Scientific Computing Facilities

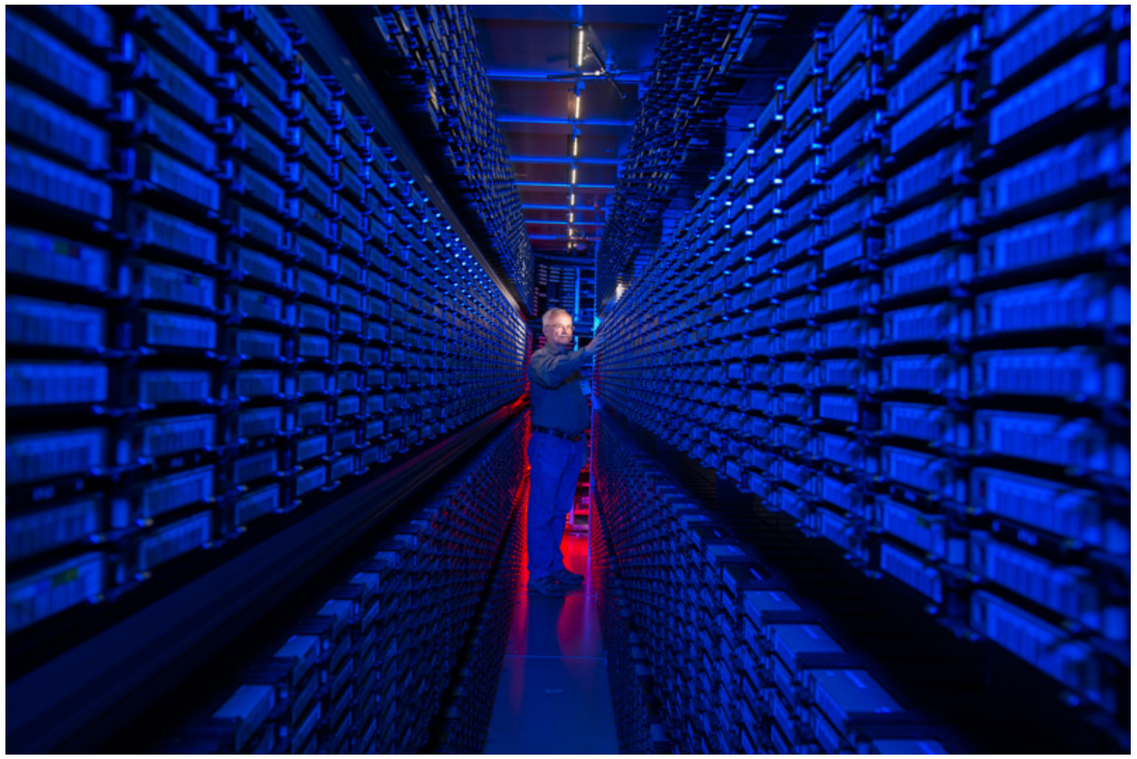
Bo Jayatilaka

Computational Science Working Group Pre-Meeting

12 April 2017

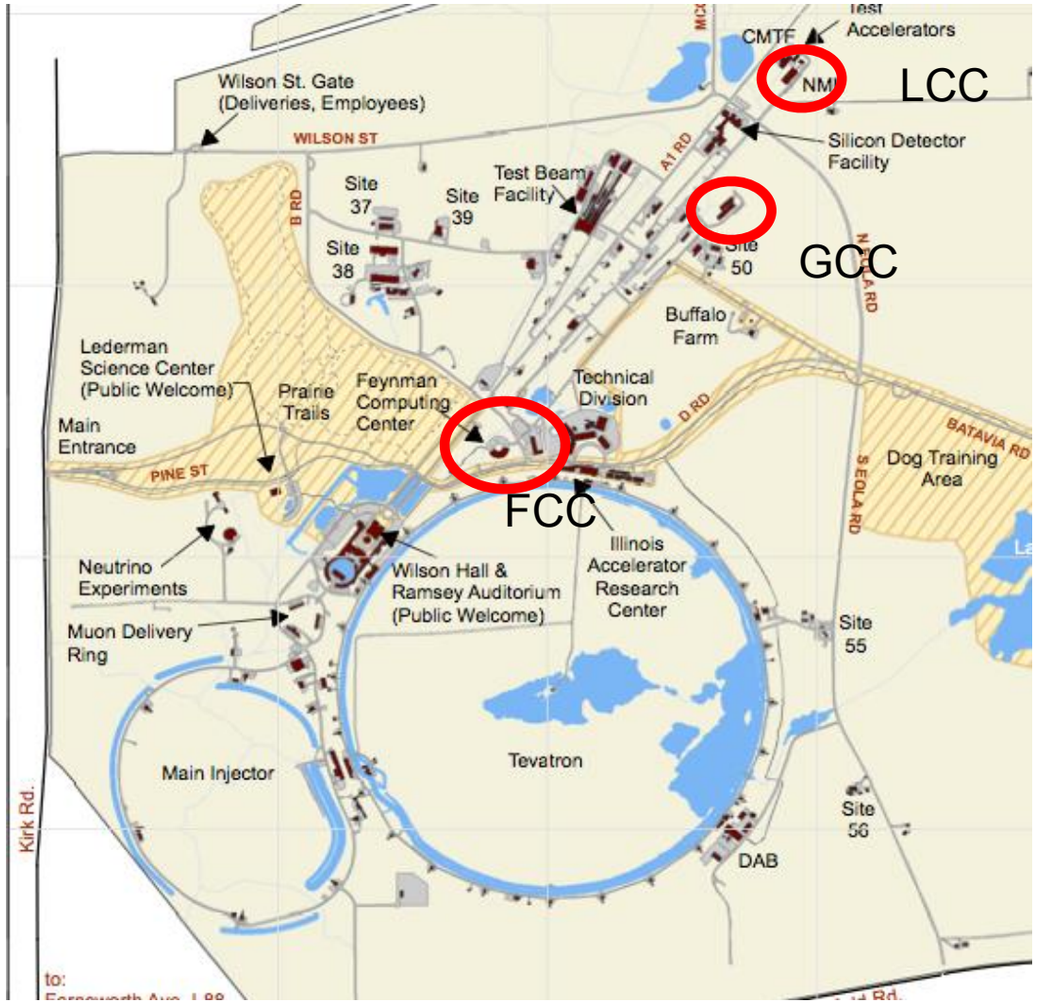
Outline

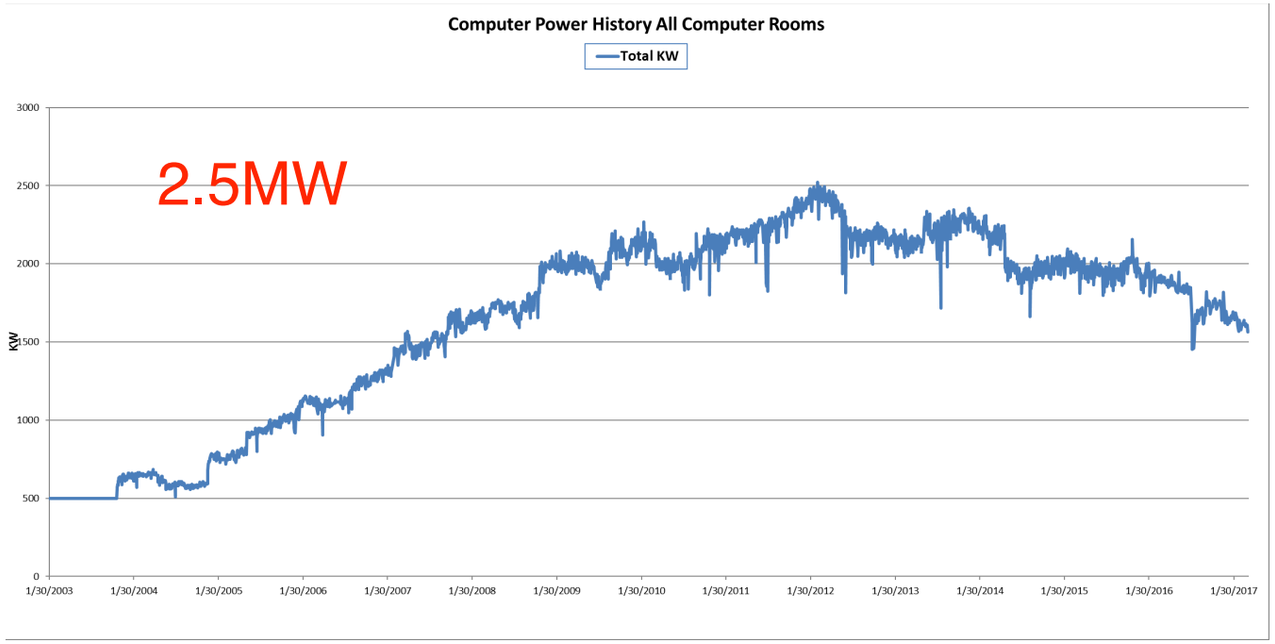
- Overview of Fermilab Computing Facilities
  - CPU
  - Tape/Disk
- External computing facilities used by Fermilab experiments
- Some thoughts on future directions



On-site computing facilities

- **Feynman Computing Center (FCC)**
  - 2 rooms with 0.75MW nominal cooling and electrical power each
    - UPS with generator backup
  - Hosts power-critical services
    - Central services (mail, web servers, etc.) and disk servers
- **Grid Computing Center (GCC)**
  - 3 rooms with 0.9MW nominal cooling and electrical power each
    - UPS with taps for external generators (no permanent generator)
  - Hosts CPUs and tape libraries
- Lattice Computing Center (LCC)
  - Being decommissioned





Fermilab

CPU

- Most scientific computing at Fermilab is done via **High-Throughput Computing** (HTC)
  - Most jobs do not require to talk to each other while running
  - Job submission almost entirely via HTCondor

- Primary HTC facilities used by experimenters
  - **Fermigrid** [~20k cores], used by 30+ Fermilab experiments
  - **CMS Tier-1** [~20k cores], used by CMS central production and global CMS community
  - **LPC** [~5k cores], used by USCMS community (primarily based at Fermilab)
  - HTC clusters are all running on **x86** architecture hardware

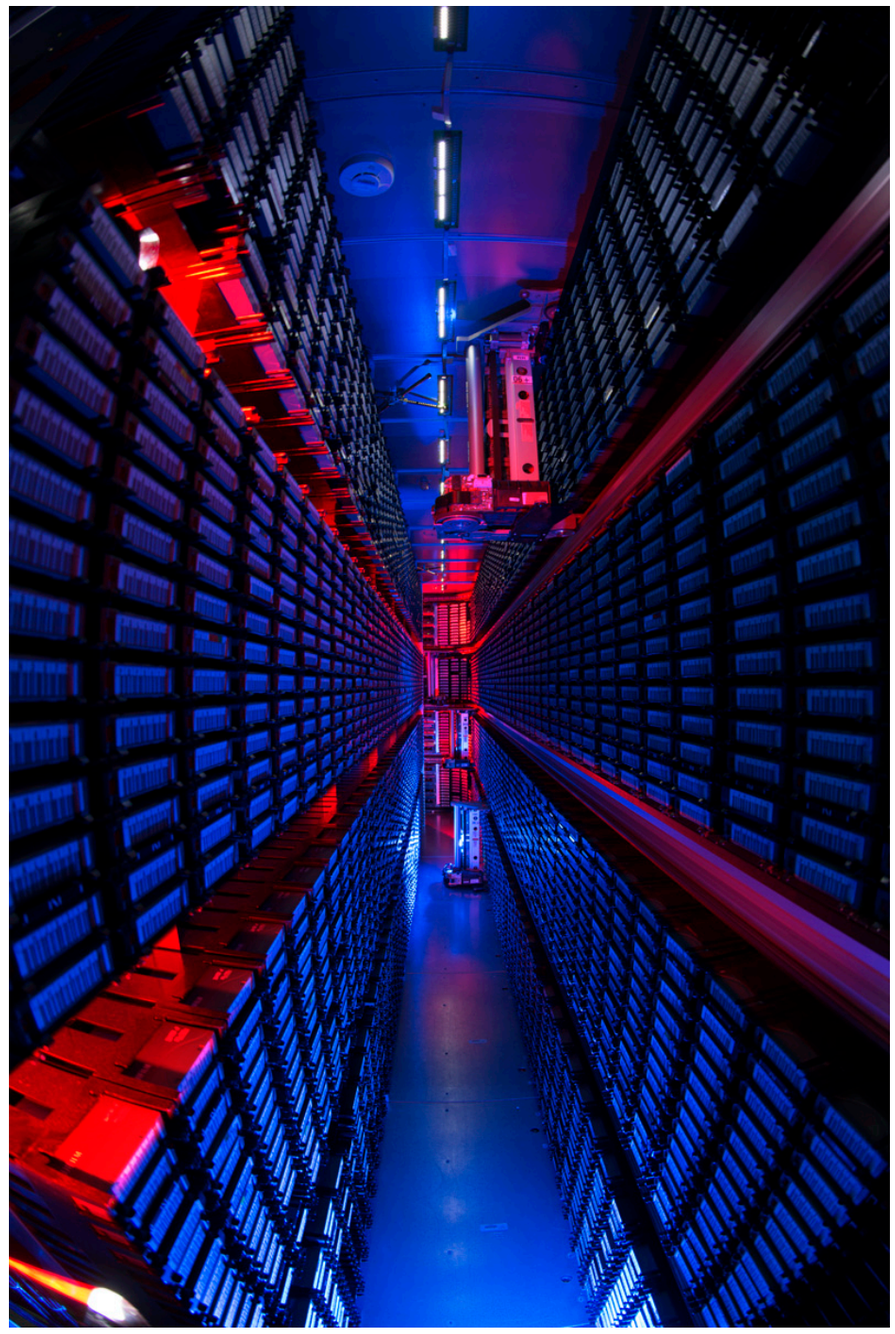
- Lattice QCD and others utilize **High-Performance Computing** (HPC)
  - ~18.5k CPU cores and ~700 GPU cores
  - HPC nodes connected via Infiniband (40Gbps) interconnect

Mass storage

- Primary storage medium is **magnetic tape**
  - Oracle SL8500 robotic libraries (10k slots each)
    - 3x for CMS, 4x for all other experiments
  - ~70 drives (mix of T10KC [5TB], T10KD [8TB], and LTO4 [800GB])
  - ~15k active media cartridges
  - Total of **93.4PB** active tape storage
    - 38.9 PB CMS, 20.5 PB CDF+D0, 33.9 PB all other expts

- **Disk** storage via dCache
  - 3.5PB caching for tape access, 1.4PB persistent space, ~20PB combined use by CMS

- Other disk storage
  - Network attached storage (NAS) on interactive nodes
  - EOS pool on LPC cluster (~5PB)



Outside of Fermilab

- FNAL GPGrid and CMS Tier1 are part of the wider **Open Science Grid** (OSG) computational fabric
  - Fermilab experiments can use **opportunistic resources** that are part of the OSG
  - Conversely, Fermilab resources, when otherwise idle, can be used by external opportunistic users from the OSG

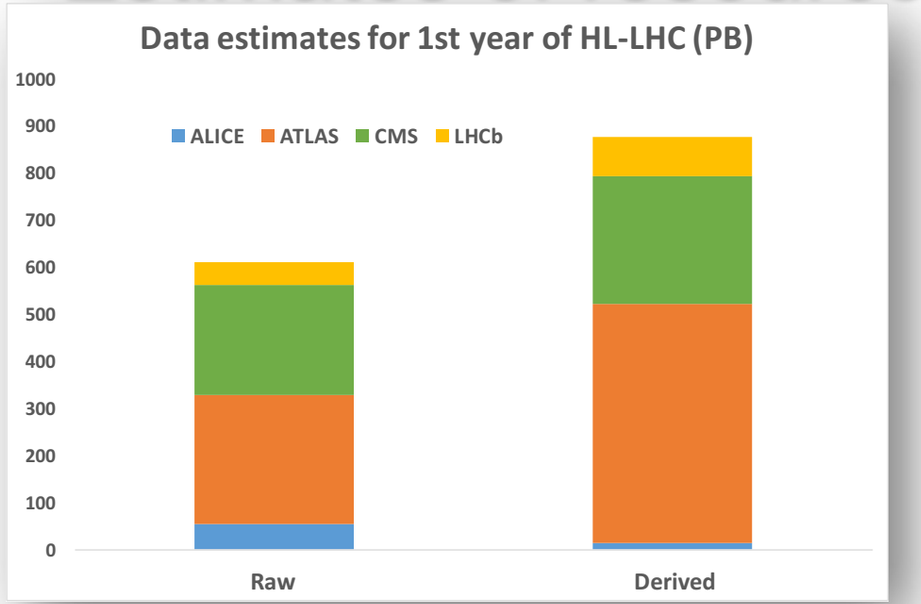
- Allocation-based **HPC** (supercomputers)
  - Some at National Labs, some (NSF-funded) at university centers
  - A number accessible via OSG

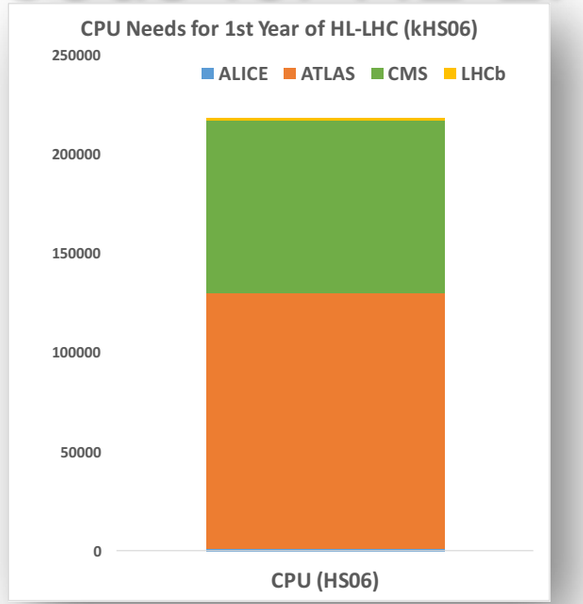
- **Commercial Clouds**
  - e.g., Amazon AWS, Google, Microsoft Azure
  - CMS and Nova have both performed large-scale production exercises on cloud resources

- In the near future: **HEPCloud**
  - Single infrastructure at Fermilab to allow access to all of the above resource types

One example of a coming challenge

Estimates of resource needs for HL-LHC





Ian Bird
WLCG Meeting 2016

Data:
- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:
- x60 from 2016

Technology at ~20%/year will bring x6-10 in 10-11 years

- Simple model based on today's computing models, but with expected HL-LHC operating parameters (pile-up, trigger rates, etc.)
- At least x10 above what is realistic to expect from technology with reasonably constant cost

Changing landscape

- HEP has enjoyed a decade of **computing resource homogeneity**
  - Intel/AMD x86(-64) architecture
  - Dennard scaling reliable for most of this period
- Data access follows **sequential paradigm**
  - Largely unchanged since late 20th century
- Resource heterogeneity is ~~coming~~ here
  - GPUs/vector processors increasingly prevalent
- Newer analysis techniques (e.g. deep learning) incredibly inefficient with sequential data access
- Shifting national cyberinfrastructure priorities
  - “Leadership class” **supercomputers** dwarf dedicated HEP computing resources

Some things to consider*

- Scaling laws seem to end across the board
  - CPU feature scaling has slowed considerably (now at 10 nm)
  - Hard drive areal density improvements have slowed
  - **Competition diminishes** across all sectors of hardware manufacturing

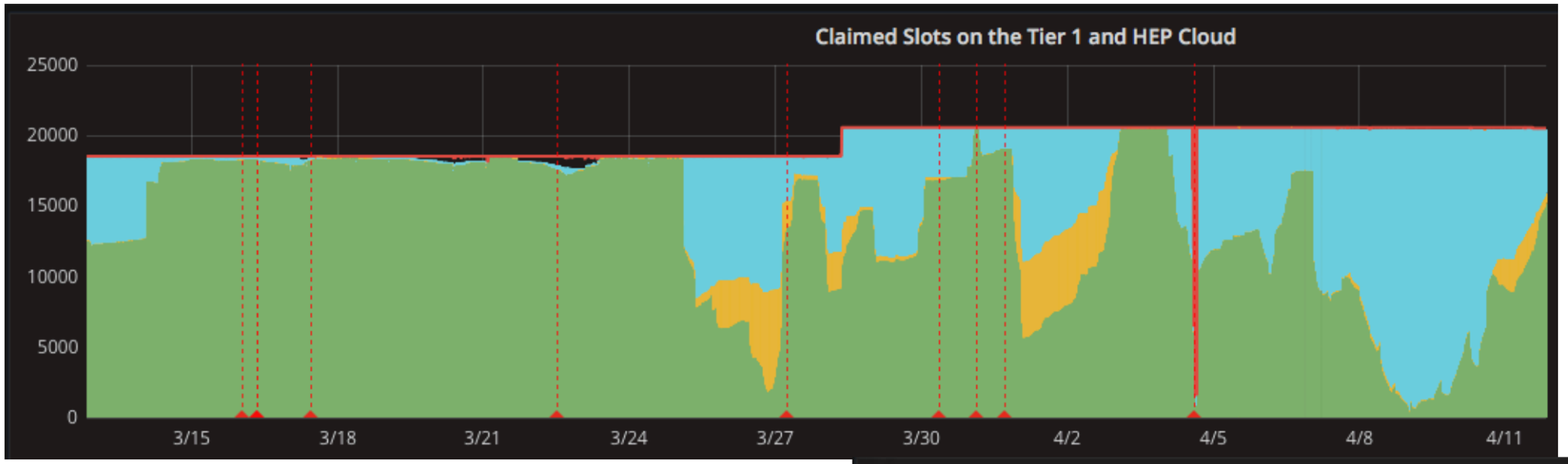
- End of “one size fits all” computing facilities?
  - Consider things such as specialized data reduction facilities
  - Do data need to always be co-located with CPU?
  - Can we optimize (a subset) of facilities for new analysis techniques?

- We need to better leverage available (external) resources
  - HEP is now one of the smaller “big data” uses in the world
  - Keep an eye on industry trends and also understand where using commercial resources makes sense
  - HEPCloud is a big step in this direction

** Views expressed here are my own*

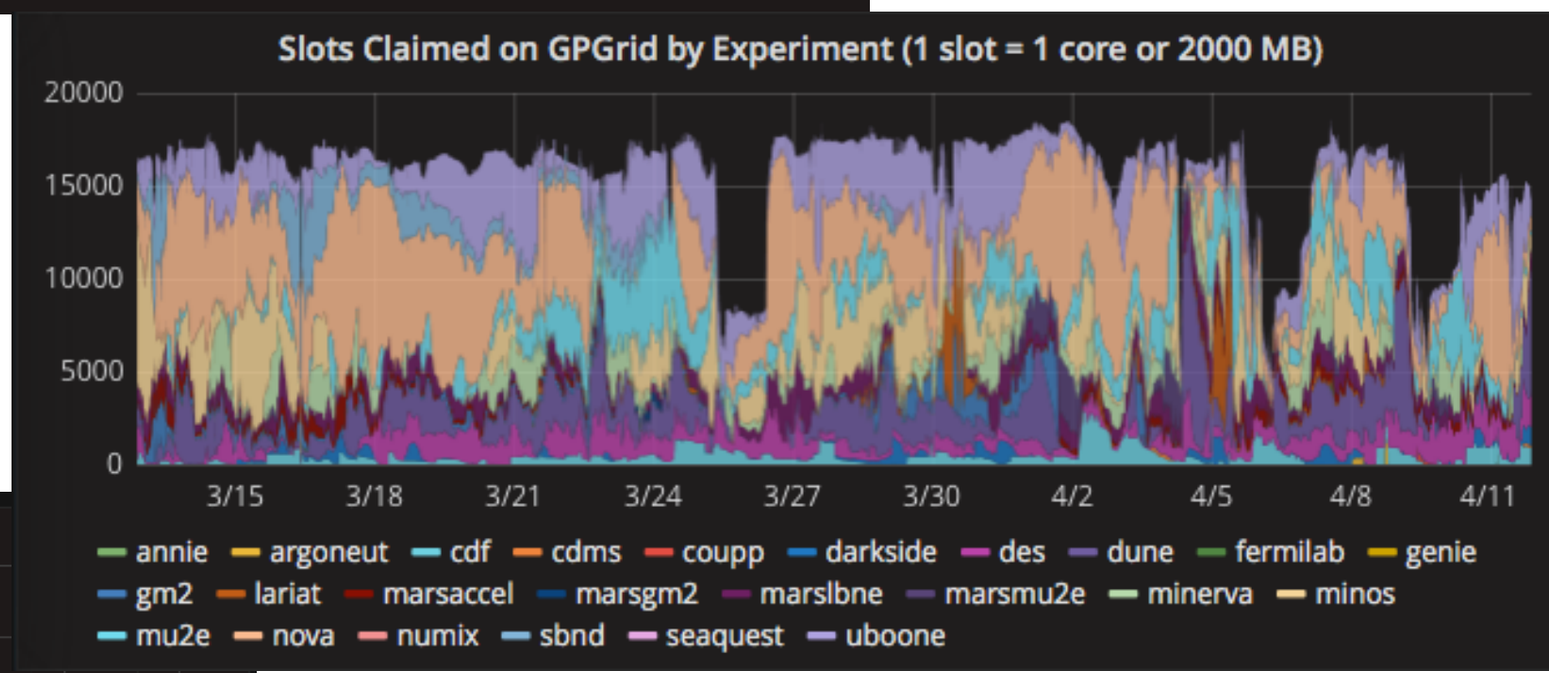
Backup

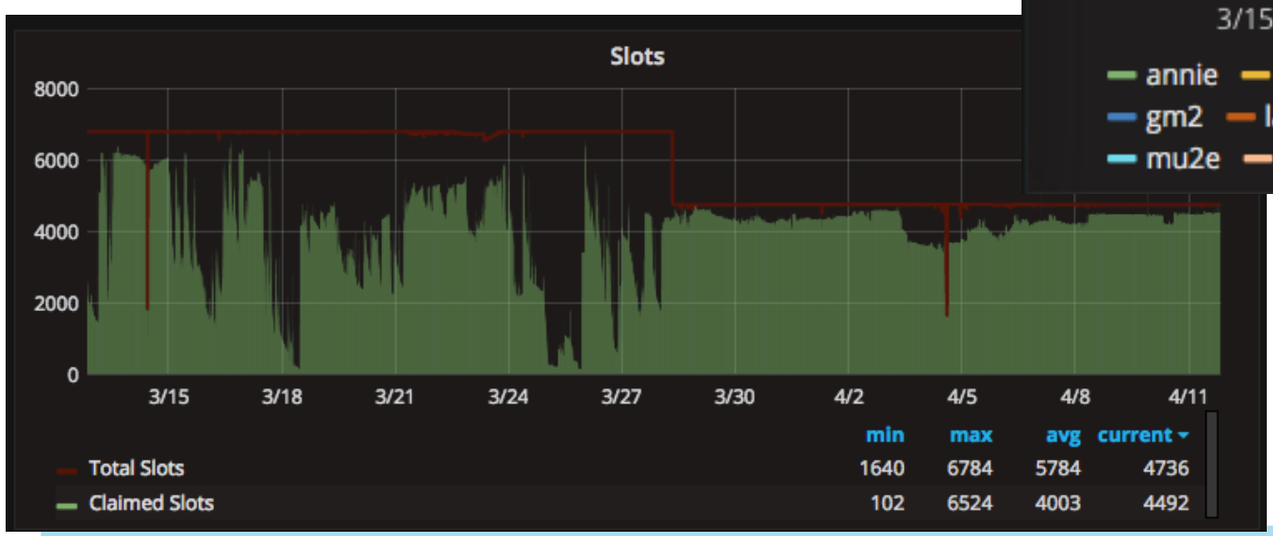
CPU: Usage (30 days)



CMS Tier1

GPGrid

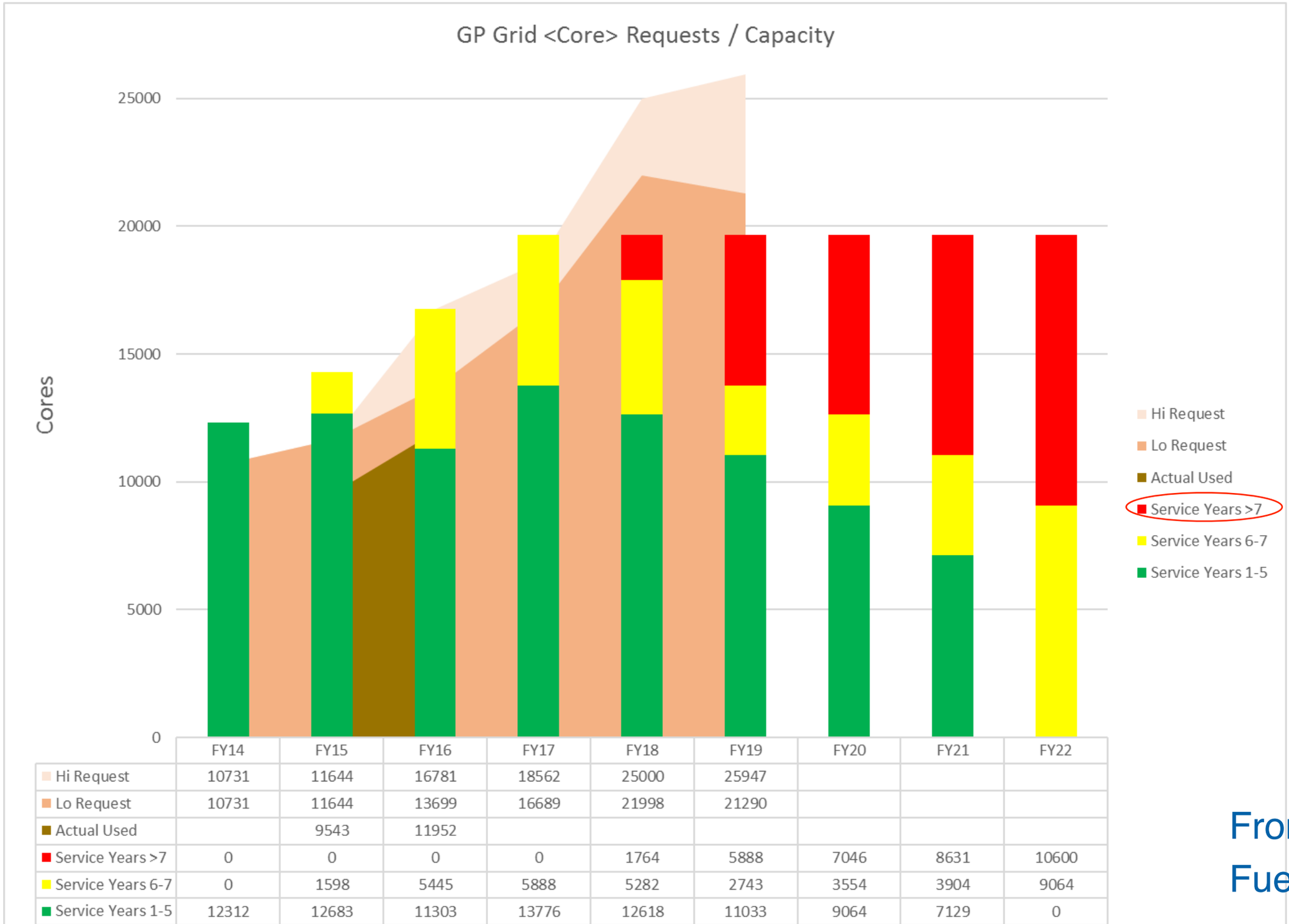




LPC

Fermilab

Facility CPU ages

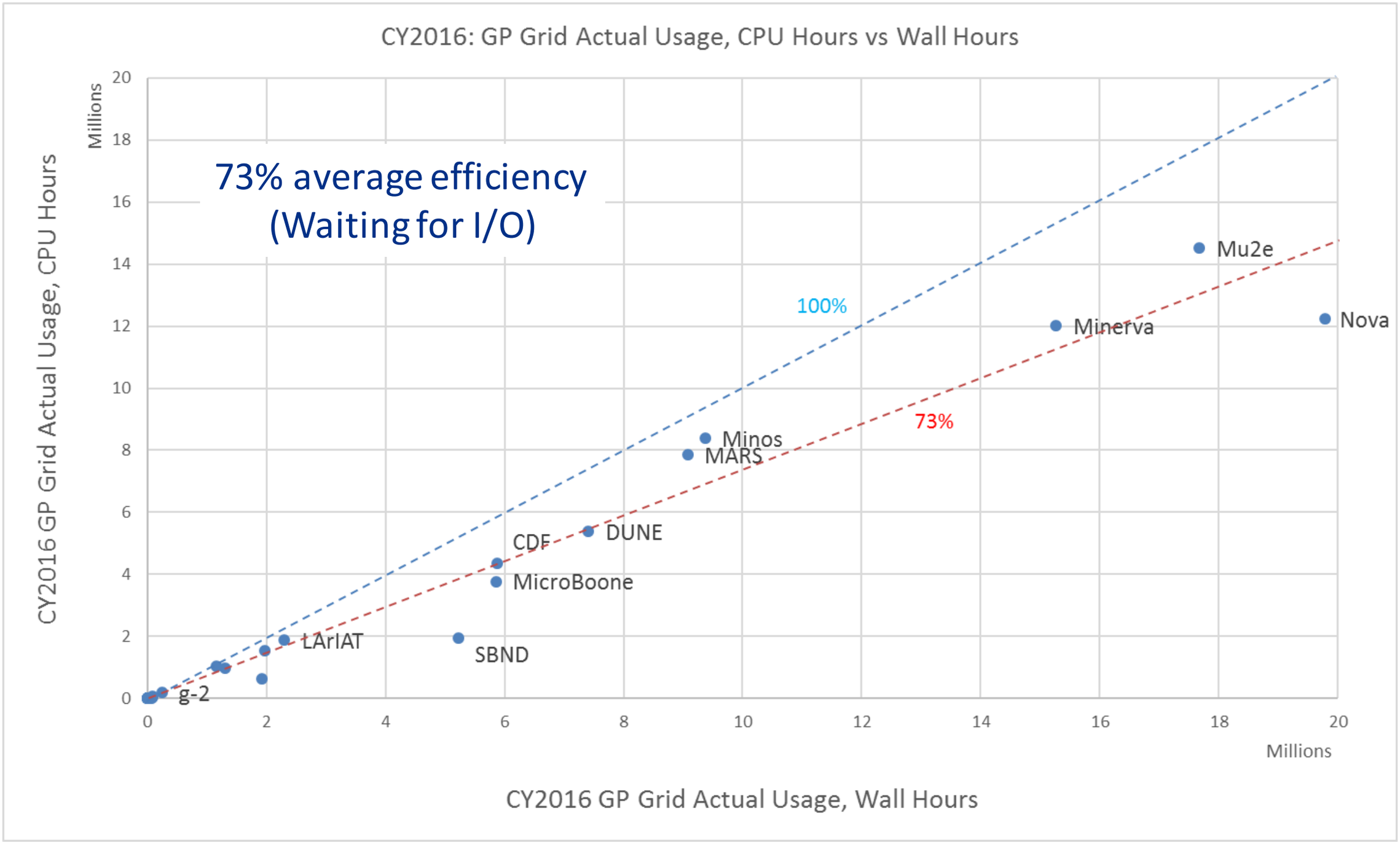


	FY14	FY15	FY16	FY17	FY18	FY19	FY20	FY21	FY22
Hi Request	10731	11644	16781	18562	25000	25947			
Lo Request	10731	11644	13699	16689	21998	21290			
Actual Used		9543	11952						
Service Years >7	0	0	0	0	1764	5888	7046	8631	10600
Service Years 6-7	0	1598	5445	5888	5282	2743	3554	3904	9064
Service Years 1-5	12312	12683	11303	13776	12618	11033	9064	7129	0

From 2017 SCPMT Fuess

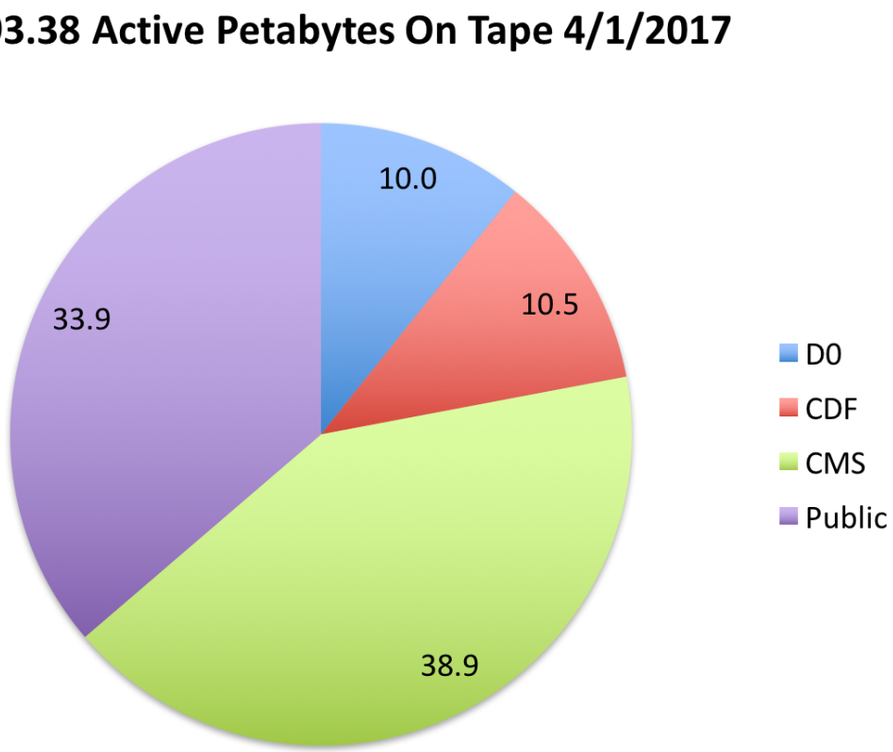
Fermilab

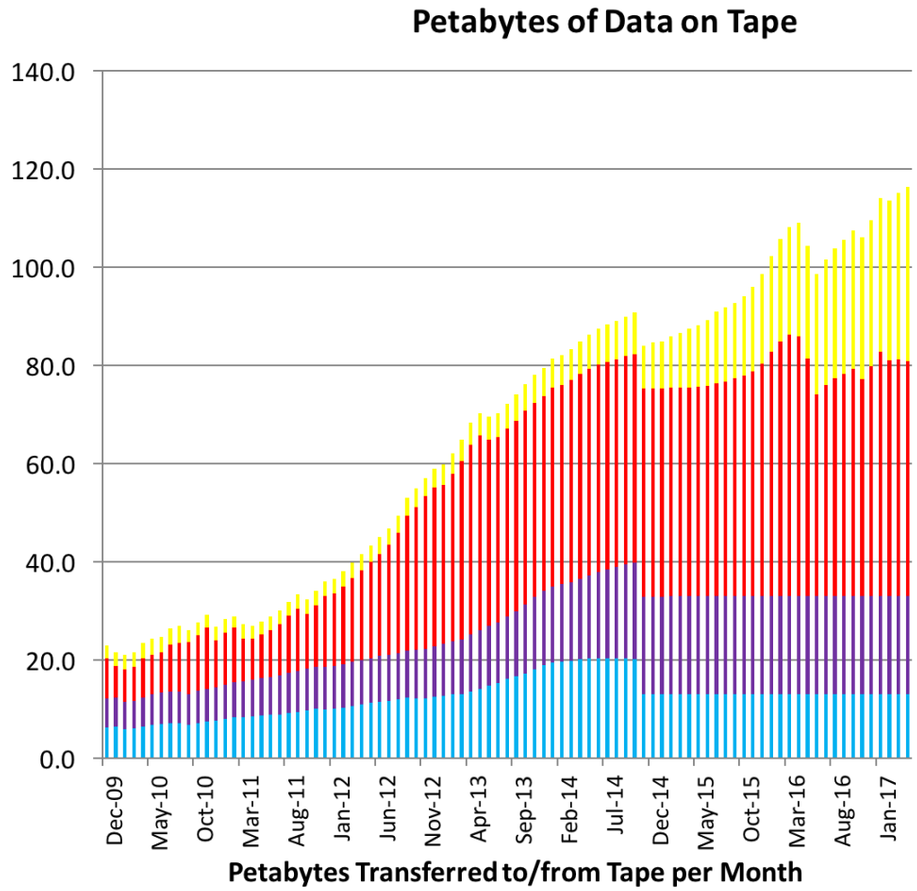
How efficiently are we using the CPU?

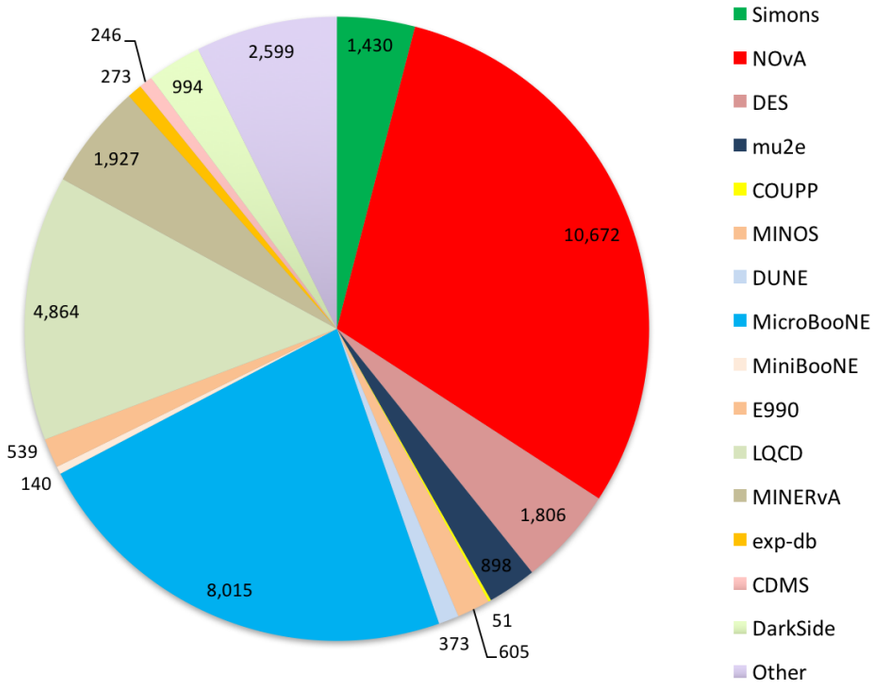


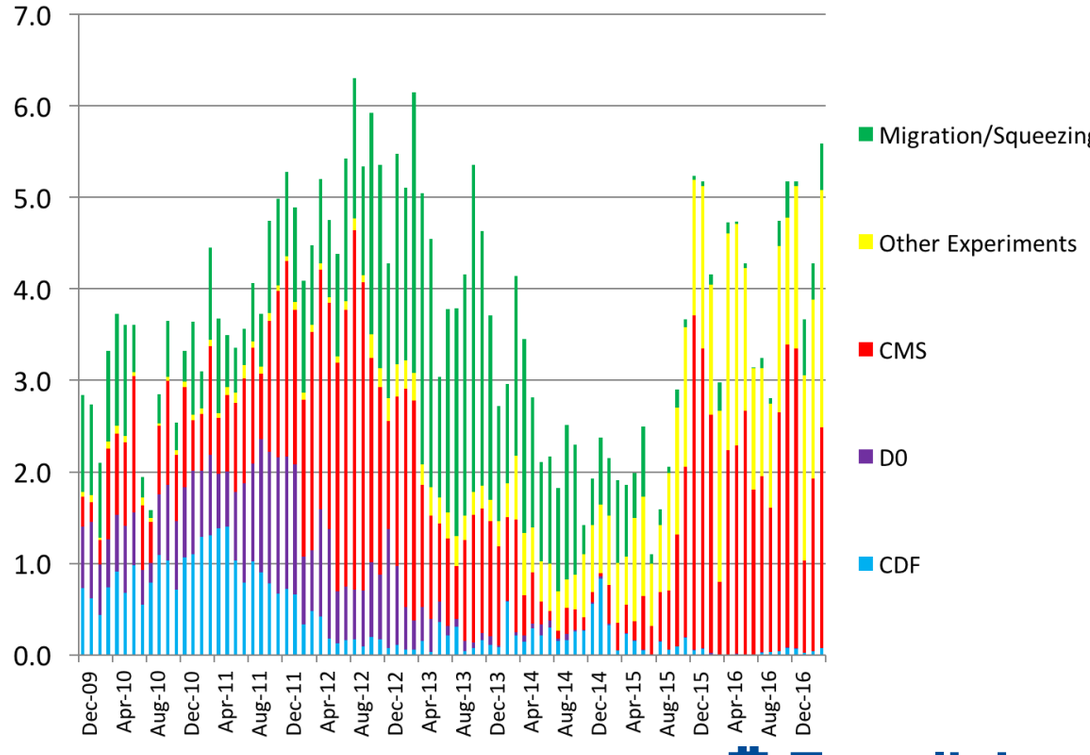
Fermilab

Storage trends









Fermilab