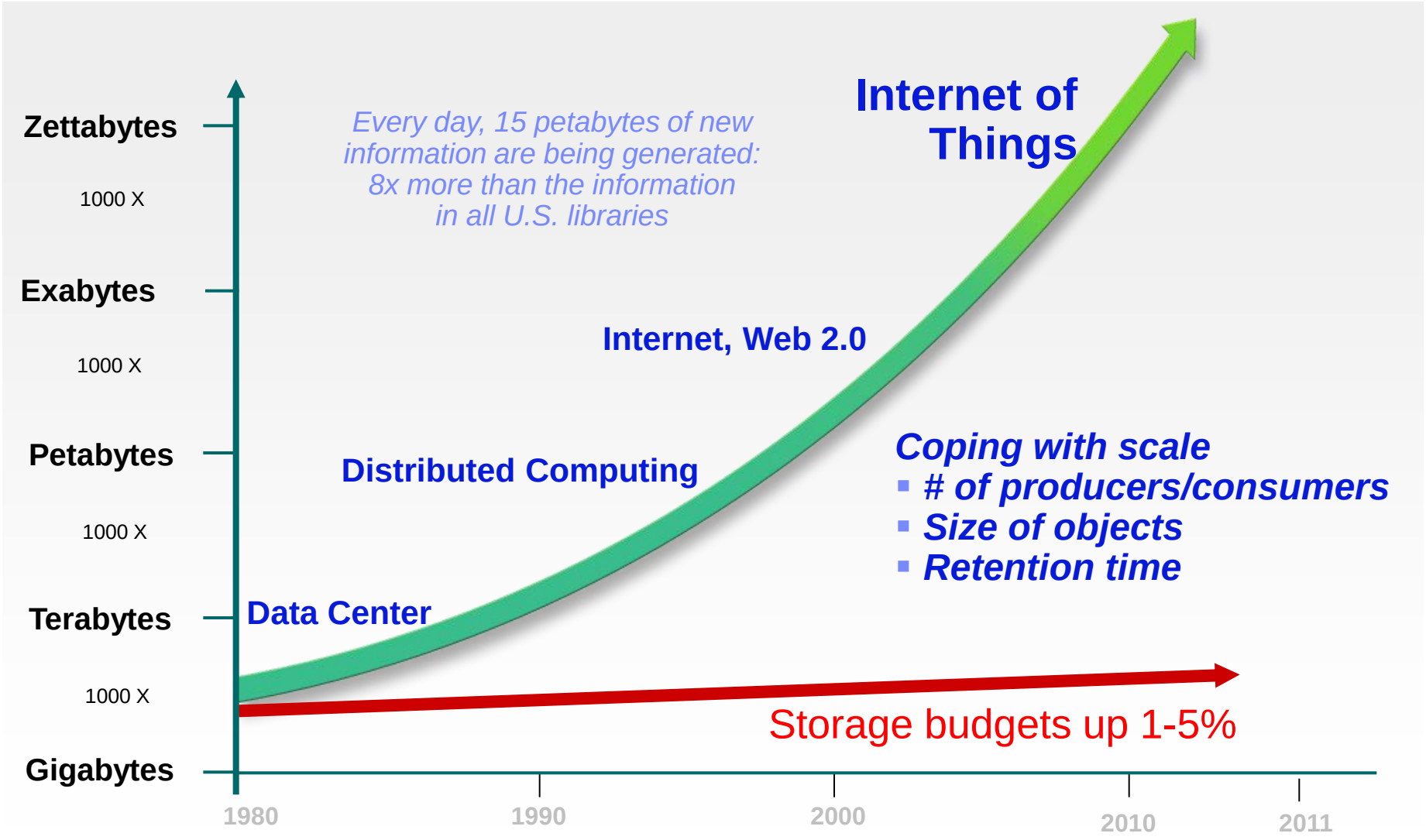
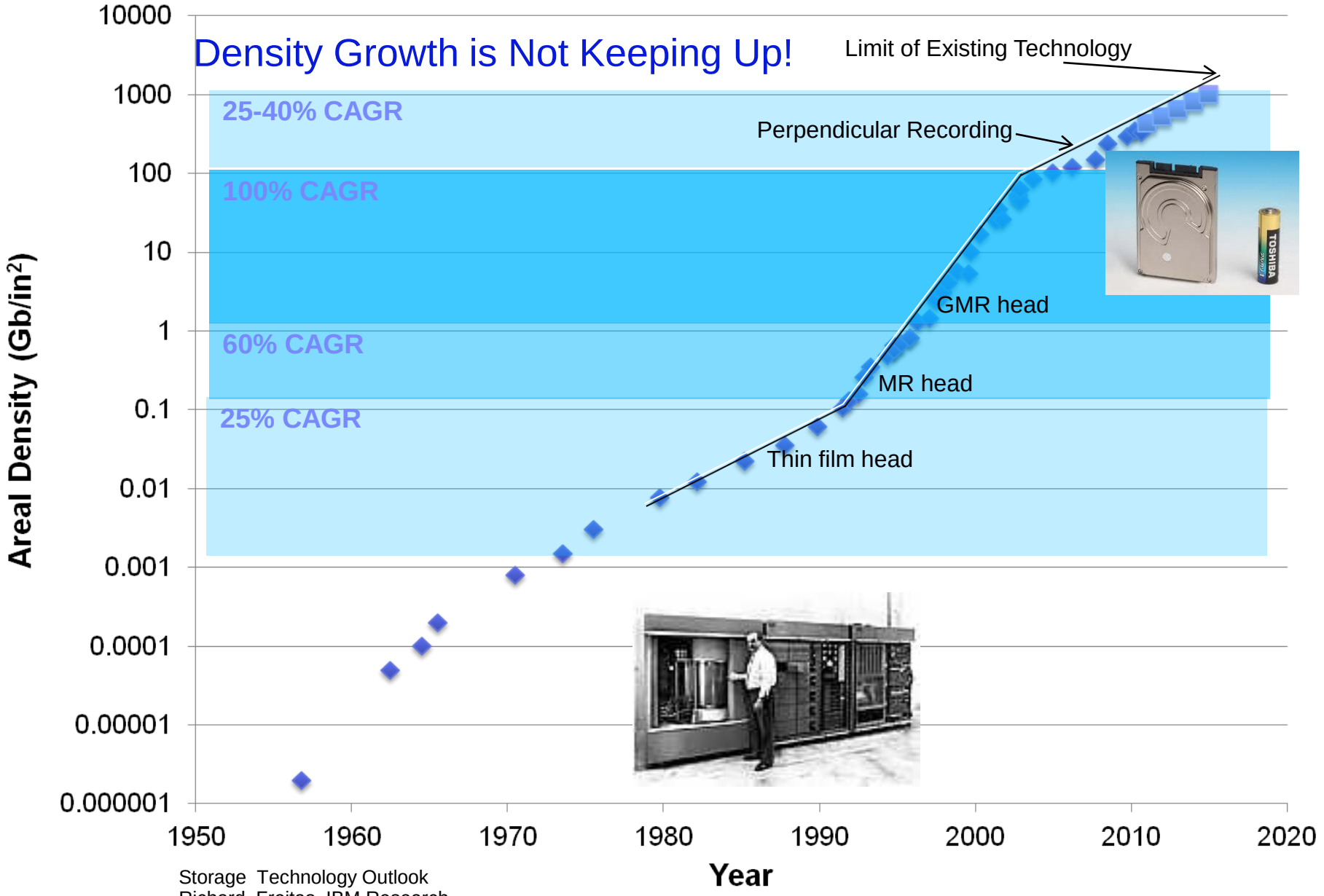
Taming the Plague of Petabytes

Matt Drahzal | IBM

Storage Requirements Devouring Resources

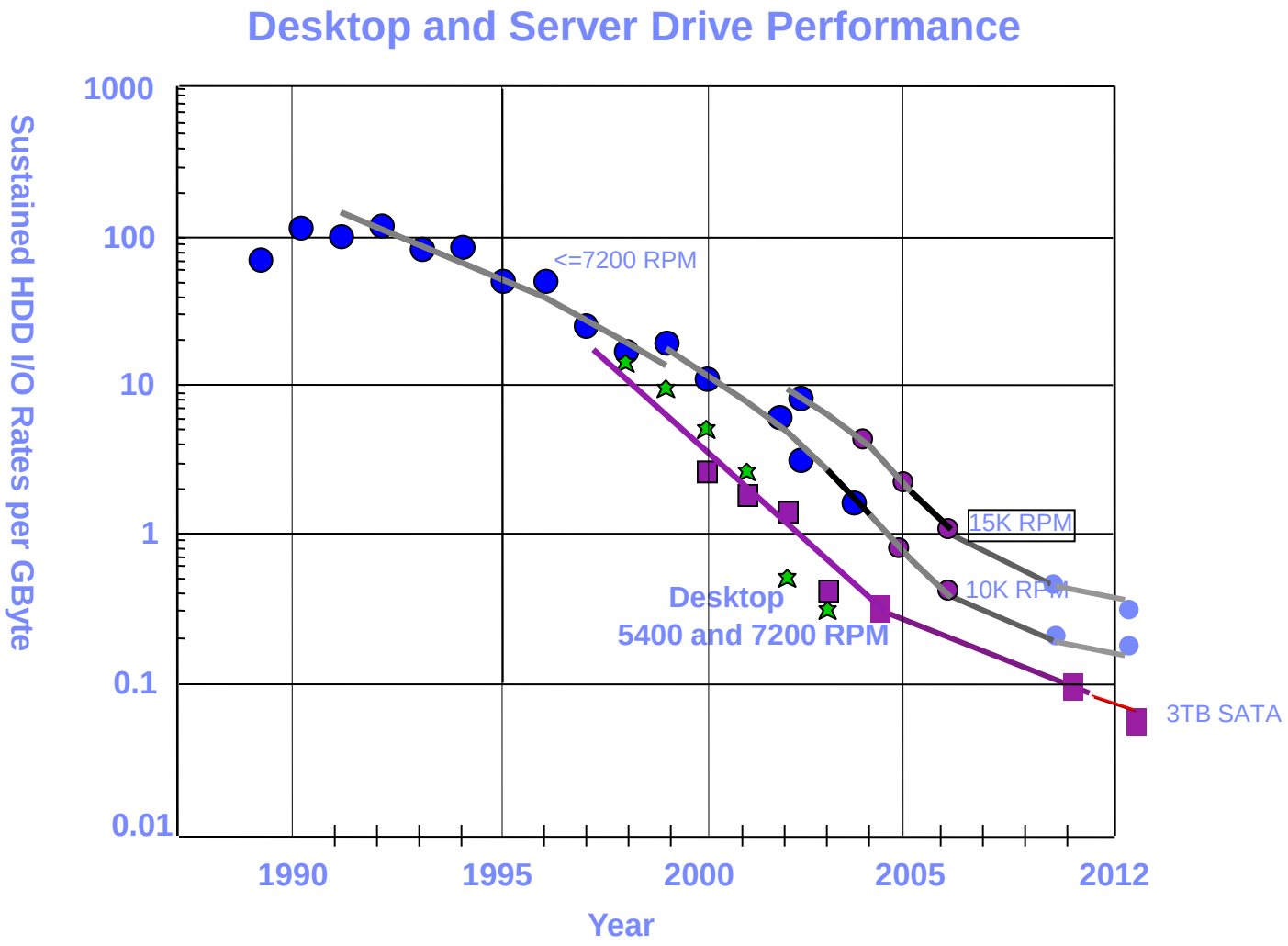


Density Growth is Not Keeping Up!



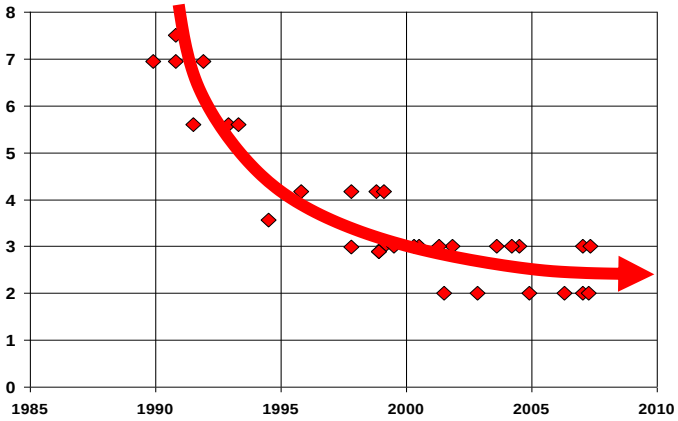
Storage Technology Outlook
Richard Freitas, IBM Research

Disk Performance Falling Behind

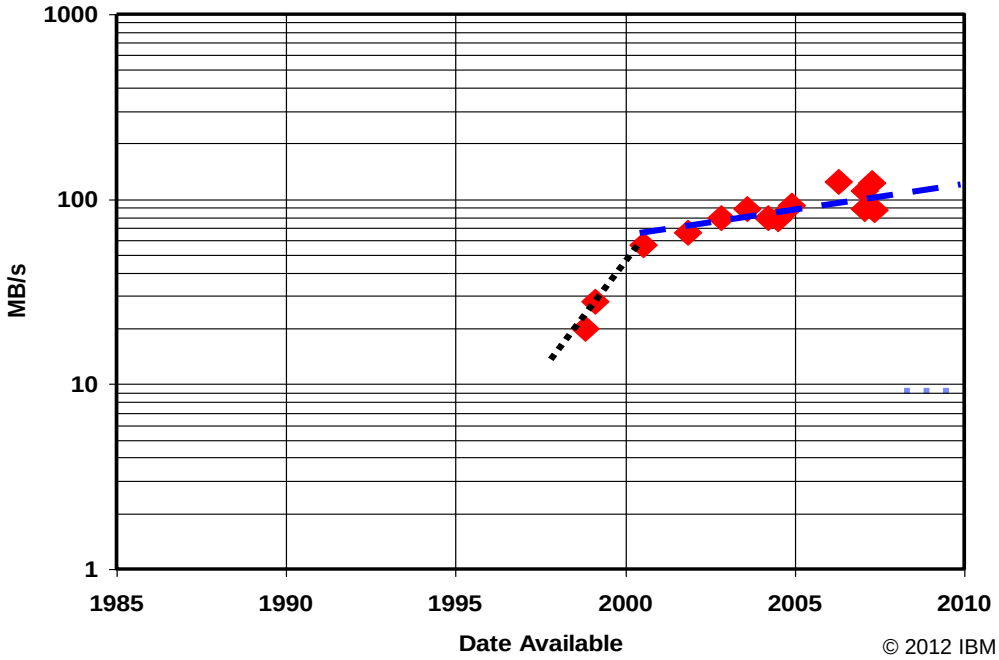


HDD Latency and Disk Transfer Speed

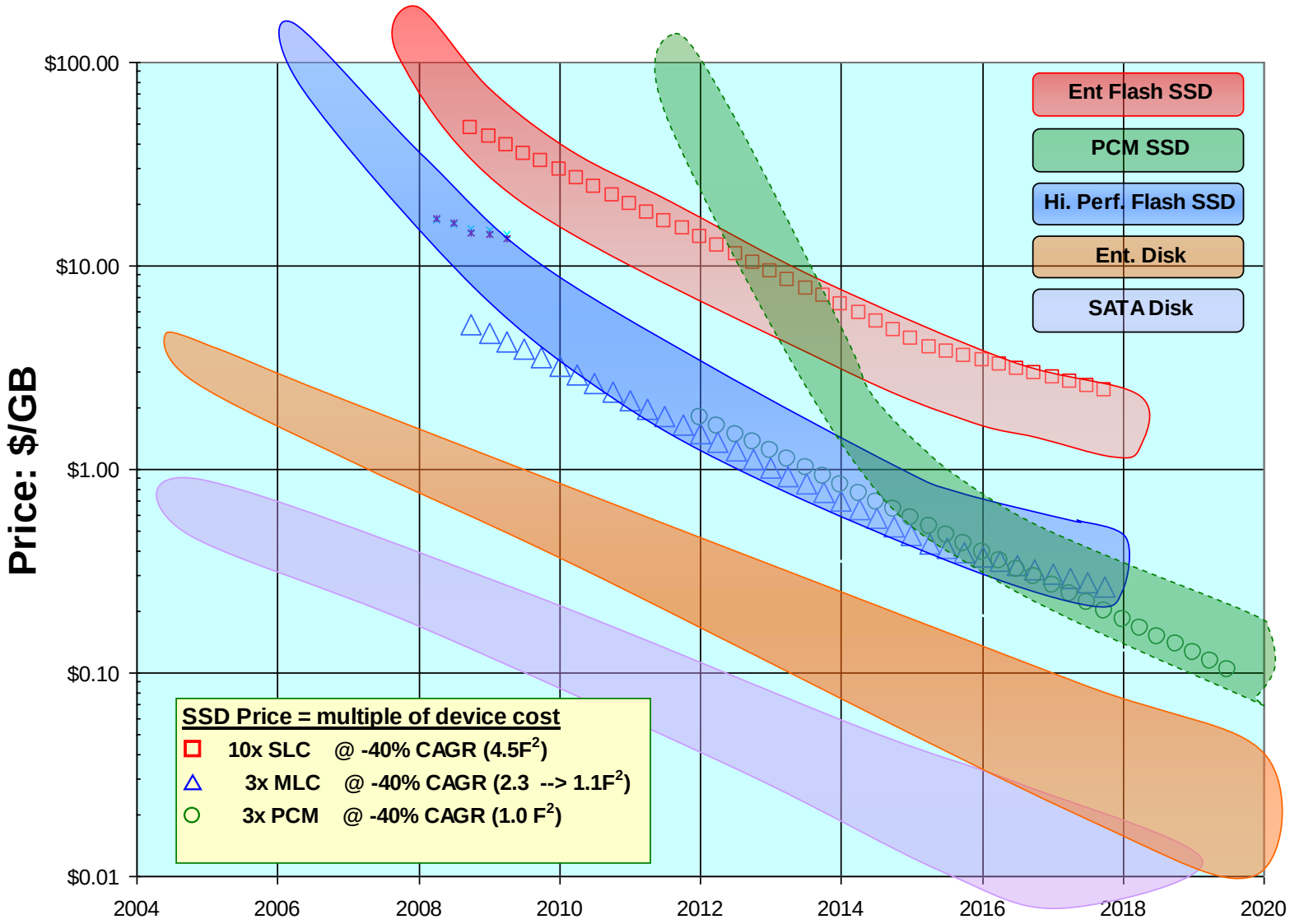
Latency
[ms]



Individual Disk Transfer Speed
[MB/s]

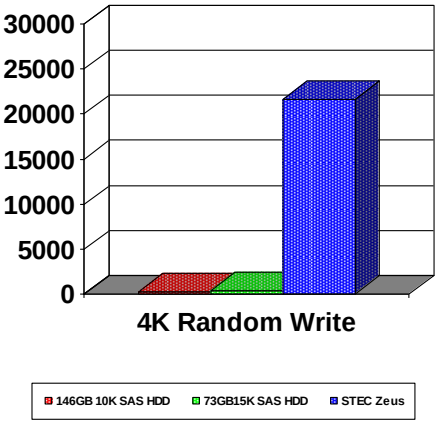


Price Trends: Magnetic disks and Solid State Disks

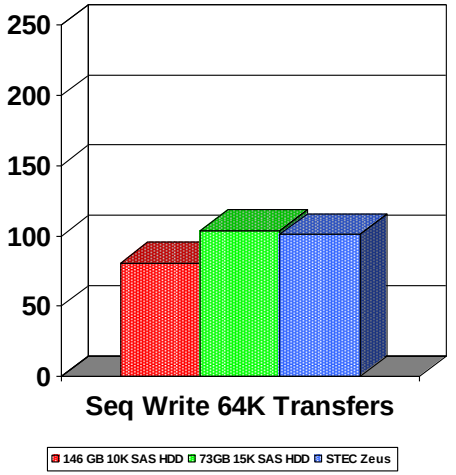


But What About Solid State Disks?

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?

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

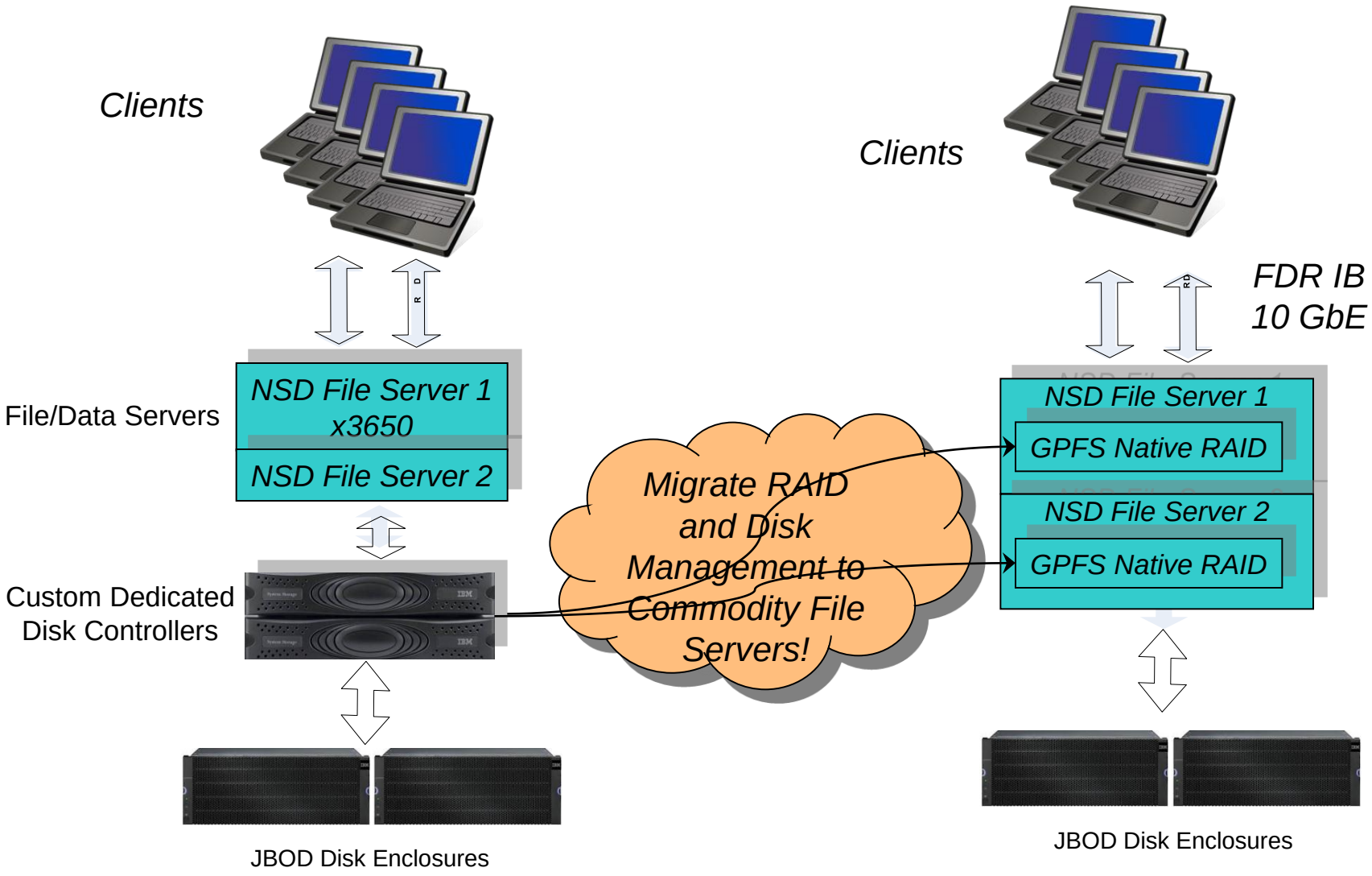
"Twin Tailed"
JBOD
Disk Enclosure

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!

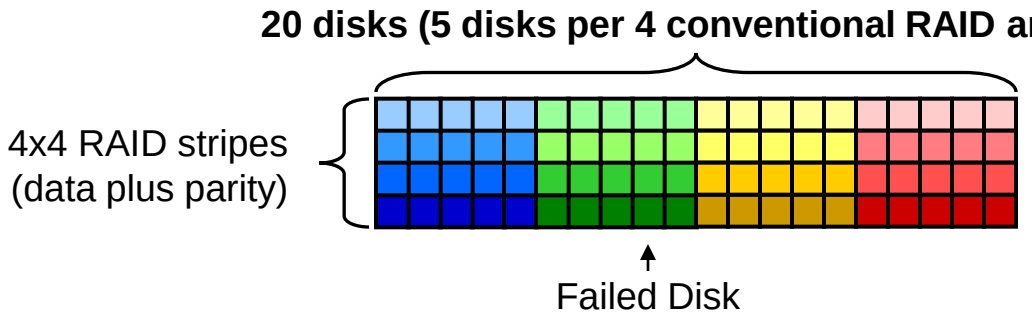


GPFS Native RAID Feature Detail

- **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

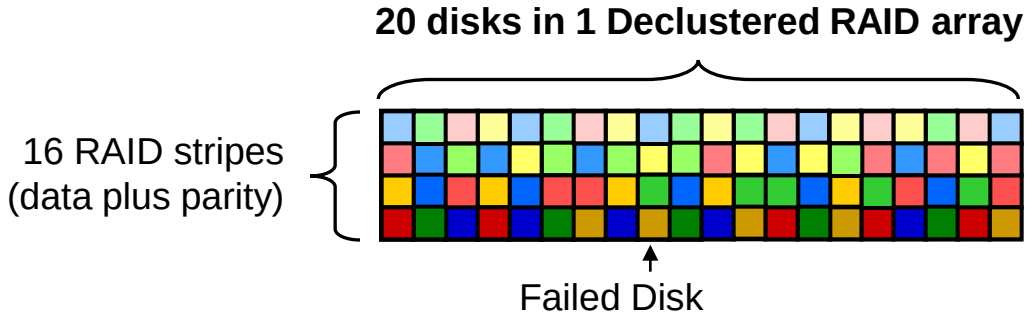
Declustering – Bringing parallel performance to disk maintenance

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



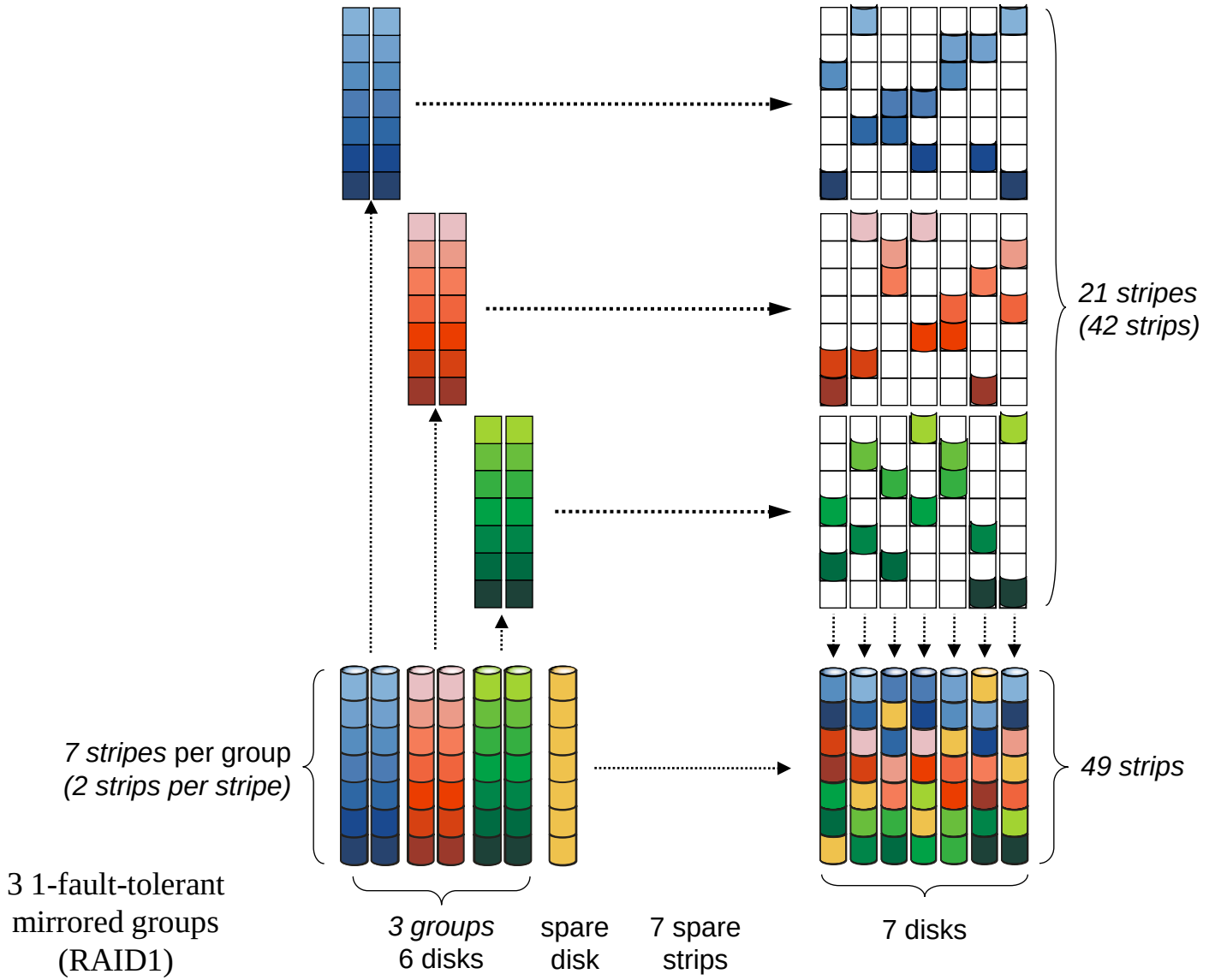
Striping across all arrays, all file accesses are throttled by array 2's rebuild overhead.

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

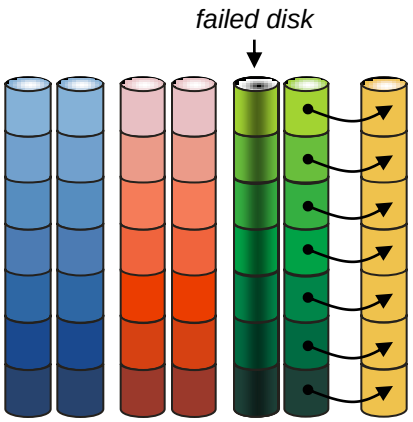


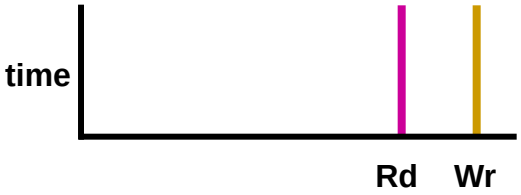
Load on files accesses are reduced by 4.8x (=19/4) during array rebuild.

Declustered RAID Example

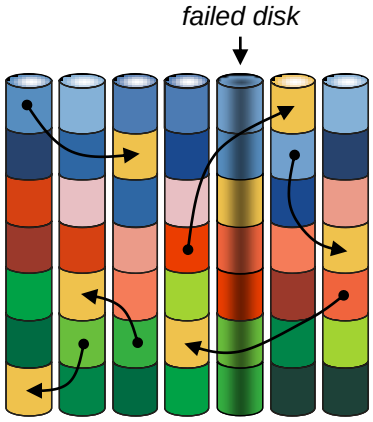


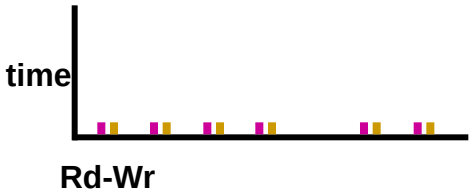
Rebuild Overhead Reduction Example





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



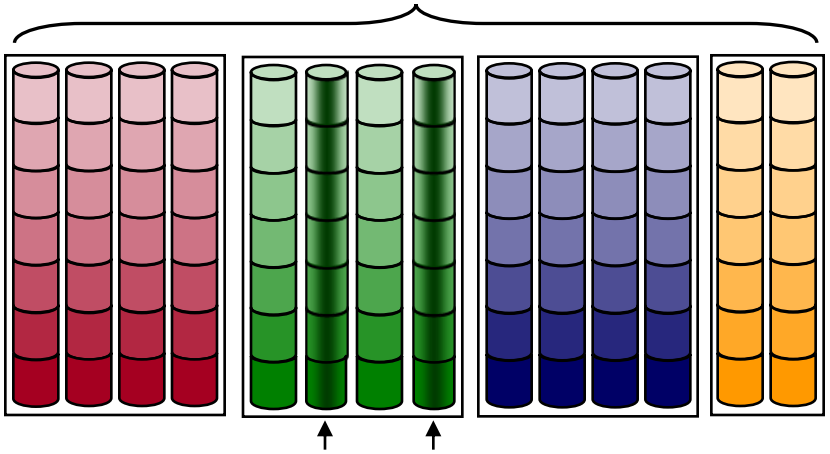


Rebuild activity spread across many disks, less disruption to user programs

Rebuild overhead reduced by 3.5x

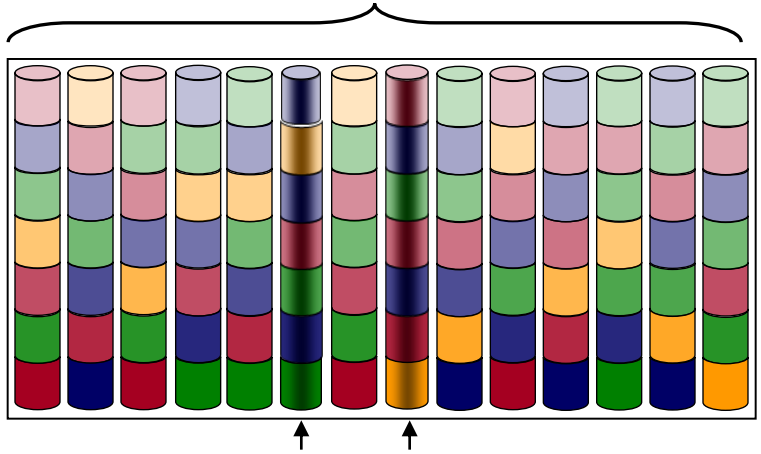
Declustered RAID6 Example

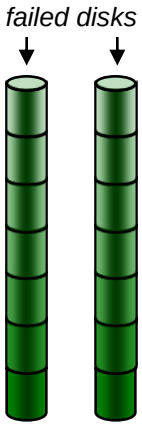
14 physical disks / 3 traditional RAID6 arrays / 2 spares



Decluster data, parity and spare

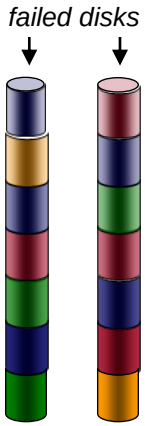
14 physical disks / 1 declustered RAID6 array / 2 spares





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



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

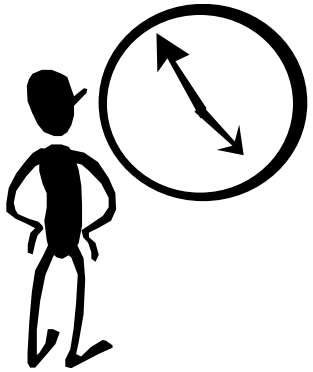
Data Protection Designed for 200K+ Drives!

- 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



- Summary
  - The Future is milliseconds away
  - Exascale storage means “THINK”ing differently
  - Using classical RAID techniques will NOT Scale
    - Disk Drives are mechanical devices
    - RAID 6 is nearing “end of applicability” as drive-count grows
  - Distance from Data will limit Analytics
    - *Keep your friends close and your important data closer*
    - ***“Again, distance matters, but often it is the cost of providing fast data access over that distance that is the root of the problem”*** (Mike Kahn, The Clipper Group)
- Tape is still with us after 50 years, disks will be with us into the distant future
  - Must begin to evolve disk storage TODAY to set the stage for the future