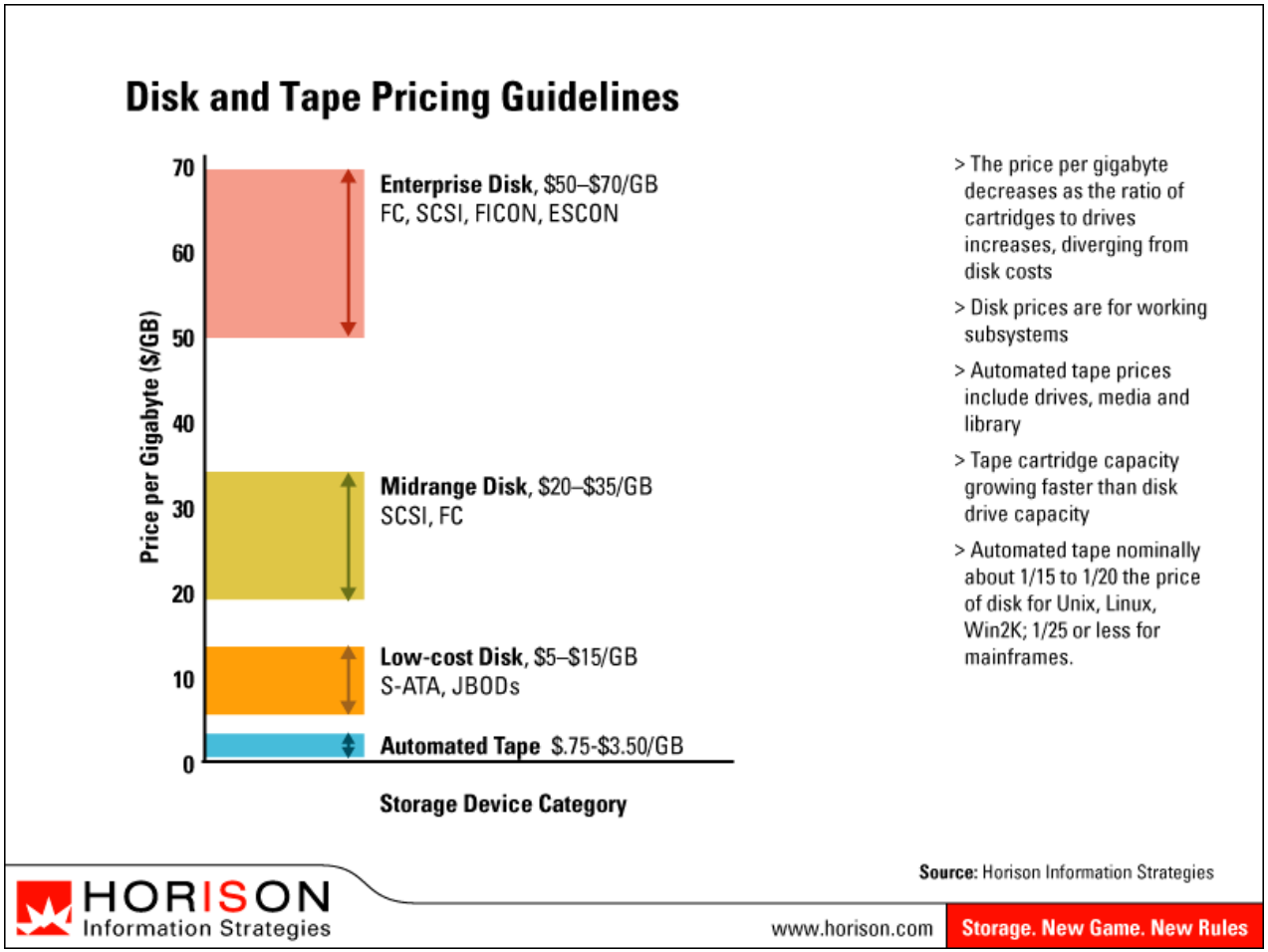
MAID and Beyond: Filling the Storage Gap

Aloke Guha, Ph.D.
CTO, COPAN Systems Inc.

Data Management Workshop
SLAC
March 16, 2004

Pricing Trends: Disk and Tape



"*The difference in price between ATA/SATA and SCSI/FC is at least 50% on a per-GB basis*"
Peter Kastner, Aberdeen Group

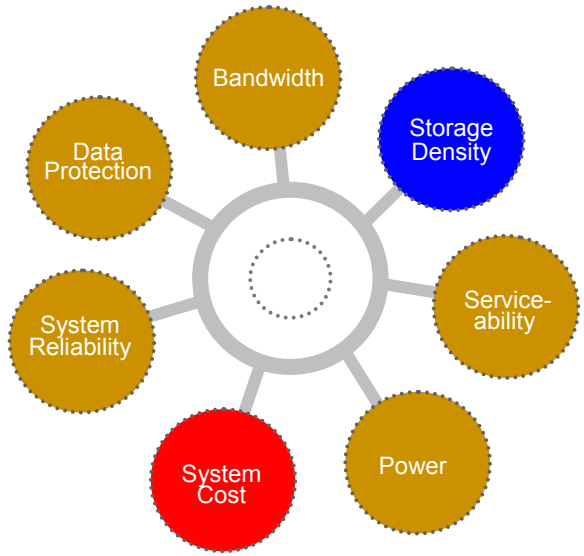
Storage System Cost vs. Media Cost

- **Storage system $/GB far exceeds media $/GB**
- **Cost Efficiency: $\frac{\text{Media Cost}}{\text{Storage System Cost}}$**
  - **Disk Example**
    - 250 GB SATA ~$1/GB vs Storage: $5-$15/GB[#1]
    - Cost Efficiency: 0.07 – 0.2
  - **Tape Example**
    - 200 GB LTO2 media ≤ $0.5/GB vs Storage: $0.75-$3.5/GB[#2]
    - Cost Efficiency: 0.14 – 0.67
- **Traditional disk systems 2x-3x higher cost than tape[#3]**

[#1] Fred Moore, Horison Strategies, October 2003
[#2] Native uncompressed capacity: cost/GB depends on ratio of tape cartridges to drives, in range of 20:1 – 80:1
[#3] Assume same compression on streaming I/O for disk or tape

Scaling and Cost Challenges

- **Replacing FC with ATA not sufficient**
  - **Density Scaling**
    - **Power**
  - **Reliable Operation**
    - **Heat**
    - **Vibration**

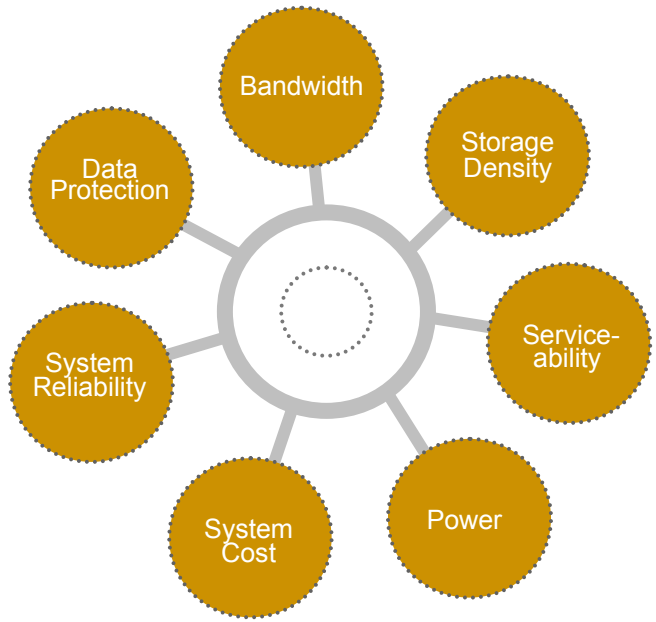


- **Need fundamentally different approach**
  - **Minimize "Slot" Cost**

$$\text{Slot Cost} = \frac{\text{(System Cost – Media Cost)}}{\text{\#Drives}}$$

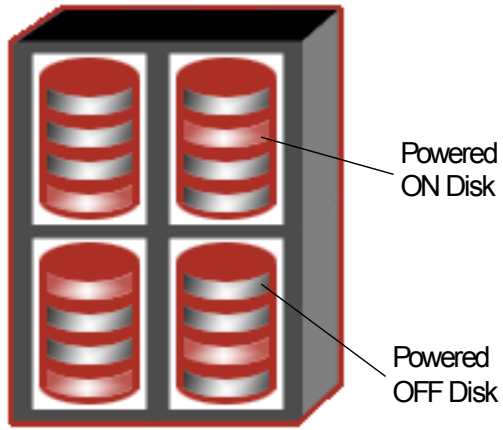
An Application-Specific Approach

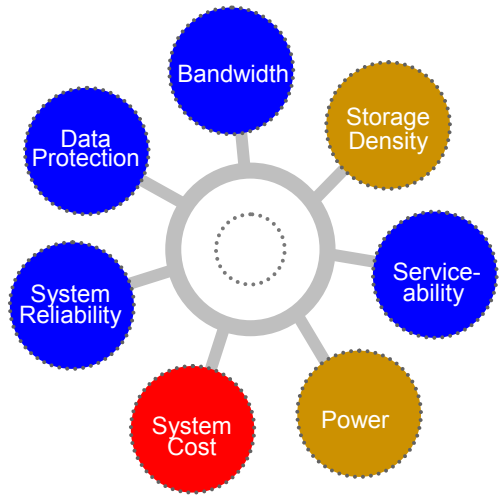
- **Needs of secondary storage**
  - Mostly sequential I/O
  - Performance: MBs not IOPs
  - Access: ms-secs
- **Optimize architecture**
  - No need for large RAM cache
  - No need for non-blocking interconnect
  - No need to access all data at all time
  - High Capacity/Bandwidth ratio
- **Constraints**
  - Data Protection
  - System Reliability
  - Serviceability
  - Power



Power-Managed Disk (MAID) and Beyond

- **Large number of power-managed drives**
  - **Infrequent access ⇒ > 50% drives powered OFF** [#4]
- **Scale Benefits**
  - **Lower heat and vibration**
  - **Higher service life**
- **Cost Benefits: Lower "Slot Cost"**
  - **Cost/GB ⇒ 1/3 to 1/4 std. RAID systems**
  - **Lower management cost from consolidation**
- **Beyond MAID: Application-Specific Design**
  - **Reduce system overhead**
  - **Optimize for required features**

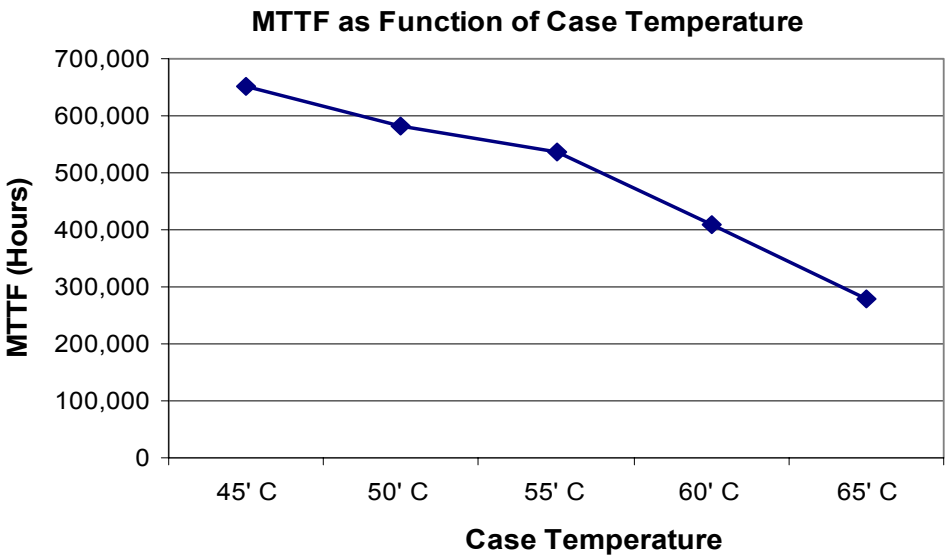


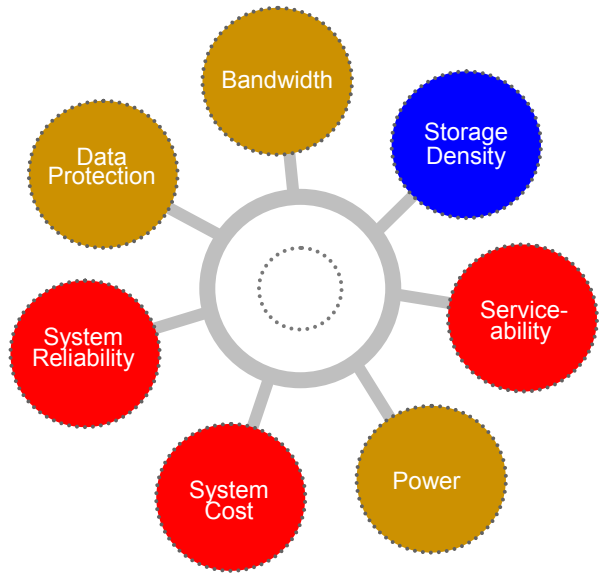


[#4] Colarelli and Grunwald, The Case for Massive Arrays of Idle Disks (MAID), Usenix FAST 2002

Storage Capacity versus Drive Life

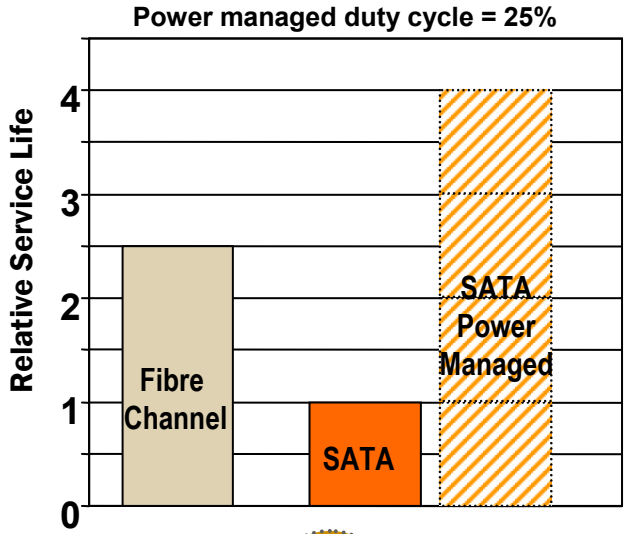
- **Drive Packing Density**
  - –O(1000) drives = 250TB+
- **Interconnect Architecture**
  - – Connectivity and Bandwidth
- **Drive Life as function of temperature**

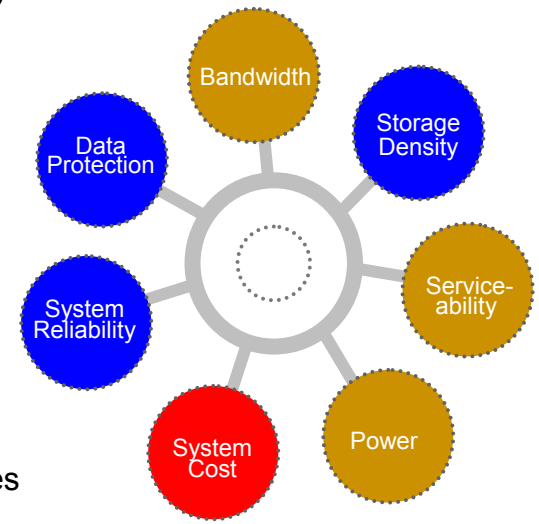




Extending Service Life

- **Effective drive life**
  - Increases with decreasing duty ratio[#5]
- **Manage start stops**
  - < 40,000 over service life
  - Power-ON cycle matched to application need
- **↑ Service life ⇒ ↑ data reliability**
- **Use storage density to increase availability**
  - Many spares to replenish failed drives
  - Rebuild data transparently





[#5] Power cycle duty ratio = # of powered-ON drives/# of powered-OFF drives

Expected Contact Start Stops

- **MAID system bandwidth/capacity limits CSS to 3% of max over 5 yrs**
- **Tape archives: ave #mounts/used volume limits CSS to <5 % of max over 5 yrs[#6]**

	Industry	Volumes Used	Daily #Mounts/Volumes Used		
			Average	Max	Median
1	Telco	373	0.0	0.1	0.0
2	Telco	1,015	0.1	0.4	0.0
3	Telco	688	0.1	0.5	0.0
4	Telco	1,189	0.0	0.0	0.0
.	.	.	.	.	.
10	Financial	835	1.6	4.2	2.1
11	Financial	10,677	0.1	0.3	0.2
12	Financial	2,061	0.1	0.2	0.1
.	.	.	.	.	.
33	Utility	278	2.6	4.9	3.5
34	Govt	3,393	0.4	0.7	0.5
35	Govt	84	0.1	0.4	0.1
.	.	.	.	.	.
	Average	1,122	0.6	1.1	0.6

Average # Start-Stops over 5 year operation

	MAID System Capacity (TB)		
MAID Bandwidth (TB/hr)	150	200	250
2	584	438	350
3	876	657	526
4	1168	876	701

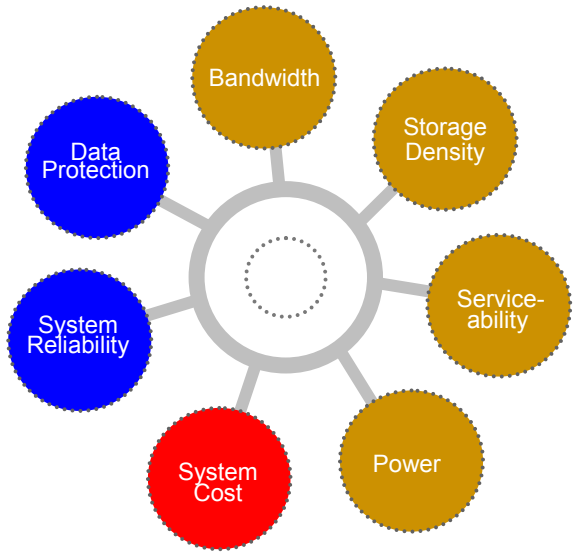
Typical Specified Limit: 40,000

[#6] Source: FileTek - data from 43 archives on tape: Volumes_Used excludes tape volumes not allocated in library

Data Reliability and Data Integrity

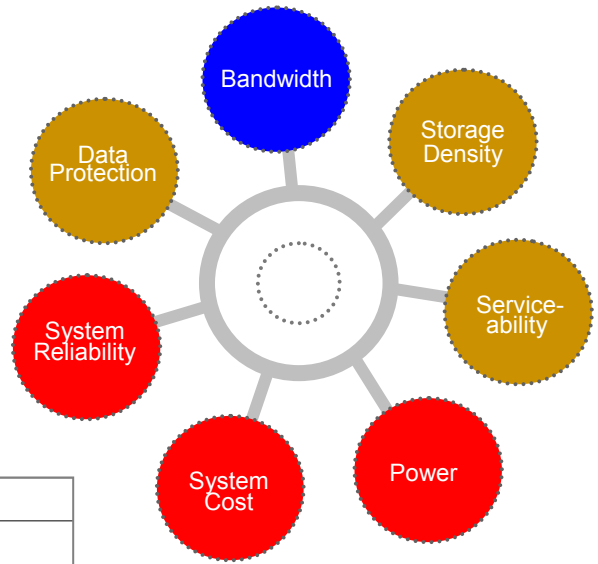
- **Proactively monitor drive health**
  - Copy data onto spare before drive failure
- **System data integrity mechanisms (ECC)**
- **Increased data reliability**
  - 50% of drive data recovered before failure
    ⇒ 4X data reliability
- **Data "revitalized" before drives fail**
  - Reduces vulnerability as disks get bigger!

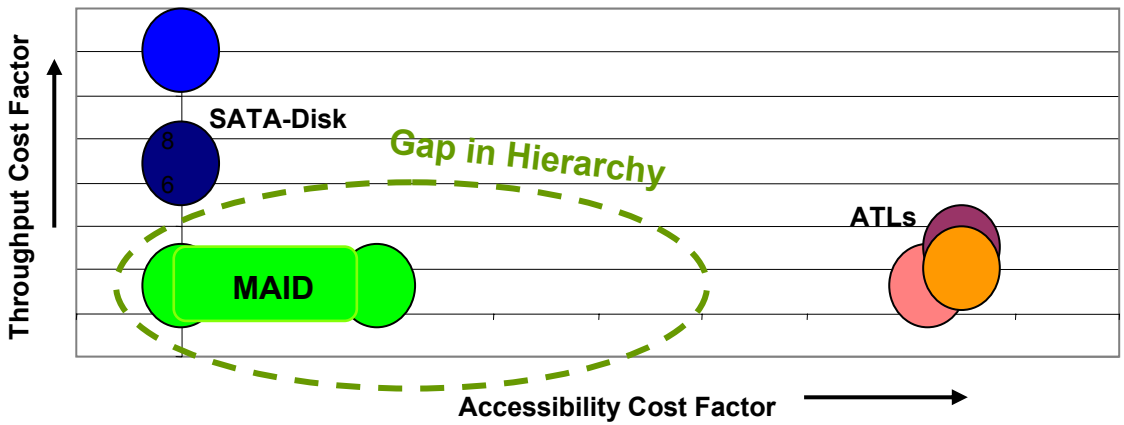
- **Revitalized Data ⇒ Longer Data Retention**



Increasing Performance

- Fraction of data on-line: 10X tape
- RAID processing and Interconnect Bandwidth
- Performance-Cost Metrics
  - Throughput Cost Factor = $\frac{\text{(Cost/GB)}}{\text{(TB/hr)}}$
  - Accessibility Cost Factor = $\frac{\text{(Cost/GB)}}{\text{(1/Time to First File)}}$

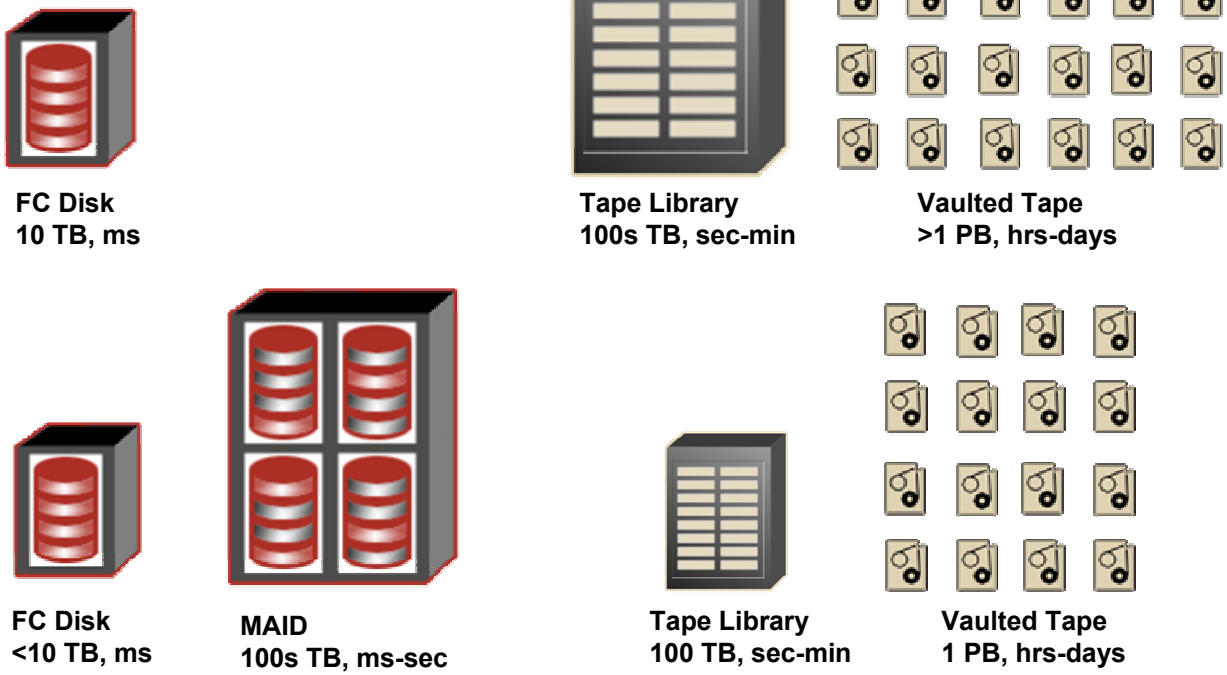




Other Cost Components

- **Technology Refresh**
  - 36 CFR 1234.30 (g) (3): annual sampling, 10-yr replacement
  - Upgrade of tape media
- **Data Protection**
  - Tape backup and duplex vs RAID
- **Storage Management**
  - Primary vs archive storage workload
- **Software and Maintenance**
- **Media management**
  - MAID as native disk has application flexibility

Filling the Storage Gap



- Increases online data by 10X
- Higher performance and data reliability
- Lower long-term data management cost

Conclusions

- Application-tuned MAID architecture for parity with tape cost
- Optimize design for storage system for minimum slot cost
  - Storage density and capacity
  - System reliability
  - Data protection and reliability
  - Performance
  - Serviceability
- MAID fills the gap in the storage hierarchy

Best of Both Worlds

Disk

- Performance
- Availability
- Data Protection

COPAN
Systems

- Performance
- Availability
- Data Protection
- Scale
- Footprint
- Cost

Tape

- Scale
- Footprint
- Cost

Massively scalable enterprise storage solutions with the reliability and performance of disk at the scale and cost of tape.

COPAN SYSTEMS

Thanks

Contact: aloke.guha@copansys.com

http://www.copansys.com