Storage Development Plan

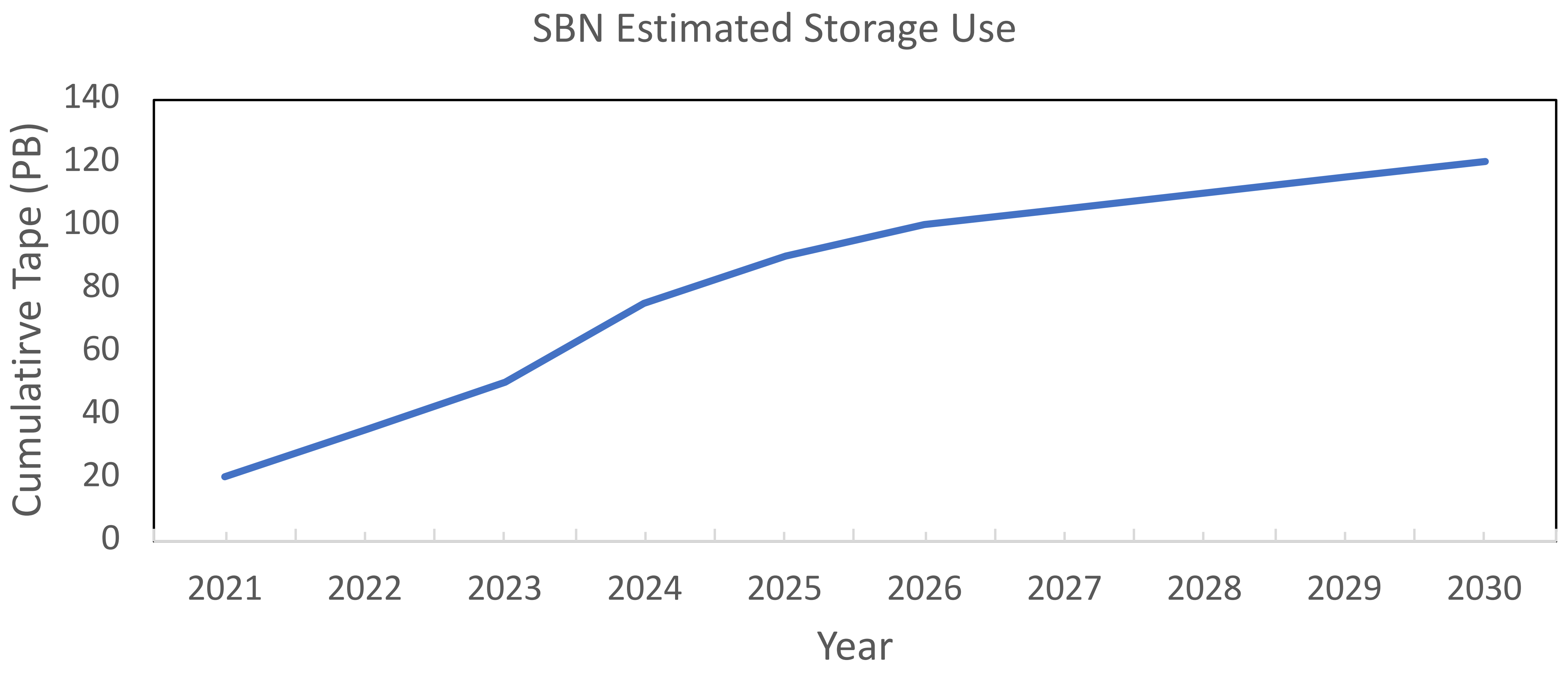
Bo Jayatilaka

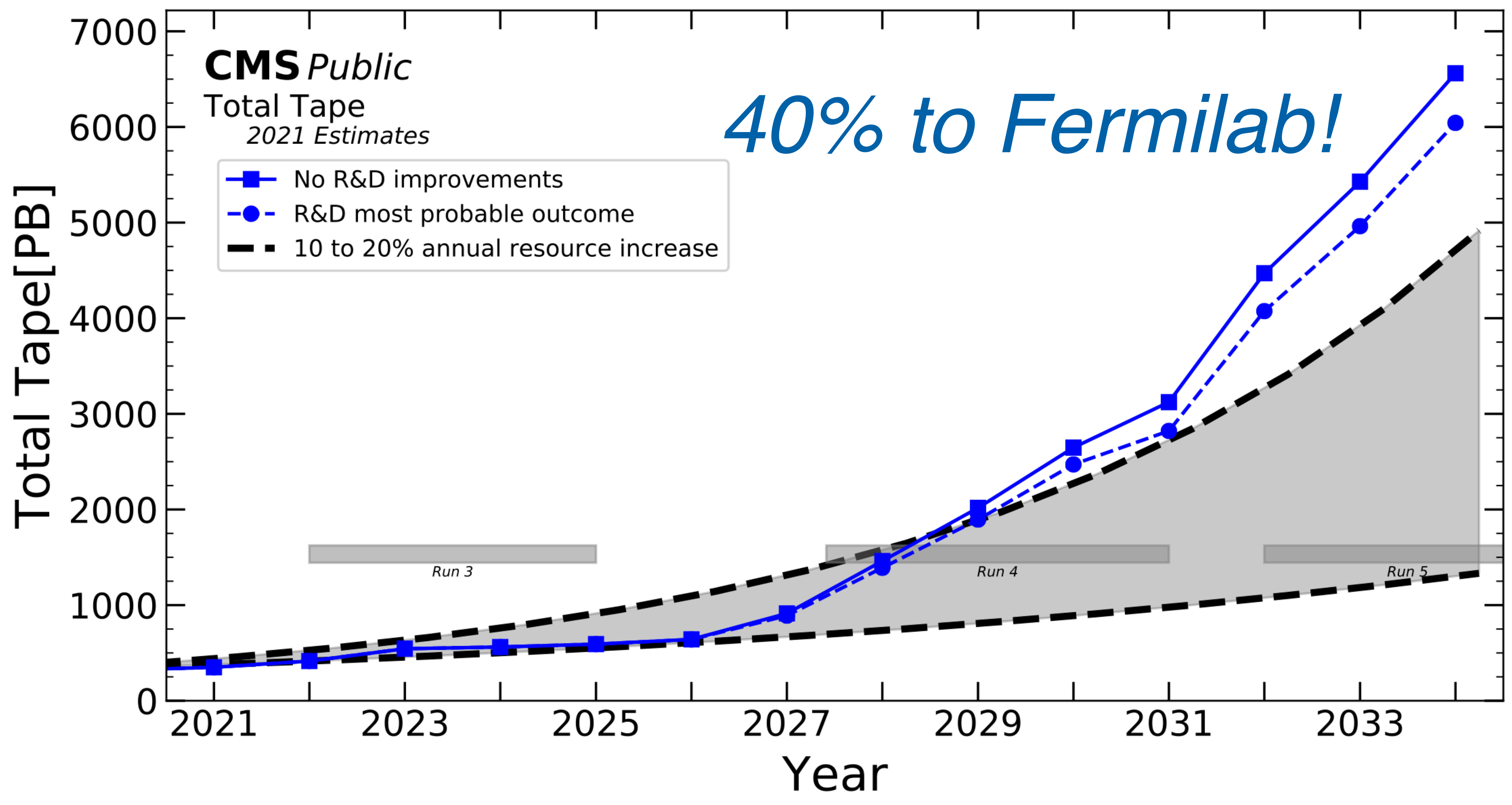
4th International Computing Advisory Committee Meeting

9 February 2022

Towards the next decade

- Fermilab will be a more “data-centric” computing site
  - CPU needs can increasingly be met by a mix of non-dedicated resources (cloud, HPC)
  - Custodial storage of data will still be necessary
    - HPC sites do not provision long-term data storage that meets experiment demands
    - Cloud storage is cost-prohibitive and results in lock-in
- A robust Exabyte-scale storage infrastructure will be needed





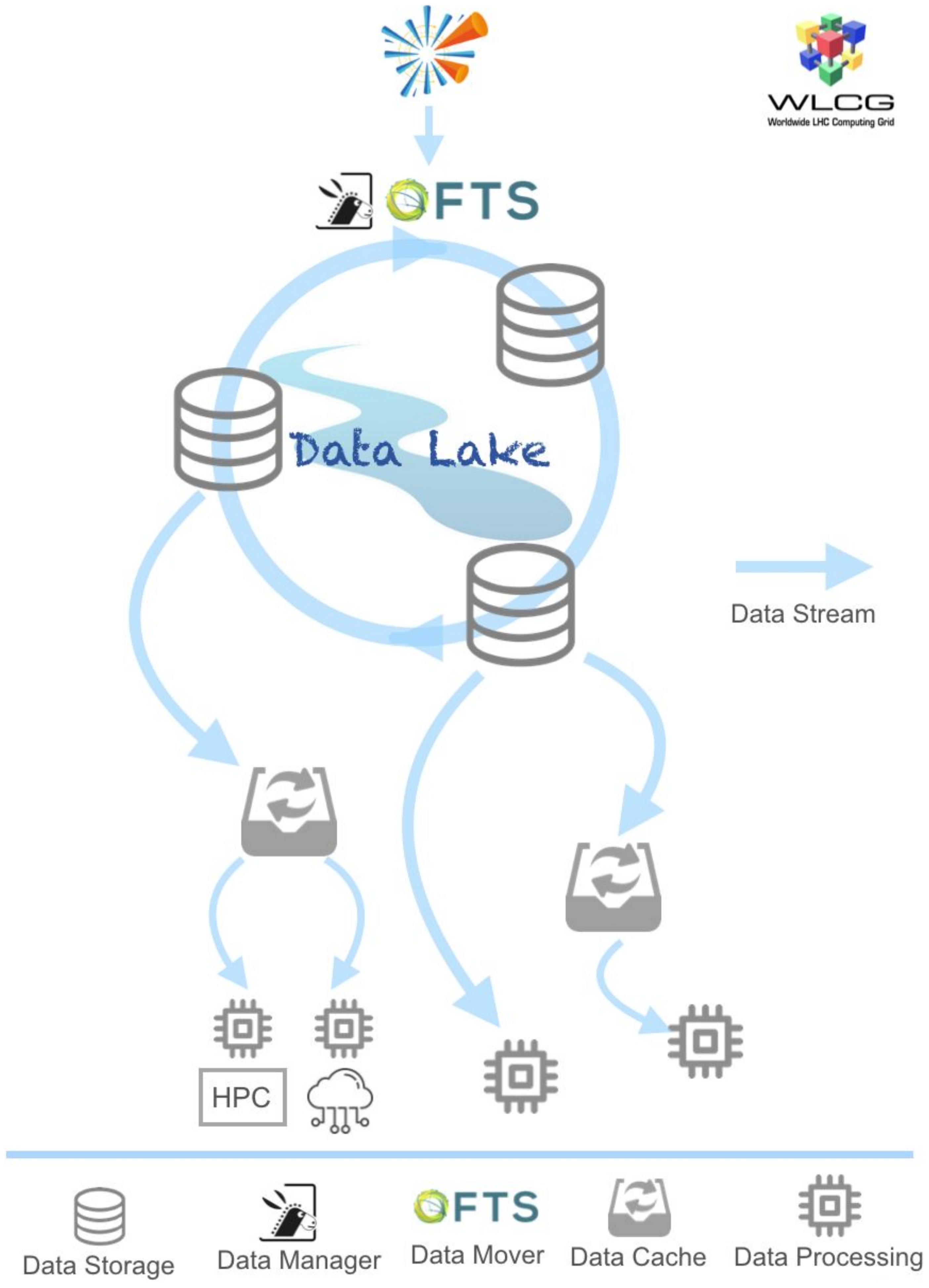
40% to Fermilab!

Key elements to address

- Global architecture
- Tape (archival storage)
- Disk (nearline storage)
- Networking (R&D plans)

Global disk/tape architecture: community-defined

- Implementation of **arbitrary QoS tiers**
  - Currently effectively with two tiers (“tape” and “disk”)
  - Future storage infrastructure must be able to map community defined QoS
- **Data lakes**
  - Data stored at Fermilab will need to seamlessly be part of defined national/global data lakes
    - Fermilab will likely be a data origin and consumer for multiple experiments
- Data **access and management** tools
  - Move to community standards (e.g. SAM->Rucio)
  - Contribute to development and support of tools
    - See Robert’s talk for more details



Global disk/tape architecture: site-defined

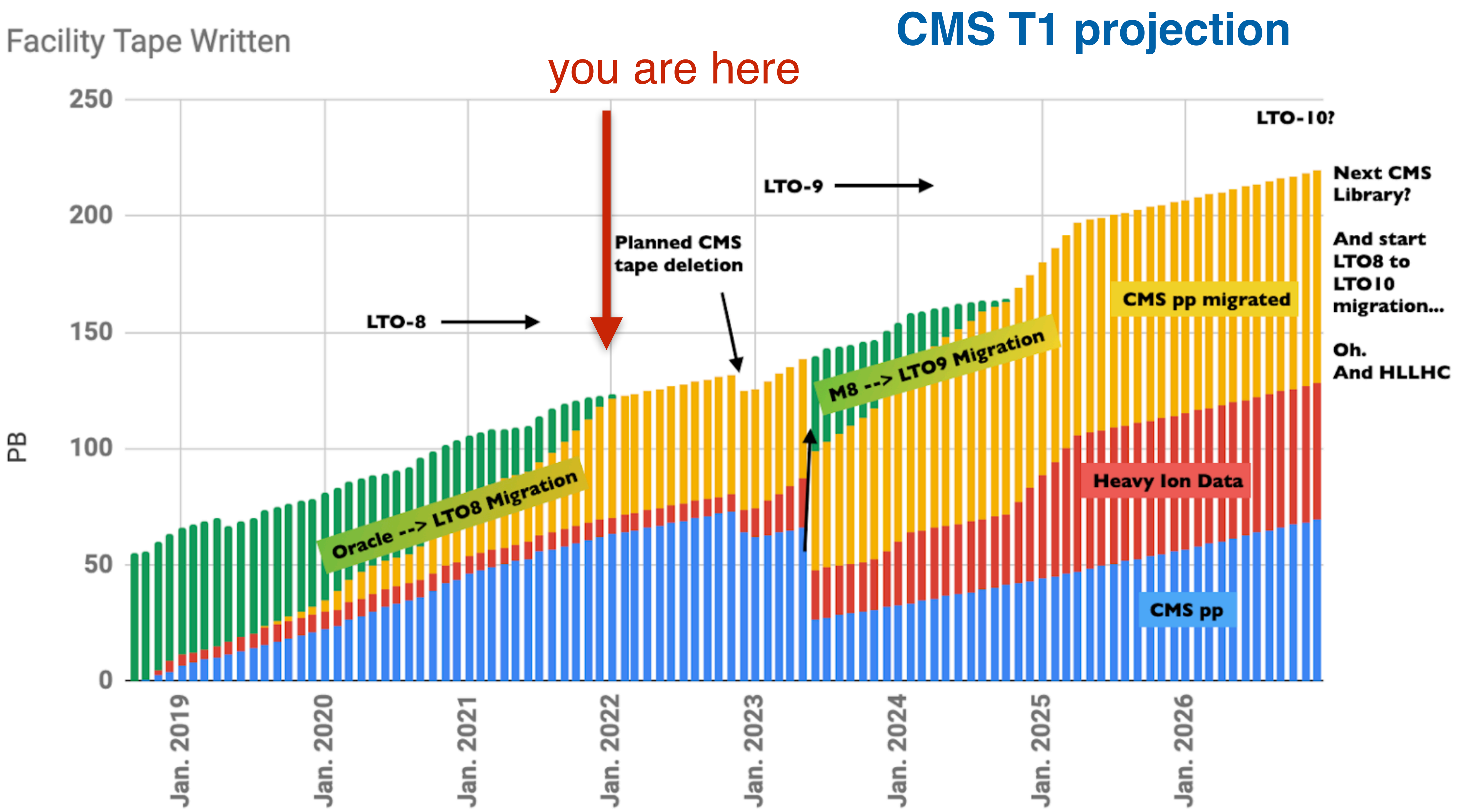
- **Separation of disk/tape nearline storage** for Fermilab experiments
  - CTA currently only supports deployment with separation
  - dCache development with CTA may also make shared deployment possible as with Enstore (see Robert’s talk)
  - As we get closer to deploying CTA, will assess the impact of making such a separation on available storage
- **Separation of storage infrastructure** for large VOs (e.g. DUNE)
  - This separation has been done for CMS and has made operations much smoother
- Fair-share/scheduling
  - Continue development efforts on dCache to help alleviate this
  - Efficient tape access continues to be an issue
    - Experiences of other **multi-VO CTA installations** will be illustrative

Tape/archival storage

- Enstore will likely not meet our needs in the HL-LHC/DUNE era
  - Imminent retirement of primary Enstore developer; community product adoption vital
- Doing a complete evaluation of **CTA** (CERN Tape Archive) to replace Enstore
  - Most Enstore tapes at Fermilab are written with the **CPIO wrapper** (see Robert’s talk)
    - Development effort will be needed to allow CTA to read CPIO tapes (and possibly write)
    - Development effort will be needed to migrate Enstore metadata to CTA
  - Enstore has a home-grown **small file aggregation (SFA)** system
    - No equivalent exists for CTA; will require development to read SFA packages and migrate
- Goal is to have a deployment **plan this summer**
  - Form an internal review team to go over this plan
  - Establish a deployment and migration timeline at this point
- Enstore will still be used until a complete migration to CTA
  - Essential to **maintain development support** for that time

Long-term archival storage strategy

- Tape migration continues to be a bottleneck
  - Migration moves from a periodic activity to a constant one
    - Do we move to a “continuous migration” model like CERN?
  - Work on current migration platform (developed for Tevatron data migration)
    - Considerable speedups; progress in Robert’s talk
  - **Default lifetimes** for data on tape may have to be introduced



Disk/nearline storage

- Three major use cases
  - **Production** compute (large data scale/serial access); currently dCache
  - **Analysis** compute (smaller data scale/more random/repeated access); currently dCache/EOS (for CMS)
  - **Interactive** compute (code development data scale); currently BlueArc/NAS
  - Similar hardware configurations across the first two
  - Interactive storage is currently a high $/TB

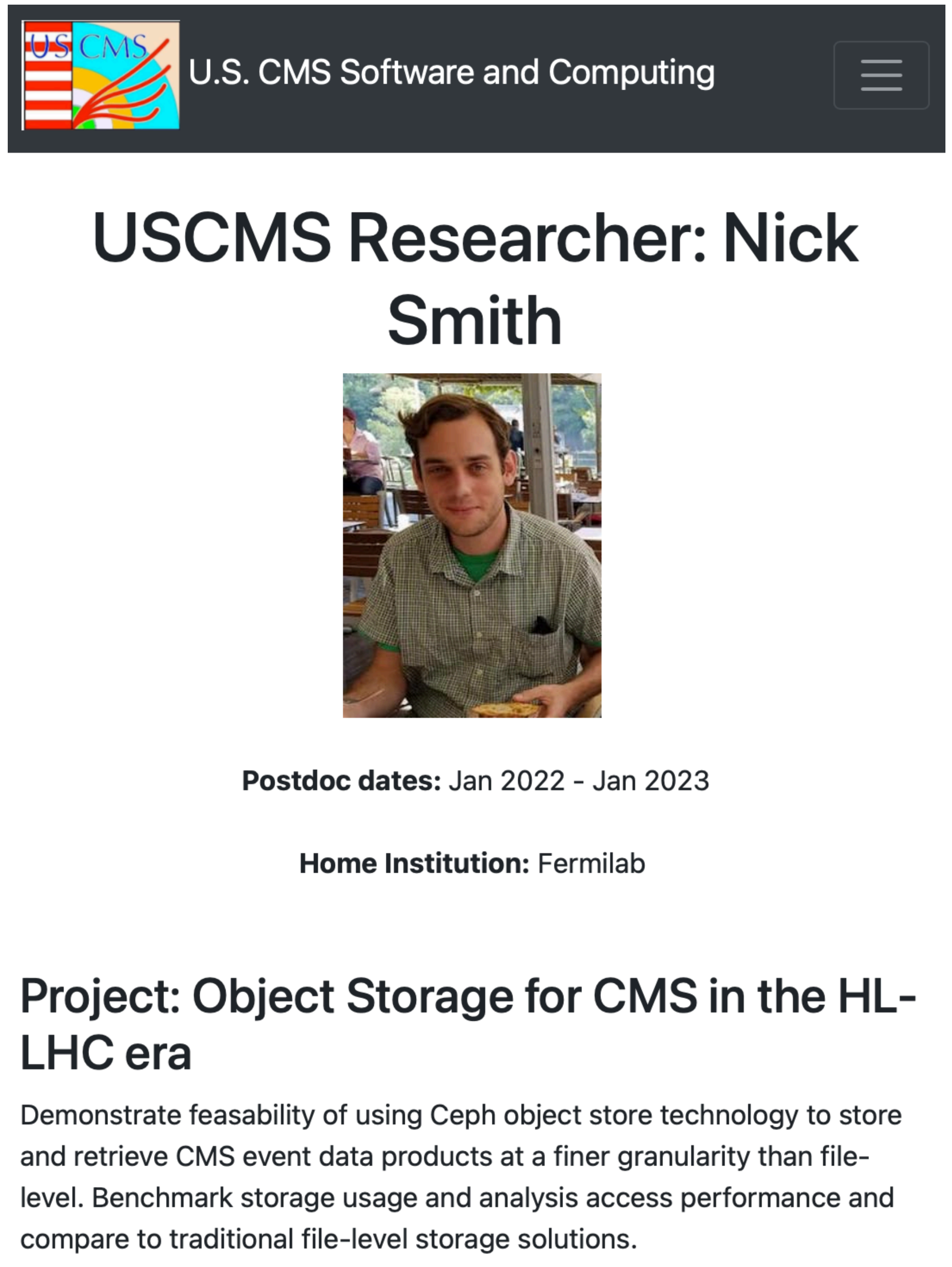
- The future
  - Move towards optimization for use cases while maintaining underlying commonality

Disk Nearline

- **Ceph** may offer multiple storage solutions in the future
  - As an **underlying storage system** with non-RAID resilience; CephFS
    - Would allow deployment of JBOD-based hardware
    - Could still run dCache on top
    - Potentially a replacement for existing NAS for interactive use
  - As a source of **object stores** for HEP use
    - Could reduce dependence on derived/reduced data formats
  - Storage system compatible with container orchestration (e.g. OKD)
  - Allows for **erasure coding** to save on raw disk space
    - CMS Tier-2s are considering Ceph, HDFS3 and EOS all in part for support of EC
- Ceph R&D efforts are a high priority in 2022
  - USCMS Operations funded project (PIs Jayatilaka and Mason) for object stores in CMS
  - Explore use of object stores for LArTPC events (particularly for DUNE)

Object storage for CMS project

- Awarded 0.5 FTE postdoc funding for one year (2022)
- Quarterly milestones (from project plan)
  - Month 1-3: Familiarization with Ceph and development of object/metadata scheme for miniAOD. Demonstrate ability to store and retrieve objects.
  - Month 4-6: Upload of collision and simulation data to Ceph as objects/metadata. Development of analysis code to retrieve objects from Ceph.
  - Month 7-9: Formulate an automatic workflow to move data in and out of this system. Benchmark performance of analysis code using object storage and compare to using analysis ntuples.
  - Month 10-12: Scale testing with multiple users. Present results at international HEP Computing meetings/workshops. Stretch goal: Work with US Tier-2 sites to establish object store data lake prototypes.



USCMS Researcher: Nick Smith

Postdoc dates: Jan 2022 - Jan 2023

Home Institution: Fermilab

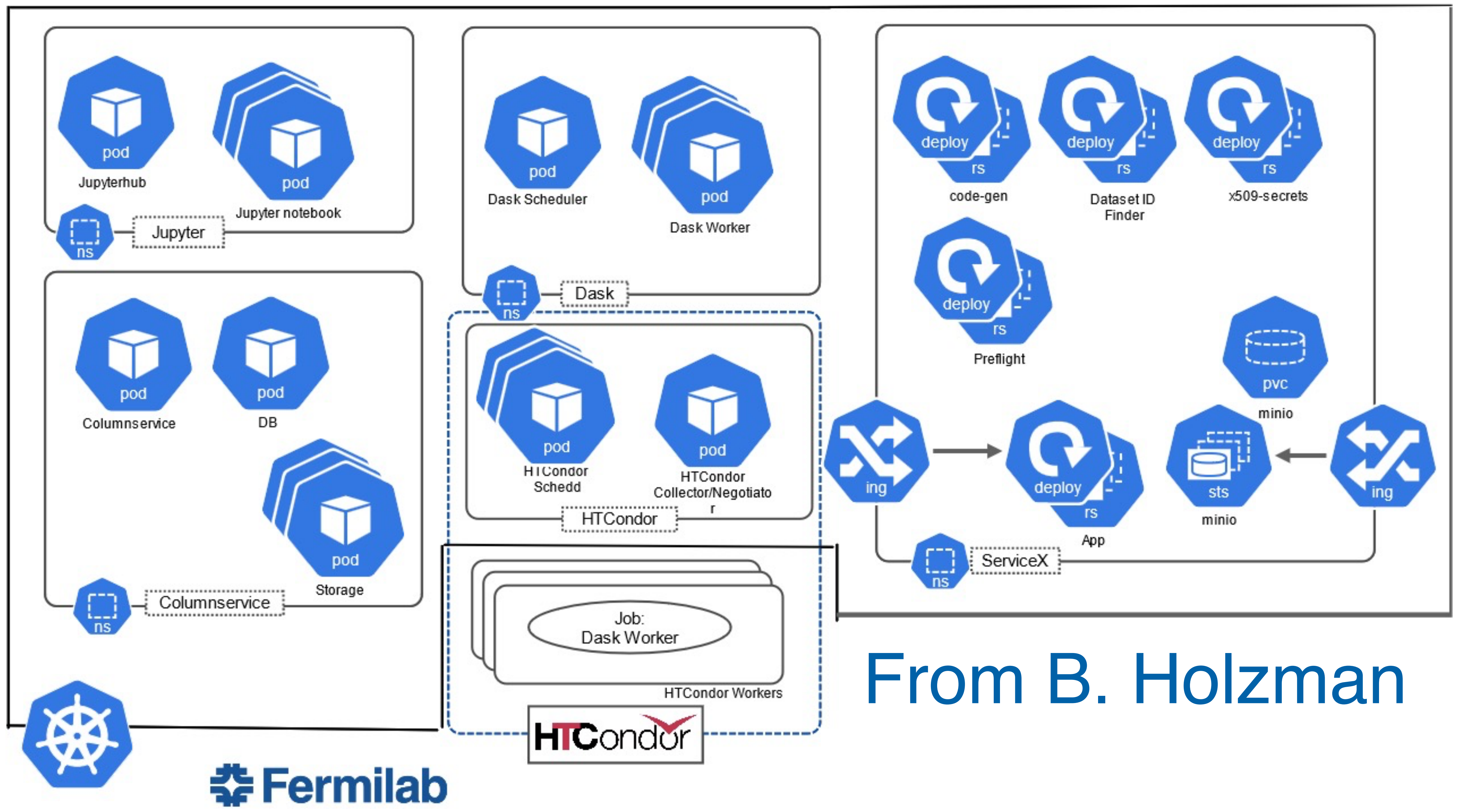
Project: Object Storage for CMS in the HL-LHC era

Demonstrate feasability of using Ceph object store technology to store and retrieve CMS event data products at a finer granularity than file-level. Benchmark storage usage and analysis access performance and compare to traditional file-level storage solutions.

https://uscms-software-and-computing.github.io/postdocs/nsmith-.html

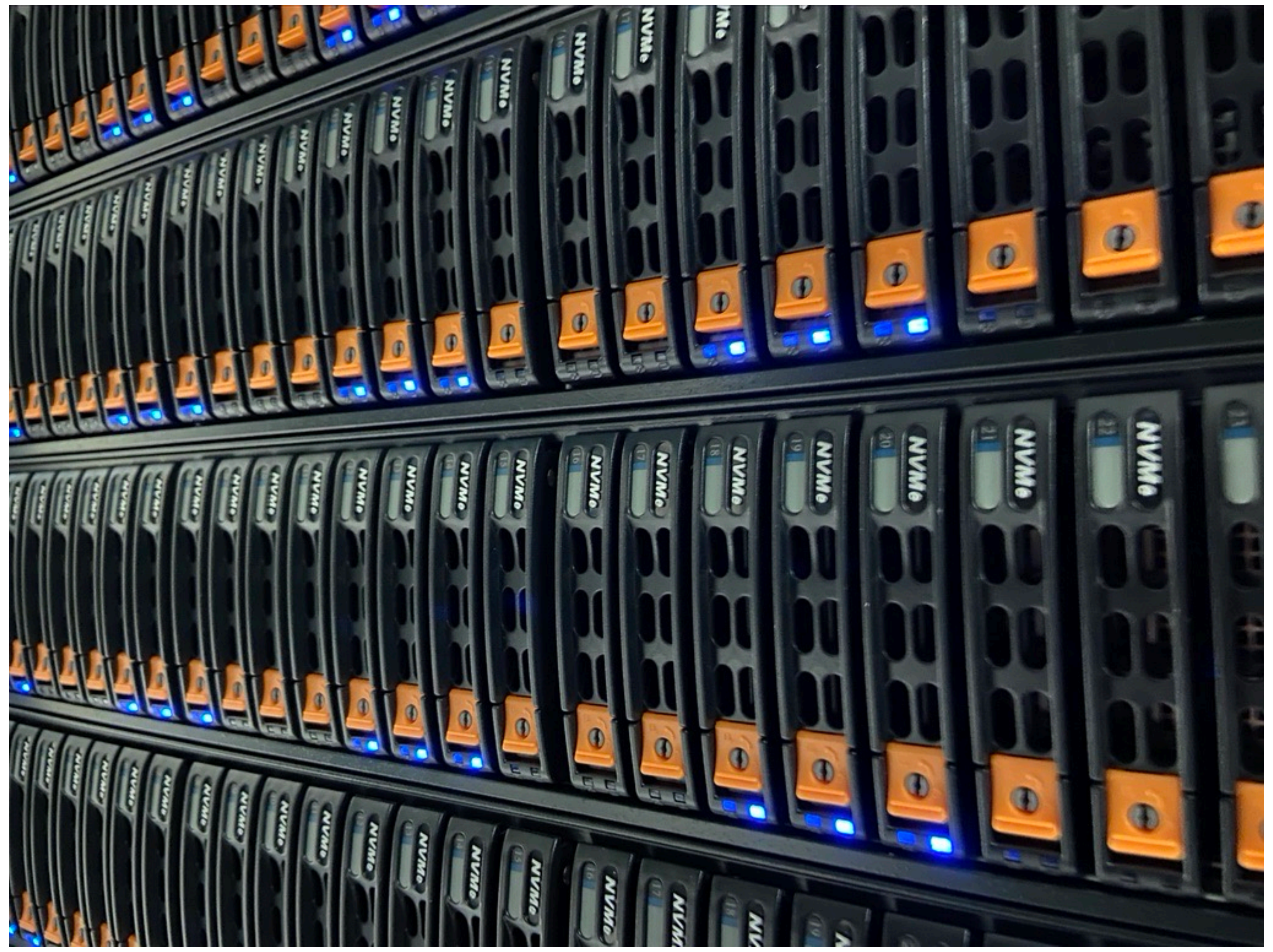
Disk

- **Elastic Analysis Facility** prototype being developed
  - Will be optimized for users to run analysis workflows
  - Storage solutions are currently existing ones
- Future task: optimize storage for analysis
  - Will need to analyze how users are accessing/using data
  - May benefit from high-speed storage and/or dedicated caches



From B. Holzman

NVMe-based storage servers at FCC2



Networking

- Fermilab being a data-centric site requires robust networking
  - Integrate networking R&D infrastructure into production infrastructure
  - Partnering with ESNet is essential
- Treat networking as a managed resource
  - Requires an end-to-end vision
  - Managing LAN connections as well as WAN may be necessary
- Goal: achieve **terabit scale** by HL-LHC start (~2029)
  - Accomplish via end-to-end managed connections
  - Effort will come from USCMS Operations as well as Fermilab
  - Closely integrated with Storage R&D efforts
  - Work with external partners including ESNet and HPC centers

Conclusions

- Fermilab’s future in computing will emphasize data storage
  - Data storage needs will be measured in EB/year
  - Will continue to be a multi-tenant environment
  - Will need to serve a greater variety of compute resources
- Future data storage architecture
  - Move towards more community solutions and away from home-grown
- Major projects in the coming year
  - Achieve readiness to transition from Enstore to CTA for tape storage
  - Explore use of Ceph for scientific data including as object store
- Continue to take part in community activities around storage
- Much progress already
  - See next talk