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

### Disk and Tape Pricing Guidelines

<table>
<thead>
<tr>
<th>Storage Device Category</th>
<th>Price per Gigabyte ($/GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterprise Disk</td>
<td>$50–$70/GB</td>
</tr>
<tr>
<td>Midrange Disk</td>
<td>$20–$35/GB</td>
</tr>
<tr>
<td>Low-cost Disk</td>
<td>$5–$15/GB</td>
</tr>
<tr>
<td>Automated Tape</td>
<td>$.75–$3.50/GB</td>
</tr>
</tbody>
</table>

- The price per gigabyte decreases as the ratio of cartridges to drives increases, diverging from disk costs.
- Disk prices are for working subsystems.
- Automated tape prices include drives, media and library.
- Tape cartridge capacity growing faster than disk drive capacity.
- Automated tape nominally about 1/15 to 1/20 the price of disk for Unix, Linux, Win2K; 1/25 or less for mainframes.

Source: Horison Information Strategies

"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 \(\leq$0.5/GB\) 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

\[
\text{Slot Cost} = (\text{System Cost} - \text{Media Cost}) \\
\#\text{Drives}
\]
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 $\Rightarrow > 50\%$ drives powered OFF

- Scale Benefits
  - Lower heat and vibration
  - Higher service life

- Cost Benefits: Lower “Slot Cost”
  - Cost/GB $\Rightarrow 1/3$ 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**
  \(-0(1000) \text{ drives} = 250\text{TB}+\)**

- **Interconnect Architecture**
  - Connectivity and Bandwidth

- **Drive Life as function of temperature**

![MTTF as Function of Case Temperature](image)
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

### Industry Volumes Used

<table>
<thead>
<tr>
<th>Industry</th>
<th>Volumes Used</th>
<th>Daily #Mounts/Volumes Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telco</td>
<td>373</td>
<td>Average: 0.0, Max: 0.1, Median: 0.0</td>
</tr>
<tr>
<td>Telco</td>
<td>1,015</td>
<td>Average: 0.1, Max: 0.4, Median: 0.0</td>
</tr>
<tr>
<td>Telco</td>
<td>688</td>
<td>Average: 0.1, Max: 0.5, Median: 0.0</td>
</tr>
<tr>
<td>Telco</td>
<td>1,189</td>
<td>Average: 0.0, Max: 0.0, Median: 0.0</td>
</tr>
<tr>
<td>Telco</td>
<td>1,068</td>
<td>Average: 0.1, Max: 0.2, Median: 0.1</td>
</tr>
<tr>
<td>Telco</td>
<td>1,118</td>
<td>Average: 0.0, Max: 0.0, Median: 0.0</td>
</tr>
<tr>
<td>Telco</td>
<td>1,189</td>
<td>Average: 0.0, Max: 0.0, Median: 0.0</td>
</tr>
<tr>
<td>Telco</td>
<td>1,189</td>
<td>Average: 0.0, Max: 0.0, Median: 0.0</td>
</tr>
</tbody>
</table>

### Average # Start-Stops over 5 year operation

<table>
<thead>
<tr>
<th>MAID Bandwidth (TB/hr)</th>
<th>MAID System Capacity (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>584 438 350</td>
</tr>
<tr>
<td>3</td>
<td>876 657 526</td>
</tr>
<tr>
<td>4</td>
<td>1168 876 701</td>
</tr>
</tbody>
</table>

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 = \( \frac{\text{Cost}}{\text{GB}} \) (TB/hr)
- Accessibility Cost Factor = \( \frac{\text{Cost}}{\text{GB}} \) \( \frac{1}{\text{Time to First File}} \)

![Diagram showing throughput and accessibility cost factors]

- SATA-Disk
- MAID
- ATLs

Gap in Hierarchy
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

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

Contact: aloke.guha@copansys.com
http://www.copansys.com