Tape Data Storage in Practice
Minnesota Supercomputing Institute

GlobusWorld 2018
Jeffrey McDonald, PhD
Associate Director for Operations
HPC Resources

- **Mesabi**
  - Cores: > 18,000
  - Memory: 67 TB
  - Accelerators: 80 K40 gpGPUs
  - Peak: 675 TF
  - 320 Gbps to Storage

- **New Technologies**
  - FPGAs
  - Nvidia GPUs
  - Intel PHI
  - Storage (Intel NVME)

- **MSI Users**
  - PI Accounts: 843
  - Users: > 4600

© 2015 Regents of the University of Minnesota. All rights reserved.
Storage

- **High Performance Storage**
  - 7.2 PB Usable
  - 48GB/s read/write
  - Available on HPC resources

- **Tier-2 Storage**
  - 3.1 PB Usable
  - Available via Amazon’s S3 interface
  - Available anywhere in the world

- **Archive Storage**
  - > 3.5 PB tape-based storage
  - offline storage

- **All available via Globus**
Who Uses MSI?

Storage Allocated by Discipline

- Biology: 19%
- Health Sciences: 14%
- Genetics: 11%
- Engineering: 12%
- Earth Sciences: 10%
- Economics: 0%
- Computer Science: 8%
- Chemistry: 8%
- Agriculture: 4%
- Astronomy: 6%
- Veterinary Medicine: 1%
- Social Science: 1%
- Physics: 2%
- Other: 3%
- Mathematical Sciences: 1%
- Informatics: 3%

© 2015 Regents of the University of Minnesota. All rights reserved.
- System: Spectra T950, three frames.
- Blackpearl Enterprise model, LTO-7 model.
- Blackpearl has two dedicated LTO-7 drives.
- Tape Library Logically partition between TSM and Blackpearl.
- Dedicated 40 GbE network link to the blackpearl.
Tape Storage as a Service

- Three PIs purchased (four) LTO-7 tape buckets.
- Each unit is for a 5-year term, is by-tape, with LTO-7 (assume 12 TB of data storage) and include redundancy.
- Purchase rate for $456.12 per unit
- Cost estimate based upon occupancy of a full frame.
Policies Considerations

• One ‘bucket’ per group.
• Tools to chunk data to adequate sizes
  – Duplicity
  – Tar
• Service model always; MSI never relinquishes the tape and we will not export.
• Globus provides data movement channel.
Tape Considerations

- LTO-7 error rates 1/200,000 tapes/1.25 Exabytes (5-9s of reliability), 1000x better than disk (LTO-consortium)
- MSI is conservative with storage, so we elected to make two copies of all user data
- Data can be deleted; space is not recovered unless bucket is re-initialized

<table>
<thead>
<tr>
<th>Storage Domains</th>
<th>Tape Used</th>
<th>Tape Allocated</th>
<th>Pool Used</th>
<th>Pool Allocated</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Storage</td>
<td>879.0 MB</td>
<td>5.2 TB</td>
<td>0.0 Bytes</td>
<td>0.0 Bytes</td>
</tr>
<tr>
<td>BackupDomain</td>
<td>24.4 GB</td>
<td>10.4 TB</td>
<td>0.0 Bytes</td>
<td>0.0 Bytes</td>
</tr>
<tr>
<td>MSI ISO primary copy</td>
<td>10.1 TB</td>
<td>15.6 TB</td>
<td>0.0 Bytes</td>
<td>0.0 Bytes</td>
</tr>
<tr>
<td>MSI ISO redundant copy</td>
<td>10.1 TB</td>
<td>15.6 TB</td>
<td>0.0 Bytes</td>
<td>0.0 Bytes</td>
</tr>
<tr>
<td>Tape Copy First</td>
<td>29.0 TB</td>
<td>36.5 TB</td>
<td>0.0 Bytes</td>
<td>0.0 Bytes</td>
</tr>
<tr>
<td>Tape Copy Second</td>
<td>21.3 TB</td>
<td>26.1 TB</td>
<td>0.0 Bytes</td>
<td>0.0 Bytes</td>
</tr>
</tbody>
</table>

- Configuration uses ‘redundant’ copy configuration.
- Placement of data is controlled by the redundant copy configuration.
- Prevents ‘mixing’ of user’s data
- Forces the redundant copy.
## Breaking down the hardware costs

<table>
<thead>
<tr>
<th>Item</th>
<th>Quantity</th>
<th>Item cost</th>
<th>Extended cost</th>
<th>% to Service</th>
<th>Term</th>
<th>Cost /Year</th>
<th>TB</th>
<th>Cost/TB/Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library Frame</td>
<td>1</td>
<td>25,000</td>
<td>25,000</td>
<td>100.00%</td>
<td>12</td>
<td>2,083</td>
<td></td>
<td>$0.3245067</td>
</tr>
<tr>
<td>Robotics from initial purchase</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>100.00%</td>
<td>12</td>
<td>0</td>
<td></td>
<td>$0.0000000</td>
</tr>
<tr>
<td>Robotics expansion</td>
<td>1</td>
<td>6,000</td>
<td>6,000</td>
<td>100.00%</td>
<td>12</td>
<td>500</td>
<td></td>
<td>$0.0778816</td>
</tr>
<tr>
<td>Library Maintenance</td>
<td>1</td>
<td>10,000</td>
<td>10,000</td>
<td>100.00%</td>
<td>1</td>
<td>10,000</td>
<td></td>
<td>$1.5576324</td>
</tr>
<tr>
<td>LTO-7 tape</td>
<td>1070</td>
<td>106.00</td>
<td>113,420</td>
<td>100.00%</td>
<td>5</td>
<td>22,684</td>
<td></td>
<td>$3.5333333</td>
</tr>
<tr>
<td>LTO-7 drives</td>
<td>2</td>
<td>10,500.00</td>
<td>21,000</td>
<td>100.00%</td>
<td>7</td>
<td>3,000</td>
<td></td>
<td>$0.4672897</td>
</tr>
<tr>
<td>BlackPearl server</td>
<td>1</td>
<td>100,000.00</td>
<td>100,000</td>
<td>100.00%</td>
<td>5</td>
<td>20,000</td>
<td></td>
<td>$3.1152648</td>
</tr>
<tr>
<td>BlackPearl maintenance contract</td>
<td>1</td>
<td>4,800.00</td>
<td>4,800</td>
<td>100.00%</td>
<td>1</td>
<td>4,800</td>
<td></td>
<td>$0.7476636</td>
</tr>
<tr>
<td>Slot activation</td>
<td>1070</td>
<td>65.00</td>
<td>69,550</td>
<td>100.00%</td>
<td>12</td>
<td>5,796</td>
<td>6,420</td>
<td>$0.9027778</td>
</tr>
<tr>
<td><strong>Total Hardware</strong></td>
<td><strong>349,770</strong></td>
<td></td>
<td><strong>68,863</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>$10.7263499</strong></td>
</tr>
</tbody>
</table>
Personnel Costing

- 10% of FTE for Technical Person.
- 10% of FTE for User Support Person.
- 5% FTE for management.

This adds a base of $4.48 cost per TB per Year.

The total cost per TB (no utilities) is $15.20 per TB per year (w/redundancy) or 456.12 per 5-years.
Power Considerations

• The archive and blackpearl use about 500W (averaged over duty cycle and share of T950). Independent of Capacity--Tapes don’t use power.

• Compare to 6 PB spinning archive, 500-12 TB drives (no redundancy) 5W/drive = 2.5kW of power. Scales with capacity, no servers included here.

• 5 year lifetime savings: $20k power + $60k cooling. Saving grow with archive size (assuming at least 5 servers for spinning).
Globus Transfers

• Data set between 2017-06-30 and 2018-04-09.
• 131 Successful transfers.
• 28 Failed transfers (various reasons).
• 40 TB transferred.
• Rates as high as 95 MB/sec and average 30 MB/sec.
• Highly directional, one in 28 from archive.
### Users/File Profiles

These are the users who purchased tapes in 2017 and their utilization. The primary concern for us was that the users aggregate their data to limit object counts on the Blackpearl system. MSI staff would spend a couple of hours with each user. Most transfers occur between our tier2 and tier3 storage for these users.

<table>
<thead>
<tr>
<th>User Dept/Type</th>
<th>Tapes</th>
<th>Files (Size)</th>
<th>Average Size</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geneticist</td>
<td>2</td>
<td>3862 (16 TB)</td>
<td>4.13 GB</td>
<td>mostly large dna sequence files</td>
</tr>
<tr>
<td>Biologist</td>
<td>1</td>
<td>4 (150 GB)</td>
<td>36 GB</td>
<td>project archives</td>
</tr>
<tr>
<td>Biologist</td>
<td>1</td>
<td>56 (600 GB)</td>
<td>11 GB</td>
<td>large dna sequence files</td>
</tr>
</tbody>
</table>
Conclusions

- MSI should spend some money on Marketing!
- Globus is providing MSI with important tools toward an HSM-like framework for storage.
- Expanding the framework for some other appropriate use cases.
- Tape continues to offer a compelling use case for storage.
Backup Slides