Tape vs Disk for Scale Out Storage

Scientific Data and Computing Center
Brookhaven National Laboratory
Shigeki Misawa
November 16, 2020

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Introduction

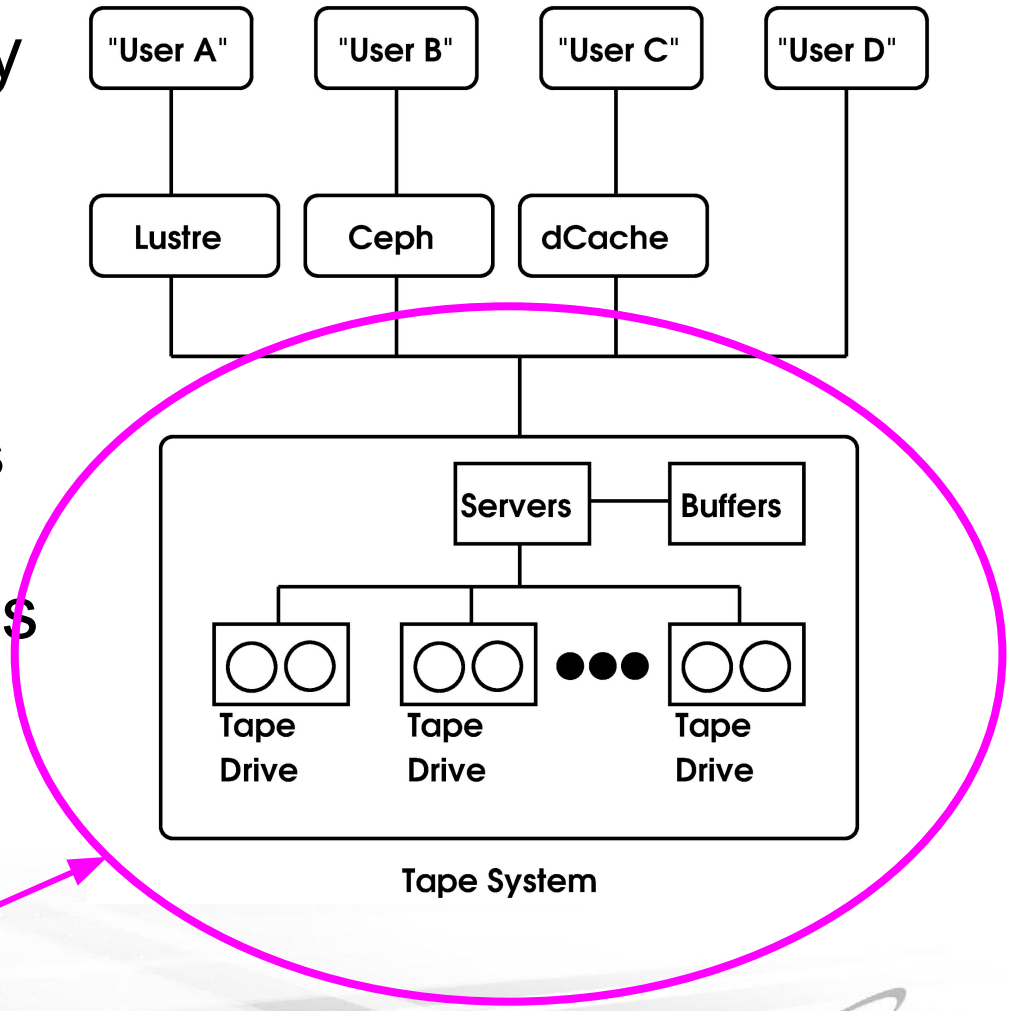
- Magnetic tape has been the default media for archival storage and “cold” storage in industry and HEP/NP
- Over the years, tape and disk has advanced at different rates and in different ways.
- New data storage systems and services have emerged utilizing disk as the underlying media
- Disk has replaced tape in many applications in the commercial world.

Tape at the SDCC

- The SDCC at BNL will need to make major investments in tape hardware
  - SDCC is moving to a new state of the art data center with more capable infrastructure, able to support higher power and weight density racks
  - The facility will use new tape libraries in the new data center
  - HEP/NP data storage requirements are changing dramatically
    - Higher bandwidth, greater data volumes, and more frequent data access, e.g. sPHENIX will collect close to an exabyte of data by 2025
  - Detailed plan necessary to optimize investments
- Now seems to be a good time to re-examine disk and tape for long term, large scale storage.

Data Center Storage Environment

- Tape system part of a storage hierarchy
  - Tape system “back end”
  - One or more disk systems as “front ends”
- Tape system utilized by multiple group
  - Different access requirements
  - Different capacity growth profiles/timelines
  - Different life cycles (periods of activity)
- Tape system used for different purposes
  - Backups
  - Active archive
  - Cold storage



Rethink this component

Evaluating Disk and Tape at the SDCC

- Estimate the total cost of ownerships for a disk based and a tape based implementation of the back end storage system
  - Provide nearline and archival storage for scientific data
  - Satisfy the storage capacity and read/write requirements over the one to two decades.
- Identify and evaluate risk and benefits of the two technologies
- Trigger investigations into alternative system architectures/implementations for disk and tape systems.
- Provide feedback to system “users” on impact of requirements
- Current scope does not investigate modifications/optimizations that can be made to the entire data center storage hierarchy. (e.g. merging front end storage systems with the back end storage system)

Storage Systems Cost

- Tape drives, media, and libraries and disks are components within the storage systems.
- Hardware costs only part of the full cost of a system
- Full cost of a storage system also depends on the following:
  - Operational environment (i.e., what requirements must be met)
  - Capabilities/limitations of the software
  - Configuration and implementation of the system, e.g.,
    - Choice of hardware and how they are assembled into a system
    - Choices made in the utilization of system features to provide the service
  - “Full lifecycle” operation of the system (component end of life policies)
  - Operational cost (i.e., manpower and infrastructure costs)

Limitations of this Cost Analysis

- Analysis may not be easily transferable to other sites
  - Tape systems are mostly unique to each site (e.g. HPSS, CTA, TSM)
  - Disk systems may be more comparable as “free parameters” are more limited compared to tape systems
  - Operating environment may vary substantially
    - # Customers
    - Usage patterns/performance requirements
    - Capacity profiles over time
  - System scale differ from site to site. Exabyte economics likely to be different from petabyte economics.
  - Available infrastructure (power, space, cooling) and their costs are site dependent
  - Procurement costs will be different from site to site

Cost Analysis Process

- Determination of costs is an iterative process
- Establish baseline assumptions on basic technologies
  - Evolution of tape media capacity and cost
  - Evolution of tape drive performance
  - Evolution of disk capacity, performance, and cost
- Establish requirements and evolution of requirements
  - System capacity and system bandwidth
  - Determine reliability, availability, and durability requirements
- Establish hardware life cycles
  - Refresh interval for hardware
  - Establish policy for tape media refresh

Cost Analysis Process (cont'd)

- Select baseline system design
- Select hardware technologies
- Flesh out the configuration and components of the entire system for "day 1"
- Evolve system over the course of the desired operational period
- Calculate costs associated with the system over the operational period
- Compare point in time and full operation period costs for the disk and tape system

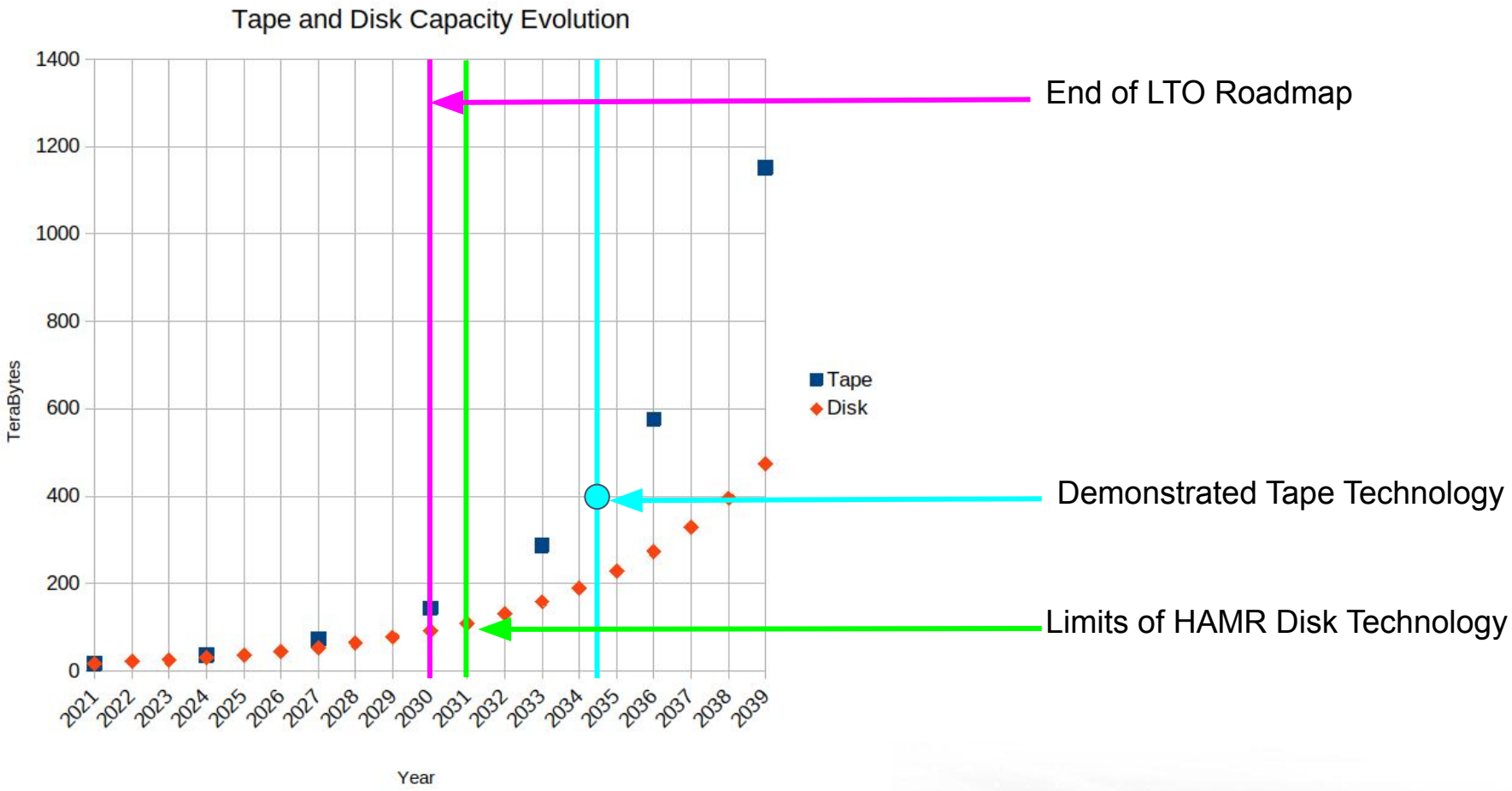
Cost Analysis Process (Iteration)

- Iterate analysis
  - Re-examine fundamental assumptions in model and system
  - Re-examine requirements, adjust as needed
  - Correct inaccuracies in model and add details ignored in previous passes to model
  - Incorporate optimizations discovered or ignored during first pass analysis
  - Consider apply optimizations to the broader system in which the disk/tape system is embedded
- Re-run cost analysis and compare results
- Iterate process as needed

SDCC Analysis Assumptions

- Tape Parameters
  - Use LTO.org capacity roadmap
    - Capacity doubles each generation
  - 3 year between generations
  - 20% BW increase per generation
  - Real world read/write 90% of max performance
    - Does not cover operational inefficiencies
  - $140/cartridge at introduction
    - 20%/yr reduction in $/TB
- Disk Parameters
  - 20%/yr HDD capacity increase
  - 20%/yr reduction in $/TB
  - 10 W single actuator HDD power
  - 15 W dual actuator HDD power
  - 5 year refresh cycle
  - Constant 250 MB/sec r/w bandwidth

SDCC Analysis Assumptions



SDCC Analysis Assumptions

- Infrastructure Parameters
  - Max floor loading - 500 lbs/sq ft
  - Power
    - Average 10 KW/rack
    - Max power 14.4 KW/rack
  - Rack
    - Standard 42U rack
    - 4000 lbs rack capacity
  - Electricity - $0.06/KWH
  - Cooling - $0.02/KWH of IT load
- "Green field" deployment
  - No existing data
  - No existing tape infrastructure
  - No "transition" cost

Baseline Tape Systems

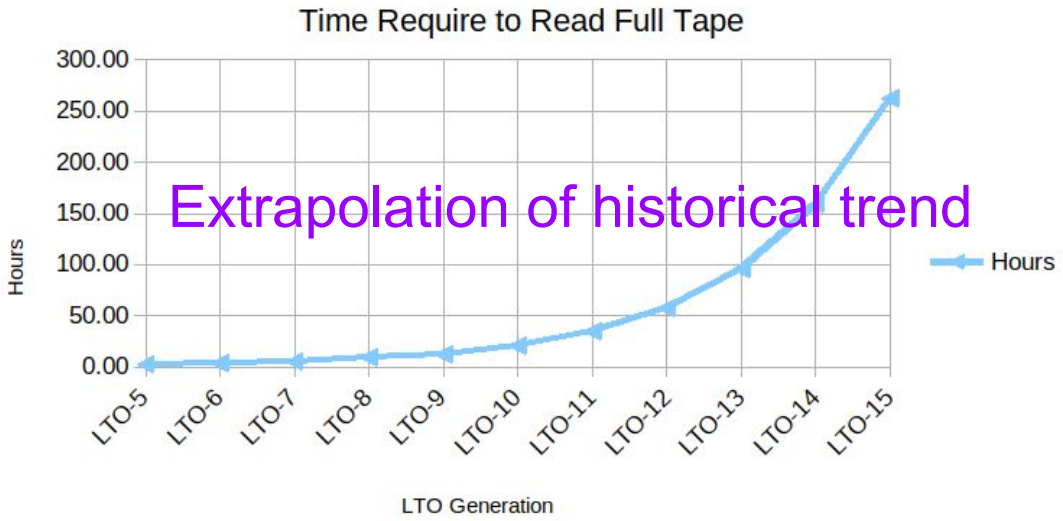
- Utilize SDCC HPSS system as the basis for the tape system
- LTO tape technology/IBM tape library
- Use simplified, regularized version of life cycle/operational model at the SDCC.
  - Switch to next generation tape technology every 3 years
  - 6 year media life cycle
    - Three year data migration window to move data from LTO-N to LTO-(N+2)
    - Implications on data migration bandwidth
  - Each ~20,000 cartridge tape library acquired in ~10,000 cartridge capacity increments
  - Maintain 5% free cartridge capacity in libraries at all times

Baseline Disk System

- Assume a single QOS, scale out dCache/Lustre/Ceph solution
- Assume software will work “at scale”
- Utilize “standard” HDD technology (i.e., no SMR/zoned disks)
- 5 year life cycle
- Size system to maintain 10% free space
- 20% capacity overhead for data protection
- Assumes 500MB/sec write performance for each 8+2 “LUN”
- Assumes server capable of 10GB/sec write performance
- Ignores cost of network infrastructure

Preliminary Analysis Results

- Relative advantage between disk and tape changes with data volumes and bandwidth
  - Disk cheaper at smaller scale, tape cheaper at larger scale
  - Tape costs increase with access bandwidth, but likely to remain constant for disks
- LTO tape drive performance is an issue
  - Speed not scaling with capacity.
    - Extrapolation of historical trends suggests 20% increase per generation
  - 128 drives per library translates to to roughly 64GB/sec tape bandwidth



Preliminary Analysis Results

- Disk power, space and cooling
  - Energy cost for disk is <10% of total cost of ownership
  - Power consumption is non-trivial
    - Peaks at ~500KW for SDCC requirements
    - Spin down may reduce energy costs but to what extent is unclear
    - Multiple architectures with spin down possible, best fit unclear
    - Impact of spin down on reliability is not quantified (COPAN MAID revisited)
  - For aggregate SDCC requirements, space and power consumed is roughly 50% of capacity allocated for critical systems requiring redundant power in the new data center
- Tape power, space, and cooling
  - For SDCC space a non issue, as there is a dedicated tape room
  - At peak, power requirements for tape roughly 10%-15% of disk solution

Open Questions/Work in Progress

- Establish “real world” performance
  - Affects server/disk ratio
  - Affects number of tape drives
  - May affect #disks
- Quantify realistic operational efficiency
  - e.g. effects of tape mount time, tape seek time,
- Incorporate equipment EOL process for disks (impact of data transfers from old to new storage)
- Re-examine data durability overhead for disk system
- Correct tape buffer implementation cost calculation
- Investigate use of enterprise tape technology instead of LTO

Possible Optimizations

- Disk
  - Merge front and back end disk 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 connectivity to required QOS
- Tape
  - Re-examine policy for migration of old data to new media
  - Re-examine details of transition to new tape drive technology
  - More precise determination of read/write bandwidth requirements
  - Migration to multi-actuator HDD or SSD tape buffers
  - Investigate enterprise tape technology

Conclusions:

- TCO is dependent on requirements, specifically
  - Accumulated data vs time. Large data volumes farther in the future benefit from advances in technology
  - Read/write requirements - Disk bandwidth naturally increases with storage capacity, tape bandwidth does not.
  - Continuous dialog with scientific experiments important to enable optimal and cost efficient use of resources
- Cost of tape difficult to calculate
  - Capacity and r/w bandwidth are decoupled
  - Resources partitioned by tape library and tape technology
  - Migration of legacy data to new media can be a complex calculation

Conclusions:

- TCO likely to be highly site dependent
  - System architecture (software capabilities)
  - Performance requirements
  - Role of tape at the site
- Strengths and weaknesses of disk and tape are different and need to be weight along with cost.
- Magnitude of transitions costs is unclear
  - If high, converting storage from disk to tape or tape to disk may not be possible, even if end state is less costly

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.