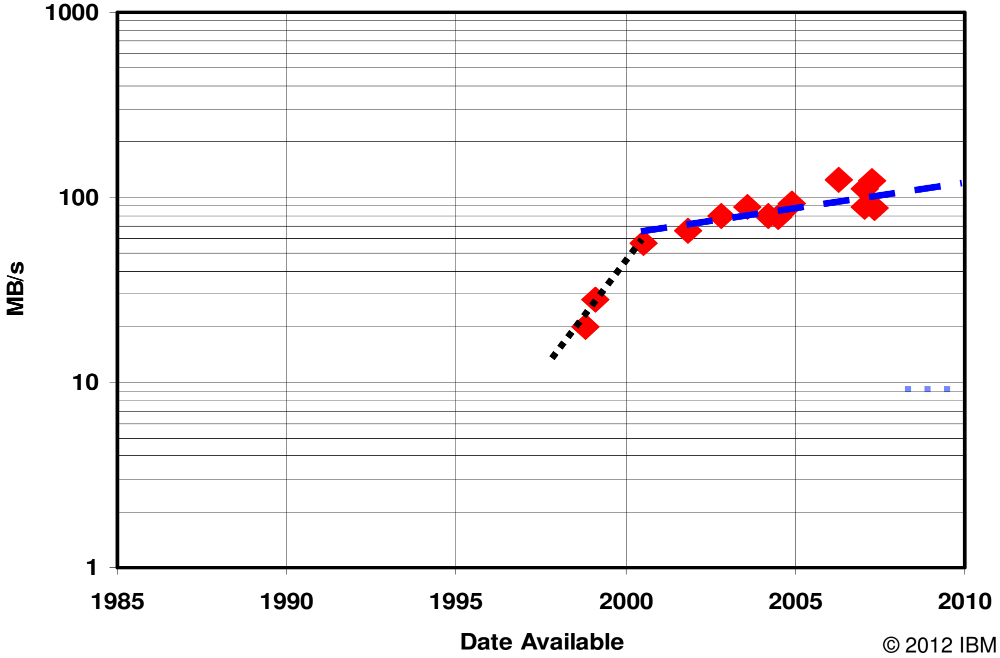


HDD Latency and Disk Transfer Speed

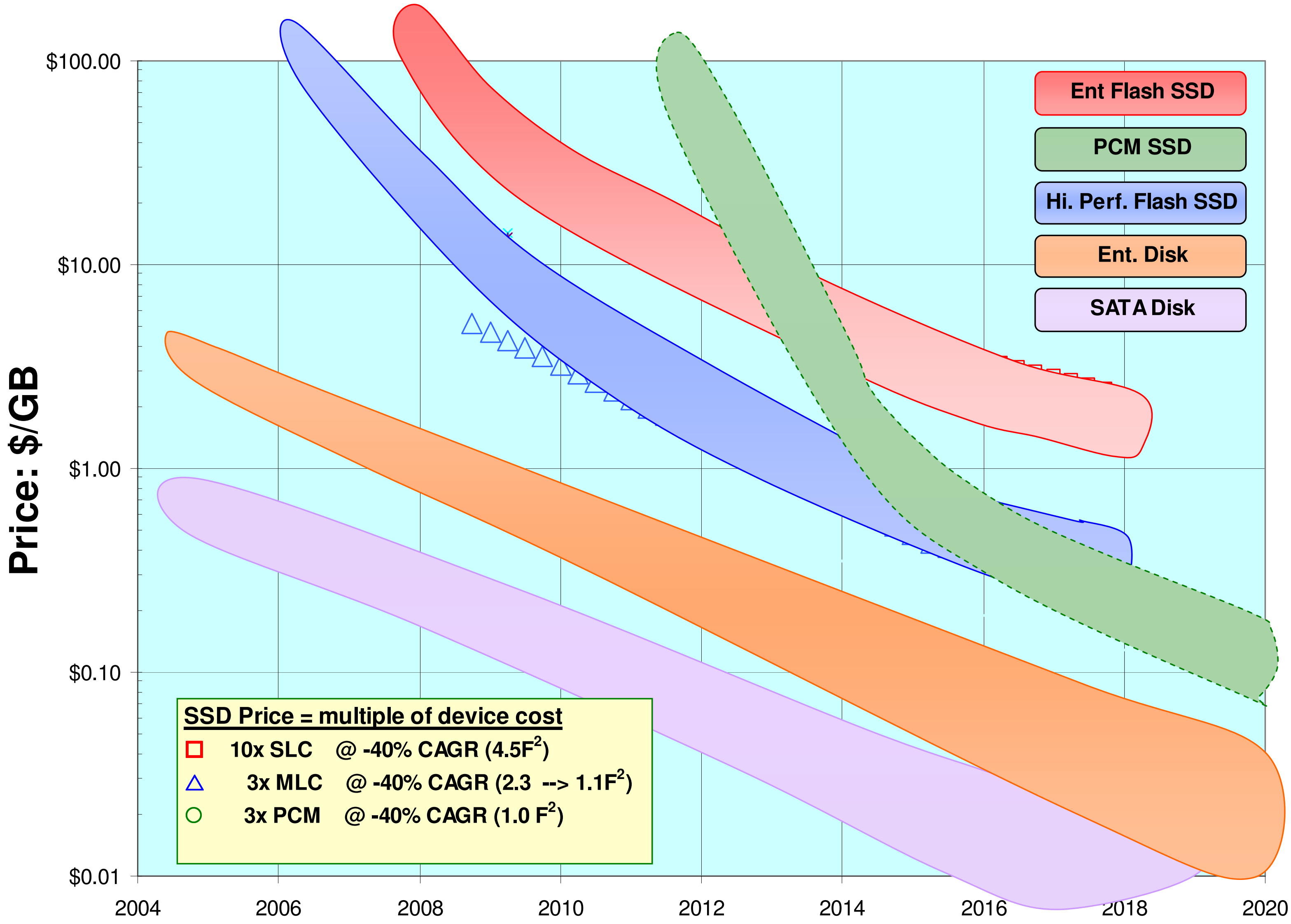
Latency
[ms]



Individual Disk Transfer Speed
[MB/s]

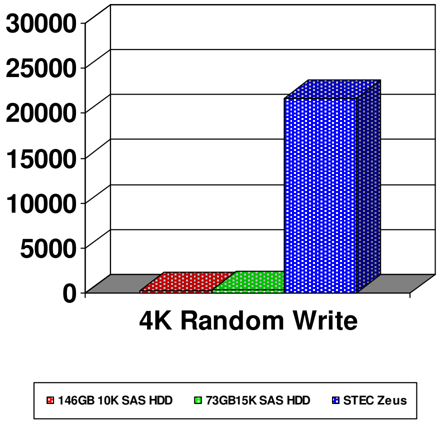


Price Trends: Magnetic disks and Solid State Disks

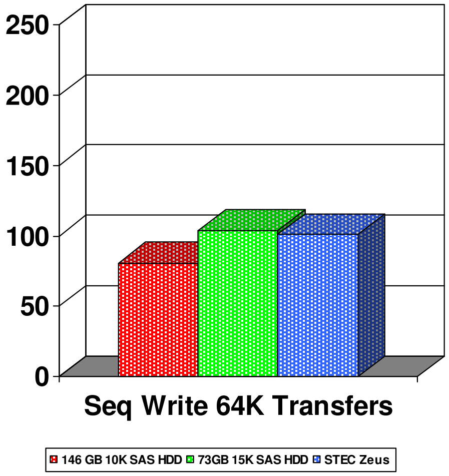


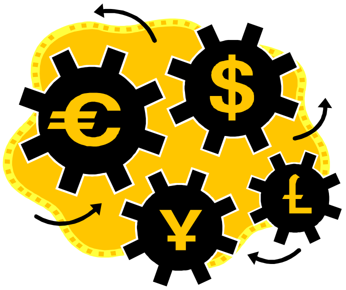
But What About Solid State Disks?

Way Faster on I/O per Second



But on Streaming Data, things are different





At 10 Times the cost per Terabyte!

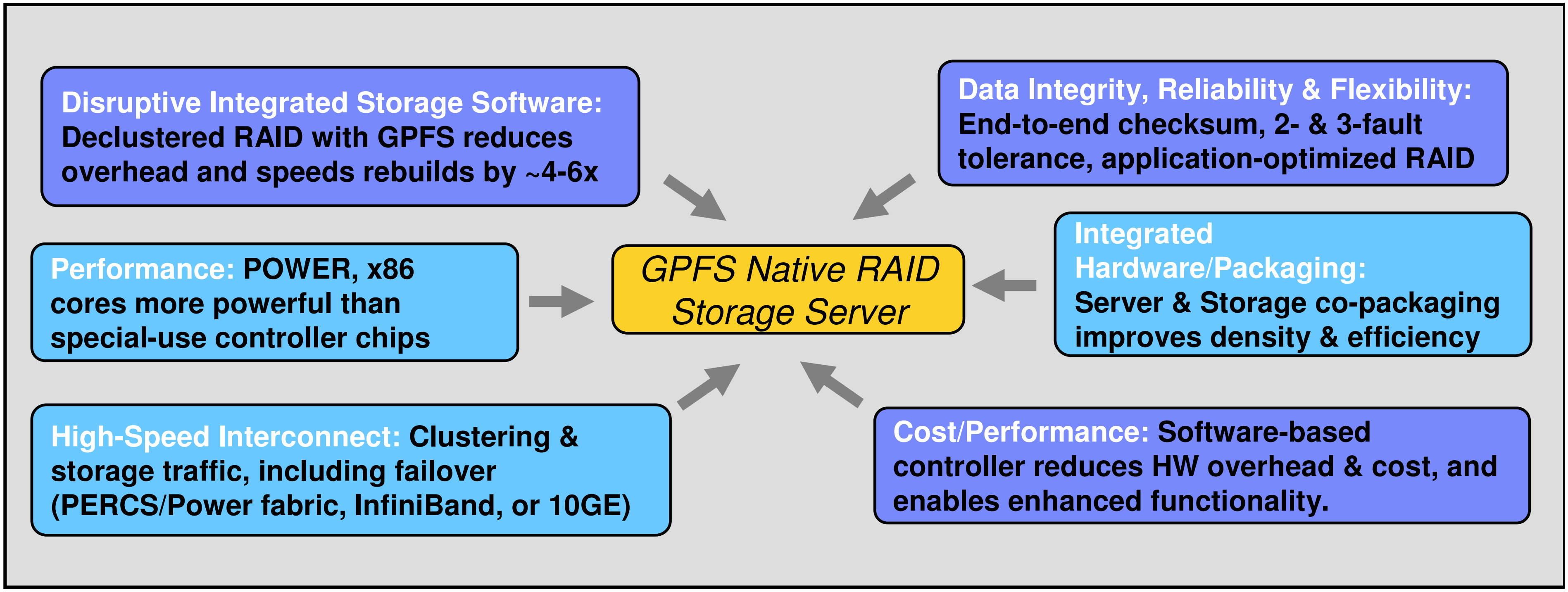
RAID Controller Evolution

- Traditional RAID has Evolved
- At one point RAID 5 was “Good Enough”
  - We now have enough disks that Mean Time to Data Loss is WAY TOO LOW

- Now, we Deploy RAID 6 everywhere
  - Is it good enough?
- Yet, Traditional External RAID controllers remain
  - Expen$ive
  - Slow to Evolve
  - Far, Far away from Processors

Where Do We Go Next?

“Perfect Storm” of Synergetic Innovations



Disruptive Integrated Storage Software: Declustered RAID with GPFS reduces overhead and speeds rebuilds by ~4-6x

Data Integrity, Reliability & Flexibility: End-to-end checksum, 2- & 3-fault tolerance, application-optimized RAID

Performance: POWER, x86 cores more powerful than special-use controller chips

GPFS Native RAID Storage Server

Integrated Hardware/Packaging: Server & Storage co-packaging improves density & efficiency

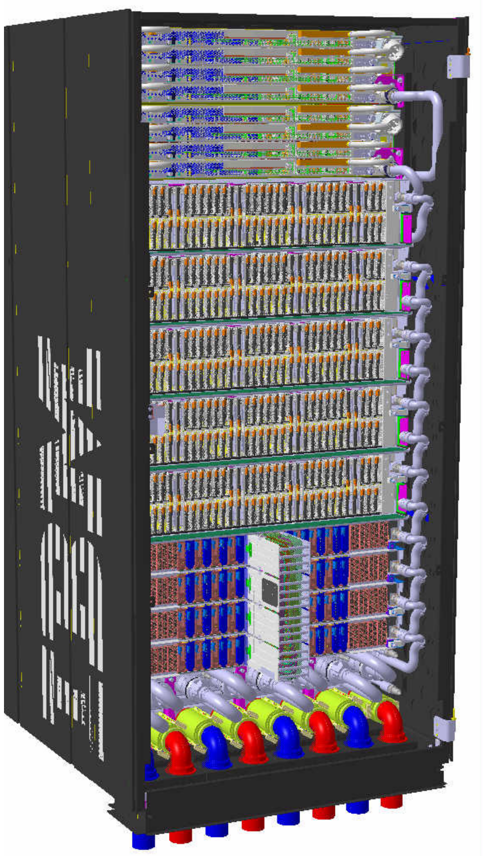
High-Speed Interconnect: Clustering & storage traffic, including failover (PERCS/Power fabric, InfiniBand, or 10GE)

Cost/Performance: Software-based controller reduces HW overhead & cost, and enables enhanced functionality.

Big Data Converging with HPC Technology

Server and Storage Convergence

Shipping NOW from POWER



1 Rack performs a 1TB Hadoop TeraSort in less than 3 minutes!

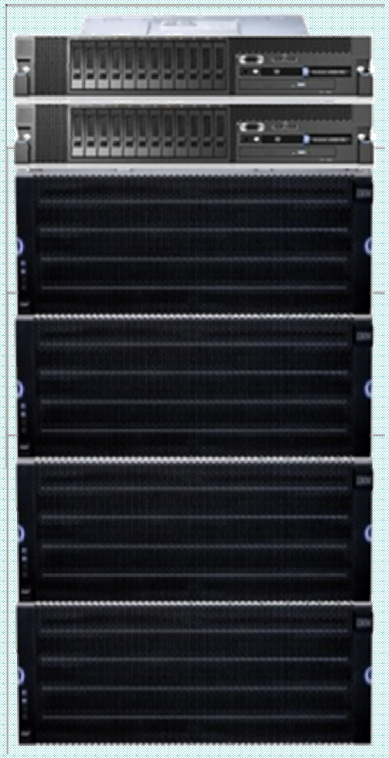
IBM GPFS Native RAID p775:
High-End Storage + Compute Server

- **Based on Power 775 / PERCS Solution**
- **Basic Configuration:**
  - 32 Power7 32-core high bandwidth servers
  - Configurable as GPFS Native RAID storage controllers, compute servers, I/O servers or spares
  - Up to 5 Disk Enclosures per rack
    - 384 Drives and 64 quad-lane SAS ports each
- **Capacity:** 1.1 PB/rack (900 GB SAS HDDs)
- **Bandwidth:** >150 GB/s per rack Read BW
- **Compute Power:** 18 TF + node sparing
- **Interconnect:** IBM high-BW optical PERCS
- **Multi-rack scalable, fully water-cooled**

Introducing IBM System x GPFS Storage Server: Bringing HPC Technology to the Mainstream

Announce 11/13!

- **Better, Sustained Performance**
  - Industry-leading throughput using efficient De-Clustered RAID Techniques
- **Better Value**
  – Leverages System x servers and Commercial JBODS
- **Better Data Security**
  – From the disk platter to the client.
  – Enhanced RAID Protection Technology
- **Affordably Scalable**
  – Start Small and Affordably
  – Scale via incremental additions
  – Add capacity AND bandwidth
- **3 Year Warranty**
  – Manage and budget costs
- **IT-Facility Friendly**
  – Industry-standard 42u 19 inch rack mounts
  – No special height requirements
  – Client Racks are OK!



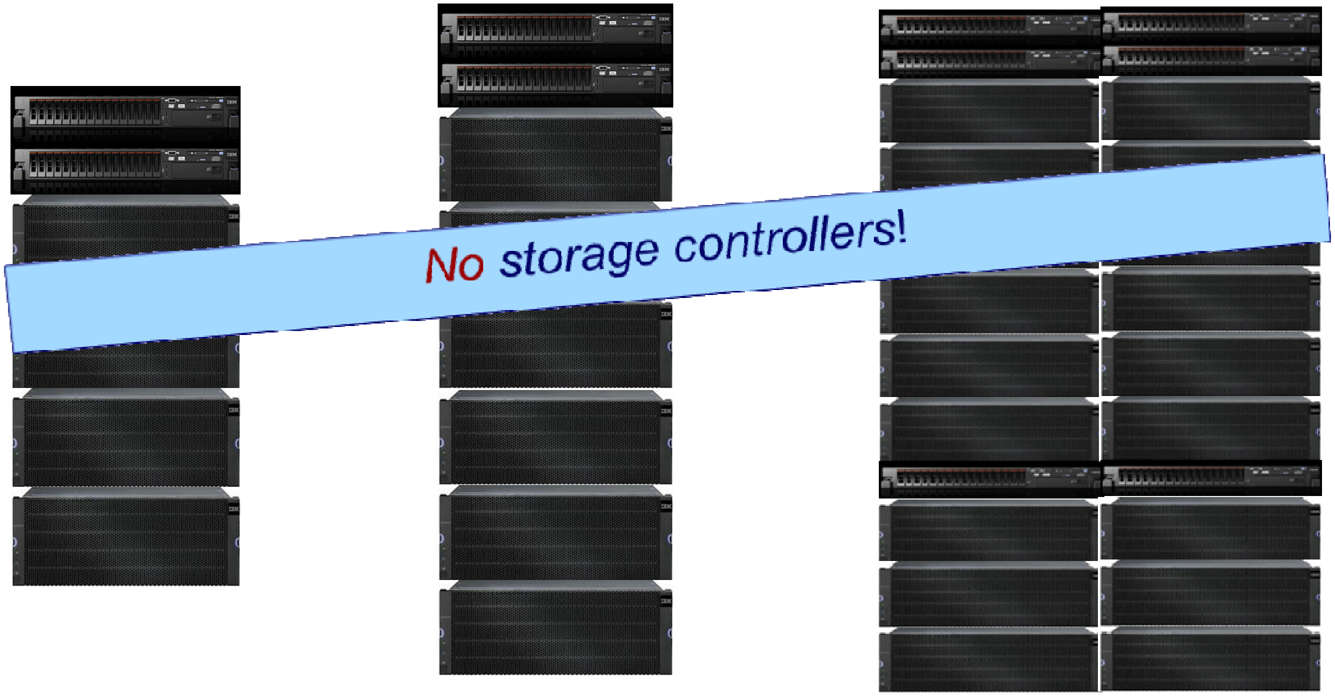
- **And all the Data Management/Life Cycle Capabilities of GPFS – Built in!**

A Scalable Building Block Approach to Storage

Complete Storage Solution
Data Servers, Disk (NL-SAS and SSD), Software, InfiniBand and Ethernet

x3650 M4

“Twin Tailed”
JBOD
Disk Enclosure

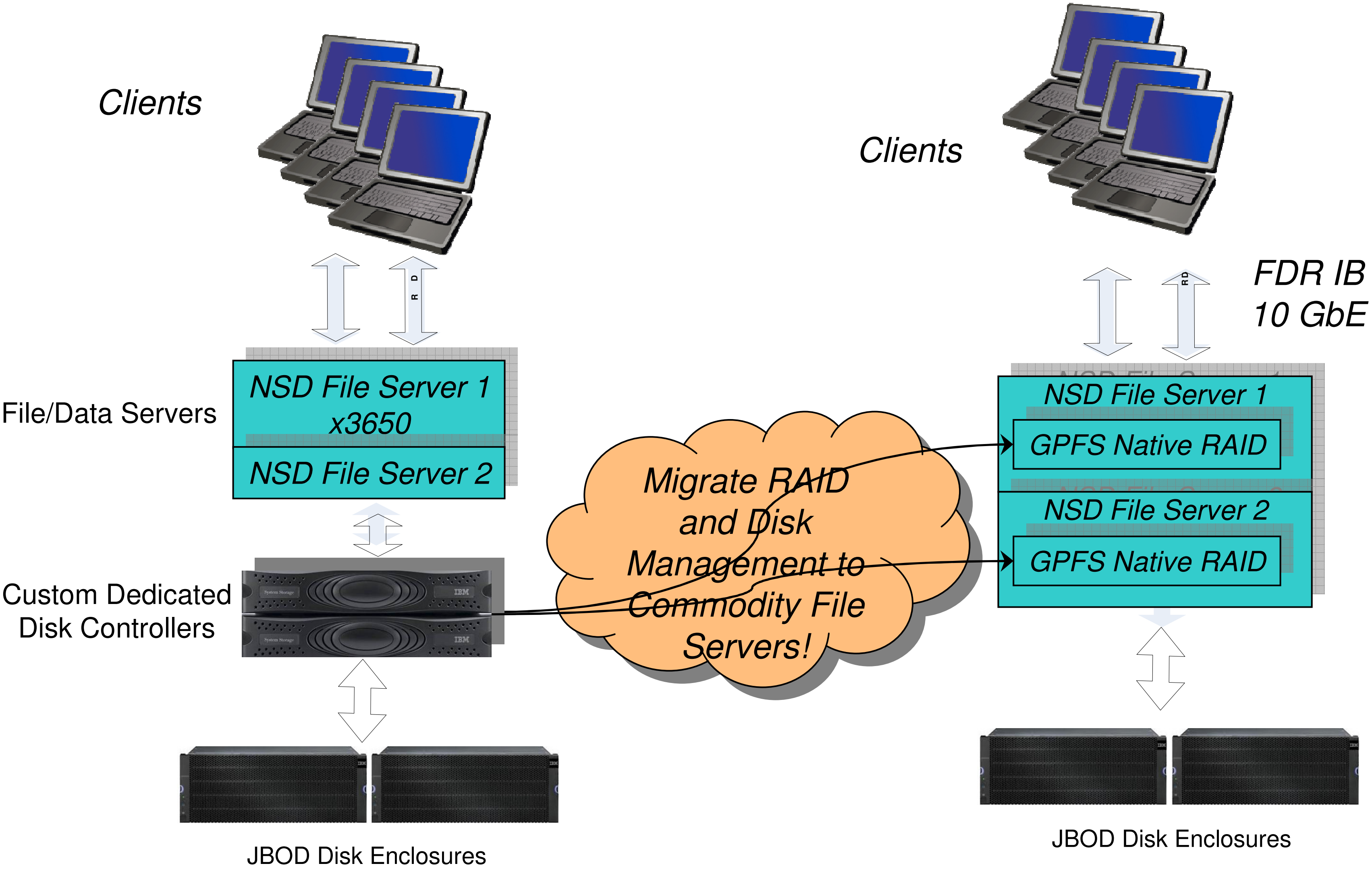


Model 24:
Light and Fast
4 Enclosures, 20U
232 NL-SAS, 6 SSD
10 GB/Sec

Model 26:
HPC Workhorse!
6 Enclosures, 28U
348 NL-SAS, 6 SSD
12 GB/sec

High-Density HPC Option
18 Enclosures
2 - 42U Standard Racks
1044 NL-SAS 18 SSD
36 GB/sec

How We Did It!



Clients

File/Data Servers

NSD File Server 1
x3650

NSD File Server 2

Custom Dedicated
Disk Controllers

JBOD Disk Enclosures

Migrate RAID
and Disk
Management to
Commodity File
Servers!

Clients

FDR IB
10 GbE

NSD File Server 1

GPFS Native RAID

NSD File Server 2

GPFS Native RAID

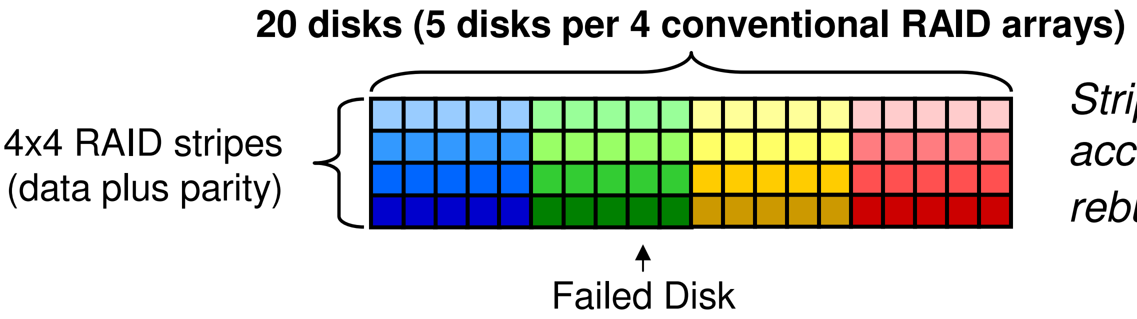
JBOD Disk Enclosures

GPFS Native RAID Feature Detail

- **Declustered RAID**
  - Data and parity stripes are uniformly partitioned and distributed across a disk array.
  - Arbitrary number of disks per array (unconstrained to an integral number of RAID stripe widths)
- **2-fault and 3-fault tolerance**
  - Reed-Solomon parity encoding
  - 2 or 3-fault-tolerant: stripes = 8 data strips + 2 or 3 parity strips
  - 3 or 4-way mirroring
- **End-to-end checksum & dropped write detection**
  - Disk surface to GPFS user/client
  - Detects and corrects off-track and lost/dropped disk writes
- **Asynchronous error diagnosis while affected IOs continue**
  - If media error: verify and restore if possible
  - If path problem: attempt alternate paths
- **Supports live replacement of disks**
  - IO ops continue on for tracks whose disks have been removed during carrier service

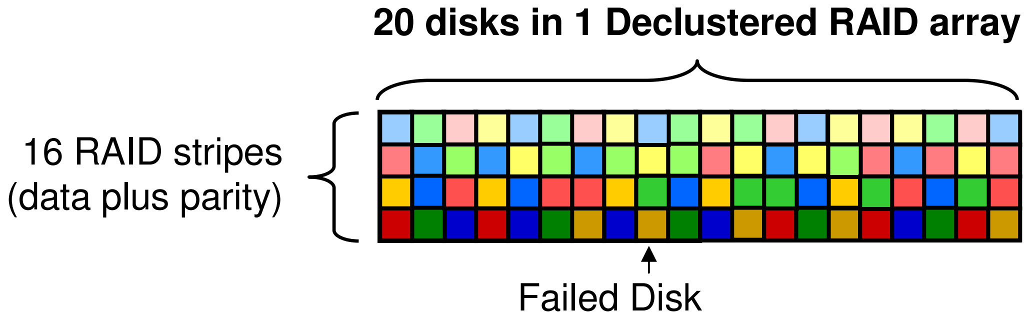
Declustering – Bringing parallel performance to disk maintenance

- Conventional RAID: Narrow data+parity arrays
  - Rebuild can only use the IO capacity of 4 (surviving) disks


20 disks (5 disks per 4 conventional RAID arrays)
4x4 RAID stripes (data plus parity)
Failed Disk

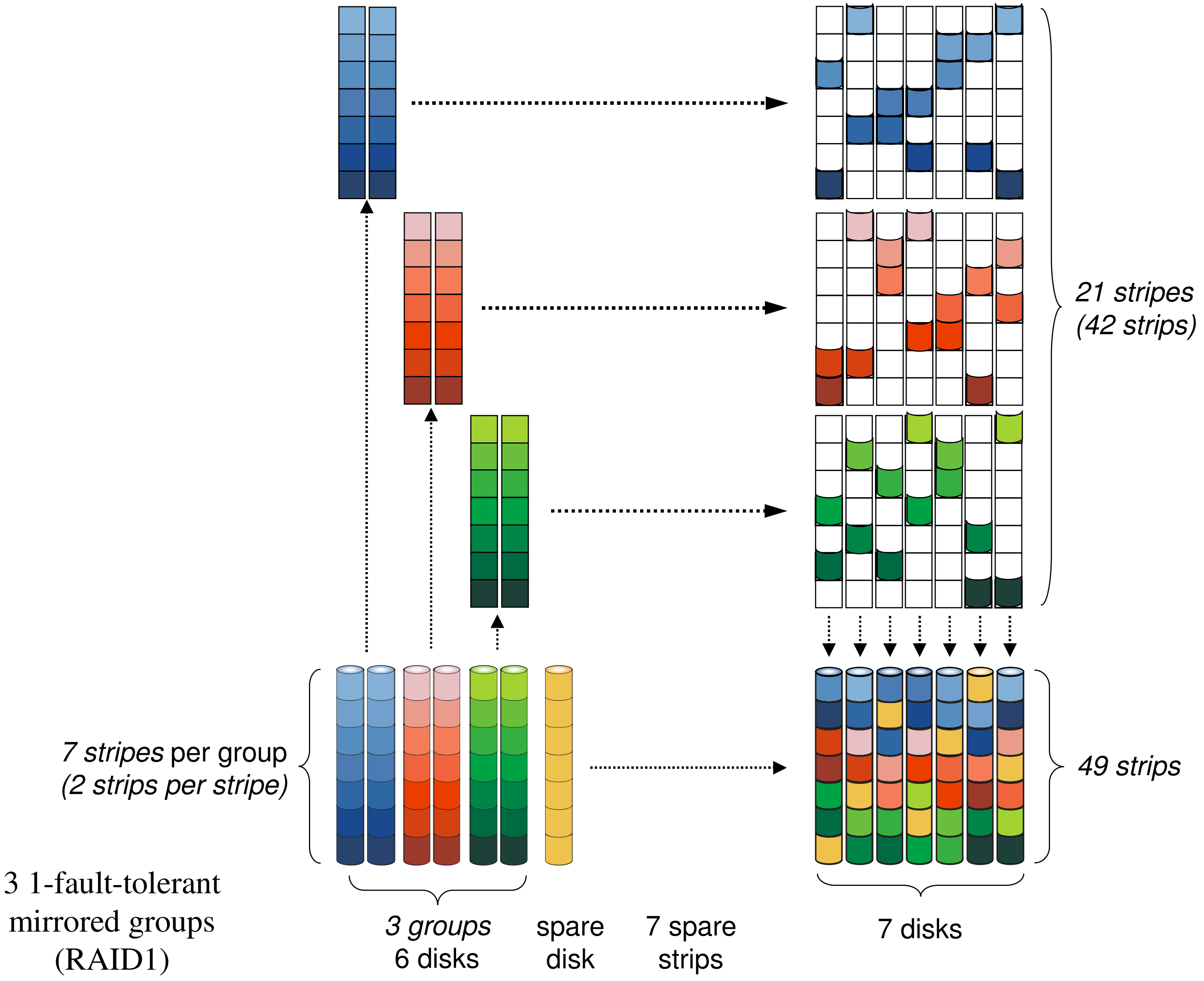
Striping across all arrays, all file accesses are throttled by array 2's rebuild overhead.

- Declustered RAID: Data+parity distributed over all disks
  - Rebuild can use the IO capacity of all 19 (surviving) disks

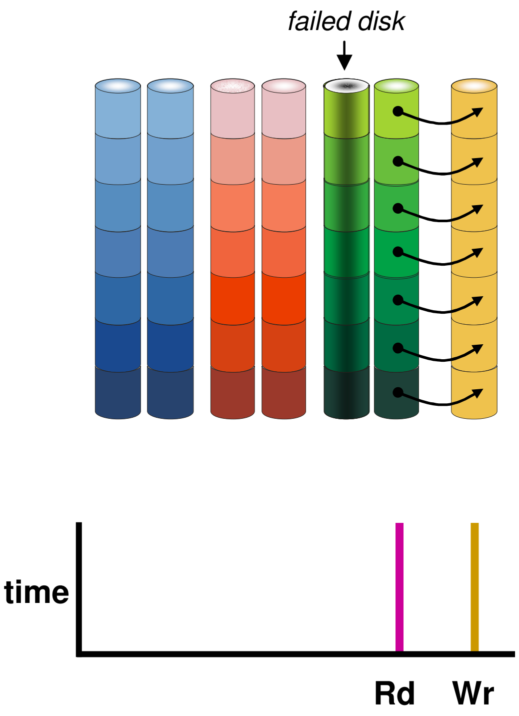

20 disks in 1 Declustered RAID array
16 RAID stripes (data plus parity)
Failed Disk

Load on files accesses are reduced by 4.8x (=19/4) during array rebuild.

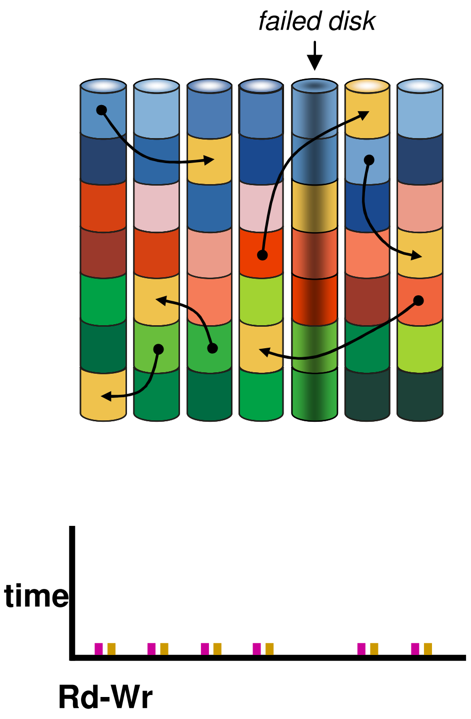
Declustered RAID Example



Rebuild Overhead Reduction Example



Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs



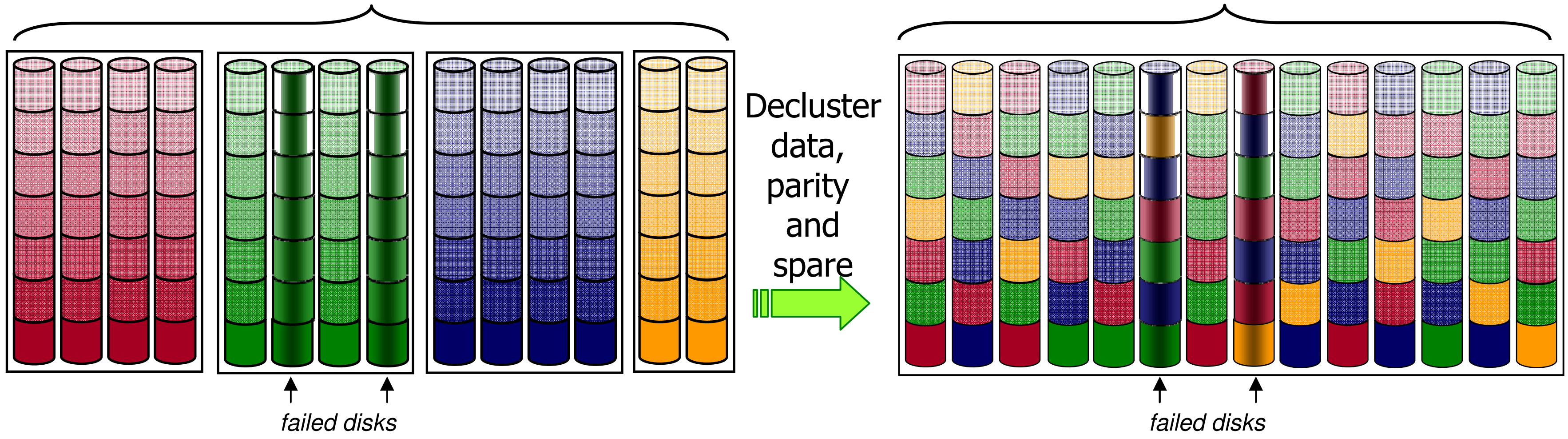
Rebuild activity spread across many disks, less disruption to user programs

Rebuild overhead reduced by 3.5x

Declustered RAID6 Example

14 physical disks / 3 traditional RAID6 arrays / 2 spares

14 physical disks / 1 declustered RAID6 array / 2 spares

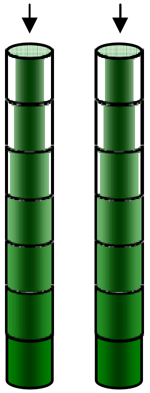


Decluster data, parity and spare

failed disks

failed disks

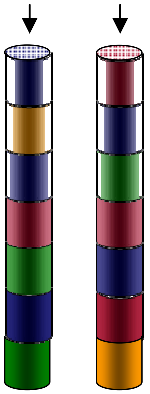
failed disks



Number of faults per stripe		
Red	Green	Blue
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0

Number of stripes with 2 faults = 7

failed disks



Number of faults per stripe		
Red	Green	Blue
1	0	1
0	0	1
0	1	1
2	0	0
0	1	1
1	0	1
0	1	0

Number of stripes with 2 faults = 1

Data Protection Designed for 200K+ Drives!

- Platter-to-Client Protection
  - Multi-level data protection to detect and prevent bad writes and on-disk data loss
  - Data Checksum carried and sent from platter to client server

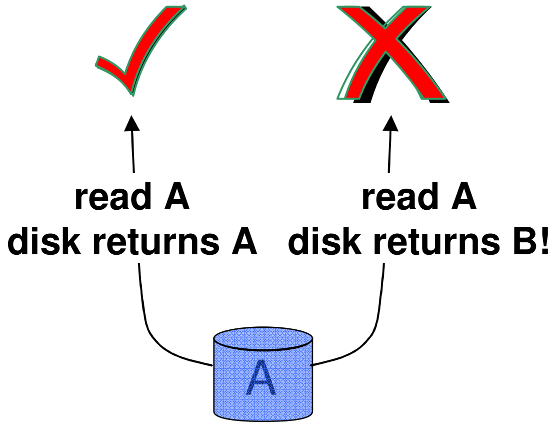
- Integrity Management
  - **Rebuild**
    - Selectively rebuild portions of a disk
    - Restore full redundancy, in priority order, after disk failures
  - **Rebalance**
    - When a failed disk is replaced with a spare disk, redistribute the free space
  - **Scrub**
    - Verify checksum of data and parity/mirror
    - Verify consistency of data and parity/mirror
    - Fix problems found on disk
  - **Opportunistic Scheduling**
    - At full disk speed when no user activity
    - At configurable rate when the system is busy

Non-Intrusive Disk Diagnostics

- **Disk Hospital: Background determination of problems**
  - While a disk is in hospital, GNR non-intrusively and *immediately* returns data to the client utilizing the error correction code.
  - For writes, GNR non-intrusively marks write data and reconstructs it later in the background after problem determination is complete.
- **Advanced fault determination**
  - Statistical reliability and SMART monitoring
  - Neighbor check
  - Media error detection and correction

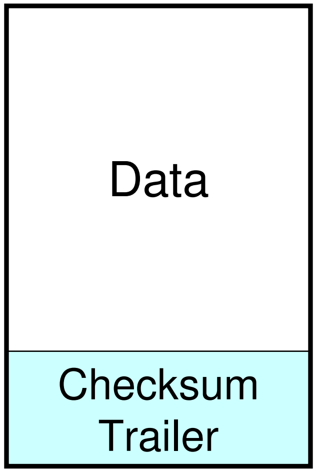
GSS Data Integrity

- Silent data corruption
  - – Caused by disk off-track writes, dropped writes (e.g., disk firmware bugs), or undetected read errors
- Old adage: "No data is better than bad data"
- Proper data integrity checking requires end-to-end checksum *plus dropped write detection.*



GSS – End-to-end Checksums and Version Numbers

- End-to-end checksums
  - Write operation
    - Between user compute node and GNR node
    - From GNR node to disk with version number
  - Read operation
    - From disk to GNR node with version number
    - From IO node to user compute node



Data

Checksum Trailer

- Version numbers in metadata are used to validate checksum trailers for dropped write detection
  - Only a validated checksum can protect against dropped writes

Extending GPFS with LTFS Enterprise Edition

LTFS EE Product Overview

- **What's new:**
  - IBM LTFS EE software enabling IBM tape Libraries to replace tier 2/3 storage

- **Client Value**:
  - Improves the efficiency and cost effectiveness of tier 2/3 storage by using IBM tape libraries in place of disk
  - LTFS EE creates "nearline" access tier 2/3 storage with tape at 1/5[1] the cost of an equivalent disk-based tier 2/3 storage environments
  - Helps reduce storage expense for data that does not need the access performance of primary disk

- 1 based on a list price comparison of a 500TB TS3310 tape library + 1 GPFS license and 1 LTFS EE license compared to a DS3700 hardware and annual maintenance

LTFS EE Product Overview

- LTFS EE Enterprise Software:
  - Based on LTFS LE
  - Supports LTFS LE supported devices
    - TS1140 Enterprise Drive
    - LTO5 or Higher Ultrium drive
  - Integrated functionality with GPFS
  - Supports GPFS Policy based migrations
  - Seamless DMAPI usage
  - Supports multiple LTFS EE nodes for scale-out capacity and I/O
  - Seamless cache controls between LTFS EE Nodes
  - Tape drive performance balancing
  - Multiple node performance balancing
  - Reliability and usability package
  - Read directly from tape functionality
    - No copy back to disk required

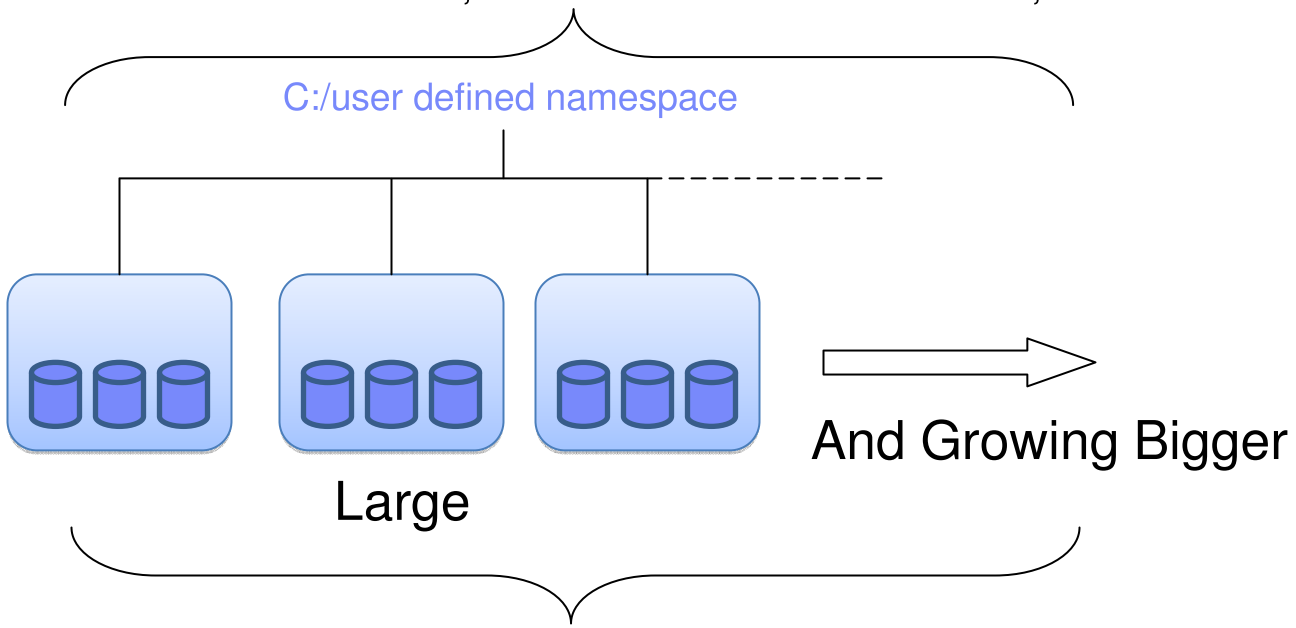
LTFS EE Use Case Categorization

	Description, Industry, Competitors	Client Example
Archive / Data Repository	•Active Archive, Data Repository, Big Data Warehouse •Large namespace, low cost file archive solution •Native GPFS, bundled/gateway platform •Healthcare, M&E, Big Data backend, Surveillance, Gov •EMC Isilon, Netapp, Amazon S3/Glacier, ISV solutions	•Media Archive
Tiered Operational Storage	•Operational NAS storage with LTFS tier •Policy based file placement on disk, LTFS, (SSD tbd) •SONAS, V7KUnified platforms •Multi-Industry IT – Finance, M&E, Scientific, Industry •EMC Isilon w/ DD, Netapp, StorNext, FileTek	•Aggregation of bus. assets •Tiering of stale data
Data Protection / Disaster Recovery	•Hyperscale NAS/Big Data backup - SONAS •Continuous DP, low RTO, policy based restore •Disaster Recovery, policy based restore •Hyperscale NAS, Big Data operational backup •Disk to NDMP (ie. none)	•SONAS @ scale Big Data backup

The Problem – Network Disk Growth…

Operational

- ➢Manageability
- ➢Cost
- ➢Data mix - Rich media & databases, etc
- ➢Uses – active, time senstive access & static, immutable data



C:/user defined namespace

And Growing Bigger

Large

Data Protection

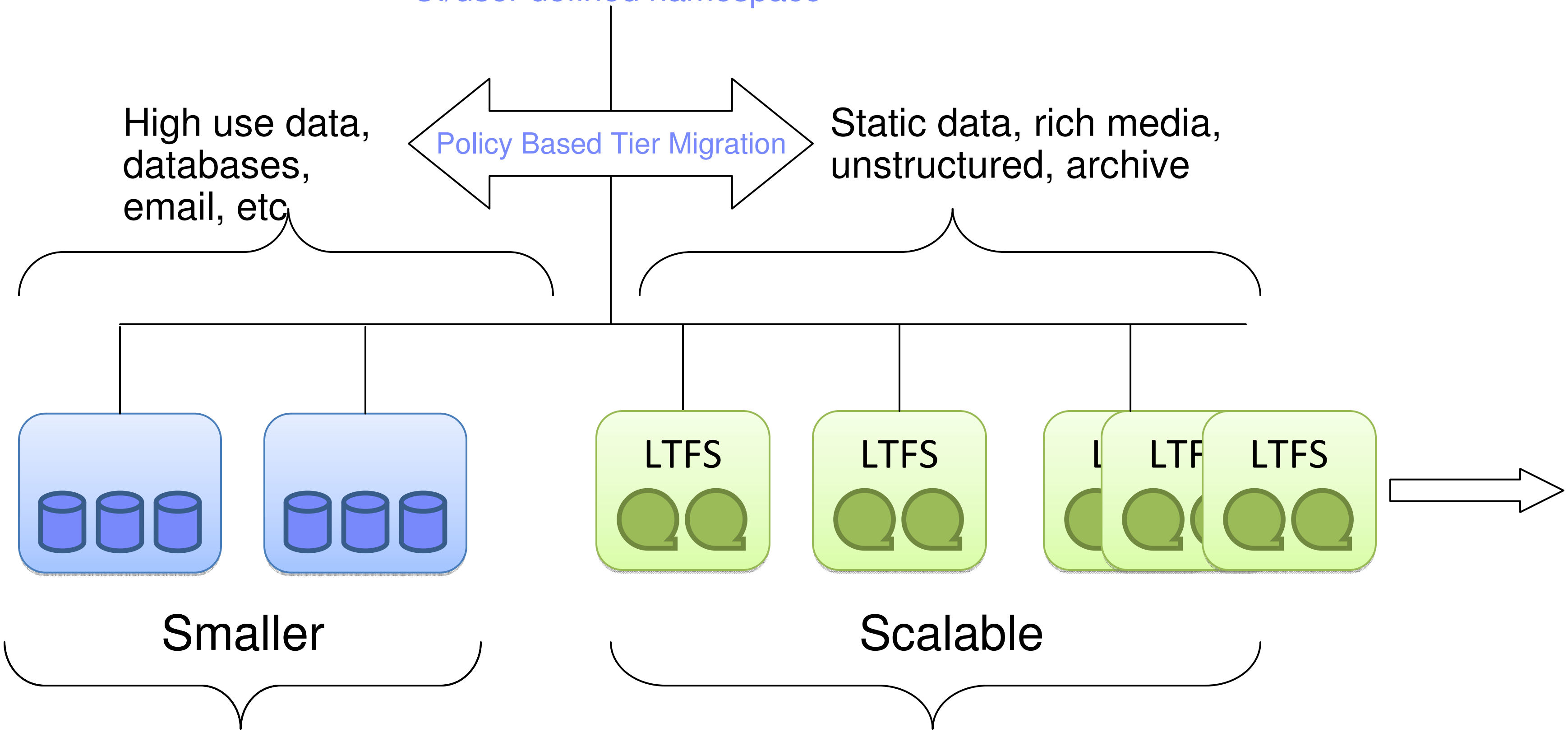
- ➢Difficult to Protect / Backup
  - ➢Cost
  - ➢Backup windows
  - ➢Time to recovery
- ➢Data mix reduces effectiveness of compression/dedupe

The Solution – Tiered Network Storage

Operational

➢Single file system view

C:/user defined namespace



High use data, databases, email, etc

Static data, rich media, unstructured, archive

Smaller

Scalable

Data Protection

➢Easier to protect

- ➢Faster Time to recovery
- ➢Smaller backup footprint

➢Time critical applications/data

➢Lower cost, scalable storage

➢Data types/uses for tape

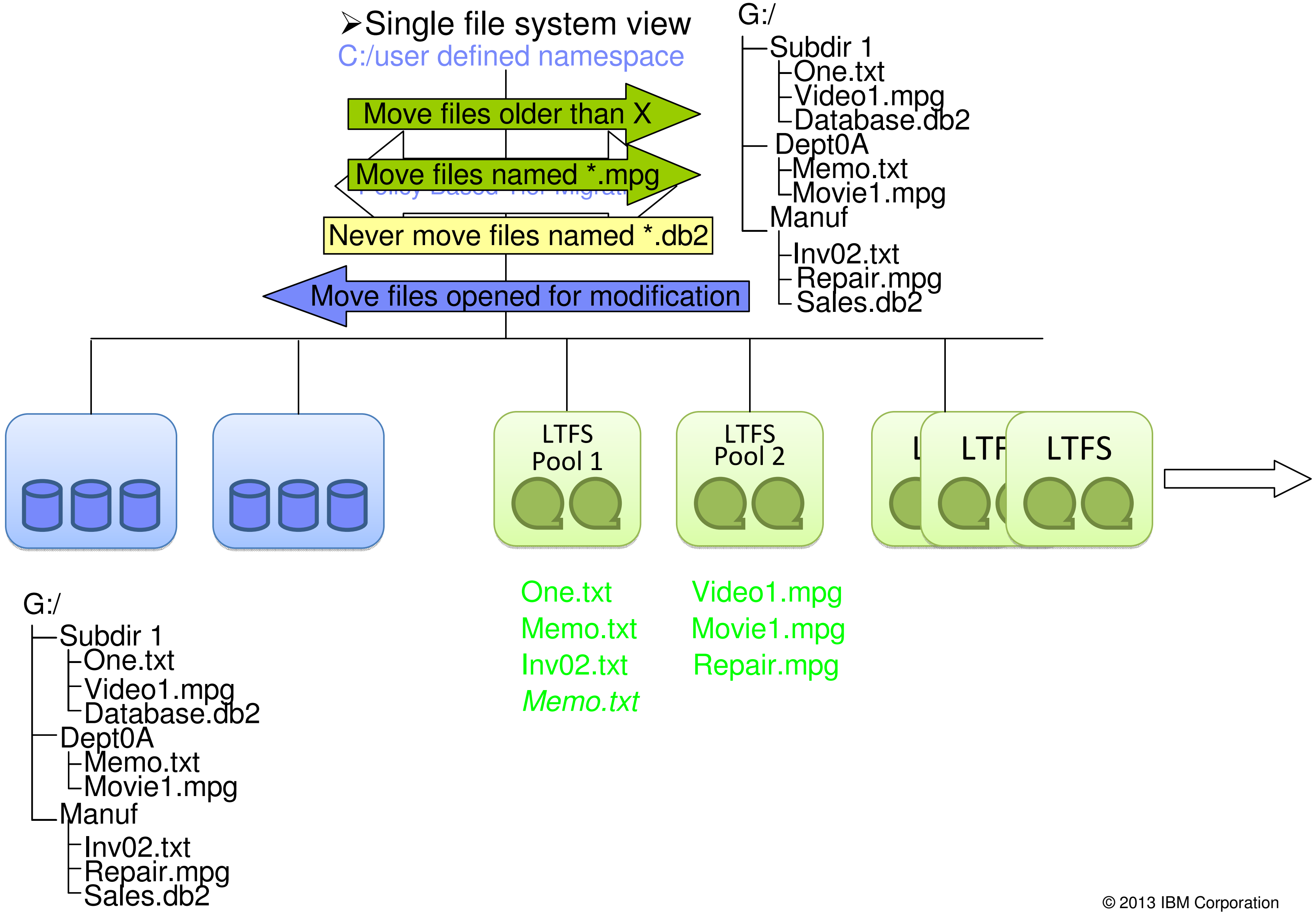
- ➢Static data, rich media, etc.

➢Replication backup strategies

GPFS- LTFS Unified Environment Storage - GLues

File Namespace

File Storage Tiers



➢Single file system view

C:/user defined namespace

```
G:/
├─Subdir 1
│  ├One.txt
│  ├Video1.mpg
│  └Database.db2
├─ Dept0A
│  ├Memo.txt
│  └Movie1.mpg
└─Manuf
   ├Inv02.txt
   ├Repair.mpg
   └Sales.db2
```

One.txt
Memo.txt
Inv02.txt
Memo.txt

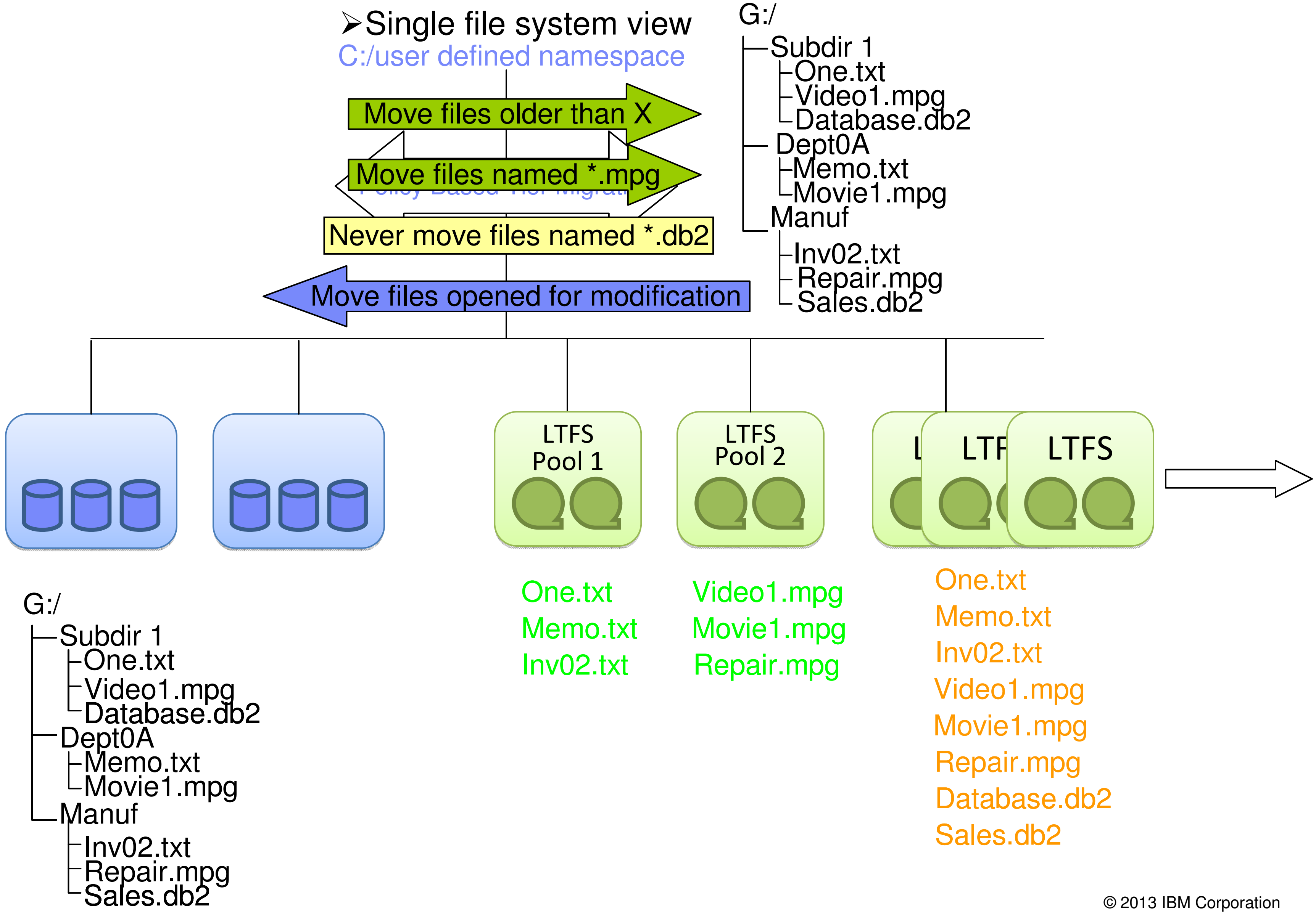
Video1.mpg
Movie1.mpg
Repair.mpg

```
G:/
├─Subdir 1
│  ├One.txt
│  ├Video1.mpg
│  └Database.db2
├─Dept0A
│  ├Memo.txt
│  └Movie1.mpg
└─Manuf
   ├Inv02.txt
   ├Repair.mpg
   └Sales.db2
```

Glues – Data Protection and Backup

File Namespace

File Storage Tiers



- Single file system view

C:/user defined namespace

Move files older than X

Move files named *.mpg

Never move files named *.db2

Move files opened for modification

```
G:/
├─Subdir 1
│  ├One.txt
│  ├Video1.mpg
│  └Database.db2
├─ Dept0A
│  ├Memo.txt
│  └Movie1.mpg
└─Manuf
   ├Inv02.txt
   ├Repair.mpg
   └Sales.db2
```

LTFS Pool 1

LTFS Pool 2

LTFS

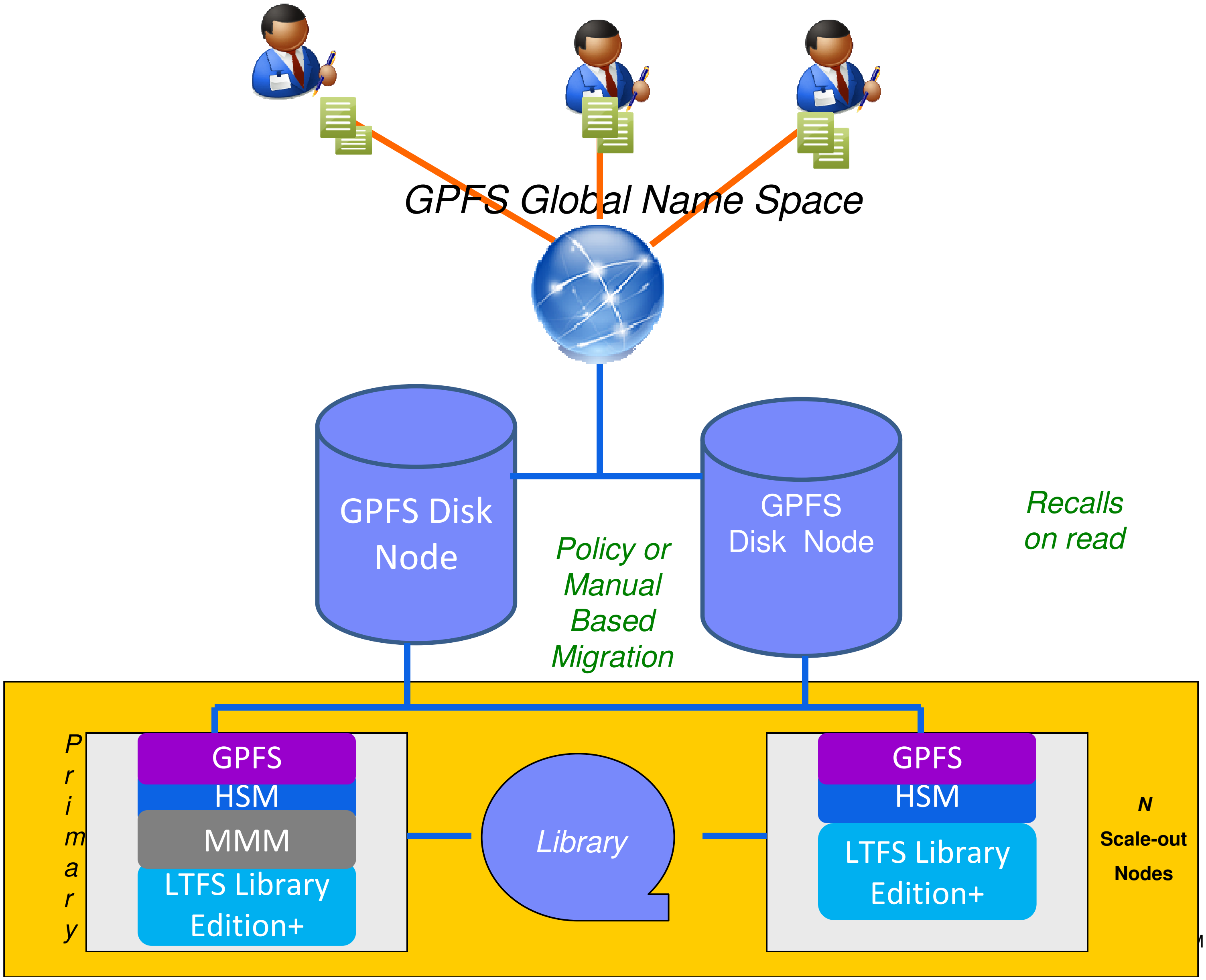
```
G:/
├─Subdir 1
│  ├One.txt
│  ├Video1.mpg
│  └Database.db2
├─Dept0A
│  ├Memo.txt
│  └Movie1.mpg
└─Manuf
   ├Inv02.txt
   ├Repair.mpg
   └Sales.db2
```

One.txt
Memo.txt
Inv02.txt

Video1.mpg
Movie1.mpg
Repair.mpg

One.txt
Memo.txt
Inv02.txt
Video1.mpg
Movie1.mpg
Repair.mpg
Database.db2
Sales.db2

Sample LTFS EE Usage Configuration



GPFS Global Name Space

GPFS Disk Node

GPFS Disk Node

Policy or Manual Based Migration

Recalls on read

Primary

GPFS

HSM

MMM

LTFS Library Edition+

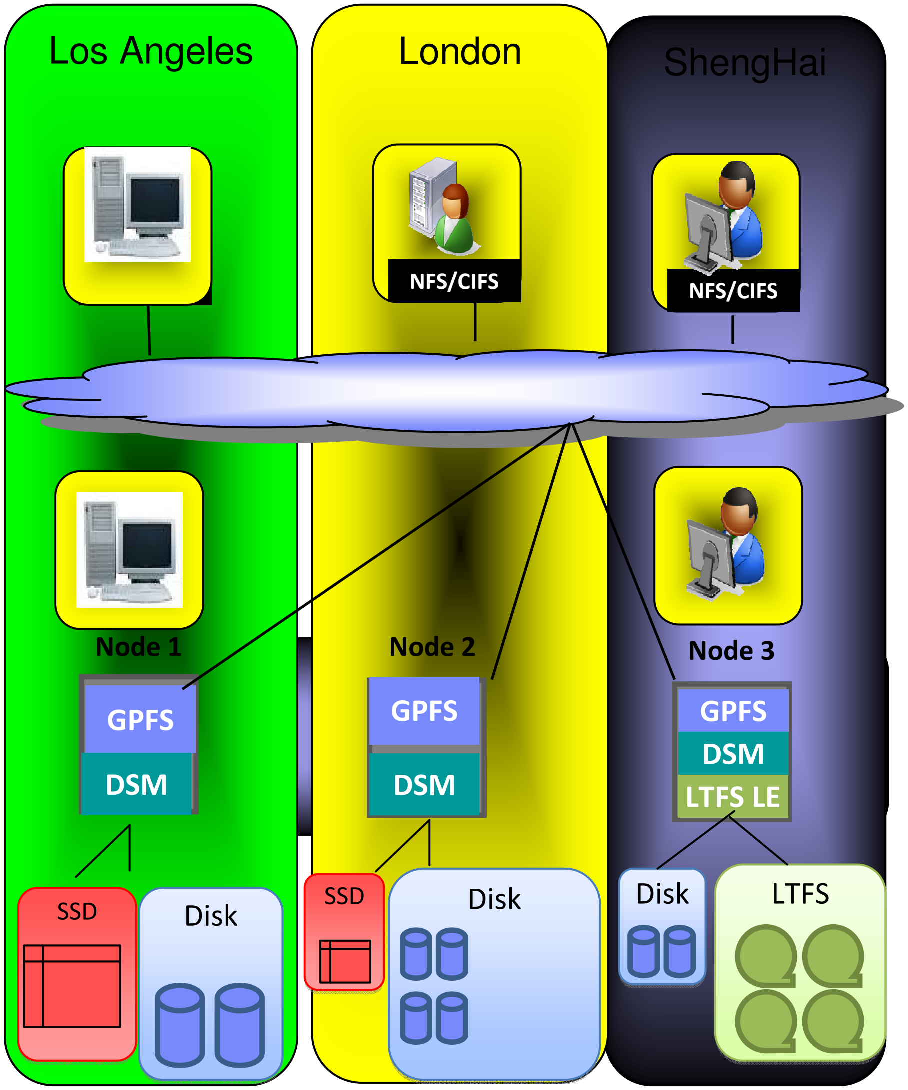
Library

GPFS

HSM

LTFS Library Edition+

N Scale-out Nodes



Smarter Storage

- Distributed Data
- Namespace file view
- Load balancing
- Policy migration
- Storage Distribution
- Reduction of cost for storage
- Data monetization

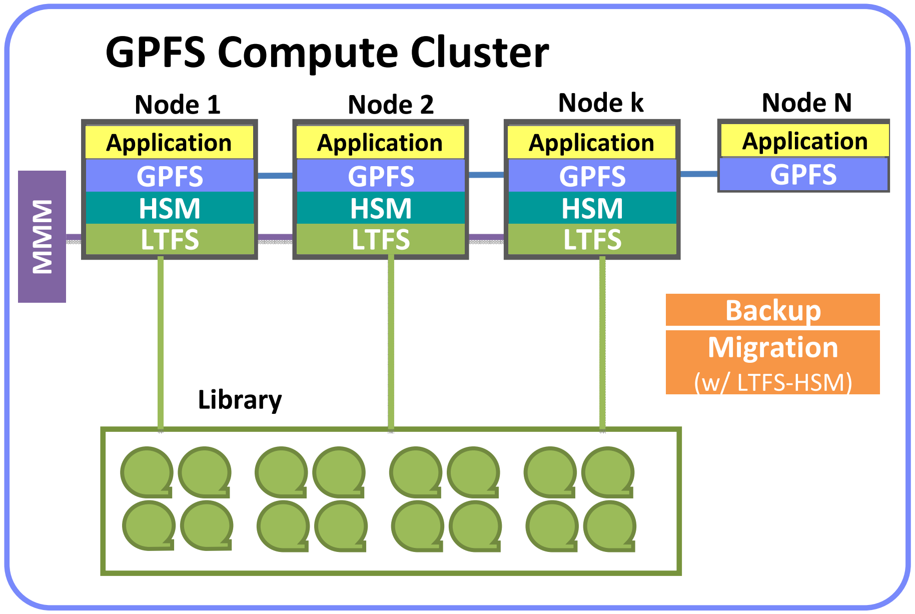
Product Usage and Customer Value

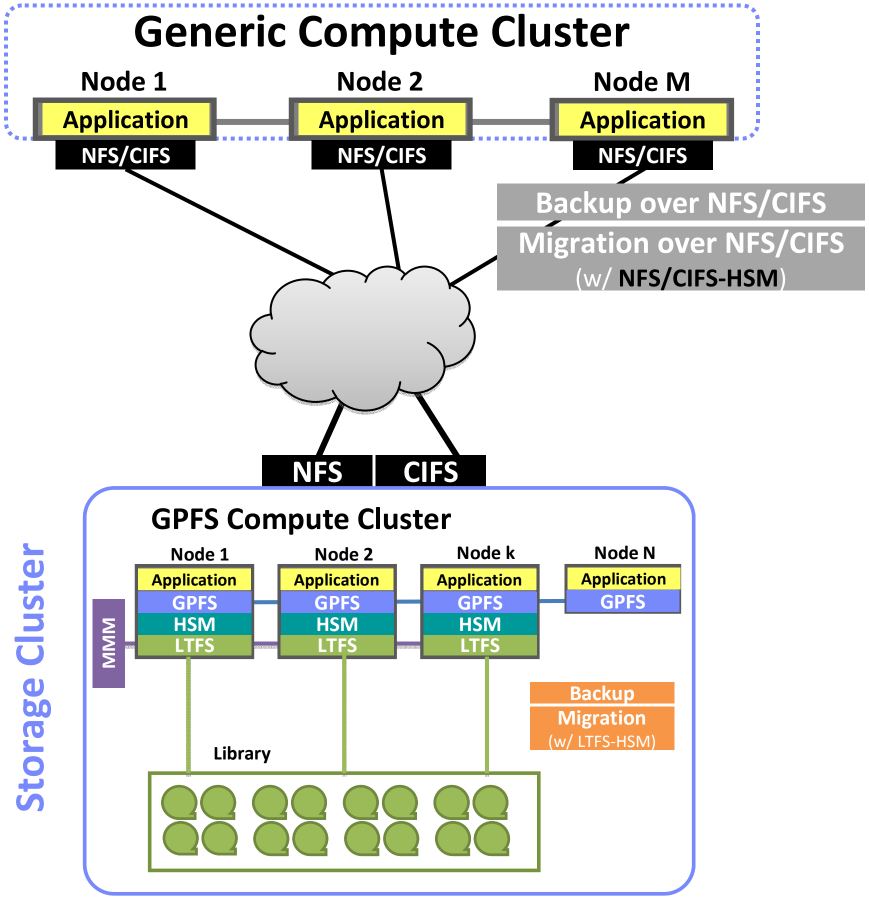
Compute Cluster using GPFS as the cluster filesystem

- GPFS disk is the main data store
- Large and/or inactive data migrate to tape
- Integrated backup & migration functions
- Suitable for HPC Big data and M&E use cases

Integrated GPFS/LTFS used as data repository

- Separate storage and computation
- Integrated GPFS/LTFS cluster used as a “Big Data Depository” and is the main data store
- All data stored on tape, GPFS used as a large cache / staging area
- Integrated backup & migration functions

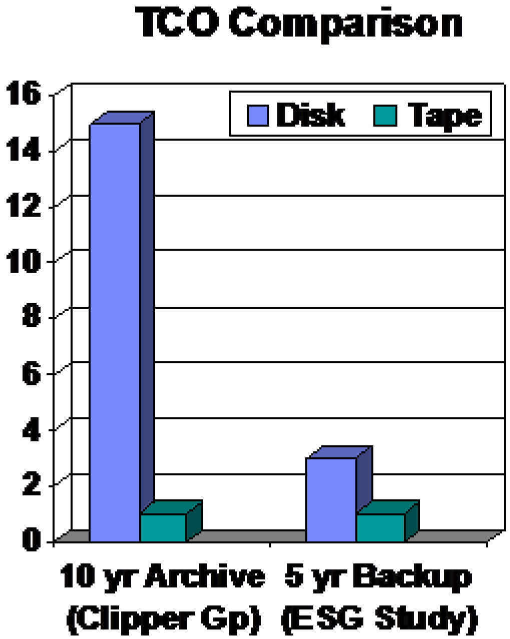




Tape saves more for Archive – Independent Studies

- Disk
  - –Storage Cost 10¢ /GB/month
  - –1 PetaByte $100K/Month
  - –5 Years $6.6 Million

- LTFS Tape
  - –Capacity Cost 0.77¢ /GB/month
  - –1 PetaByte $7.7K/Month
  - –5 Years $462 Thousand



How competitive is tape?

Case Study: “Big Bank Inc.” Financial Archive

- **Assumptions**
  - 3 year retention before technology refresh
  - 3PB Near-line Long Term retention
  - Continuous long term I/O of 120TB per day
  - Software layer managing the disk created by the customer
  - The SAN is a wash needed for either
  - That “Big Bank Inc.” is willing to try to create their own rack system for Disk

Tape cost $189/TB

270 JC4 carts per PB = $44/TB, 1 time cost
12 Drives for 120TB/Day I/O = $46/TB
6 Servers for I/O direct connect = $20/TB
LTFS EE 6 server licenses = $34/TB
Library 80K first PB, 50K ever 2.5PB after = $42/TB)

44+46+20+42+34 = $186/TB

Power

889 watts * 8760 hours * $1.12/KwH
= $8721/year operating cost
= $3/TB operating per year

***For every PB beyond 3PB remove an average of $15/TB

Disk cost $400/TB Build with Controller/s

6 Servers for I/O direct connect = $20/TB
In House build Back Blaze 3.0 (assume 18 month out price for 4TB disk drives) = $55/TB
Dual power supply, RAID6
40% disk fall out of 3 years = $22/TB
Logical volume manager = $70/TB

20+55+22+70 = $167/TB

Power

600 watts *8760 * 16 * $1.12/KwH
= $94187
= $31/TB operating per year

****For every PB beyond 3 PB add $3/TB

RAID based controller = $200/TB (fixed cost for this analysis)

Disclaimers and Trademarks

- The performance data contained herein was obtained in a controlled environment based on the use of specific data. Actual results that may be obtained in other operating environments may vary significantly. These values do not constitute a guarantee of performance.

- Product data is accurate as of initial publication and is subject to change without notice.

- No part of this presentation may be reproduced or transmitted in any form without written permission from IBM Corporation.

- References in this document to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM program product in this document is not intended to state or imply that only IBM's program product may be used. Any functionally equivalent program may be used instead.

- The information provided in this document has not been submitted to any formal IBM test and is distributed "As Is" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into their operating environment.

-

- While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

- The following terms are trademarks or registered trademarks of the IBM Corporation in either the United States, other countries or both:

- IBM, S/390, ES/3090, ES/9000, AS/400, RS/6000, MVS/ESA, OS/390, VM/ESA, VSE, TPF, OS/2, OS/400, AIX, DFSMS/MVS, DFSMS/VM, ADSTAR Distributed Storage Manager, DFSMSdfp, DFSMSdss, DFSMShsm, DFSMSrmm, FICON, ESCON, Magstar, Seascape, TotalStorage, Tivoli, AIX, OS/400, TSM, BRMS

- Other company, product, and service names mentioned may be trademarks or registered trademarks of their respective companies.

- Linear Tape-Open, LTO, LTO Logo, Ultrium and Ultrium Logo are trademarks in the United States and/or other countries of Hewlett-Packard, IBM, Seagate.

References

- http://public.dhe.ibm.com/common/ssi/ecm/en/pos03096usen/POS03096USEN.PDF
- http://www-03.ibm.com/systems/resources/IBM-GPFS-Use-Cases-April_2011_Final.pdf
- http://www.ibm.com/developerworks/wikis/display/hpccentral/General+Parallel+File+System+%28GPFS%29
- http://www-03.ibm.com/systems/resources/introduction-to-gpfs-3-5.pdf

smarter together
IBM