A New Approach to Disk-Based Scalable Mass Storage System

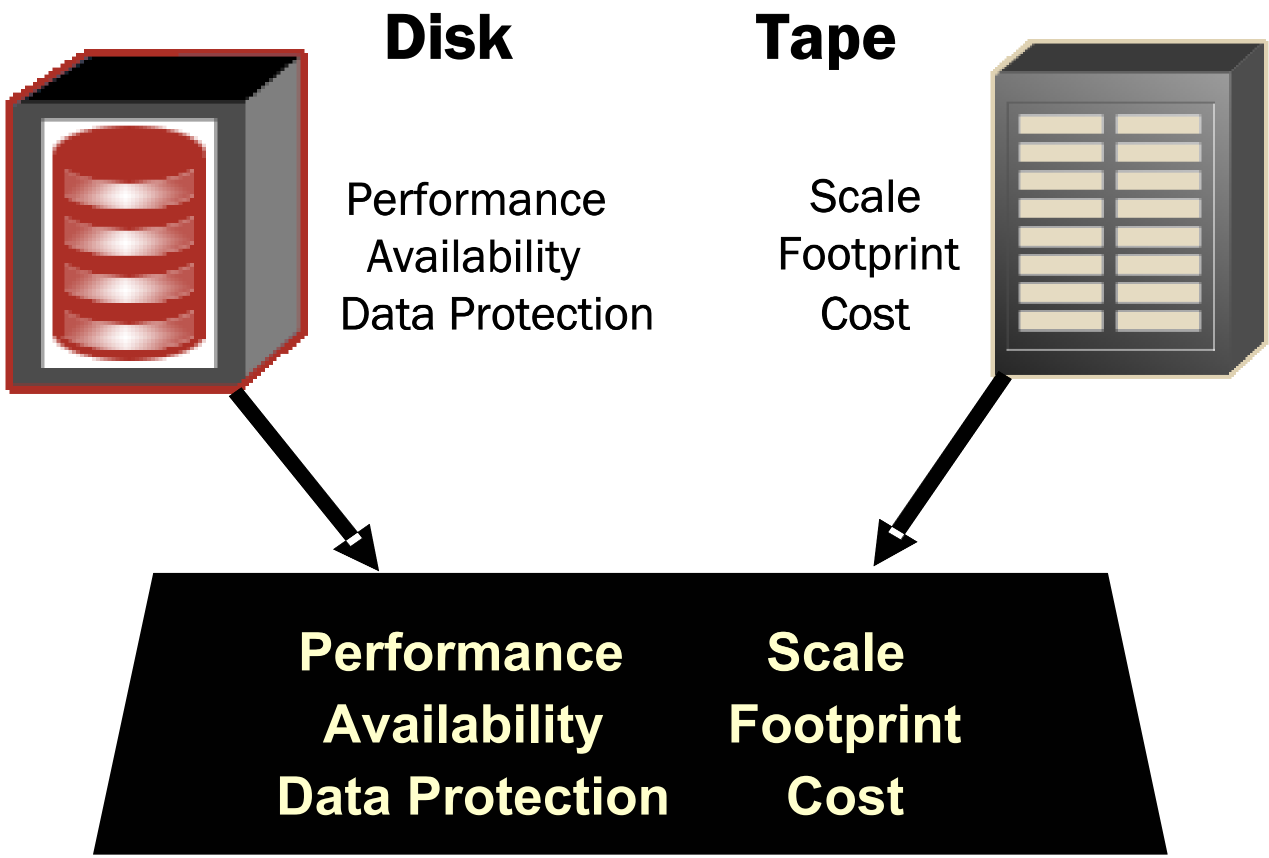
Dr. Aloke Guha

CTO, COPAN Systems

aloke.guha@copansys.com
(303) 827 2500

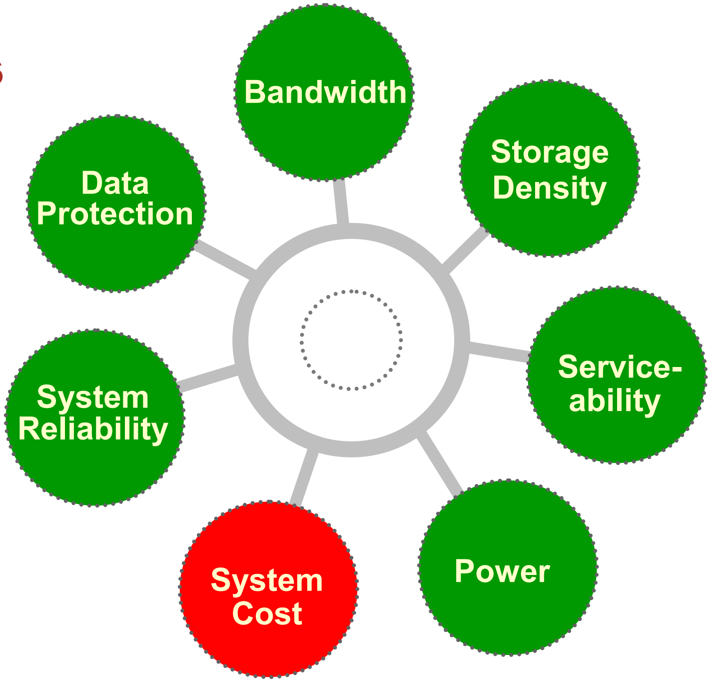
NASA/IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004

Wishing Well: Best of Both Worlds



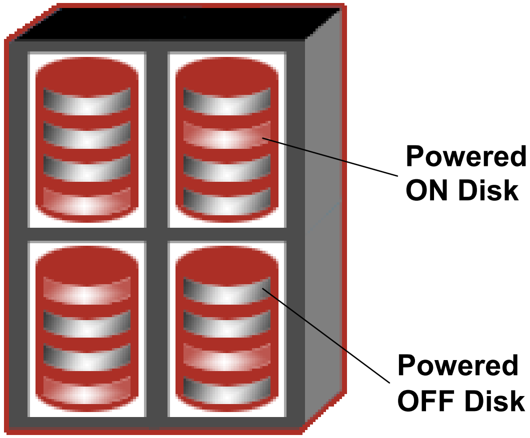
Application-Driven Approach

- **Secondary Storage Needs**
  - I/O: Sequential or Predictable Access
  - Performance: Mbytes/sec, not IOPs
  - Latency: msecs — seconds
- **Design Guidelines**
  - No need for large RAM cache
  - No need to access all data at all time
  - No need for host-disk Non-blocking Interconnect
  - High Capacity/Bandwidth ratio
  - Data Availability/Integrity
  - Serviceability



Power-Managed Disk . . . MAID

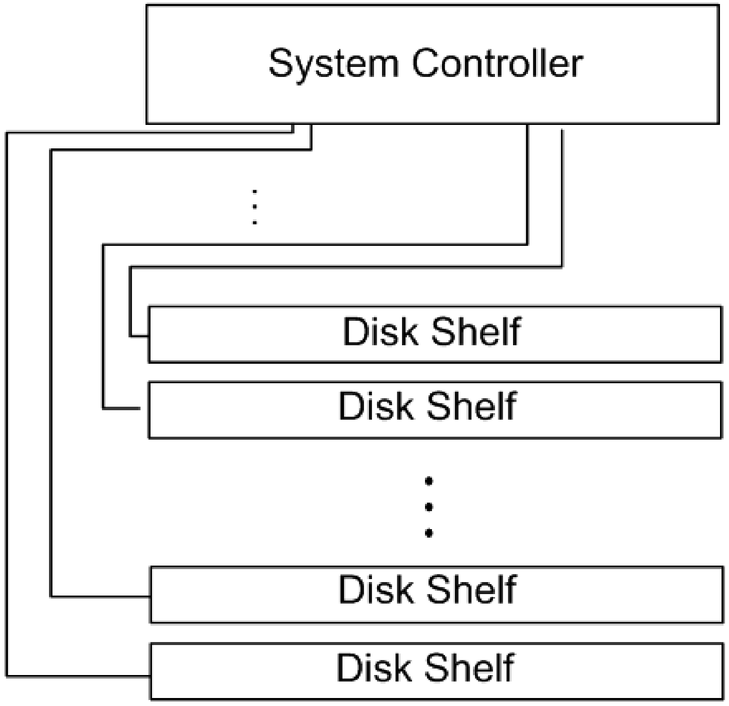
- Large number of power-managed drives
  - More than 50% drives powered OFF*
  - Power-cycling by policy for application
- Benefits: Scale, Cost, Service Life
- Cost Benefits: Lower Cost/Drive
  - 1/3 to 1/4 of typical RAID systems
  - Lower management cost from consolidation
- Beyond MAID
  - *Optimize scale and cost for RAS and Performance*

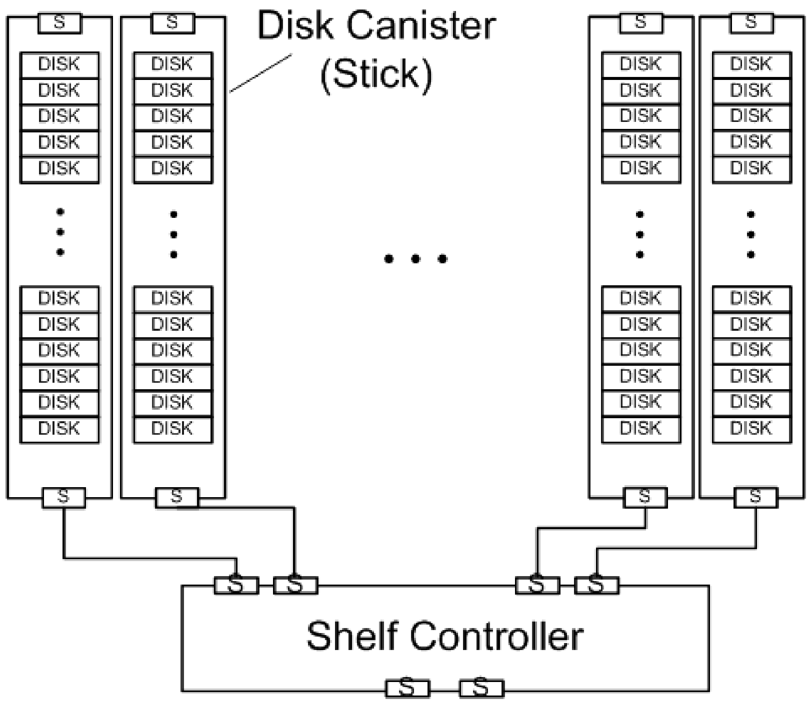


*Colarelli and Grunwald, The Case for Massive Arrays of Idle Disks (MAID), Usenix FAST 2002

3-Tier Architecture

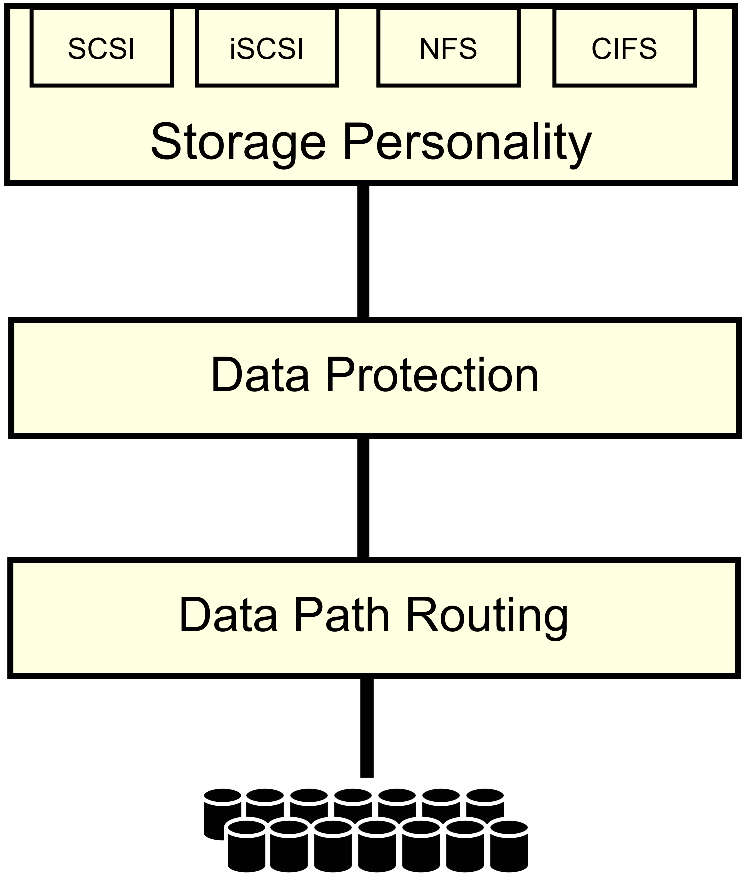
- Capacity scaling: optimize tier dimensions
- Parallel modular RAID: bandwidth scales with capacity
- Interconnect to support I/O to and control large # of drives
- Flexibility to present different presentations: file/disk/tape





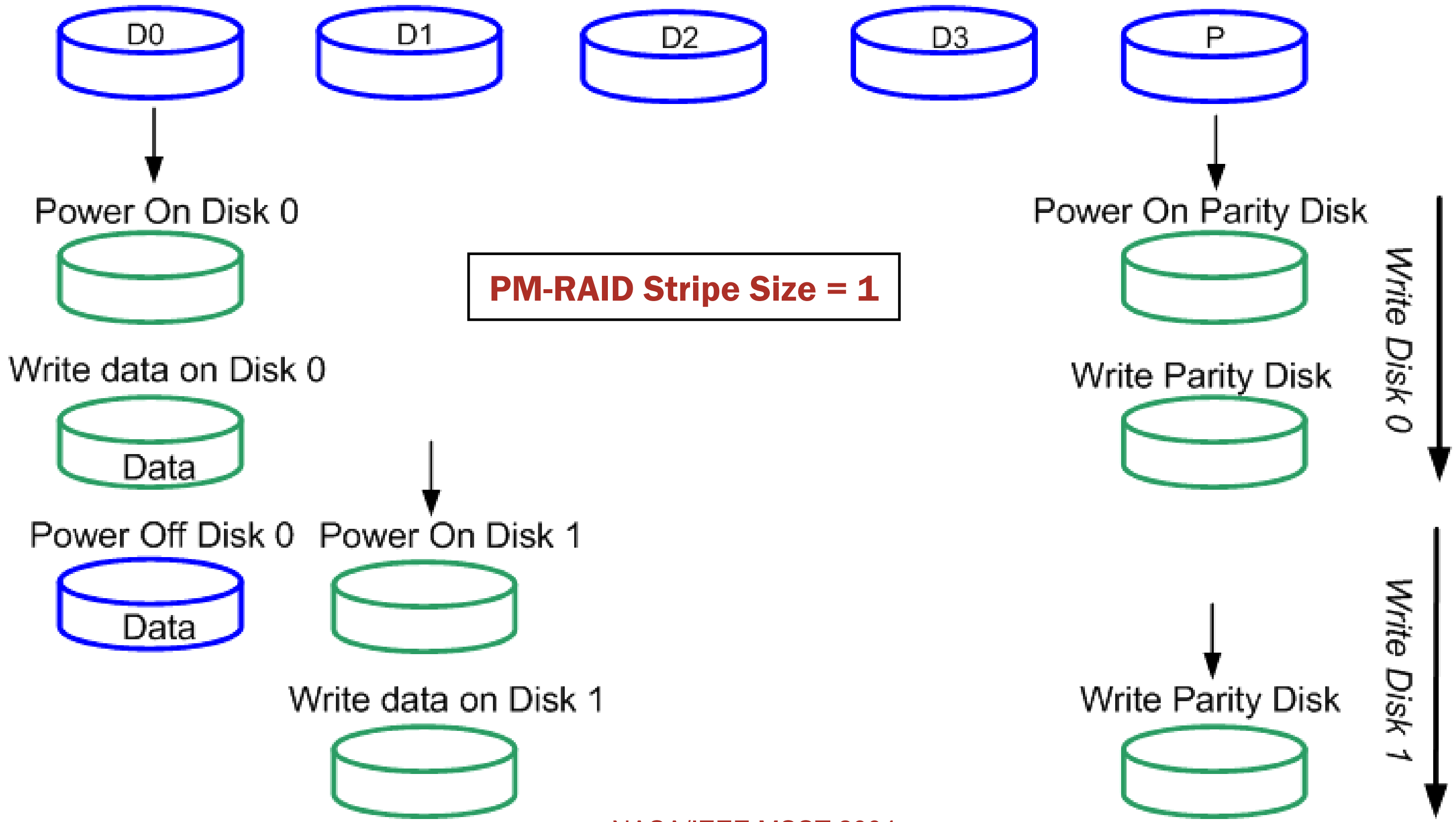
Functional Architecture

- Layer 2: Storage Personality: System
  - Storage Network Protocols
  - Logical Data Object Management
  - Load Balancing
- Layer 1: Data Protection: Shelf
  - RAID Acceleration
  - Power Management
  - Device Management
- Layer 0: Data Path Routing: Canister
  - Protocol Router
  - Monitoring
  - Manage Environmental Attributes



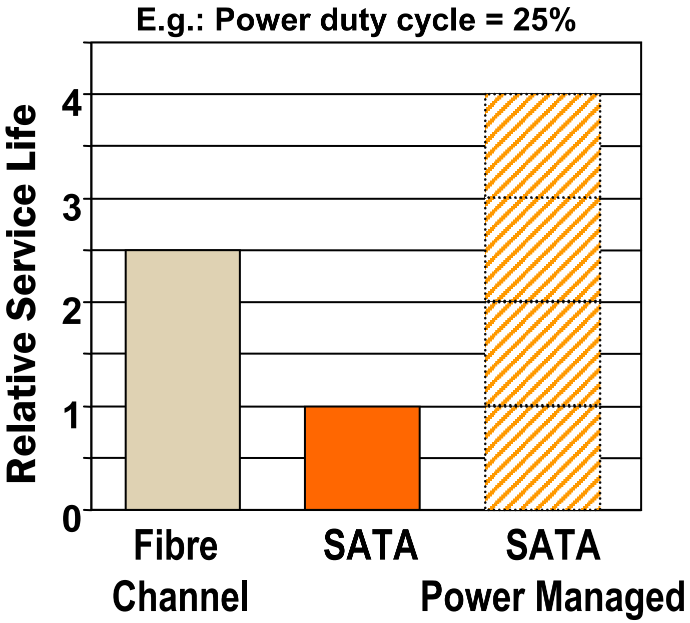
Power-Managed RAID (PM-RAID)™

- Data protection with only subset of drives powered in RAID group
- Number of drives powered dictated by application needs
- Multiple options on data organization to support application



Device and System Reliability

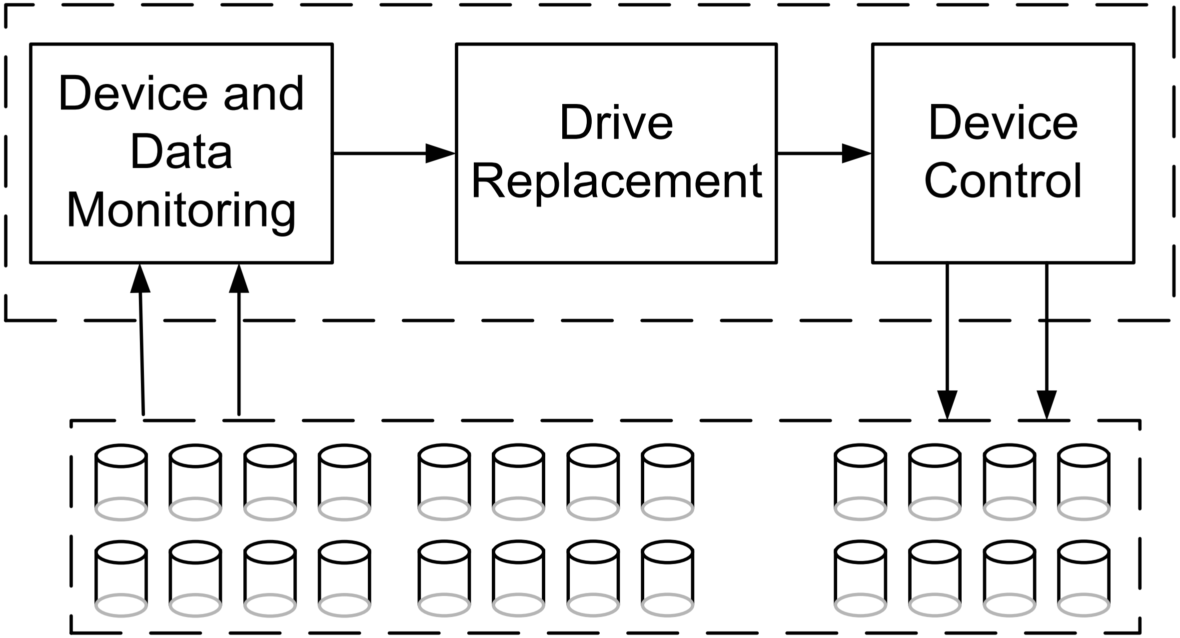
- **Effective drive service life, serviceability period**
  - **Improves with decreasing duty ratio***
- **Manage start stops**
  - **≤ 50K over service life**
  - **Match to application need**
- **Use disk density for availability**
  - **Spares to replenish failed drives**
  - **Rebuild data transparently**



*Power duty cycle ratio = # of powered-ON drives/# of powered-OFF drives

Increasing Data Reliability

- Device health monitoring
- Proactive data management: closed-loop control
- Revitalize data on disk for long-term data retention
- System data integrity mechanisms: ECC, CRC, Parity



Media Life versus Replacement Period

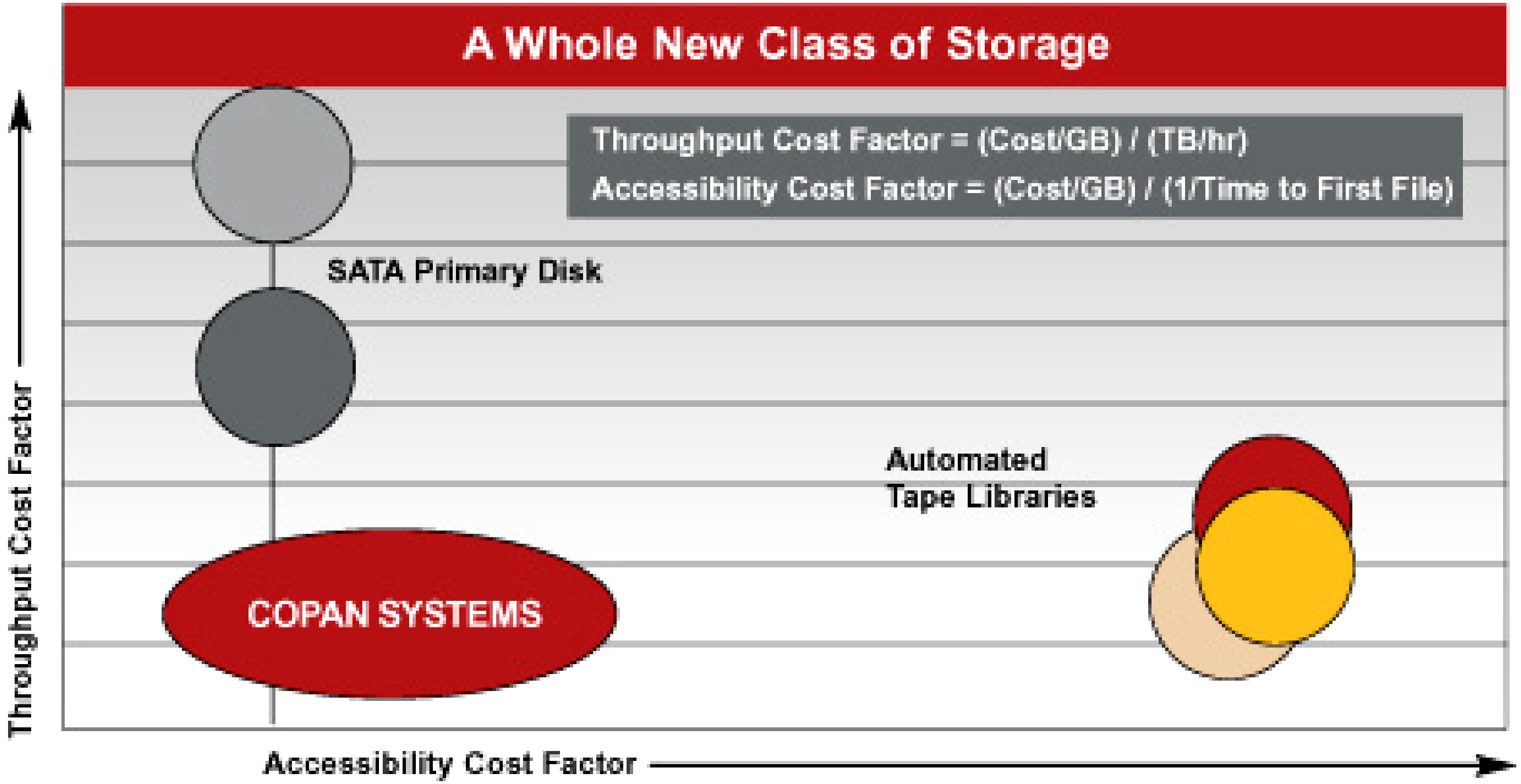
	Tape	Disk
Media Life	20+ yrs (40% RH, 20°C)[1]	5-7 yrs[2]
Drive Life	< 300 Khrs	400K-600 Khrs
Replacement Compatibility	Usually not backward compatible	NA
Replacement Period	5+ yrs	3-5 yrs

- **Replacement in Practice**
  - Frequency driven by need for upgrading capacity and data rate
  - Regulatory: 36 C.F.R. 1234.30(g)(4) – replacement in 10 years
- **Data reliability**
  - RAID on disks: 15% better recovery with disk than tape[3]

[1]Source: Enterprise Storage Group, Dec.'03; [2]Source: National Media Laboratory, '95; [3] Source: IBM

Increasing Performance

- Fraction of data on-line: ~10X tape
- Design: RAID processing, Interconnect Bandwidth, Disk Cache



Early Results: Data Rate

- Disk Drive bandwidth
  - 40 MBs+ media; 150 MBs SATA interface
- Power-managed RAID in shelf
  - Bandwidth increases with stripe size
  - I/O rate increases with block size
- Multiple streams/shelf: limited by interface
- Aggregate streams with multiple shelves

- Early Results
  - ~90 MBs/single stream uncompressed/shelf
  - Further Improvements: Tuning, Compression

Early Results: Access Time

- **Access Time: leverage HDD access time**
  - **Powered ON Drive: access time is 10s of millisecs**
  - **Powered OFF Drive: spin-up time, data access 10s-15s**

Random Access of File/Drive: uncompressed 100 MB#

9940B TAPE: streaming @ 30MB/sec

Load	Ave. Time to 1st Byte*	File Xfer	Unload	
18 sec	41 sec	3.3 sec	18 sec	**Total: 80 sec**

SATA 7200 RPM Disk: streaming @ 40 MB/sec – increases with RAID

Spin up	Ave. Time to 1st Byte	File Xfer	Spin down
ms-6 sec	0.1 sec	2.5 sec	0.1 sec

Total (power-off AND cache miss): 8.7 sec*

Total (power-on OR disk cache): 2.7 sec

*COPAN uses 256K I/O size; ave. time to first byte on tape depends on location of file (0 - 90 s)

Conclusions

- Application-tuned, Optimized MAID
  - Storage Capacity and Cost
  - Reliability: Power-Managed RAID
  - Performance: Bandwidth, Access Time
  - Serviceability
- Early results meeting goals and expectations
- Filling the Gap in the Storage Hierarchy!