



Fermilab Storage Experience (and planning)

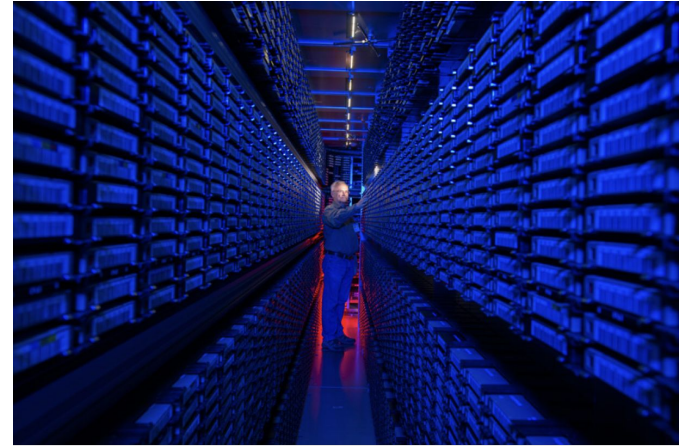
Bo Jayatilaka and David Mason

HSF-WLCG Virtual Workshop

19 November 2020

Current storage landscape at Fermilab

- Custodial and active storage for all Fermilab experiments' scientific data
 - This includes considerable storage for “external” experiments/projects (e.g. CMS and DES)
- Utilizing a tape+(spinning) disk HSM
 - Tape managed by Enstore (Fermilab)
 - Disk managed by dCache (DESY+Fermilab+NDGF)
- **264 PB** of tape in use (226 PB active)

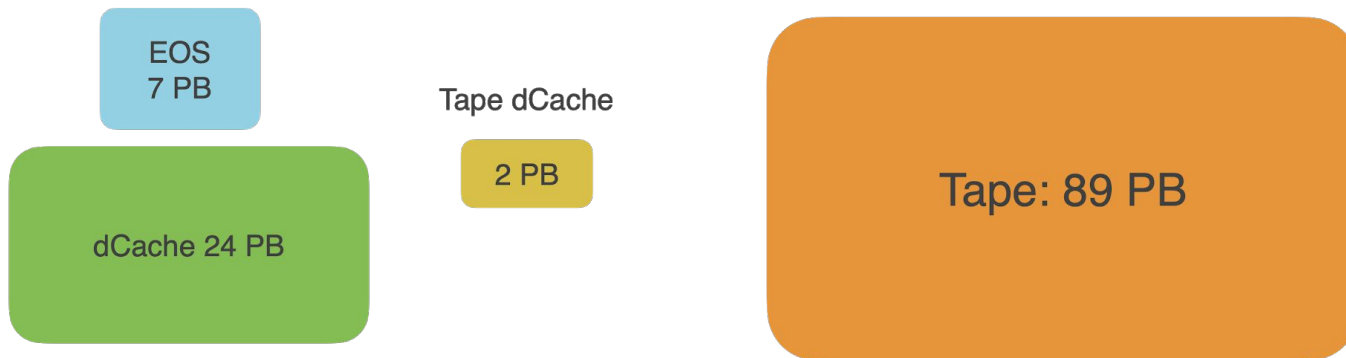


Storage infrastructure

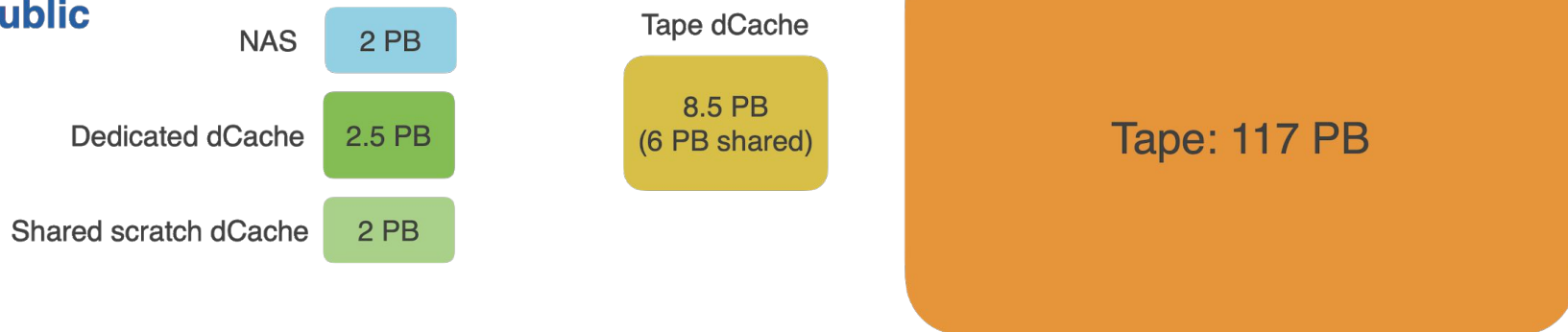
- Two major storage “instances”
 - *Public*: neutrino (DUNE, SBN, etc), muon (g-2, mu2e), DES, LQCD, etc
 - Generally all projects/experiments except CMS and Tevatron
 - *CMS*: dedicated for CMS Tier 1 storage
 - Also managed: analysis-only EOS pool
- Dedicated hardware for each instance
 - Tape library complexes
 - Multiple 10k slot libraries (details later)
 - Multiple dCache pools
 - Commodity SAN configuration (storage servers+disk arrays)
- **Resource allocation** and **use cases** differ considerably between the two

Storage infrastructure in a nutshell

CMS



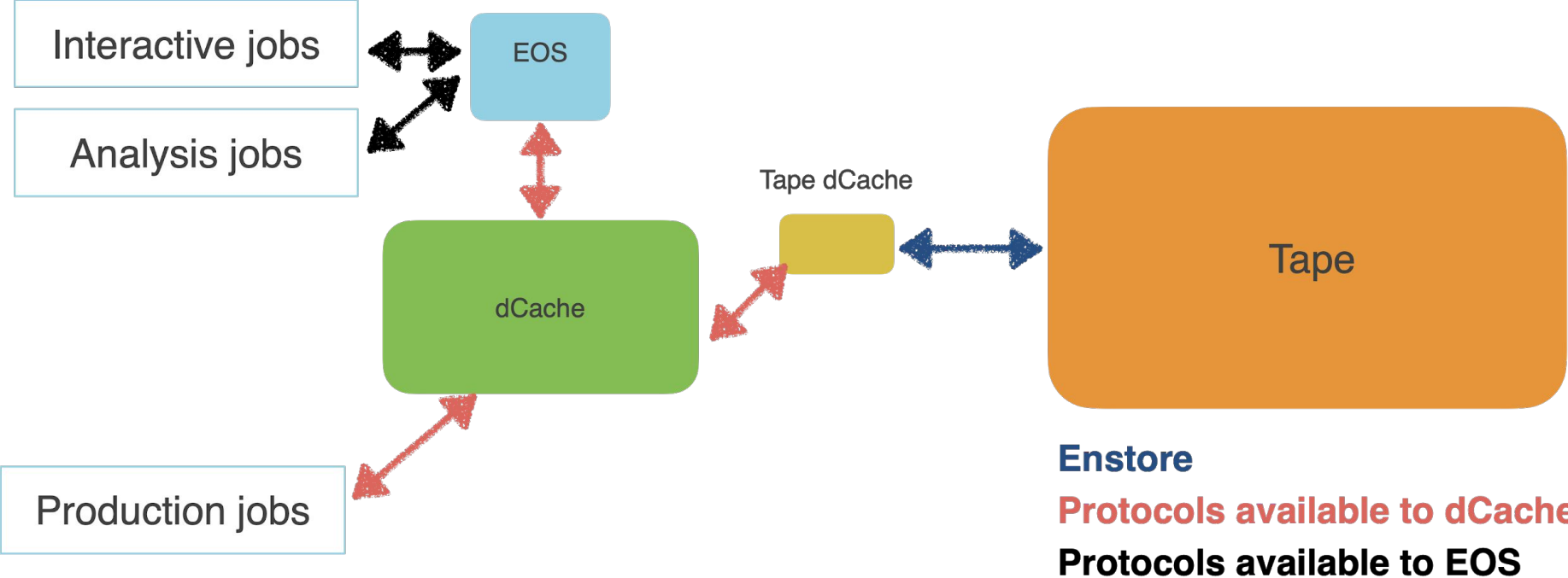
Public



Archival storage at Fermilab today

- Current tape formats in use
 - **LTO-8**: 12 TB capacity, 360MB/s maximum drive speed (~4100)
 - **M8** (special LTO-7 format): 9 TB capacity, 300MB/s max drive speed (~10400)
 - **T10000C/D**: 5.4/8.5 TB capacity, 250MB/s max drive speed (~18500)
 - Being **migrated** to LTO-8
- Current tape infrastructure in use
 - **3 IBM TS4500**: ~10k slots per library
 - 96 LTO-8 drives
 - **2 Oracle/Storagetek SL8500**: ~10k slots per library, 2 per library complex
 - About 90 10kC and D drives
 - To be retired
 - **Preparing for a sixth library in 2021**

CMS storage at Fermilab (today)



Ongoing issues/concerns/inefficiencies

- Files currently recalled and written essentially at random
 - Hundreds of datasets being transferred to write to tape simultaneously at any given time from dozens of sites across the world
 - Placed in tape families, but data arrives when it arrives.
 - Data recalled as needed by jobs
 - Data rarely streaming from tape
 - Swapping cartridges frequently, often to only read single files
- Currently in process of migrating about **50 PB** of data from SL8500 library purchased in 2010 to newer TS4500 library.
 - Process takes years.

Archival storage at Fermilab for the HL-LHC

Assumptions drawn from ESNet requirements document

- Annual data volumes for Run 4 to tape
 - 364 PB RAW, 240 PB AOD, 30(3) PB MiniAOD, 0.24(5) PB NanoAOD: **695 PB**
 - Assume 40% to Fermilab/US: **279 PB**
- Required data throughput: **400 Gbps**
 - read/write total?

Assumptions about future formats

- **LTO-9** (available now)
 - 18 TB tape capacity
 - 400MB/s max throughput
- **LTO-10** (extrapolating, ~3 years)
 - ~36 TB tape capacity
 - ~500MB/s max throughput
- Similar exercise can be done for IBM enterprise

Trying to meet HL-LHC requirements

- Data: 279PB annually
 - 23k LTO-8, 15.5k LTO-9, 7.8k LTO-10
 - **1-2 current-sized libraries per year**
- Throughput: 400Gbps (50GB/s)
 - Current Enstore efficiency achieves an average of 60% (~200MB/s) max throughput
 - LTO-8: 250 drives at avg, 140 at maximum
 - LTO-9: 208 drives at avg, 125 at maximum
 - LTO-10: 167 drives at avg, 100 at maximum
- *nb* maximum drives per library (regardless of frame count)
 - 125 for IBM TS4500, 144 for Spectra TFinity
 - Given number of drives required, min of 2 independent libraries will be required

Resources needed for Run 4 (~3 year run)

- Tape libraries (compare to current TS4500)
 - **2.5-4.5x** current slot capacity (2-3 independent libraries)
- Tape drives (compare to current LTO-8)
 - **7x** current average throughput
 - 4.5-6x current total drives
- Media (compare to current Tier-1 storage)
 - ~**10x** current storage media
- Disk buffer (for tape access only)
 - Currently a 1:50 disk:tape ratio
 - At 400Gbps 4PB fills in a day; assume a need for 4-7 days worth
 - **7.5-15x** current tape disk buffer
 - Not ready to assume a storage type for this
- Single (3 year) HL-LHC run will require **an order of magnitude more** archival storage infrastructure

Summary and Outlook

- Fermilab supplies scientific data storage for a wide range of HEP uses
 - **264 PB** of tape media, **48 PB** of disk
- Oracle exit from tape drive market dealt significant blow
 - O(100PB) of data to be migrated
- **CMS** storage needs will dominate by the end of the decade with HL-LHC
 - Storage footprint/capacity will be an order of magnitude larger than today
- Not discussed today in detail
 - Storage needs evolution of Fermilab experiments
 - See S. Timm's talk on DUNE yesterday
 - Analysis storage needs evolution
 - See analysis facility talks on Tuesday