MAID (Massive Array of Idle Disks)  
Meeting the Long-Term Data Challenge

Dr. Aloke Guha
CTO, COPAN Systems
2605 Trade Center Drive, Ste D
Longmont, CO 80503
Email: aloke.guha@copansys.com
Phone: (303) 827 2500 FAX: (303) 827 2504

Presented at the THIC Meeting at the National Center for Atmospheric Research
1850 Table Mesa Drive, Boulder CO 80305-5602
June 29, 2004
Traditional Storage Hierarchy

**Primary Disk**
- Fibre Channel
- $30-$100/GB
- 15%
- ms
- RAID

**Nearline Tape**
- $0.75 - $3.5/GB
- 85%
- 100s TB - PB
- sec – min . . . hrs

**Backup/Migrate**
- Offline Tape

**Recover/Restore**
Today’s Storage Hierarchy

- Bulk of data stays on tape
- Bulk of data is unprotected
- Poor recall/restore performance

Primary Disk
- Fibre Channel
- $30-$100/GB
- 10%
- ms
- RAID

Secondary Disk
- SATA
- $5-$15/GB
- 5%
- ms
- RAID

Nearline Tape
- $0.75 - $3.5/GB
- 100s TB – PB
- 80%
- sec – min . . . hrs
Information Management Complexity

- Frequent Data Movement
- Physical tape handling
- Network, Server Utilization
- CAPEX, Labor Cost
Application-Driven Approach to Storage

- **Secondary Storage Needs**
  - I/O: Sequential or Predictable Access
  - Performance: Mbytes/sec, not IOPs
  - Latency: msecs - secs

- **Design Guidelines**
  - No need for large RAM cache
  - No need to access all data at all time
  - No need for non-blocking interconnect
  - High Capacity/Bandwidth ratio
  - Data Availability/Integrity
  - Data Retention
  - Serviceability
MAID: Power-Managed Disks

- Large # power-managed disks
  - > 50% drives powered OFF*
  - Power-cycling by policy
  - Defined in SNIA

- Scale, Cost, Service Life

- Lower Cost/Drive
  - 1/4 - 1/3 of typical RAID systems
  - Lower management cost

- Extending MAID
  - Performance and scale
  - Reliability
  - Cost

*SNIA ILM TWG Definition
Refers to a storage system comprising a very large array of disk drives where a majority of the drives are powered off. The goal of a MAID storage system is to reduce the energy consumed by a large-scale storage array while increasing storage density and maintaining performance similar to conventional disk arrays or tape libraries.

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

System Controller

Storage Shelf

Storage Shelf

Storage Shelf

Scalable Architecture

Fibre Channel

Serial ATA

Shelf Controller

Stick Controller

Stick Controller

Stick Controller

Storage Personality

Policy Engines, Performance/Load Balancing, Object Management

Storage Shelf

Data Protection

RAID Support, Data Caching, Power/Device Management

System Controller

SCSI

iSCSI

NFS

CIFS

Data Path Routing

Protocol Router, Device Path Configuration, Device/Environment Monitoring

Canister

Serial ATA

Serial ATA

Storage Controller

Powered ON Disk

Powered OFF Disk
Extending MAID: Power-Managed 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

PM-RAID Stripe Size = 1
Increased Disk and System Reliability

- Effective drive service life
  - Increases with decreasing duty ratio*

- Increases Data Reliability

- Explicitly manage start stops
  - $\leq 50K$ over service life
  - Match to application need

- Use disk density for availability
  - Spares to replenish failed drives
  - Rebuild data transparently
  - Data Revitalization for Long-Term Data

*Power duty cycle ratio = # of powered-ON drives/# of powered-OFF drives
Extending MAID: Data Reliability

- Device health monitoring
- Proactive data management: closed-loop control
- Revitalize data on disk for long-term data retention
- System data integrity mechanisms
Filling the Performance Gap

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

- Aggregation Benefits
  - Multiple streams/shelf
  - Multiple shelves

- Results
  - 90 MBs/single stream uncompressed/shelf
  - Over 720 MBs for 8-shelf system
  - Further Improvements: Tuning, Compression
Exploiting Disk Performance: Access Time

- Access Time: 10X better than tape
  - Powered ON Drive: access time is in ms
  - Powered OFF Drive: 6s spin-up time, 10s data access

### Random Access of File/Drive: uncompressed 100 MB

<table>
<thead>
<tr>
<th>9940B TAPE: streaming @ 30MB/sec</th>
<th>Load 18 sec</th>
<th>Ave. Time to 1st Byte* 41 sec</th>
<th>File Xfer 3.3 sec</th>
<th>Unload 18 sec</th>
<th>Total: 80 sec</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>SATA 7200 RPM Disk: streaming @ 40 MB/sec – increases with RAID</th>
<th>Spin up ms-6 sec</th>
<th>Ave. Time to 1st Byte 0.1 sec</th>
<th>File Xfer 2.5 sec</th>
<th>Spin down 0.1 sec</th>
<th>Total (power-off AND cache miss): 8.7 sec*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (power-on OR disk cache): 2.7 sec</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Ave. time to first byte on tape depends on location of file (0 - 90 s)*
Simplifying Information Management

- CAPEX
- Performance
- Data Protection, Longevity
- Minimize Data Movement
- All Data Accessible All Time
- Reduced Management Cost
First Commercial MAID: Revolution 200T

Purpose Built Architecture
- Optimized cost, density, performance and reliability
- Spin disks only when necessary
- Performance for bandwidth, not IOPS

Enterprise Reliability
- Long term reliability with SATA
- Validation and revitalization of data
- RAID protected

8 shelves,
8 canisters each,
14 drives each
896 drives
224 TB in a single rack! (uncompressed)

Performance
2.4TB / HR
~22TB/sq. ft.
Conclusions

▪ MAID: exploits best of disk and tape
▪ Extensions meet secondary storage needs
  - Capacity and Cost
  - Reliability: Power-Managed RAID
  - Performance: Bandwidth, Access Time
  - Serviceability
  - Retention
▪ Filling the Gap in the Storage Hierarchy
▪ Simplifying Long-Term Data Management