

Storage R&D updates

- Towards a Storage R&D roadmap by the end of the summer
 - Critical items on tape software
 - Can we enter a collaborative agreement with CERN on CTA? (LS-K and JA inquiring)
 - Do we have development effort continuity? R_004706 is out
 - Disk questions
 - Can we move away from hardware RAID?
- Object stores
 - Kevin and Vito are standing up a CEPH test instance
 - Retired dCache pool nodes; set up as “single disk arrays” to emulate JBODs
 - Aim to evaluate for use by DUNE
 - Submitting CMS S&C Ops RA funding proposal to evaluate object storage for CMS
- HEPIX this week
 - Erasure coding working group formed (following this activity)
 - Other storage-related highlights to follow (for reference)

EOSCTA infrastructure for Run3

- Run3 constraints:
 - >60GB/s of bandwidth
 - 8 hours of cache **We assume much more than this is needed!**
- 64 buffer servers installed:
 - 200GB of RAM, 500GB-1TB NVMe (OS + logs)
 - 16x2TB SSDs, 25Gb/s each
 - total: 2PB at 200GB/s simplex
- 100Gb/s Router uplinks (no stacking)
 - $\sim 2/3$ blocking factor



LHC Run 3 – boundary conditions

Physics requirements: **Similar to HL-LHC CMS+DUNE for Fermilab (3 year period)**

- Data amount per year: 150 PB LHC + 30 PB non-LHC = 180 PB
- Total data amount for 3 years: 540 PB
- Transfer rate requested: 10 GB/s for *each* of the 4 LHC experiments

Infrastructure limits:

- 2 LTO libraries, 2 IBM 3592 enterprise libraries
- 24900 free LTO slots; 9300 free IBM 3592 enterprise slots
- 48 tape drive slots per each library



LHC Run 3 – tape infrastructure plans

1/2

Space requirements – Tape Media:

- Fill free 9300 IBM enterprise slots with 3592JE (20TB) media = 186 PB
- Fill free 24900 LTO slots with LTO-9 (18TB) media = 448 PB
- = 634 PB of total available tape capacity will be sufficient for next 3 years

Throughput requirements – Tape Drives:

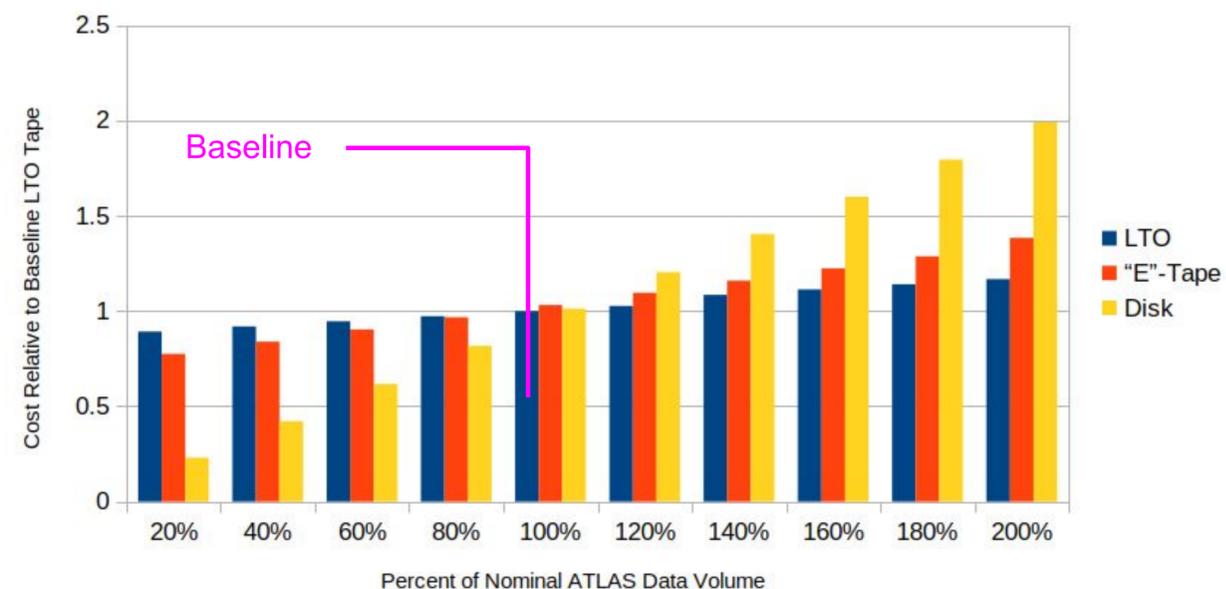
- IBMLIB3:
 - Total 38 x IBM TS1160 and 10 x IBM TS1155
- IBMLIB4:
 - Total 38 x IBM TS1160 and 10 x IBM TS1155
 - Total 162 (new) drives: 76 x IBM TS1160 and 86 x LTO-9
- Expected throughput per tape drive ~300 MB/s
- Total >45 GB/s which should be sufficient for LHC Run 3
- IBMLIB1:
 - Total 38 x LTO-9 and 10 x LTO-8
- SPECTRALIB1:
 - Total 48 x LTO-9



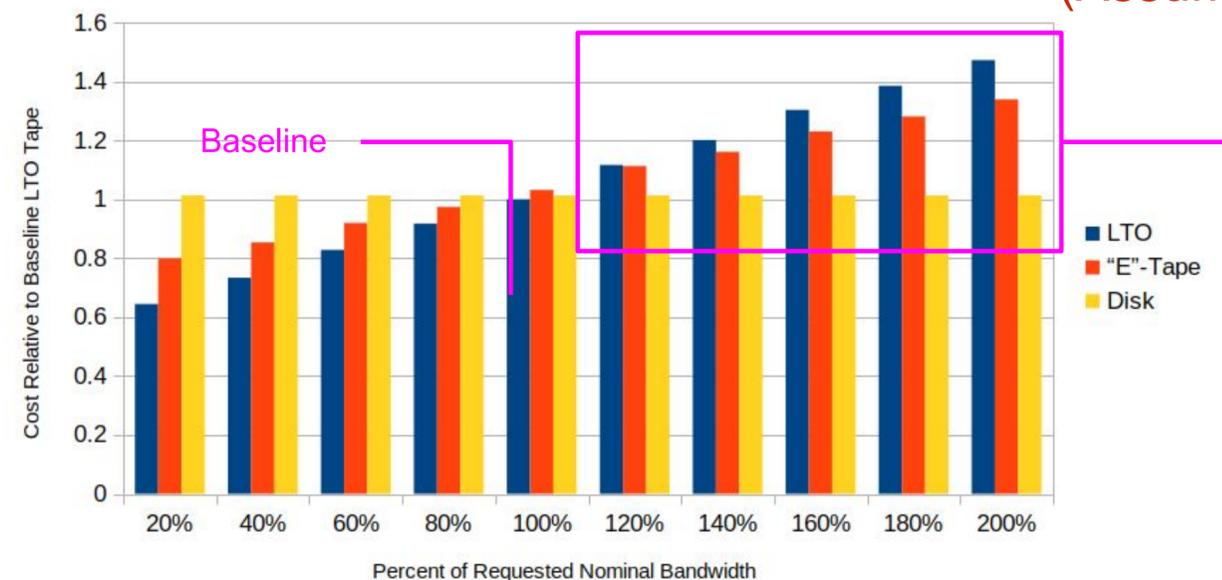
Relative Cost Comparison for ATLAS

BNL estimates that disk and tape would have similar costs for ATLAS Run 3+Run 4 (Assumes no legacy data)

Tape/Disk Cost vs Data Volume (ATLAS)



Tape/Disk Cost vs Bandwidth (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



Cost?

Erasure coding experience at RAL (using CEPH)

- First order:
 - EC 8 + 3 means 72.7% usable space.
 - 2 x Replication means 50% usable space. } Need to purchase 45% more capacity with Replication
- Erasure Coding has higher CPU and Memory requirements compared to Replication. Assume:
 - 2 x CPU
 - 1.5 x Memory } Adds 3 - 5% to cost of hardware for Erasure Coding
- Assume Erasure Coding requires larger overhead ~5%
- Upfront costs for same amount of usable storage with Replication ~25% more than Erasure Coding.
- Power cost over 5 years ~40% upfront cost.
 - Additional ~20% cost over the lifetime of the hardware for Replication .
- For RAL, Total Cost of ownership: Erasure Coding is about 70% the cost of Replication.