CREST: Computing Resources Evolution Strategy

CREST editor team

CSAID Roadmap Meeting

November 30, 2023

CREST Editor Team

- 10-year facility size planning ➜ Lisa Goodenough
- Facility machine rooms ➜ Atiq Siddiqui
- Processing ➜ Farrukh Khan
- Storage ➜ Rafael Rocha, Lorena Lobato Pardavila, Dmitry Litvintsev
- Networking ➜ Phil Demar
- User-facing facilities ➜ Burt Holzman
- Cybersecurity ➜ Mine Altunay
- Databases ➜ Steve White
- Monitoring ➜ Kevin Retzke

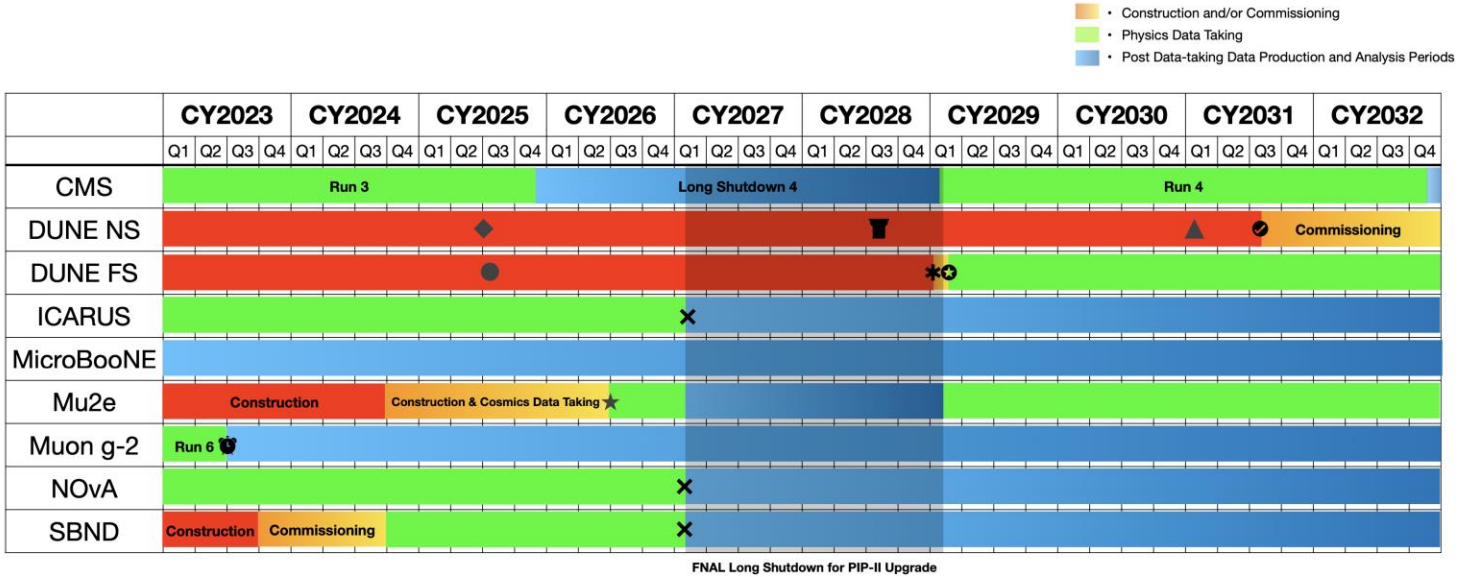
Strategy discussions

- Every Friday (if not canceled) 2 PM central in FCC1 and on zoom
- Indico category: https://indico.fnal.gov/category/1465/
  - Every strategy discussion comes with its own indico agenda and linked notes document
- Announcement and discussion:
  - Listserv: crest-discussions@listserv.fnal.gov
    - Indico sends event invitations to this list
  - FNAL computing slack: #crest-discussions

- Past and future strategy discussions:
  - 7/20 and 9/7: Storage
  - 9/15: Networking
  - 9/29: Processing
  - 10/6: 10-year resource projections
  - 10/20: Processing and GPUs
  - 11/3: User-facing facilities
  - 11/10: Databases and monitoring
  - **12/1: Stakeholder decision matrix and guiding principles**
  - **12/8: Cybersecurity**

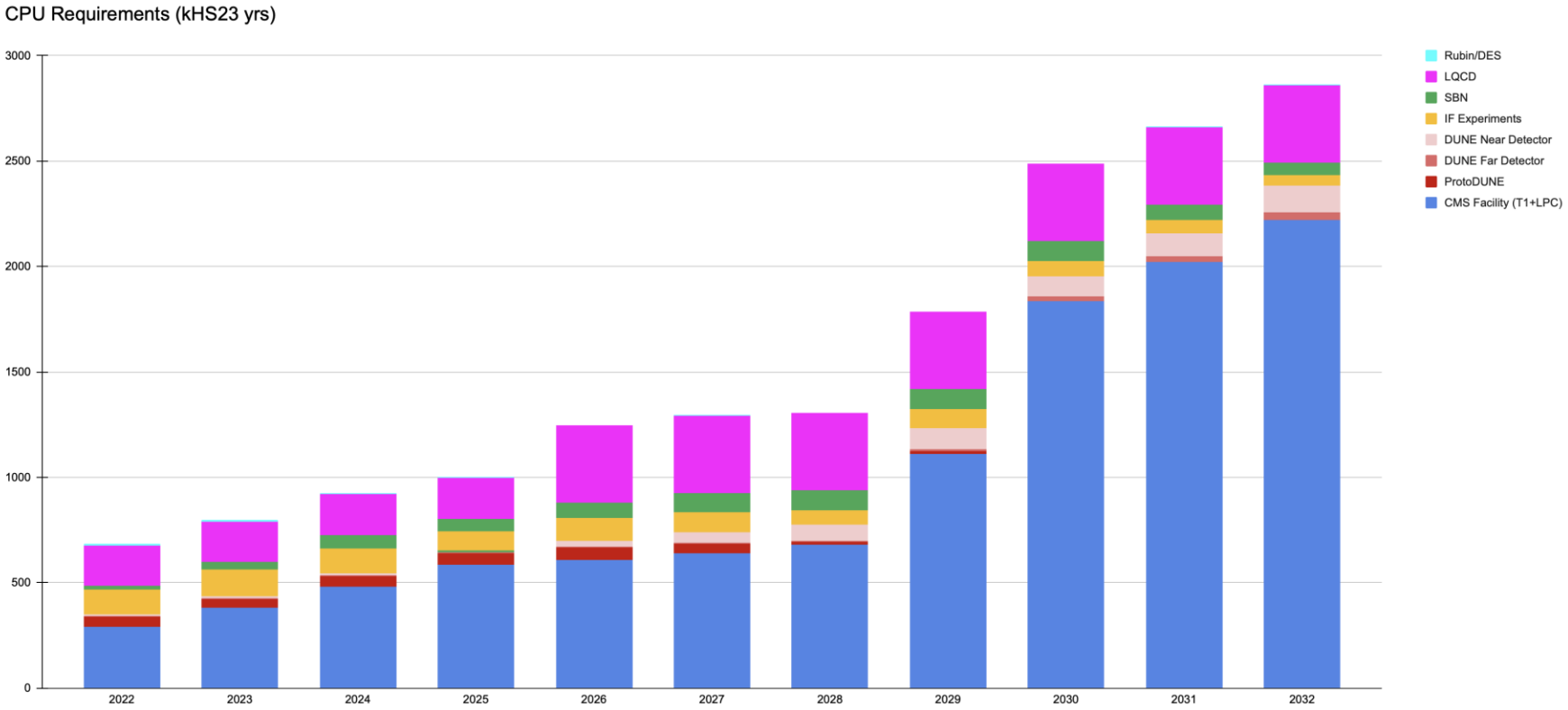
10-year facility size planning

10-year facility size planning: Current status Schedule



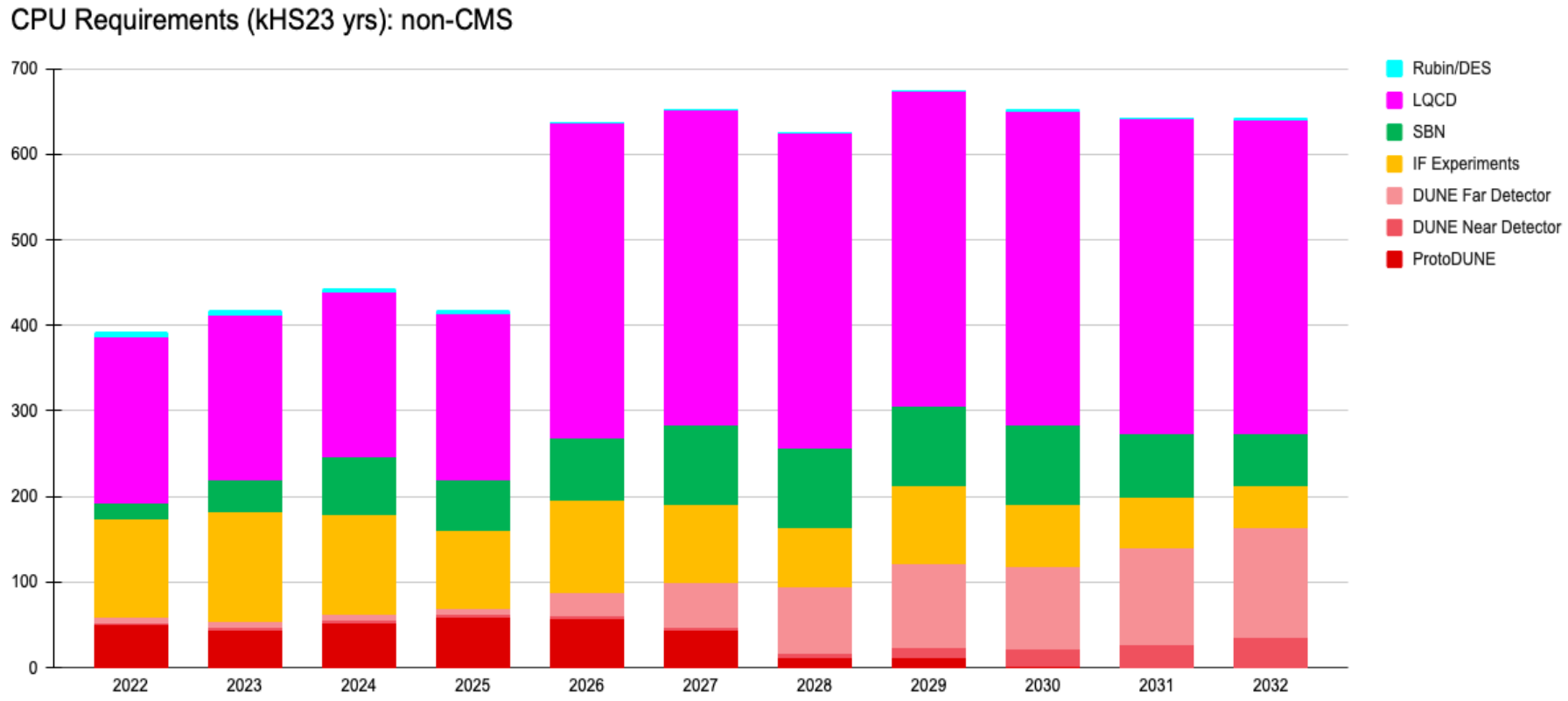
- Muon g-2: End of Beam Data Taking
- Mu2e: Start of Commissioning with Beam & Physics Data Taking
- ICARUS, NOvA, & SBND End of Beam Data Taking
- DUNE Start of Near Site Facilities Construction
- DUNE Near Site Hall Construction Complete
- DUNE Near Site Start of Beam Operations
- DUNE Near Detector Online
- DUNE Far Site Building and Site Infrastructure Complete
- DUNE Start of Far Detector #1 Commissioning
- DUNE Start of Far Detector Science

10-year facility size planning: Current status All CPU



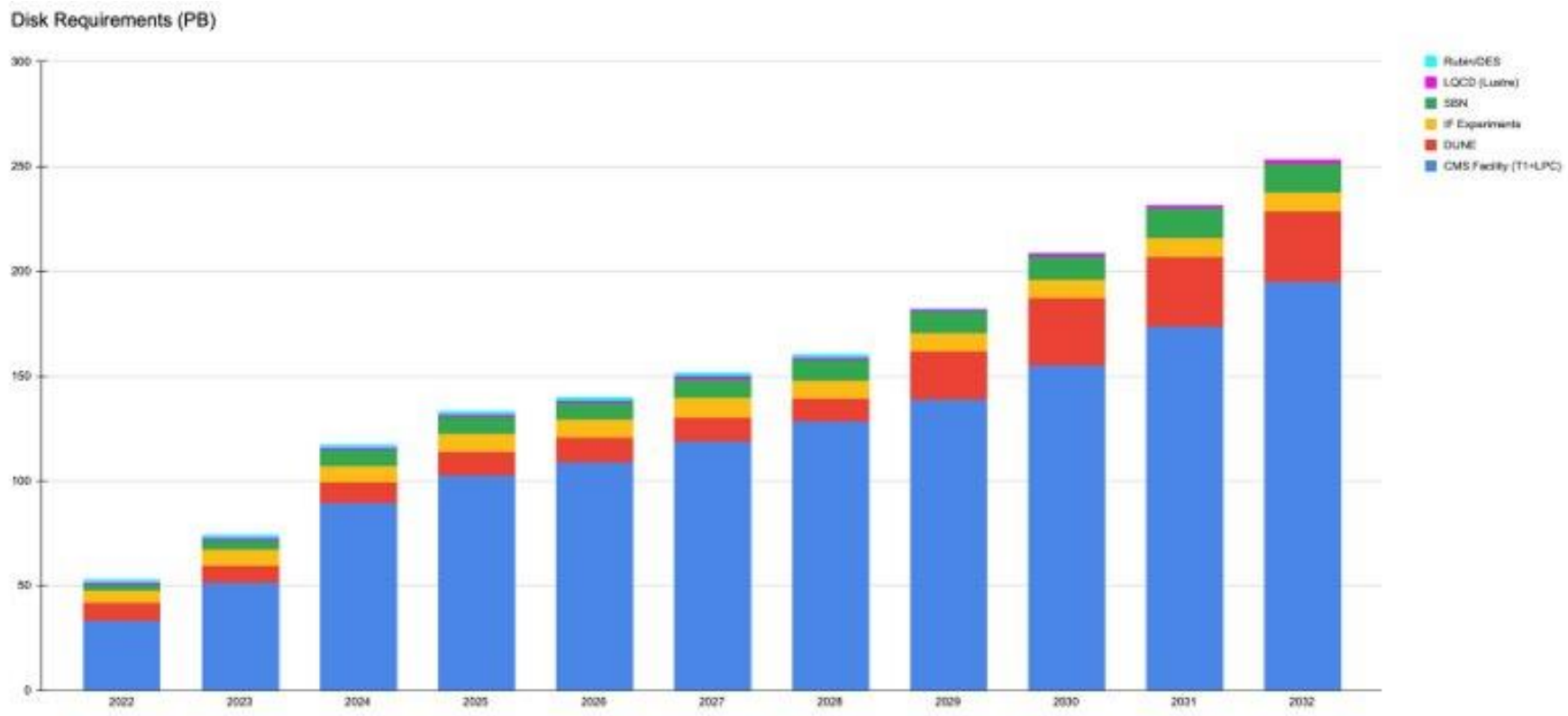
Fermilab

10-year facility size planning: Current status non-CMS CPU



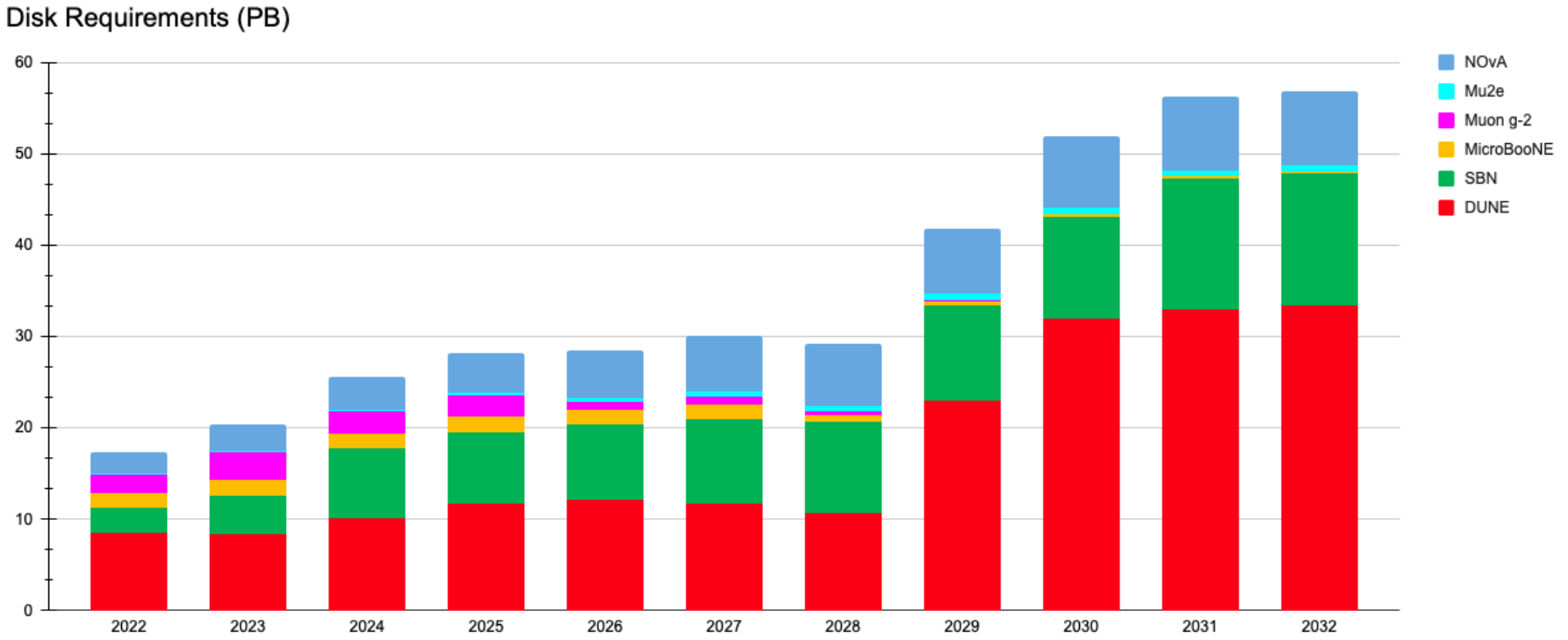
Fermilab

10-year facility size planning: Current status All Disk

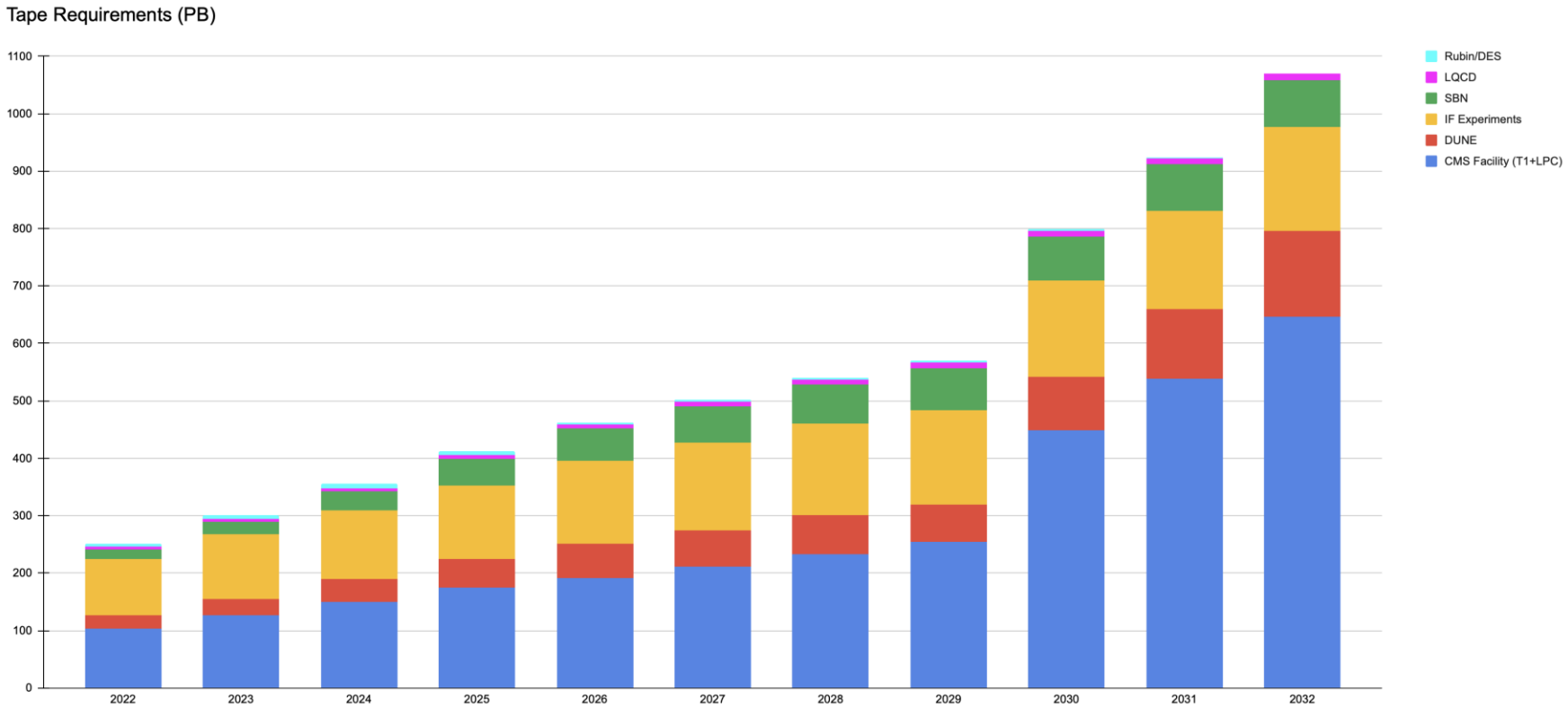


Fermilab

10-year facility size planning: Current status IF Disk



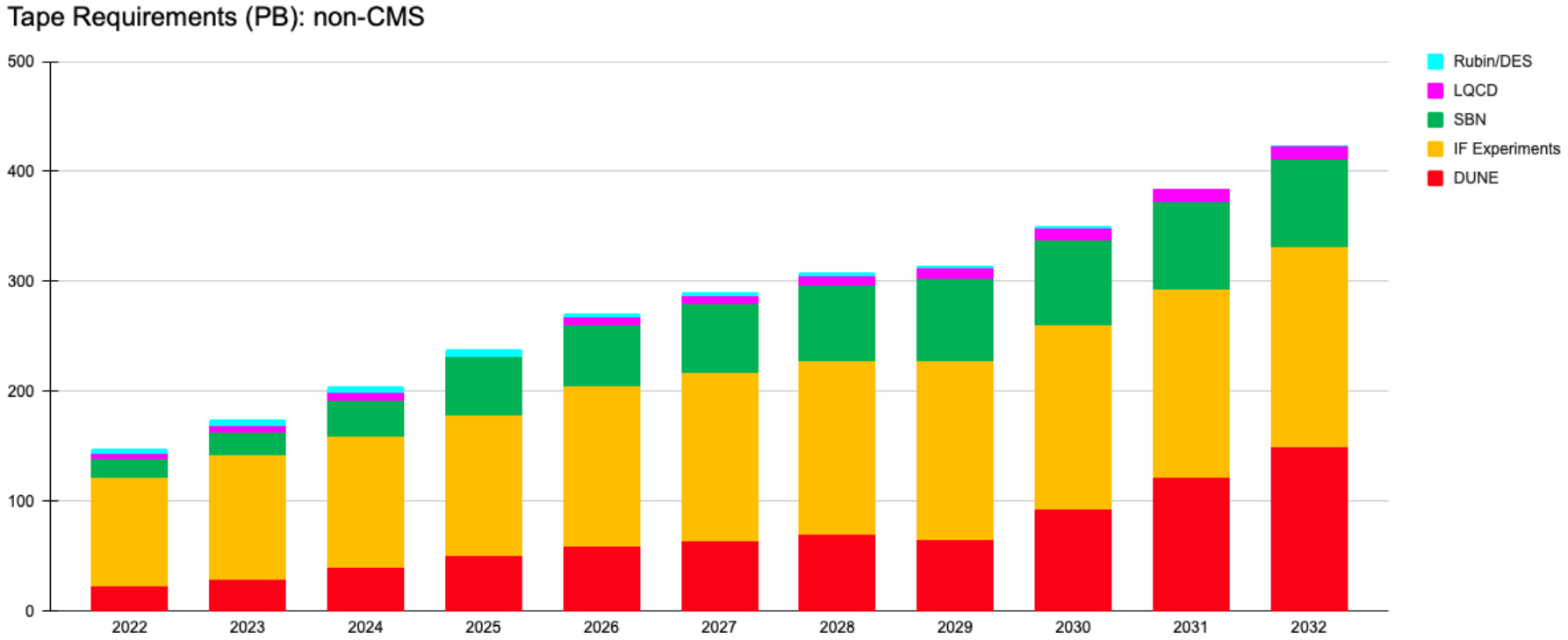
10-year facility size planning: Current status All Tape



Dune enhancement factor 1.5 2029-2032

Fermilab

10-year facility size planning: Current status non-CMS Tape



No DUNE enhancement 2029-2032

Fermilab

Facility machine rooms

Facility machine rooms: Current status

- Two Primary Data Centers Facilities
  - Grid Computing Center (GCC)
    - Compute, Networking and Tape Libraries
    - No generator backup - End of Life: 2030
  - Feynman Computing Center (FCC)
    - Networking, Misc. IT, Storage and Tape Libraries
    - Generator backup available - End of Life: 2035

Total Capacity	Power (UPS)	Cooling	Racks	Rack Density
GCC (CRA, CRB, NRA, NRB, TR-A)	1.62 MW	~1.7 MW	193+3 TL	10 KW / Rack
FCC (FCC2 & FCC3)	1.2 MW	~1.1 MW	253+4 TL	5 KW / Rack

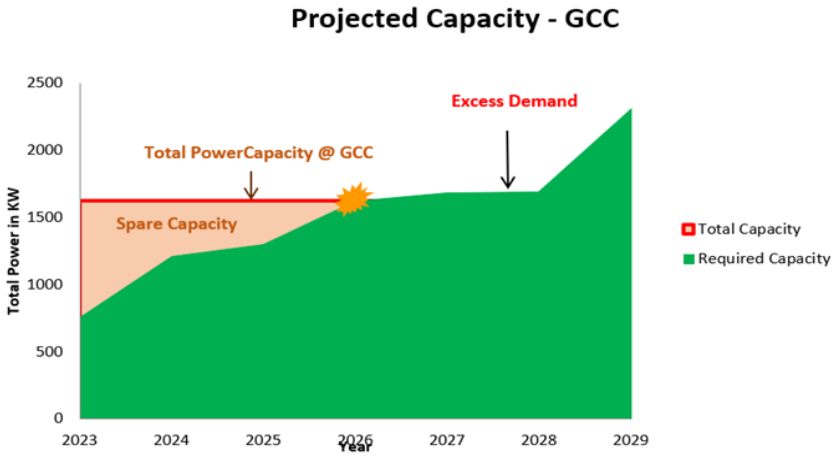
Facility machine rooms: Strategy

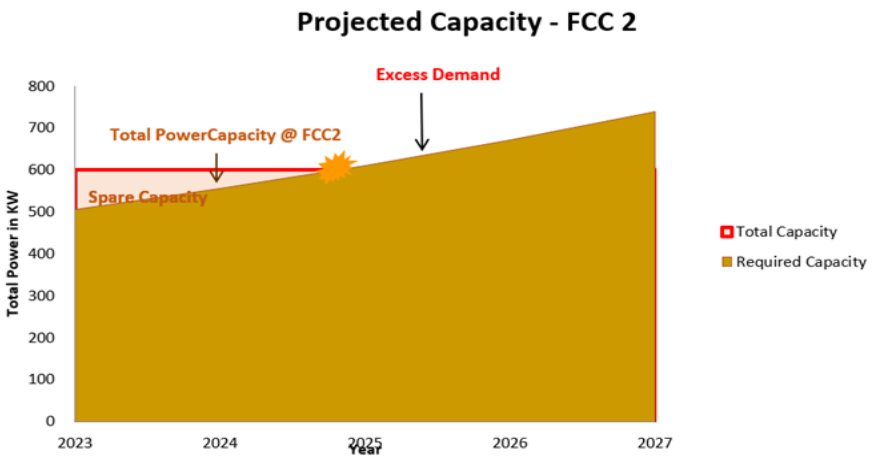
Based on the resource usage projection from experiments

- GCC will be reaching its full power capacity in 2026
- FCC2 will be reaching its full power capacity in 2025

Strategic Initiatives

- Repurpose Computer Room C to meet projected requirements for 2024-2030
- Design/Build a net-zero, scalable, and modular brownfield / greenfield computing facility to address laboratory computing needs from 2030 to 2050





Processing

Processing: Current status

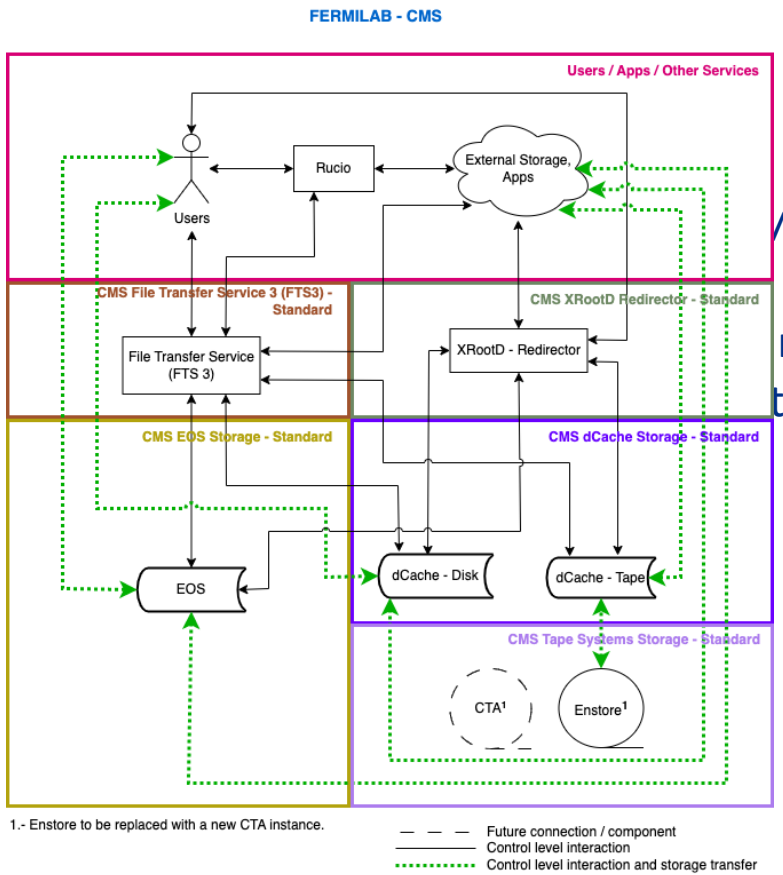
- There are two major processing platforms used across processing facilities:
  - HTCondor – developed at UW Madison and is used to schedule high throughput computing workloads
  - Slurm – a popular HPC processing platform developed by SchedMD
- There are four major facilities on-site:
  - FermiGrid Compute Facility – uses HTCondor and primarily serves Fermilab VOs other than CMS
  - CMS Compute Facility – uses HTCondor and primarily serves CMS
  - Wilson Compute Facility – uses Slurm and primarily serves Fermilab users, projects and VOs
  - LQCD Compute Facility – uses Slurm and primarily serves LQCD users
- HEPCloud Decision Engine is used in FermiGrid and CMS compute facilities to provision and access HPC and cloud resources
- In addition to the on-site processing facilities, we also run GlideinWMS pilot pools for CMS and DUNE. GlideinWMS is also used to augment FermiGrid compute capacity through pilots requested across grid sites
- FermiCloud is an OpenStack based service that was historically used for processing but is now largely used by CSAID developers as their playground

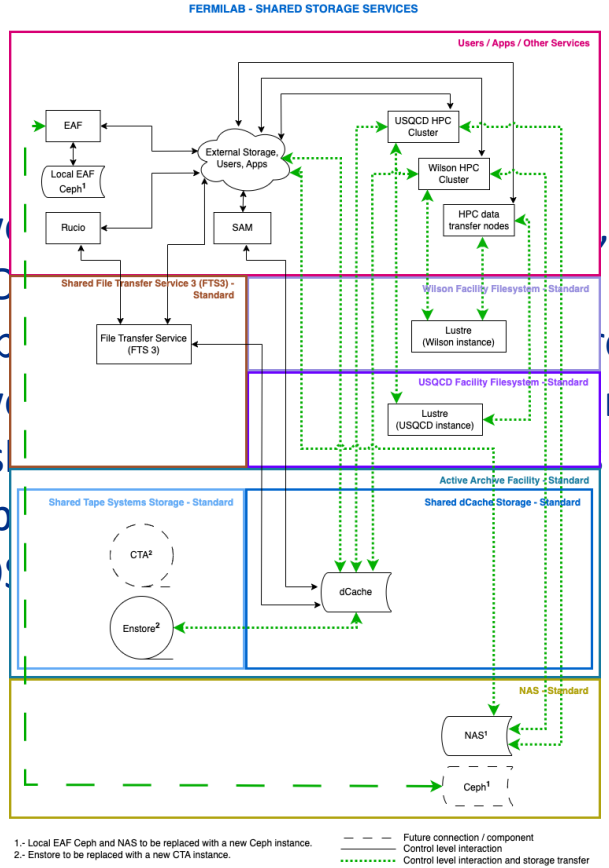
Processing: Strategy

- Processing platform should be able to handle a total of 10 million cores (on-site and off-site)
- Processing platforms and facilities should follow zero trust architecture principles and provide adequate traceability and isolation
- Processing platform should be able to reasonably accommodate evolving processing units and architectures (PPC, GPUs, DPUs, TPUs etc.)
- Processing platform should have the scheduling capabilities to handle direct acrylic graph and step/task chain workflows
- Processing platform should provide reasonable flexibility handle competing priorities and deadlines from experiments/VOs
- We should run a community adopted HPC processing platform (like Slurm) to provide our users a playground to develop their software and workflows to run at larger DOE HPC sites
- Processing platform services should leverage virtualization or containerization to maximize resource usage and minimize data center hardware footprint
- It is desirable, not strictly required, for the processing platform to provide power saving features and capabilities
- We are still investigating different use cases and storage/network access patterns for GPU workloads. This will help us recommend a strategy for future GPU deployment across our processing facilities and platforms
- FermiCloud is a useful service that needs sufficient effort and direction to evolve

Storage

Storage: Current status



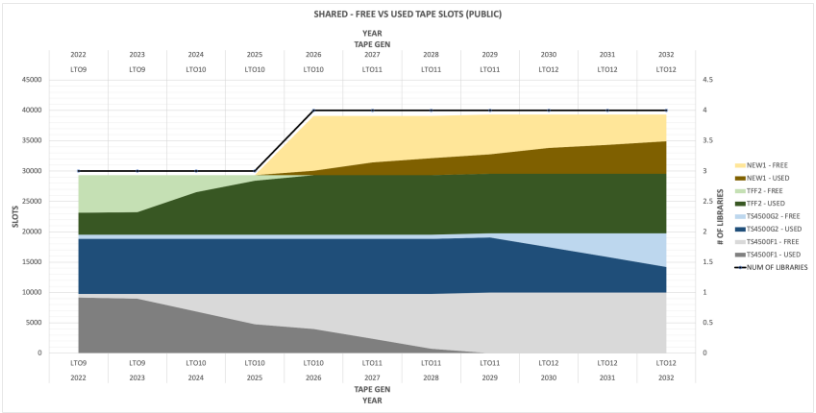


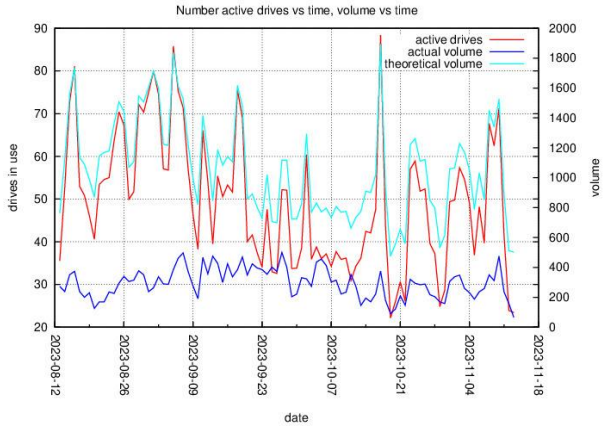
CMS

- Tw
- Tap
- Tw
- EO

Storage: Strategy

		Commercial (Example: IBM, HPSS)	Freeware (Example: Lustre, OSM)	In-House (Example: Enstore)	Collaborative (Example:CTA , Enstore*)
GUIDING PRINCIPLES	**Implementation-Independence:** Strive to describe the facility strategy without referring to specific implementations. Where necessary, we will give suggestions about R&D projects that need to be conducted to determine a path towards implementation	Product roadmap is not under control	There is not influence on development support	Product roadmap under control	We can suggest about R&D to other collaboration + adapt to our needs
	Stakeholder: Develop and update the facilities strategy to serve the scientific use cases of the Fermilab user community and the experiments forming the scientific mission of the lab. Also, strive to be closely connected to the planning of the scientific mission of Fermilab and try to anticipate future needs.	New features or modifications depend on the vendor	May have broad support community	We can add features modifications based on our needs	May face priorities differences with other members of the board
	Commonality: Develop the facility strategy striving to minimize the number of systems providing the same functionality with different implementations. The aim is to reduce operations costs by minimizing the number of systems that need to be supported and maintained.	Allow one solution for all, same functionality with different implementations + May provide a complete solution	There might be influence on ? development support, but we won't be the product owner. It might not be easy to be supported and maintained. *Example: Lustre. We pay a small membership monthly but we are not actively collaborating*	May evolve into endemic product used only by us	Minimize services to reduce costs + broaden - collaborative support = services are maintained and supported cost effectively
	Future-Capability: Strive to prepare the facility to add new capabilities easily to support and even enable future parts of Fermilab's scientific mission.	Full support by vendor while the product is alive. *An example would be Intel buying McAfee.* It may lead to vendor on both software & hardware (you need to request the addition of new capabilities and always depending on the vendor)	Uncertain long-term commitment	Sometimes we cannot add features because of lack of support. Mostly, support relies on us	Collaborative support and we can add or request implemntenation based on our needs
	Holistic-Cost: Aiming for a "Quality of Service-oriented architecture/strategy". The goal is to obtain the best quality of service for our facilities by optimizing cost, time and scope of the activities involved with it.	Expensive. Difficult to find a commercial solution that fulfills all our requirements, bridging the scope gap make come at additional cost. Most commercial solutions are backups oriented trying to be more profitable.	Free (only the software is)	Costly in support resources	Reduces and optimize costs + optimize personnel (shared resources). May will require resources investment depending on the MOU
	Environmental-Awareness: We will optimize the facility strategy to maximize scientific support while minimizing the environmental footprint of operating the facility.	Whatever the software-ish selected, needs to take advantage of the HW			
	Re-Usability: Will develop a facility strategy that maximizes the reusability of capabilities for different experiments and user groups.	Sometimes it is not possible to re-use capabilities for different experiments and user groups	Re-usability of capabilities for different experiments and user groups		We assume other collaboration has similar cases, so we can anticipate the capabilities to be (re)used
	Isolation: The facility strategy needs to guarantee users and user groups to be isolated from each other while operating on resources at the facility. This not only concerns cybersecurity but also~~performance related~~ issues. The performance of applications of user and user groups should not be impacted by the performance of applications of others.	The performance of applications of user and user groups might be impacted by the performance of the solution does not fulfill our needs. Cost dependency also. NOTE: It's yellow because of the probability of finding a solution of something that covers everything	The performance of appliations fits the purpose when the solution matches our use cases	The performance of appliations should not be impacted. User groups and users to be isolated	The performance of appliations should not be impacted. User groups and users to be isolated
	Increment: The facility strategy should follow an incremental approach of changes and additions and avoid disruptions where possible.	Dependent on vendor so the changes and additions do not follow incremental approach	The changes and additions + avoid disruptions if solution matches our use cases	Product roadmap under control, being able to follow an incremental approach of changes and additions	Changes added following the roadmaps of the other collaborations or we can do it but we might need support(Product road map may not be under our control)
	Accountability: Part of the facility strategy is to track the cost of providing services to the scientific community including both hardware and person power aspects	We pay, we have the track of the cost of services	Opensource or in-house resources , we track our own cost of services		We track our own costs since this solution is mostly also open-source + shared resources

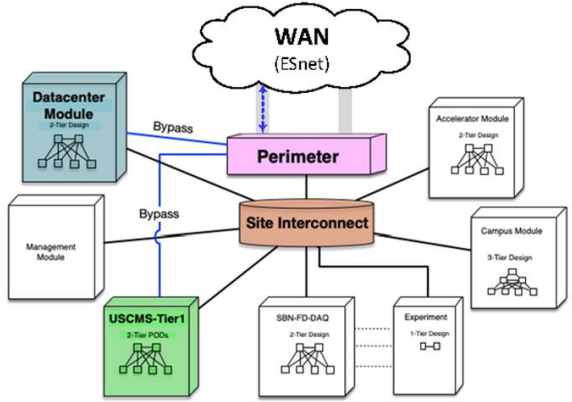


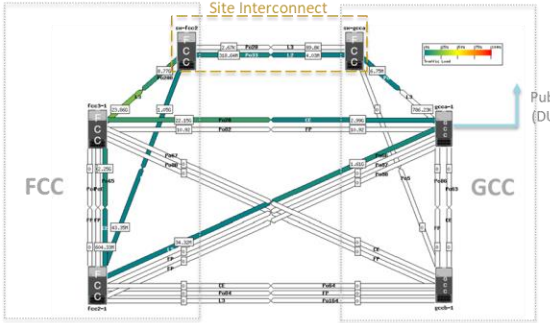


Networking

Networking: Current status (I)

- FNAL campus network has modular design:
  - General Data Center module
  - US-CMS Tier-1 module
  - Site Interconnect for inter-module connectivity
  - Perimeter module delineates off-site
- General Data Center & US-CMS T1 modules:
  - Large distribution switches in FCC2, FCC3, GCC-A, & GCC-B
  - Top-of-rack (ToR) access switches in server racks
  - ToR switches dual-homed to FCC & GCC distribution switches





General Data Center module

Fermilab

Networking: Current status (II)

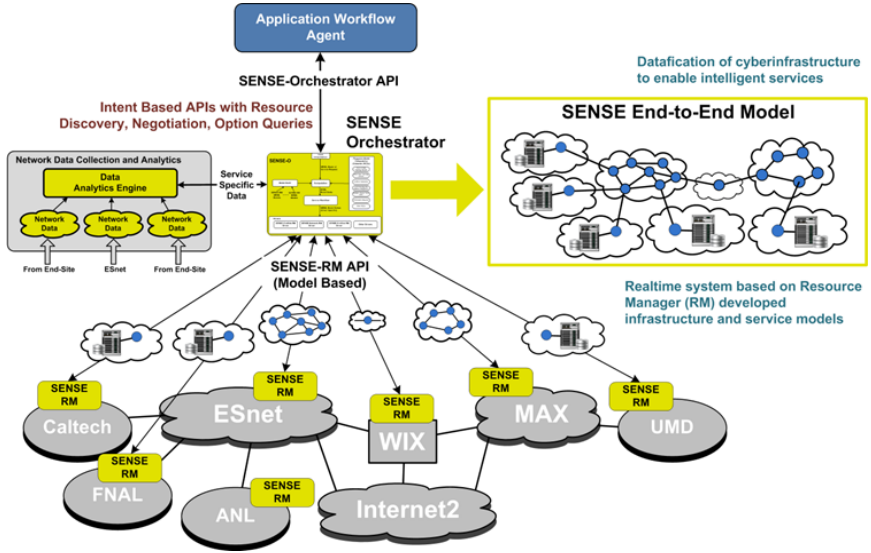
- Network technology bandwidth trends:
  - Historically characterized by 10x increases bandwidth every ~6-8yrs
    - 10M/100M/1GE/10GE/100GE
    - Switches/routers with corresponding increases in fabric capacity
    - Starting with 400GE, bandwidth increase is only 4x
    - Next step-up will be 800GE - only 2x
    - But bandwidth step-ups will come faster, with flatter cost curves
      - Leads to greater use of Link Aggregation Groups (LAGs) for high bandwidth environments:
      - LAG = N x 100/400/800GE (where N = 2, 4, 8, etc.)
  - May necessitate high fiber demand within & between data centers

Networking: Strategy (I)

- Physically, network support needs shouldn't be significantly different than today
  - Anticipate just tracking the bandwidth technology curve
    - Replace distribution & ToR switches currently in place as requirements, technology, and budgets allow
  - Data Center (DC) network infrastructure footprint isn’t expected to change
    - Should not require additional DC rack space, power, cooling
  - Caveats:
    - May be increased demand for intra-DC and inter-DC fiber
    - Emerging network services & monitoring capabilities may need supporting computing resources

Networking: Strategy (II)

- Support for software-defined networks (SDN) will be needed::
  - SDN = virtualized overlay network paths
  - Intended to provide specific service needs (ie., QoS) and/or isolation
  - Currently evaluating SENSE with Rucio
- Biggest planning uncertainty is how networks will be used:
  - Networks are key for distributed computing, connecting processing with storage
  - How data is moved & requirements on that data movement will determine what networks need to be capable of



Fermilab

User-facing facilities

User-facing facilities: Current status

Note: user-facing facilities are primarily composed of processing, network, and storage services, significant overlap with other CREST areas

- Shell + Batch:
  - **GPVM**: Virtual machines tailored to experiment, interactive use, NAS storage
  - **CMS LPC**: Virtual machines tailored to CMS, interactive use, CMS NFS
  - **LQ1/LQ2**: Batch with interactive access (USQCD only), separate storage
  - **Wilson Cluster**: Batch with interactive access, separate storage plus NAS /home
  - **DES Cluster**
- Web + Batch + Other:
  - **Elastic Analysis Facility**: Jhub + ancillary services + access to batch
- Misc:
  - **FermiCloud**: VMs mainly for developers
  - Small clusters all over the lab

User-facing facilities: Strategy

- Our strategy should emphasize flexibility and agility
  - We do not know how far the pendulum will swing between the "old" and "new"
  - Developments in this ecosystem occur in a much more rapid timeframe than our ability to strategize about it
    - Agility in hardware
      - Keep modularity in mind at procurement
      - De-silo / consolidate facilities where possible
    - Agility in software
      - Deploy/use flexible infrastructure (when it makes sense)
      - Continue to rely on generalized provisioning tools
    - Note this comes at a cost!

Cybersecurity

Cybersecurity: Strategy

- **Federation will be a significant part of our future strategy and handled centrally by SNOW.**
  - Moving our data sources to SNOW. Centralized data processing. Ferry and DocDb will be early adopters. SNOW will process non-Fermilab and international users faster and provide the list to all applications.
- **Moving away from IGTF**
  - User certificates will become obsolete by 2026. This should reduce our dependency on IGTF since we can get host certificates from the same issuers as the rest of our lab services. We will also move to ACME based systems.
- **Impact of the new cyber security requirements on the science.**
  - We expect **MFA** will have no impact on non-interactive batch job submission and data processing. **Zero trust** will have a smaller impact, micro segmentation of network, some systems may need new IP addresses. **Disk encryption** is complete and implemented, it will have no impact on scientific systems as long as they have no sensitive data and all physical security requirements are in place. **Logging enhancements** will impact scientific services, all services will be required to send real time data to central logging facility. All **web services** will go through a web application firewall. **Vulnerability management** requires timely fix of all vulnerabilities, otherwise service will be blocked from the network. Fermilab employees will not be allowed **BYOD**. Unclear how collaborators will be impacted, only collaborators with Fermilab emails will be impacted.

Cybersecurity: Current status

- Federated identity project is our top priority
  - Job submission has been moving to tokens.
  - Local experiments, not dependent on storage, has been moving to token-only.
  - Mu2e, Icarus, SBND has been moving to tokens successfully.
  - Storage (FTS and Rucio) expected move to tokens in the next year.
  - CILogon Token Provider has implemented changes to be compatible with Rucio and FTS.
- Moving away from user certificates, while providing host certificates from InCommon
  - Some operational issues indicating that we should reduce our dependency on IGTF.

Databases

Databases: Current

- Database are supported jointly by ITD and SCD.
  - Hardware and VMs are provided by both departments, providing the resource limits are followed by the experiments.
  - ITD provides database system support, SCD provides Database application support with some joint effort for experiment issues.
- There are currently 1150+ PostgreSQL/MariaDB relational databases in use, all of whom must be backed up, administered and updated.
  - The growth in number and size of databases increases over time.
    - We have some DBs over 1TB and IFBeam is approaching 20TB (this DB, required for analysis it would take 1 week to restore).
  - At last count we have one NoSQL database in production. This is a growing area, more are expected.

Databases: Strategy

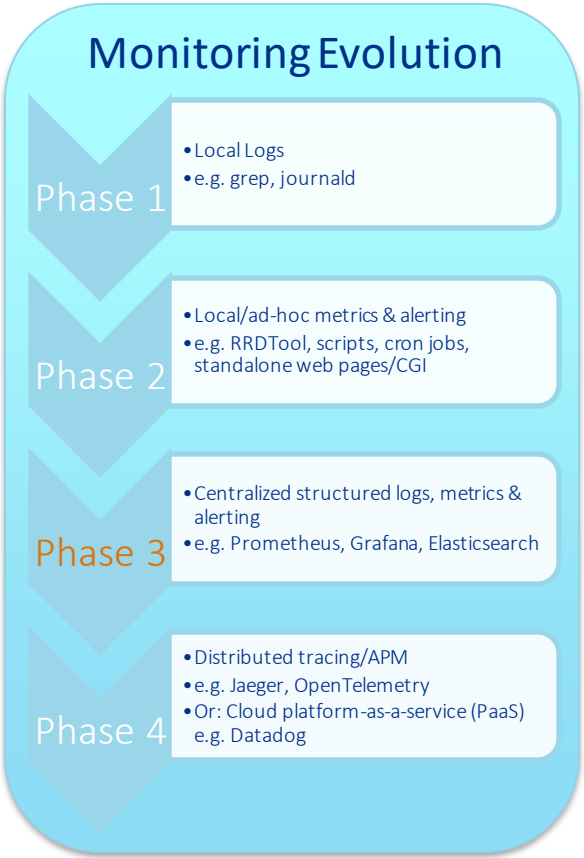
- Database maintenance workload and resources are on an increasing slope. They are easy to create, near impossible to shutdown.
  - Investigate if there is a better way to handle their maintenance for backups and upgrades. Can we determine what databases can be retired as this number grows?
    - For maintenance, containerizing databases is being looked at.
  - As experiments age the size of their databases also grows. Will current backup/recovery strategies handle increases over the next 10 years?
    - DUNE alone is expected to produce 25 years of data which will affect various databases. (IFBeam, which DUNE will use is currently 20TB with only under 10 years data.)
  - Size estimates for new databases are consistently low. How can we get accurate estimates?
- Create a joint SC & ITD team to provide a comprehensive strategy.
  - Would it be beneficial to manage databases as a facility?
  - Should an experiment's database schemas be reviewed before databases are created?

Monitoring

Monitoring: Current status

Definition: **Monitoring** ("observability") is the collection, storage, analysis, and dissemination of information about **operational** systems and services, and includes logging, metrics, tracing, alerting/paging, and reporting/dashboards.

- Landscape is largest centralized monitoring platform in CSAID, similar to MONIT at CERN, integrating data from:
  - Batch systems (e.g. HTCondor, SLURM, jobsub, HEPCloud)
  - Storage systems (e.g. dCache, enstore, NFS, SAM, EOS, RCDS/CVMFS)
  - Transfer services (e.g. IFDH, FTS)
  - Internal services (e.g. FERRY, POMS, Redmine, Landscape itself)
- There is also substantial monitoring outside Landscape, e.g.
  - SSImetrics (metrics, reporting)
  - CheckMK (alerting, paging via Service Now)
  - Various service/area-specific solutions, e.g. network throughput (metrics)
  - ITD, cybersecurity have their own platforms and ecosystems, e.g. Splunk (logging), CrowdStrike (endpoint protection)



Monitoring: Strategy

- Centralized observability platform provides many advantages
  - Integrating data from multiple services has a synergistic effect, allowing correlation of data between services - "single pane of glass"
  - Robust data storage & backups
  - Reduce duplication of effort & possibility to leverage components elsewhere, e.g. message queues
- … but has many considerations
  - Who needs to access the data, how will it be used?
  - Security? (especially with logs)
  - Integration friction and onboarding time?
  - Platform team effort, platform operations and maintenance overhead
  - Requires buy-in from developers and operators (along with other stakeholders)
- Challenge: service and system developers and operators may have existing observability solution, or platforms may provide built-in solutions
  - Lack of cohesive vision/silos lead to independent solutions
  - May be substantial effort to migrate, but sometimes trivial thanks to standardization efforts

Discussion