Tape Storage at BNL

Pre-GDB - Tape Evolution

Scientific Data and Computing Center
Brookhaven National Laboratory
Shigeki Misawa
February 9, 2021

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Tape Mass Storage at BNL

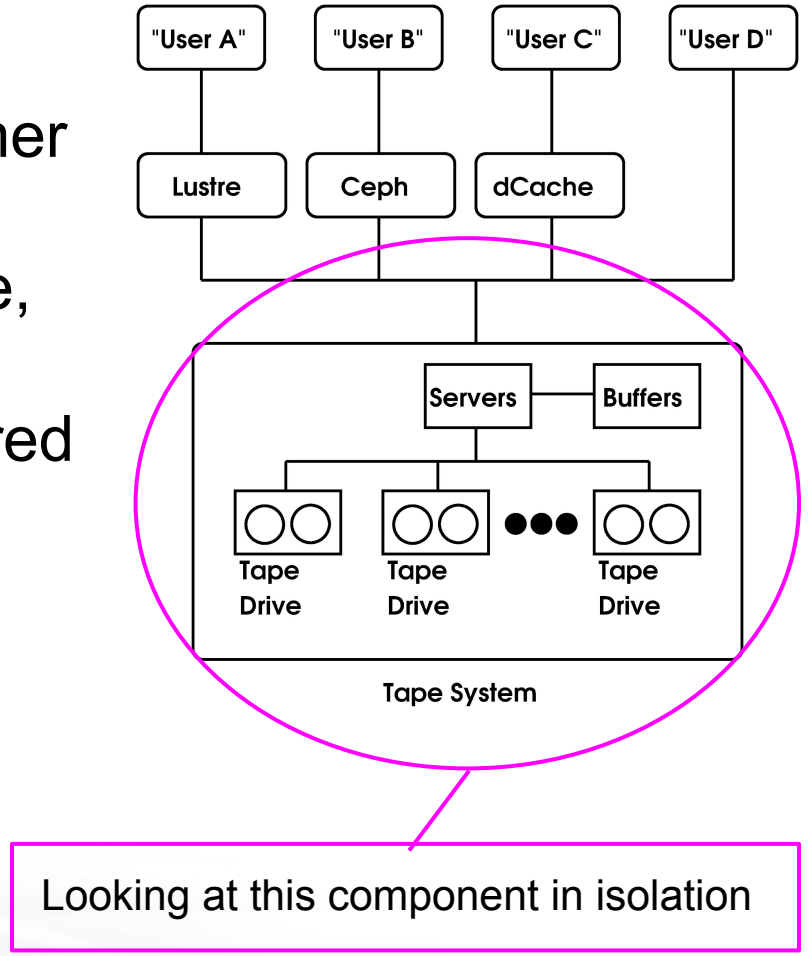
- Used for near-line and archival storage of NP/HEP data
- Multiple factors driving closer look at mass storage
  - Significantly higher bandwidth, larger data volumes and greater read access for ATLAS in the HL-LHC era and sPHENIX at RHIC.
  - Storage technologies evolving at different rates.
  - Migration to new data center, with no migration of existing EOL equipment
  - Optimizing future investments requires detailed plans
- This presentation analyzes the effects of data volume and access bandwidth on the cost of tape and disk mass storage solutions

Estimating Cost of Disk vs Tape

- This cost analysis focuses primarily on the system and assumes or includes the following:
  - "Greenfield" deployment - No migration cost to switch between tape and disk. No legacy data.
  - Evolution of technologies 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
  - Network costs are included

Costs Not Included in Analysis

- Ignored factors include
  - Organizational - Manpower costs, multi-customer cost sharing opportunities
  - Infrastructure - Analysis assumes power, space, cooling infrastructure are available
  - Alternate system implementations not considered
    - e.g. Tiered disk storage, drive spin down, etc
    - Alternate tape software and hardware
  - Inter storage hierarchy optimization
    - Analysis looks only at the mass storage system.
    - Cost savings of collapsing storage hierarchies not investigated
  - Inefficient utilization of resources

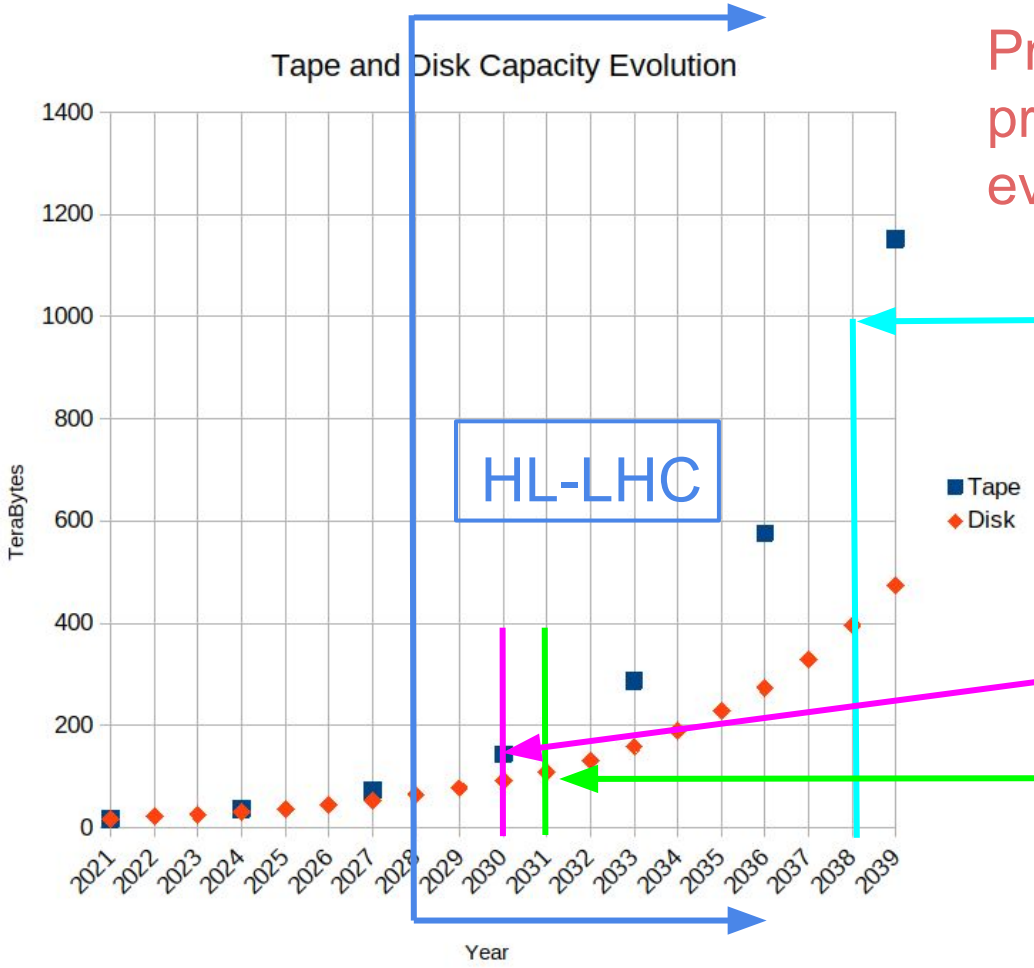


Technology Evolution

- Tape Parameters
  - Use LTO.org capacity roadmap
    - Capacity doubles each generation
  - 20%/yr reduction in $/TB for media
  - Utilization of 90% of max tape drive bandwidth [1]
  - 3 years between generations
  - 9 year media refresh cycle
    - LTO-N copied to LTO-(N+3)
  - 20% tape drive BW increase per generation
- 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 sparse reads of tape media, i.e., assumes no tape head seeks

Disk and Tape Roadmap Limitations



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

Demonstrated Tape Technology

End of LTO Roadmap

Limits of HAMR Disk Technology

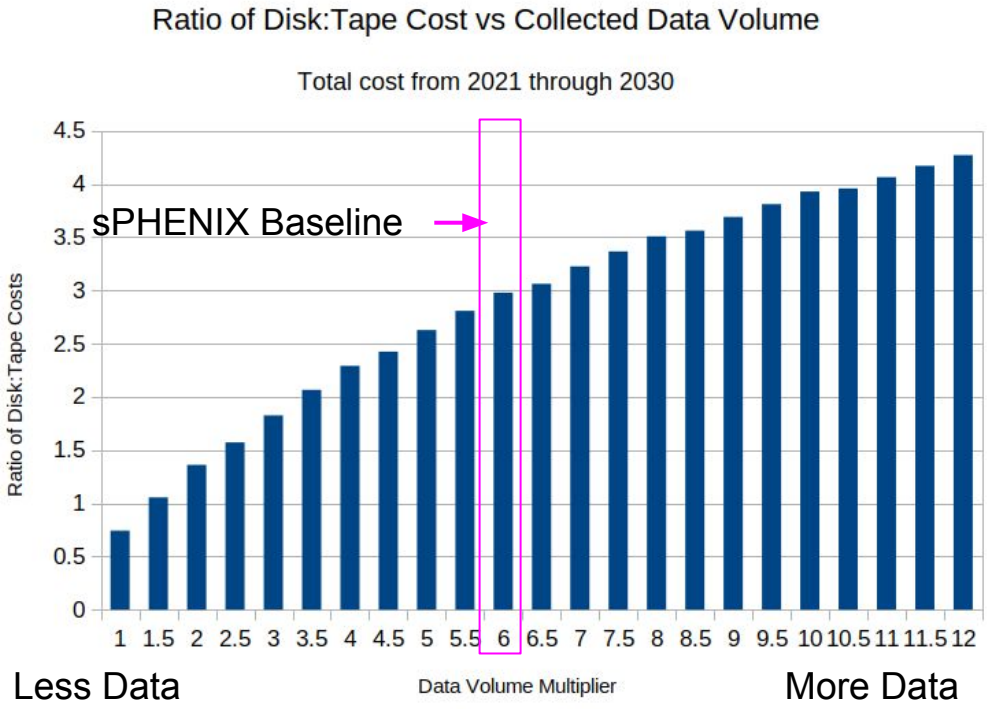
Disk/Tape System Assumptions

- Tape System
  - HPSS-like solution
  - Library w/ 20K cartridge capacity
  - Library deployed in 10K cartridge capacity increments
  - Maintain 5% free slot capacity at all times
  - 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 overhead
  - 500MB/sec “LUN” write performance
  - 10GB/sec capable servers
  - 400 disks per server

Comments on Disk/Tape

- Disk and tape are fundamentally different
  - Differences in data durability need to be acknowledged
- Disks are an “online” media
  - Disks are electrically energized at all times
  - Disk systems are online at all times
  - “Disk copies aren’t backups”
- Tapes are an “offline” media
  - Tapes only exposed to electrical issues when mounted
  - Potentially safer from ransomware and accidental deletion
  - Tape media life is substantially longer than disk

Cost Comparison for sPHENIX at RHIC



	Injested Data Volume Per Year (petabytes)								
	2021	2022	2023	2024	2025	2026	2027	2028	2029
Multiplier									
1	10	20	30	60	0	0	0	0	0
2	20	40	60	120	0	0	0	0	0
3	30	60	90	180	0	0	0	0	0
4	40	80	120	240	0	0	0	0	0
5	50	100	150	300	0	0	0	0	0
6	60	120	180	360	0	0	0	0	0
7	70	140	210	420	0	0	0	0	0
8	80	160	240	480	0	0	0	0	0
9	90	180	270	540	0	0	0	0	0
10	100	200	300	600	0	0	0	0	0
11	110	220	330	660	0	0	0	0	0
12	120	240	360	720	0	0	0	0	0

sPHENIX baseline data ingest ramp

Ratio of total cost as a function of collected data volume

Data Volume = 120 PB x Data Volume Multiplier

Peak BW = 30 GB/sec

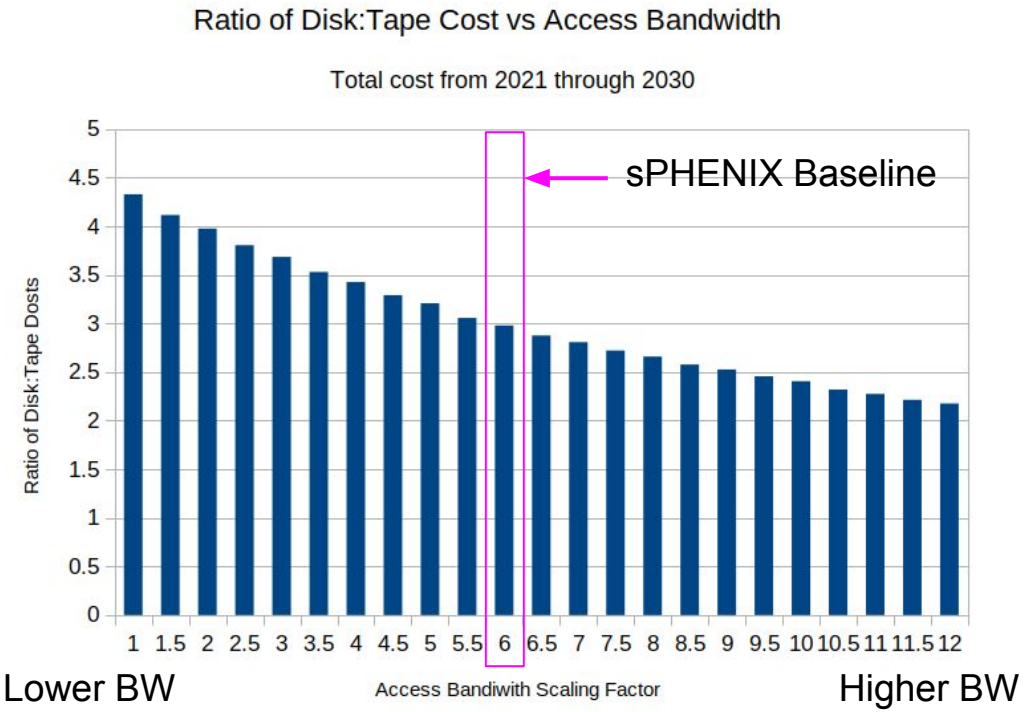
Data collection period - 2021 thru 2024

Vary sPHENIX data ingest requirements

Multiples of 10:20:30:60 ingest ramp

sPHENIX baseline is 6 x (10:20:30:60) PB/yr

Cost Comparison for sPHENIX at RHIC



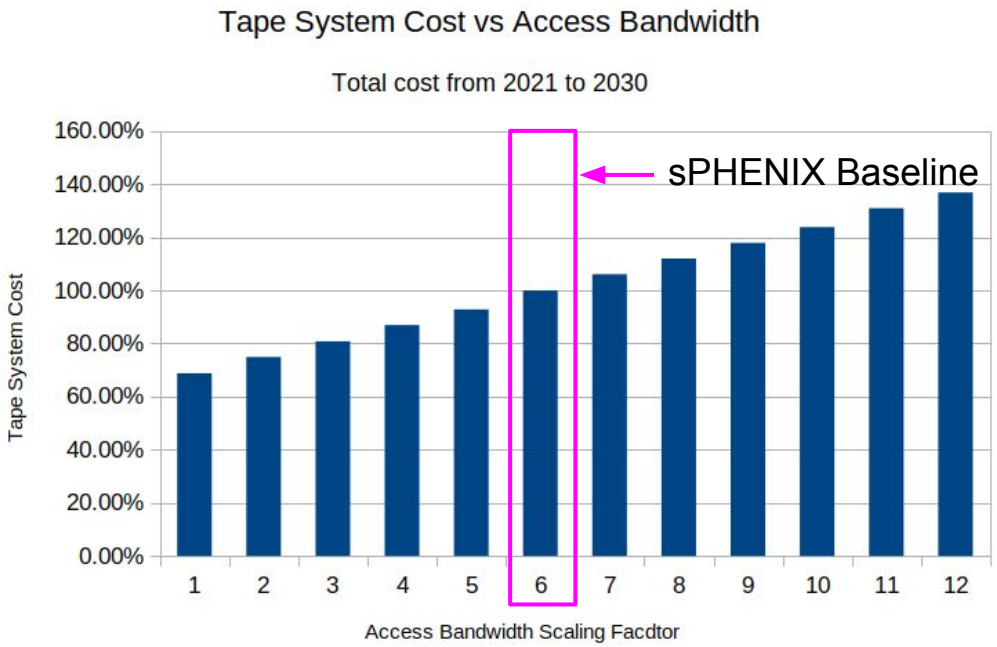
	Access Bandwidth Per Year (gigabytes/sec)								
	2021	2022	2023	2024	2025	2026	2027	2028	2029
Multiplier									
1	1	5	5	5	5	5	0	0	0
2	2	10	10	10	10	10	0	0	0
3	3	15	15	15	15	15	0	0	0
4	4	20	20	20	20	20	0	0	0
5	5	25	25	25	25	25	0	0	0
6	6	30	30	30	30	30	0	0	0
7	7	35	35	35	35	35	0	0	0
8	8	40	40	40	40	40	0	0	0
9	9	45	45	45	45	45	0	0	0
10	10	50	50	50	50	50	0	0	0
11	11	55	55	55	55	55	0	0	0
12	12	60	60	60	60	60	0	0	0

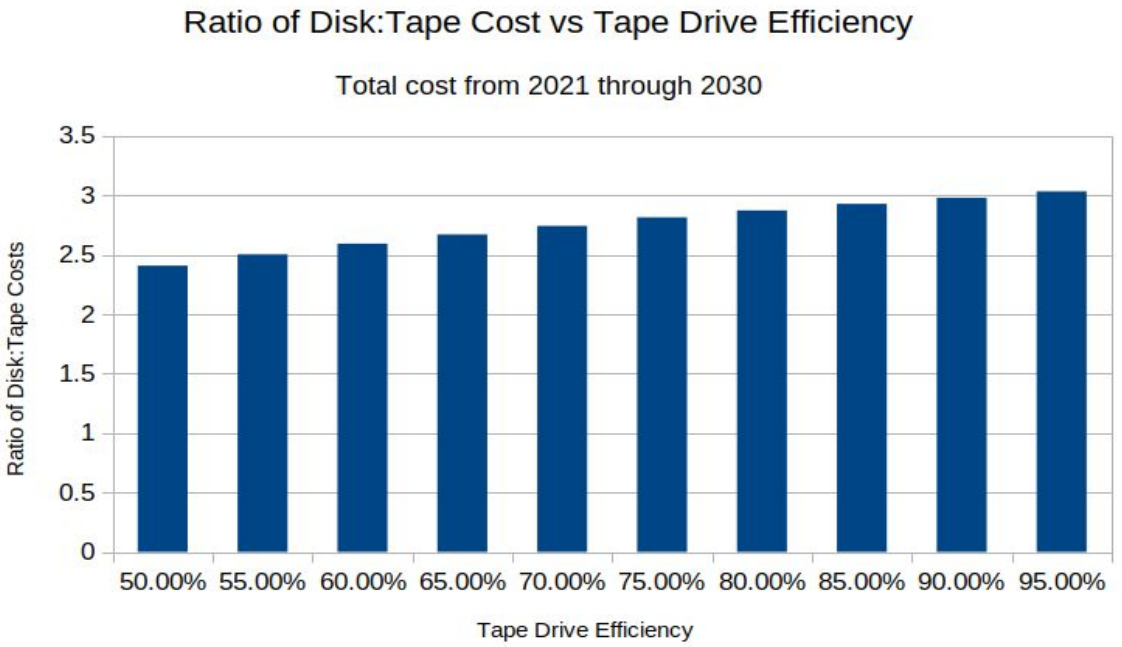
sPHENIX baseline bandwidth ramp

Ratio of total cost as a function of access bandwidth
Peak BW = 5 GB/sec x BW Scaling Factor
Total Collected Data Volume = 720 PB
Data collection period - 2021 thru 2024

Vary sPHENIX access bandwidth requirements
Multiples of 1:5:5:5:5:5 access BW ramp
sPHENIX baseline is 6 x (1:5:5:5:5:5) GB/sec

Cost Comparison for sPHENIX at RHIC





Tape system cost as a function of access Bandwidth

Ratio of total cost as a function of tape drive efficiency

Peak BW = 30 GB/sec

Total Collected Data Volume = 720 PB

Data collection period - 2021 thru 2024

Analysis Results

- Relative advantage between disk and tape changes with data volumes and bandwidth
  - Ratio of tape/disk cost decreases with increasing data volume
  - Ratio of tape/disk cost increases with increasing access BW
- Timing is important
  - Higher disk capacity makes disk more competitive at a given data volume
  - Tape/disk cost crossover point dependent on details

Analysis Results

- To first order, inefficient use of tape drive resources (non-sequential and non-contiguous access of data) can be modeled as lower performance tape drives.
  - Results in higher tape system costs.
- Cost of migrating from tape to disk likely to be high
  - Increases initial data volume
  - Requires supporting both tape and disk during transition period

Areas For Further Investigation

- Disk
  - Merge front end and back end disk mass storage systems
  - Hierarchical system
    - Multiple QOS partitions
  - Utilize SMR drives
    - ~20% cost savings
    - Requires software
  - Spin down disks
    - Requires software (e.g. FreeNAS)
    - Reliability ?
  - Tailor network to required QOS
- Tape
  - More precise accounting of read/write inefficiencies
  - Migration from multi-actuator HDD to SSD tape buffers
  - Investigate enterprise tape technology
- Analysis of transition costs
  - Cost of parallel infrastructure
  - Cost of moving legacy data

Conclusions:

- TCO is dependent on requirements, specifically:
  - Accumulated data vs time - Large data volumes further in the future benefit from higher density disks
  - Read/write requirements - Disk bandwidth naturally increases with storage capacity (more HDDs), tape bandwidth does not.
  - Continuous dialog with scientific experiments important to enable optimal and cost efficient use of mass storage resources

Conclusions:

- Predictions beyond 10 years are problematic due to technology and economic uncertainties
  - HDD - ~2029 transition from HAMR to Bit Patterned Media (BPM)
  - Tape - Read/write performance an issue.
    - LTO-9 12.5 hours to read full tape
  - Tape/HDD - Economics of the business: Are they viable ?
  - Role of SSD in capacity storage is unclear.
    - Cost /TB for SSDs has been dropping but remains 5x-10x higher than HDD.