Magnetic Tape for Mass Storage in HEP

Scientific Data and Computing Center
Brookhaven National Laboratory
Shigeki Misawa
17 March 2021

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN SCIENCE ASSOCIATES

Tape Mass Storage at BNL

- Used for near-line and archival storage of NP/HEP data
- Multiple factors driving closer look at mass storage
  - Significantly higher bandwidth, larger data volumes and greater read access for ATLAS in the HL-LHC era and sPHENIX at RHIC.
  - Storage technologies evolving at different rates.
  - Optimizing future investments requires detailed plans

- Is tape still the media of choice for HEP/NP mass storage ?

Evaluating Options at the SDCC

- Estimate the cumulative cost of ownerships through 2030 for disk and tape based mass storage solutions
  - Provide feedback to system “users” on impact of requirements on cost of ownership and type of system deployed
  - Preliminary evaluation of system risks and benefits
  - Scope of the investigation does not cover compute models, data formats, or other user access optimizations
  - Scope of the investigation does not examine opportunities for backend/frontend cost optimizations
- Requirements taken from sPHENIX and ATLAS requirements through 2030

Estimating Cost of Disk vs Tape

- This cost analysis focuses on the system (not manpower) and assumes or includes the following:
  - "Greenfield" deployment - No legacy data/equipment
  - Evolution of LTO tape and hard disks taken from roadmaps, public vendor comments, or historical projections.
  - Assumes specific implementations of a tape and disk systems
  - Operational power and cooling costs
    - $0.06/KWH for "Industrial Electric Power" costs in NY
    - Estimated facility PUE (1.25) used to calculate cooling costs
  - Assumes 24x7 availability and operation of equipment (100% utilization)
  - Local area network costs are included

Disk/Tape System Configuration

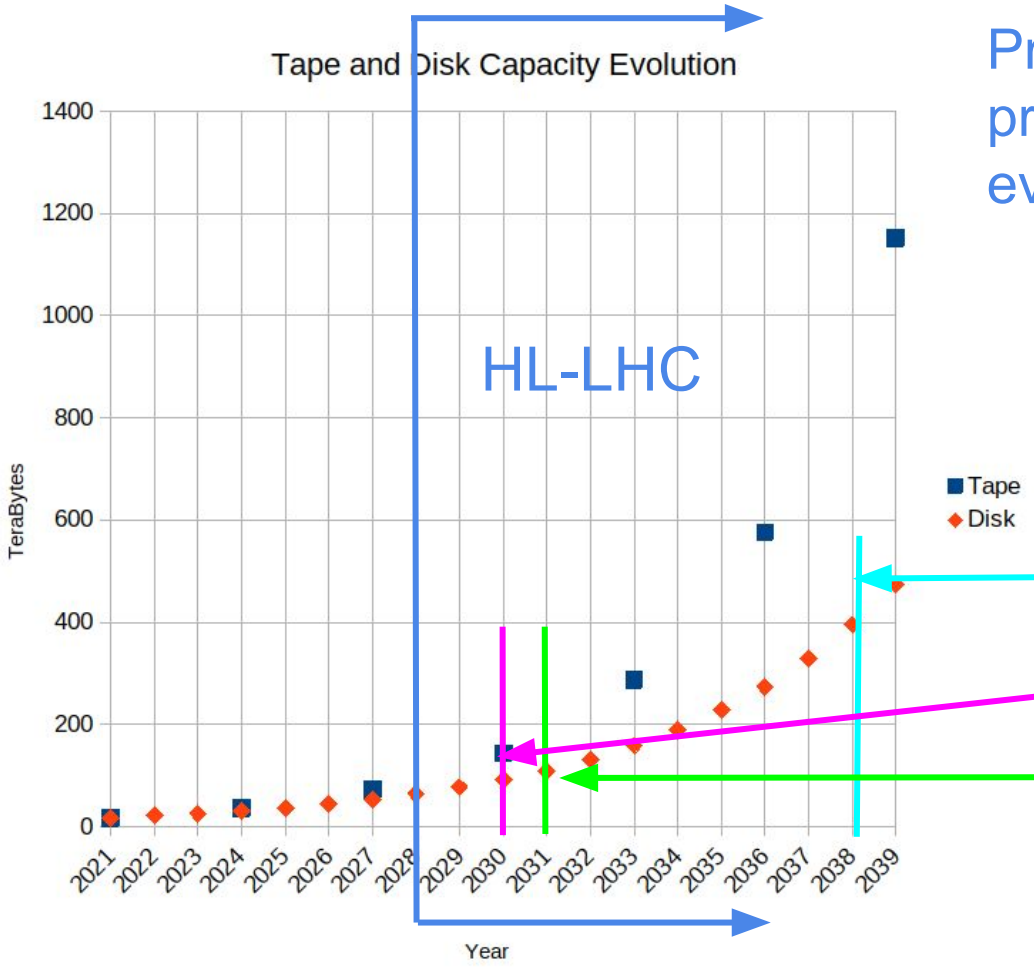
- Tape System
  - HPSS solution
  - Library w/ 20K cartridge capacity
  - Library deployed in 10K cartridge capacity increments
  - Maintain 5% free slot capacity at all times
  - 9 year media refresh cycle
    - LTO-N copied to LTO-(N+3)
  - Tape drives needed for media migration included
  - 20 year library life
- Disk System
  - Single QOS system
  - dCache/Lustre/Ceph solution
  - Maintain 10% free space
  - 20% EC/ECC space overhead
  - 500MB/sec “LUN” write performance
  - 10GB/sec capable servers
  - 400 disks per server

Technology Evolution

- Tape Parameters
  - Use LTO.org capacity roadmap
    - Capacity doubles each generation
  - 20%/yr reduction in $/TB for media
  - 20% tape drive BW increase per generation
  - Assume at best 90% of max tape drive bandwidth is achievable [1]
  - 3 years between generations
  - 20 year tape library life
- Disk Parameters
  - 20%/yr HDD capacity increase
  - 20%/yr reduction in $/TB
  - 5 year refresh cycle
  - Constant 250 MB/sec r/w bandwidth (single actuator)
  - Power Consumption
    - 10W - single actuator
    - 15W - dual actuator
  - PMR/HAMR disks (no SMR)

[1] Does not account for real world access and operational inefficiencies, e.g. sparse reads of tape media (skipping over files, random access of files)

Disk and Tape Roadmap Limitations



Projections beyond 2030 are problematic as predicting technology evolution is fraught with uncertainty

Tape technology demonstrated in the laboratory

End of LTO roadmap

Limits of HAMR disk technology

Tape/Disk Data Durability Differences

- Disk and tape are different and are not completely interchangeable
- Conventional disks are an “online” media
  - Disks are electrically energized and online at all times
  - “Disk copies aren’t backups”
  - 8+2 erasure code likely to be insufficient protection from data loss
- Tapes are an “offline” media
  - Tapes only exposed to electrical issues when mounted
  - Potentially safer from ransomware and accidental deletion
  - Theoretical tape media life is substantially longer than disk

Cost Comparisons

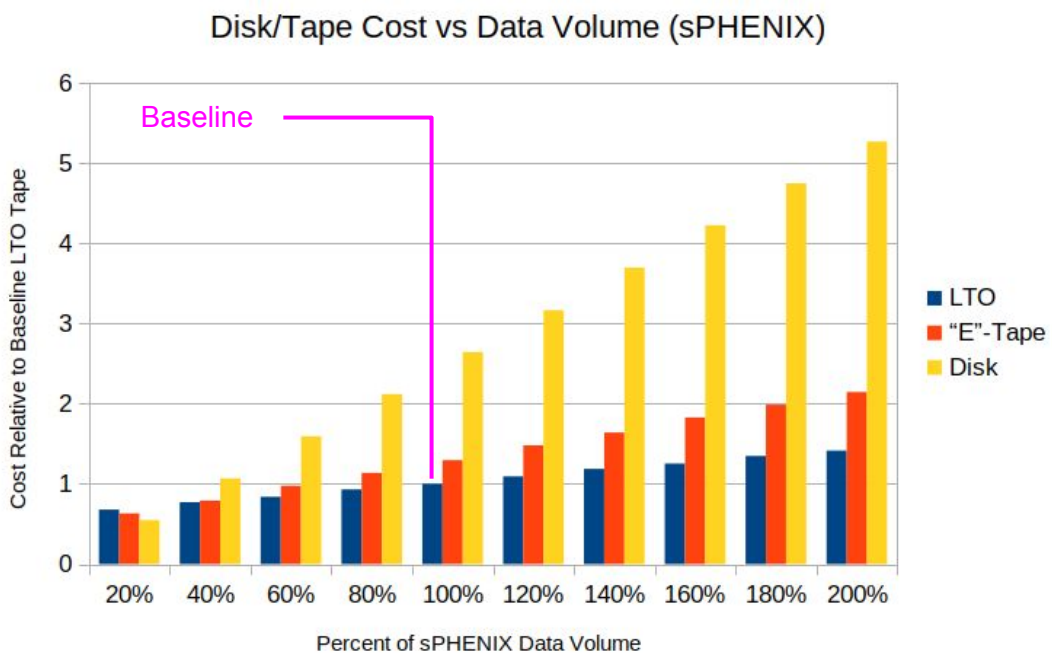
- LTO tape, disk, and “E” tape systems
  - “E” tape - Proxy for enterprise tape. 2x drive bandwidth, 2x media capacity, 2x $/TB and higher drive costs compared to LTO.
  - Insufficient information to evaluate real enterprise tape technology
- sPHENIX and ATLAS requirements vs time:

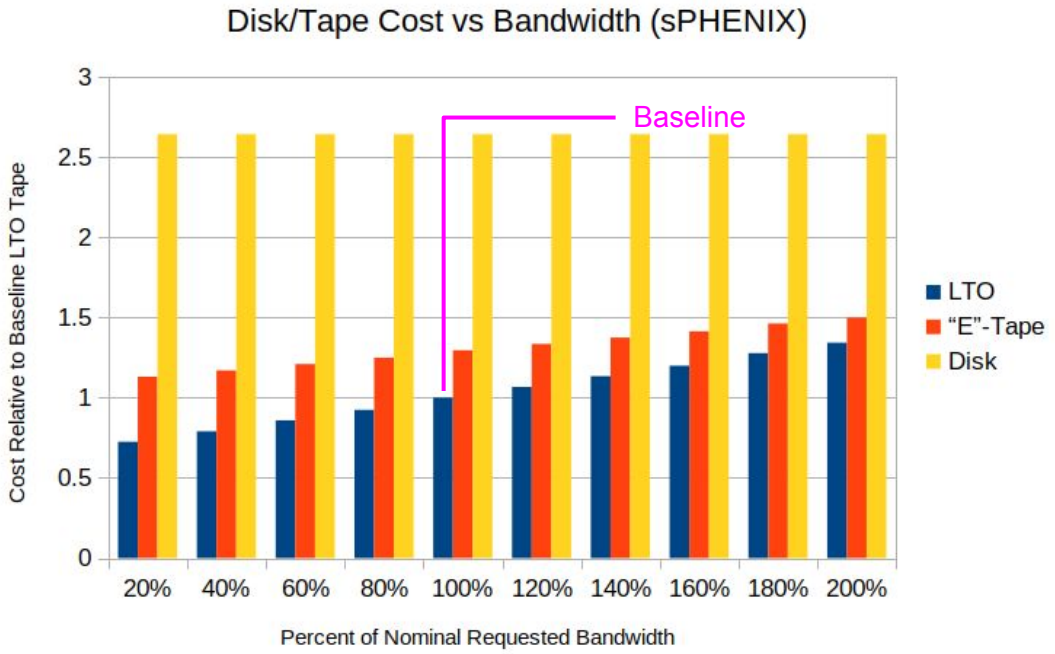
		2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
ATLAS	BW (GB/s)	10	10	10	10	10	10	30	30	30	30
	Data Volume (PB)	4	-2	13	11	14	11	58	112	168	149
sPHENIX	BW (GB/s)	1	16	31	31	31	31	0	0	0	0
	Data Volume (PB)	1	135	240	320	0	0	0	0	0	0

Volume of data added per year

- Variations in tape drive read efficiency (30% to 90%)
- Variations in BW and data volume requirements (20% to 200%)
- Analysis assumes only one experiment exists. (Mass storage system dedicated to one experiment, no cost sharing)

Relative Cost Comparison for sPHENIX



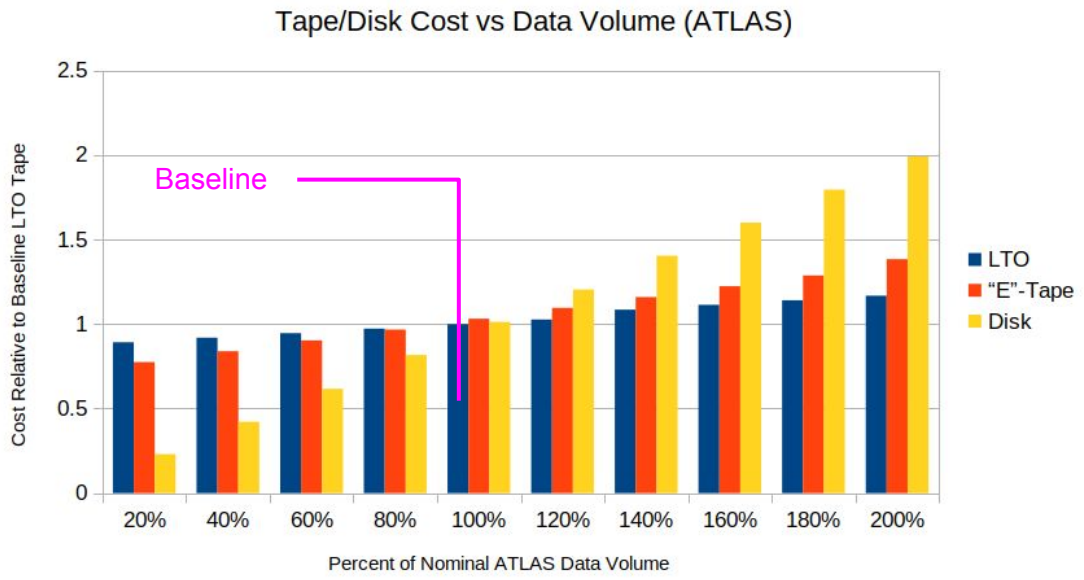


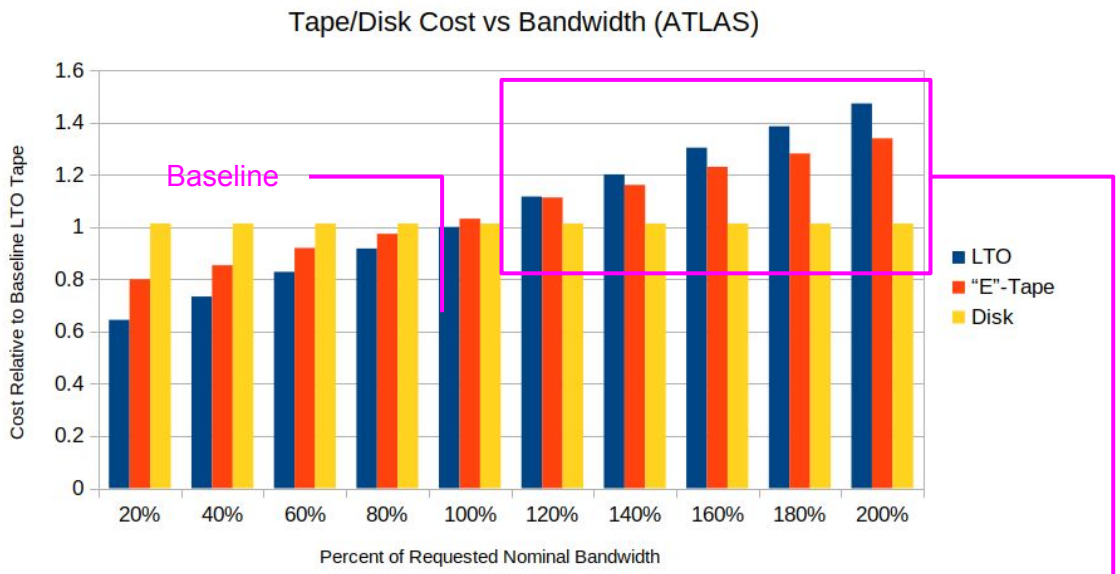
Cost increase as data volume increases, with “E” tape and disk costs rising faster than LTO costs. Higher cost of “E” tape and dIsk caused by higher $/TB for media

Tape costs increase with bandwidth, with LTO costs rising faster than “E” tape costs. High cost of disk caused by high cost of media ($/TB)

Analysis assumes NO legacy data

Relative Cost Comparison for ATLAS



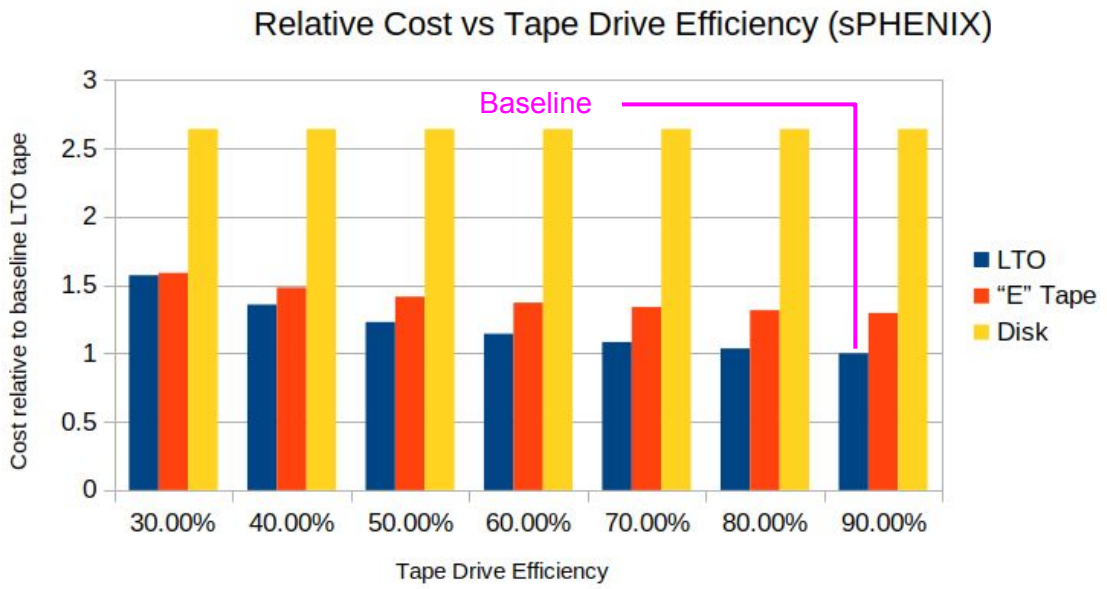


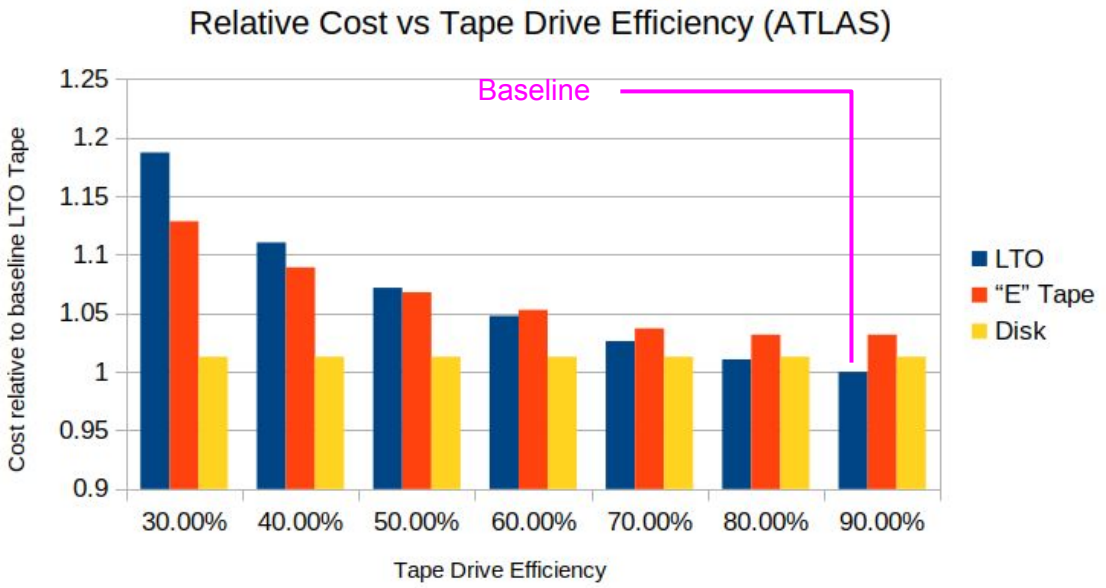
Disk and “E” tape costs increase relative to LTO tape system as data volume increases. Disk costs rise rapidly with increased data volume. Disk more competitive for ATLAS as HDD media cost are lower compared to sPHENIX time period

Disk and “E” tape costs decrease relative to LTO tape system as data rate increases. High access bandwidth makes tape more expensive than disk

Analysis assumes NO legacy data

Relative Cost vs Tape Drive Efficiency





Tape system costs decrease with increased tape drive efficiency. Read bandwidth is assumed to reach 100% of the bandwidth requirement. High cost of disk for sPHENIX caused by high data volumes relative to HDD capacity. (sPHENIX capacity requirements occurs earlier than ATLAS and are greater over the 2021-2030 time period)

Tape system costs decrease with increased tape drive efficiency. Note that changing inefficiency is assumed to affect only reads. Write inefficiency is assumed to be 90%. Reads do not reach 100% of the aggregate bandwidth requirement.. Lower cost of "E" tape and disk result of lower data volumes relative to media capacity. (ATLAS capacity requirements occur later in time than sPHENIX)

Analysis assumes NO legacy data

Comments

- Legacy data is a significant barrier to transitioning between disk and tape.
  - Upfront migration of data entails substantial investment in hardware and would take considerable time and effort. It will also result in periodic spikes in costs over time as migration hardware is refreshed.
  - Transition via normal life cycle requires supporting two systems for years, but avoids huge upfront costs and long term, cyclical spikes in costs associated with an upfront migration
- Cost sharing of tape infrastructure will reduce cost of tape for each user.

Conclusions

- High bandwidth access makes tape significantly less attractive compared to disk in out years.
- Poor utilization of tape drives (low efficiency) can noticeably increase the cost of tape.
- Continuous dialog with scientific experiments important to enable optimal and cost effective use of resources