The Outlook for Archival Storage at CERN

Michael Davis and Germán Cancio

Storage Group • IT Department • CERN
19 October 2017
Overview

- Roadmap for Tape
- Roadmap for Disk
- Alternatives to Tape
- Archival at CERN
Tape Market

Market dominated by LTO consortium (≈95%)

- **Drives**: IBM, HP, Quantum. Oracle resells LTO drives from IBM, HP.
- **Media**: Fujifilm, Sony. TDK exited the tape media market in 2014.

**Enterprise Tape**: IBM, Oracle (≈4%)

- **Latest IBM Drive**: TS1155 (introduced May 2017; 15 Tb, 350 Mb/s)
- **Latest Oracle Drive**: T10KD (introduced September 2013; 8TB, 250 Mb/s)

**Large-scale libraries** (≥ 10K slots)

- Oracle, IBM, Spectra Logic, now also Quantum
From Giant MagnetoResistive (GMR) to Tunnel MagnetoResistive (TMR)

- TMR is $6 \times$ more sensitive than GMR
- HDDs have been using TMR since 2005
- GMR has reached its density limits for tape
- IBM TS1155 uses TMR
- LTO-8 will use it
Tape Track Density
Source: IBM DACH (2015)

- Scaled bit cells:
- Magnified 25x:

- Tremendous potential for future scaling of tape track density
- Key technologies: improved track follow servo control, improved media, reader, data channel
Aereal Density Trends
Source: INSIC (2016)
Aereal Density Trends
Source: INSIC (2016)

04/2015 Fuji/IBM Demo (using BaFe)
123 Gb/in²
220 Tb tape
Aereal Density Trends
Source: INSIC (2016)

04/2015 Fuji/IBM Demo (using BaFe)
123 Gb/in²
220 Tb tape

08/2017 Sony/IBM Demo (using CoPtCr)
201 Gb/in²
330 Tb tape

IBM TS1155
9.6 Gb/in²
Aereal Density Trends
Source: INSIC (2016)

04/2015 Fuji/IBM Demo (using BaFe) 123 Gb/in² 220 Tb tape
08/2017 Sony/IBM Demo (using CoPtCr) 201 Gb/in² 330 Tb tape
IBM TS1155 9.6 Gb/in²

30% CAGR seems realistic
# LTO/IBM Enterprise Tape Roadmap

**Source:** IBM DACH (2015)

## LTO Generations

<table>
<thead>
<tr>
<th>LTO Generations</th>
<th>LTO-5</th>
<th>LTO-6</th>
<th>LTO-7</th>
<th>LTO-8</th>
<th>LTO-9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New Format Capacity (Native)</strong></td>
<td>1.5 TB (L5)</td>
<td>2.5 TB (L6)</td>
<td>6 TB (L7)</td>
<td>Up to 12.0 TB (L8)</td>
<td>Up to 25 TB (L9)</td>
</tr>
<tr>
<td><strong>Other Format Capacities (Native)</strong></td>
<td>800 GB (L4) (400 GB L3 R/O)</td>
<td>1.5 TB (L5) (800 GB L4 R/O)</td>
<td>2.5 TB (L6) (1.5 TB L5 R/O)</td>
<td>6 TB (L7) (2.5 TB L6 R/O)</td>
<td>Up to 12 TB (L8) (6 TB L7 R/O)</td>
</tr>
<tr>
<td><strong>Native Data Rate</strong></td>
<td>140 MB/s</td>
<td>160 MB/s</td>
<td>300 MB/s</td>
<td>Up to 472 MB/s</td>
<td>Up to 708 MB/s</td>
</tr>
</tbody>
</table>

## TS1100 Generations

<table>
<thead>
<tr>
<th>TS1100 Generations</th>
<th>TS1130</th>
<th>TS1140</th>
<th>TS1150</th>
<th>TS1155</th>
<th>TS1160</th>
<th>TS1165</th>
<th>TS1170</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New Format Capacity (Native)</strong></td>
<td>1 TB (JB) 640 GB (JA)</td>
<td>4 TB (JC) 1.6 TB (JB)</td>
<td>10 TB (JD) 7 TB (JC)</td>
<td>15 TB (JD)</td>
<td>18-20TB (JE) 15-17 TB (JD) 10-12 TB (JC)</td>
<td>~30 TB (JE) 15-17 TB (JD) 10-12 TB (JC)</td>
<td>30-40 TB (JE)</td>
</tr>
<tr>
<td><strong>Other Format Capacities (Native)</strong></td>
<td>700 GB (JB) 500 GB (JA) 300 GB (JA)</td>
<td>1 TB (JB) 700 GB (JB) (All JA R/O)</td>
<td>4 TB (JC)</td>
<td>TBD (JC)</td>
<td>10 TB (JD) 7 TB (JC) 4 TB (JC)</td>
<td>TBD</td>
<td>18-20 TB (JE) 15-17 TB (JD) 10-12 TB (JC)</td>
</tr>
<tr>
<td><strong>Native Data Rate</strong></td>
<td>160 MB/s</td>
<td>250 MB/s</td>
<td>360 MB/s</td>
<td>360 MB/s</td>
<td>Up to 500 MB/s</td>
<td>Up to 500 MB/s</td>
<td>Up to 1000 MB/s</td>
</tr>
</tbody>
</table>

---

Any statements regarding IBM’s future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
## LTO/IBM Enterprise Tape Roadmap

**Source:** IBM DACH (2015)

### LTO Generations

<table>
<thead>
<tr>
<th>LTO Generations</th>
<th>LTO-5</th>
<th>LTO-6</th>
<th>LTO-7</th>
<th>LTO-8</th>
<th>LTO-9</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Format Capacity (Native)</td>
<td>1.5 TB (L5)</td>
<td>2.5 TB (L6)</td>
<td>6 TB (L7)</td>
<td>Up to 12.0 TB (L8)</td>
<td>Up to 25 TB (L9)</td>
</tr>
<tr>
<td>Other Format Capacities (Native)</td>
<td>800 GB (L4) (400 GB L3 R/O)</td>
<td>1.5 TB (L5) (800 GB L4 R/O)</td>
<td>2.5 TB (L6) (1.5 TB L5 R/O)</td>
<td>6 TB (L7) (2.5 TB L6 R/O)</td>
<td>Up to 12 TB (L8) (6 TB L7 R/O)</td>
</tr>
<tr>
<td>Native Data Rate</td>
<td>140 MB/s</td>
<td>160 MB/s</td>
<td>300 MB/s</td>
<td>Up to 472 MB/s</td>
<td>Up to 708 MB/s</td>
</tr>
</tbody>
</table>

### TS1100 Generations

<table>
<thead>
<tr>
<th>TS1100 Generations</th>
<th>TS1130</th>
<th>TS1140</th>
<th>TS1150</th>
<th>TS1155</th>
<th>TS1160</th>
<th>TS1165</th>
<th>TS1170</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Format Capacity (Native)</td>
<td>1 TB (JB) 640 GB (JA)</td>
<td>4 TB (JC) 1.6 TB (JB)</td>
<td>10 TB (JD) 7 TB (JC)</td>
<td>15 TB (JD)</td>
<td>18-20 TB (JE) 15-17 TB (JD) 10-12 TB (JC)</td>
<td>~30 TB (JE) 15-17 TB (JD) 10-12 TB (JC)</td>
<td>30-40 TB (JE)</td>
</tr>
<tr>
<td>Other Format Capacities (Native)</td>
<td>700 GB (JB) 500 GB (JA) 300 GB (JA)</td>
<td>1 TB (JB) 700 GB (JB) (All JA R/O)</td>
<td>4 TB (JC)</td>
<td>TBD (JC)</td>
<td>TBD</td>
<td>TBD</td>
<td>18-20 TB (JE) 15-17 TB (JD) 10-12 TB (JC)</td>
</tr>
<tr>
<td>Native Data Rate</td>
<td>160 MB/s</td>
<td>250 MB/s</td>
<td>360 MB/s</td>
<td>360 MB/s</td>
<td>Up to 500 MB/s</td>
<td>Up to 500 MB/s</td>
<td>Up to 1000 MB/s</td>
</tr>
</tbody>
</table>

**Note:** Released Q2 2017

---

Any statements regarding IBM’s future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
LTO/IBM Enterprise Tape Roadmap
Source: IBM DACH (2015)

<table>
<thead>
<tr>
<th>LTO Generations</th>
<th>LTO-5</th>
<th>LTO-6</th>
<th>LTO-7</th>
<th>LTO-8</th>
<th>LTO-9</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Format Capacity (Native)</td>
<td>1.5 TB (L5)</td>
<td>2.5 TB (L6)</td>
<td><strong>6 TB (L7)</strong></td>
<td><strong>Up to 12.0 TB (L8)</strong></td>
<td>Up to 25 TB (L9)</td>
</tr>
<tr>
<td>Other Format Capacities (Native)</td>
<td>800 GB (L4) (400 GB L3 R/O)</td>
<td>1.5 TB (L5) (800 GB L4 R/O)</td>
<td>2.5 TB (L6) (1.5 TB L5 R/O)</td>
<td>6 TB (L7) (2.5 TB L6 R/O)</td>
<td><strong>8</strong></td>
</tr>
<tr>
<td>Native Data Rate</td>
<td>140 MB/s</td>
<td>160 MB/s</td>
<td>300 MB/s</td>
<td>Up to 472 MB/s</td>
<td>Up to 700 MB/s</td>
</tr>
</tbody>
</table>

TS1100 Generations

<table>
<thead>
<tr>
<th>TS1100 Generations</th>
<th>TS1130</th>
<th>TS1140</th>
<th>TS1150</th>
<th>TS1155</th>
<th>TS1160</th>
<th>TS1165</th>
<th>TS1170</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Format Capacity (Native)</td>
<td><strong>1 TB (JB)</strong> 640 GB (JA)</td>
<td><strong>4 TB (JC)</strong> 1.6 TB (JB)</td>
<td><strong>10 TB (JD)</strong> 7 TB (JC)</td>
<td><strong>15 TB (JD)</strong></td>
<td>18-20TB (JE) 15-17 TB (JD) 10-12 TB (JC)</td>
<td>~30 TB (JE) 15-17 TB (JD) 10-12 TB (JC)</td>
<td>30-40 TB (JE)</td>
</tr>
<tr>
<td>Other Format Capacities (Native)</td>
<td>700 GB (JB) 500 GB (JA) 300 GB (JA)</td>
<td>1 TB (JB) 700 GB (JB) (All JA R/O)</td>
<td><strong>4 TB (JC)</strong></td>
<td>TBD (JC)</td>
<td>TBD</td>
<td>18-20 TB (JE) 15-17 TB (JD) 10-12 TB (JC)</td>
<td></td>
</tr>
<tr>
<td>Native Data Rate</td>
<td>160 MB/s</td>
<td>250 MB/s</td>
<td>360 MB/s</td>
<td>360 MB/s</td>
<td>Up to 500 MB/s</td>
<td>Up to 500 MB/s</td>
<td>Up to 1000 MB/s</td>
</tr>
</tbody>
</table>

Any statements regarding IBM’s future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Tape Roadmap</th>
<th>Disk Roadmap</th>
<th>Alternatives to Tape</th>
<th>Archival at CERN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>oooooo</td>
<td>oooo</td>
<td>oooooo</td>
<td>ooooo</td>
</tr>
</tbody>
</table>

Oracle Enterprise Tape Roadmap
Tape Drive Head Manufacturing
Source: Spectra Logic (2017)

Supply Chain update

Drive Head Manufacturers

- **ORACLE**
  - 123,000 Heads Per Year

- **IBM**
  - 108,000 Heads Per Year

Tape Drive Manufacturers

- **ORACLE® T10K Drives**
  - 3,000 Per Year

- **HP LTO Drives**
  - 120,000 Per Year

- **IBM LTO Drives**
  - 100,000 Per Year

- **IBM TS Drives**
  - 8,000 Per Year
Supply Chain update

Drive Head Manufacturers

- **Oracle**: 2,500 Heads Per Year
- **IBM**: 228,000 Heads Per Year

Tape Drive Manufacturers

- **Oracle® T10K Drives**: 2,500 Per Year
- **HP LTO Drives**: 120,000 Per Year
- **IBM LTO Drives**: 100,000 Per Year
- **IBM TS Drives**: 8,000 Per Year

**Combined IBM Heads**: 228,000

Source: Spectra Logic (2017)
Media is ≈50% of tape TCO (besides drive/library hardware and maintenance)

Media costs ≈10–15 CHF/Tb, but price decline has slowed (−20% per year over last four years)

LTO media shipments decreasing for 10 years:
Media is ≈50% of tape TCO (besides drive/library hardware and maintenance)

Media costs ≈10–15 CHF/Tb, but price decline has slowed (−20% per year over last four years)

LTO media shipments decreasing for 10 years:
Tape Roadmap: Summary

- Outlook very positive in terms of projected improvements in tape capacity

**BUT:**

- Only one remaining major manufacturer/R&D of tape drive technology (IBM)
- Only two remaining media manufacturers (Fujifilm, Sony)
- Will this market continue to sustain tape research (new heads/new media) and production?
Spinning Disk Market
Source: CERN (2017)

- **Manufacturers:** WD (41%), Seagate (37%), Toshiba (22%)
- ≈600 Eb/year, decreasing since 2010. 10% of sales are “Nearline” drives, as used in HEP
- Increased competition from Cloud/SSD for enterprise disks
Spinning Disk Technology

Technology Evolution

- Shingled Magnetic Recording (SMR) disks: 2013
- Helium-filled disks (more platters): 2013
- Heat-Assisted Magnetic Recording (HAMR) disks: ≈2018

Capacity/Pricing Evolution

- Next 3–12 months: 14–16 Tb
- By 2020: 20 Tb
- By ≈2025: 100 Tb with HAMR/HDMR
- “Nearline” disks cost ≈35–40 CHF/Tb (−14%/year)
- Very shaky price evolution
Spinning Disk Technology
Source: ASTC (2016)

- PMR = Perpendicular Magnetic Recording
- PMR+ = PMR with Two-Dimensional Magnetic Recording (TDMR) and/or Shingled Magnetic Recording (SMR)
- HAMR+ = Heat Assisted Magnetic Recording with TDMR and/or SMR
- HDMR = Heated-Dot Magnetic Recording (BPMR+HAMR+TDMR)

Graph showing the roadmap for areal density (Tb/in\(^2\)) from 2013 to 2025 with future technology milestones.
Spinning Disk Technology
Source: ASTC (2016)

PMR = Perpendicular Magnetic Recording
PMR$^+$ = PMR with Two-Dimensional Magnetic Recording (TDMR) and/or Shingled Magnetic Recording (SMR)

HAMR$^+$ = Heat Assisted Magnetic Recording with TDMR and/or SMR

HDMR = Heated-Dot Magnetic Recording (BPMR+HAMR+TDMR)

30% CAGR—realistic?
SSD Market

- Shipped capacity in 2016: ≈45 Eb
- SSD shipped capacity is ≈7.5% of HDD shipped capacity. Expected to grow to 20% by 2021.
- Large investments required for SSD manufacturing ($200–300 Bn)
- SSD price/Tb expected to be $O(10) \times$ HDD price/Tb price for the foreseeable future
Disk Servers for Archival

Current CERN EOS disk servers

- One CPU node and $2 \times 24$ enterprise-class disks
- JBOD with 2 replicas
- Disk cost is 75%
- TCO is $\approx 3 \times$ tape CHF/Tb

Investigation into “Monster” disk servers to optimise disk-to-infrastructure cost ratio

- Use commodity disks to optimise CHF/Tb
- à la BackBlaze (30–40% cheaper) [BackBlaze (2015)]
Massive Array of Inexpensive Disks (MAID)

Source: CERN (2016)

- Testing 192 disks on one server (2 trays with $8 \times 24$ HDDs each)
- Up to 1.1 Pb raw (using 6 Tb disks)
- Evaluate different file system and redundancy layouts (RAID, erasure coding,...) → $f$(capacity, reliability, performance)
Massive Array of Inexpensive Disks (MAID)
Source: CERN (2016)

- No warranty/certification for our use case
- Measure reliability of different models/vendors (integrity, failure rates) using SMART and EOS monitoring
- Compensate lower reliability with higher redundancy
- Review operating procedures: let disks die rather than replacing
- Two servers in production for ALICE (successfully so far)
Optical

Inside Facebook’s Blu-Ray Cold Storage Data Center
BY RICH MILLER - JUNE 30, 2015 — 14 COMMENTS

Archival Disc Roadmap

<table>
<thead>
<tr>
<th>Capacity</th>
<th>Signal Processing Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>300GB</td>
<td>High Linear Density</td>
</tr>
<tr>
<td></td>
<td>(Multi Level Recording Technology)</td>
</tr>
<tr>
<td>500GB</td>
<td>High Linear Density</td>
</tr>
<tr>
<td></td>
<td>(Inter Symbol Interference Cancellation Technology)</td>
</tr>
<tr>
<td>1TB</td>
<td>Narrow Track Pitch (Crosstalk Cancellation Technology)</td>
</tr>
</tbody>
</table>

Basic Specification

Double-Sided Disc Technology
λ=405nm, NA=0.85, Layer Structure: 3Layers/side
Optical Libraries
Source: Sony

Base Unit ADZ-DRS1
1100kg (2425lb)
Without drives, drive servers & power supply

Robotics Unit ADZ-TRS1
600kg (1323lb)

Expansion Unit ADZ-EXS1

Optical Array Drive ADZ-D3
2.5kg (5.5lb)

Media Tray ADZ-M3
1.52kg (3.4lb)
incl. 64 Archival Discs

1,770mm (69.7")
930mm (36.6")
937 mm (36.9")
2,400 mm (94.5")
1,876 mm (73.9")
2,100 mm (82.7")
Holographic
Source: Akonia Holographics (2015)

- Record information across media volume, not just surface
- High densities using different recording angles, wavelengths, position on single media location
- Potentially, $O(Gb)/mm^3$, fast R/W rates
- Prototypes, even ECMA standards
- No products on the market, nor any signs of upcoming ones
CERN Tape Archive
Source: CERN/HEPiX (2017)

One EOS instance per experiment

- EOS namespace and redirector
- EOS disk server
  - Tape files appear in EOS namespace as replicas.
  - EOS workflow engine glues EOS to CTA

Central CTA instance

- CTA front-end
  - Archive and retrieve requests
- CTA metadata
  - Schedule information and tape file locations
- CTA tape server
  - Files

Tape file catalogue, archive / retrieve queues, tape drive statuses, archive routes and mount policies
CERN Tape Archive
Source: CERN/HEPiX (2017)

- **Q4 2017**
  - Additional and redundant backups of AFS/NFS and LEP data

- **Q2 2018**
  - Scaling tests ($10^6$ files runs, both direction, 13 real + 90 virtual drives tests)
  - Ready for small experiments

- **Q4 2018**
  - Ready for LHC experiments

CTA and CASTOR use the same tape format. Only metadata will need to be migrated. No files will need to be copied between tapes.
Recommended Access Order
Source: Cristina Moraru (2017)

Recall Files in Sequential Order

Recall Files in Recommended Access Order
Recommended Access Order
Source: Cristina Moraru (2017)
Recommended Access Order
Source: Cristina Moraru (2017)
Summary

- Tape is still the most cost-effective archival solution for HEP

**BUT:**

- Concerns about the long-term sustainability of a tape market dominated by a single technology provider

**Mitigating Market Risks**

- Investigate massive commodity disk setup (MAID) for archival
- Keep an eye on optical storage
References

Josef Weingand/IBM. The Future of Tape! Presentation by IBM DACH (Data Protection and Retention Group), 2015.


References II

Advanced Storage Technology Consortium (ASTC).
Disk Technology Roadmap, 2016.

Tim Nufire.
The Hardware Inside B2 Cloud Storage/Storage Pod 5.0, November 2015.
BackBlaze Blog.

Roberto Valverde Cameselle/CERN.
“Monster” Node Testing @ CERN IT. 2016.

Sony.
Sony Everspan Library Specifications.

