

# HEPiX Site Report: SLAC Shared Science Data Facility (S3DF)



Yemi Adesanya  
TID Scientific Computing Systems Division

November 5th, 2024

# S3DF is modern infrastructure, services, capabilities

---

- SLAC Shared Science Data Facility (S3DF)
- Modern hardware for massive scale analytics in a modern data center designed ground-up for research computing
- High-speed networking storage solutions for data-intensive pipelines/applications
  - Scale up to Tb/s links with 100s of PBs online
- Support for the entire data pipeline from detector input through to results publication
- Funding plans that incorporate hardware lifecycle
- Raising the bar for baseline SLAC Scientific Computing - everybody benefits from the shared resource model

# S3DF is centrally managed and integrated

---

- TID Scientific Computing Systems division is responsible for delivering S3DF
- Lab indirect funding for both core staff and core infrastructure
- We engage with all SLAC science directorates
- Strategic goal of effectively replacing legacy siloed infrastructure
- Storage in DAQ areas essential for high-throughput pipelines
- Each experiment/project has a defined S3DF “Facility” footprint with the ability to fund and prioritize resources
- Lifecycle plans developed and communicated to PIs and leaders



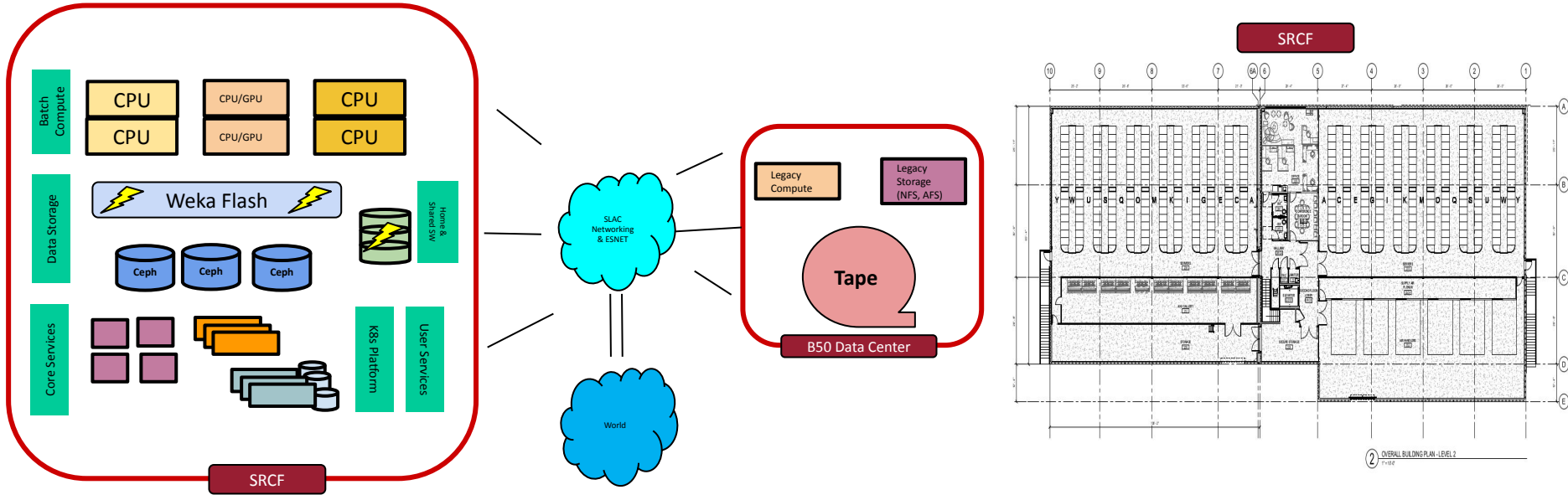
# Stanford Research Computing Facility (SRCF)

---

- A purpose built datacenter that is located on the SLAC campus
- Designed for research computing based on S3DF requirements and capacity projections
- Consisting of two 3MW modules (SRCF-I, SRCF-II) to provide combined 6MW across 300 racks
- SLAC leasing agreement for up to 2.5MW by 2030
- SRCF has a **resilient** but not redundant **power infrastructure**
  - UPS and generator protected, providing significant assurance should there be a regional power outage
  - **Flywheel technology** allows for glitch-free power at all times and for smooth transitions to generator power in case of a power cut.
  - Cooling design is non-traditional and especially energy efficient: ambient air fan systems for 90% of the year (for the hotter days and for equipment needing chilled water, high-efficiency air cooled chillers are available)

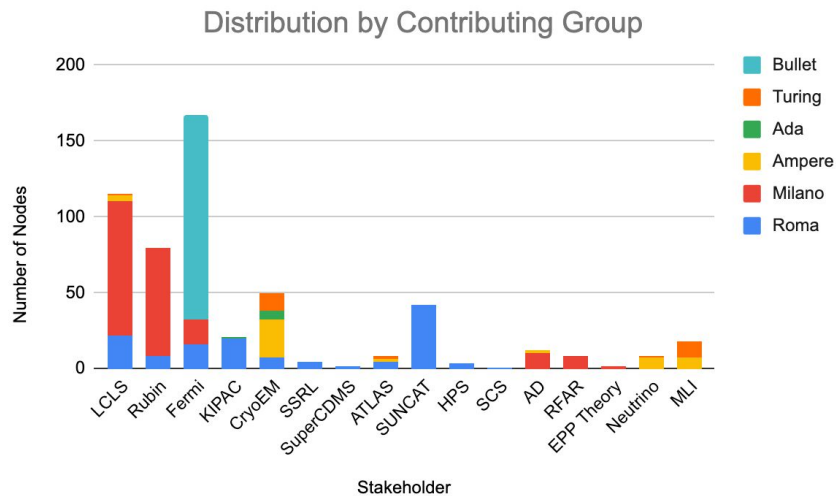
# S3DF Architecture

S3DF - An evergreen, heterogeneous system for data-intensive scientific computing

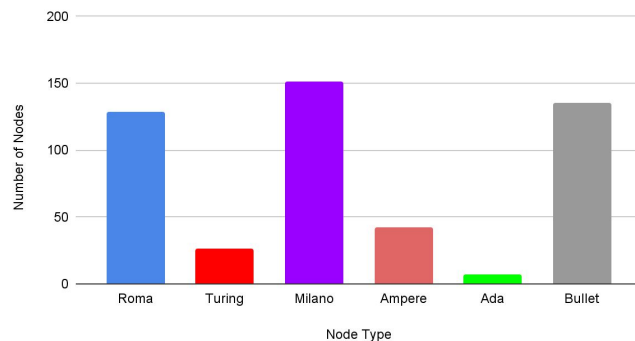


# S3DF Hardware - Compute

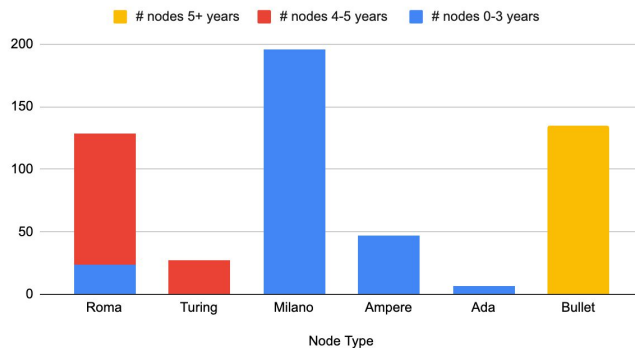
## Compute HW - CPU and GPU-accelerated compute nodes



Number of Nodes vs. Node Type

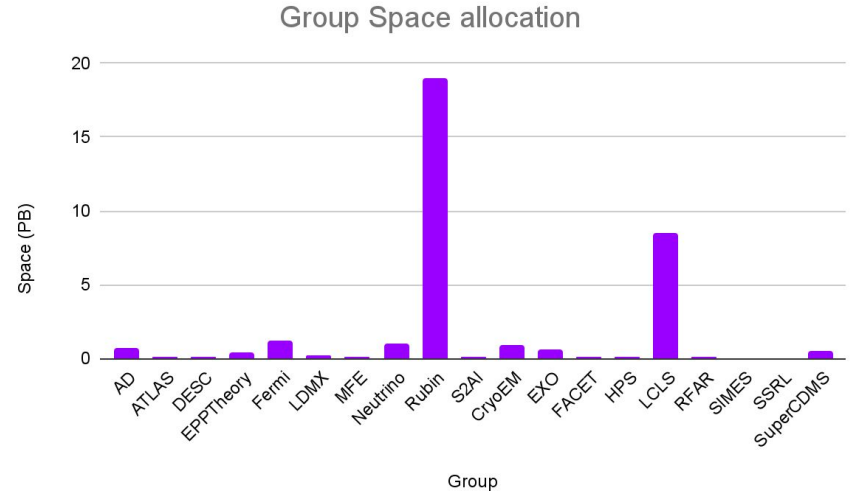
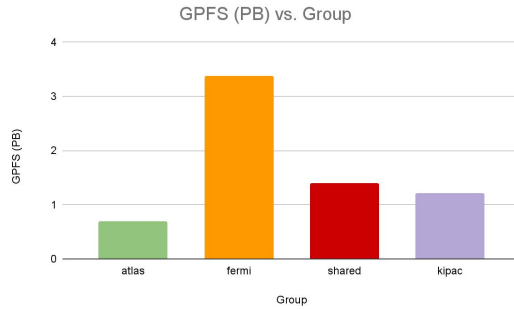
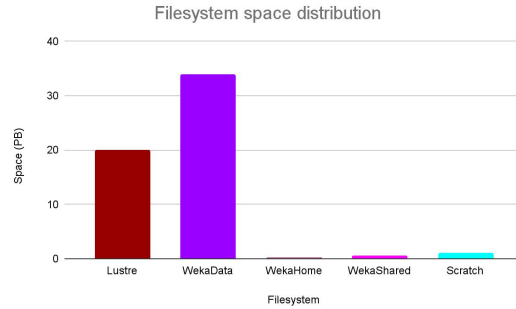


Age distribution of Nodes



# S3DF Hardware - Storage

Heterogeneous storage - Weka/Ceph; Lustre; GPFS; NFS



# WekaFS + Ceph

---

- S3DF online capacity is projected to exceed 250PB by 2026, reaching 750PB by 2031
- Rubin USDF requires S3 object storage in addition to traditional POSIX NFS
- NVMe flash clusters are implemented using WekaFS
  - Provides data protection via distributed erasure coding
  - WekaFS IOPS and capacity can be scaled by expanding clusters while they remain online (an improving feature)
  - Supports data access via native POSIX, NFS, and S3 interfaces
- Cost-effective capacity storage via Ceph NL SAS spinning disk
  - 4RU JBOD building blocks with 80x16TB
- WekaFS provides automatic storage tiering to external S3 stores
- Experiments can choose/tune the NVMe-to-spinning disk ratio for their filesystem



# Spectra Logic Tape Library

---

- TFinity tape library entered production in May 2022
- 20,000 tape slots provide over 220PB of capacity, up to 2500PB in 2032
- 12TB LTO-8 media now
- Projects/experiments will pay for their media
- IBM HPSS is our primary archive software -- developers plan to provide S3 support



# S3DF Filesystem Tree

---

## *Weka HOME cluster (100% flash, automatic backups) LAB FUNDED*

- `/sdf/home/<u>/<username>`: Home directories. Disk quotas imposed for all users.
- `/sdf/sw/<package>/<version>`: For general purpose (broad usage) software not installed on each node
- `/sdf/group/<organization>/<groupname>` or `/sdf/group/<groupname>`: For group/project specific software (e.g., lcls/psdm, ad/hla, etc.)

## *Weka SCRATCH cluster LAB FUNDED*

- `/sdf/scratch/<facility>/...`: 3 months retention on a best effort basis (actual retention can be shorter or longer depending on usage)

## *Weka DATA cluster (Flash layer with large scale disk-based object tier, manual backups) DIRECT FUNDED*

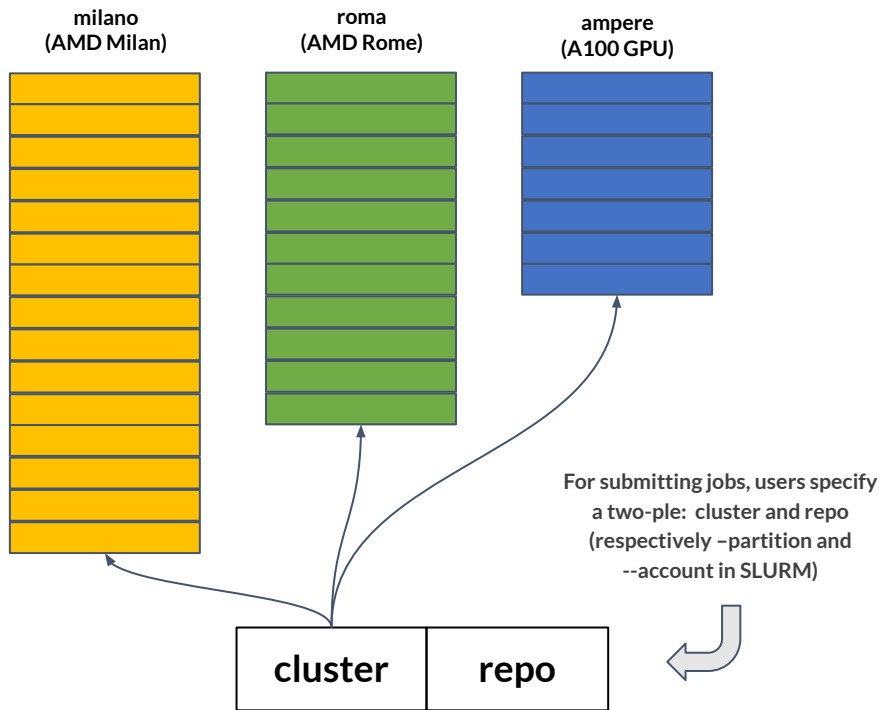
- `/sdf/data/<facility>/...`: For actual science data (not code or docs). Some examples:
  - LCLS experimental: `/sdf/data/lcls/<instrument>/<experiment>`
  - LCLS accelerator: `/sdf/data/lcls/accel/<bld|bsa|ca|...>`
  - FACET experimental: `/sdf/data/facet/<instrument>/<experiment>`
  - FACET accelerator: `/sdf/data/facet/accel/`
  - CryoEM: `/sdf/data/cryoem/<YYYYMM>/<experiment>`

# Kubernetes

---

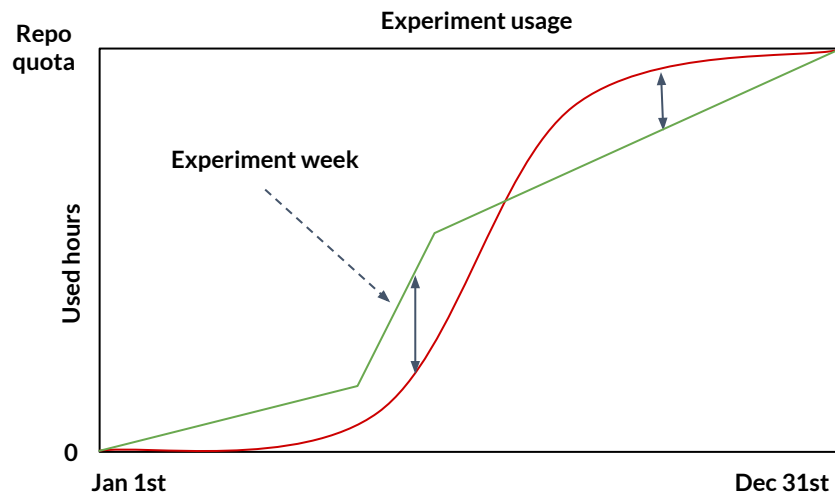
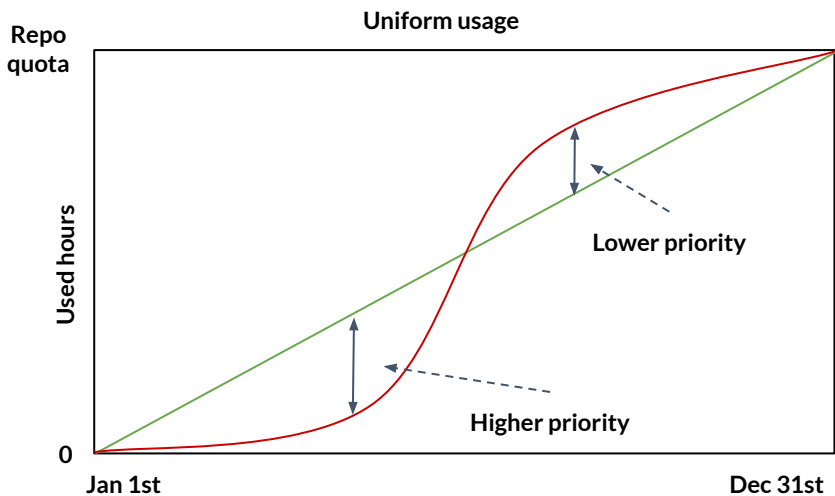
- k8s = modern cloud architecture for container orchestration
- Provide science services with high-availability, flexibility, scalability
- Minimize traditional baremetal hosting
- S3DF k8s worker cluster node = 64 cores, 256GB, 100Gb ethernet
- Dedicated WekaFS storage for k8s
- Rubin Science Platform was designed for cloud and containerized from day 1
- We must help developers migrate science workflows from baremetal and VMs
  - Recent onsite Google k8s training was well received
- Partner with IT teams to address any networking and Cyber challenges

# Slurm Compute Clusters & Repos



- All nodes within a **cluster** have the same hardware specs **and** same access to the storage
- A **facility** is a program/project which can buy resources (eg LCLS, Rubin, SUNCAT, SuperCDMS, etc)
- A **repo** is a set of resources associated with a group of people (eg an LCLS or a cryo-EM experiment)
- Not all repos will have access to all clusters

# Slurm Fairshare for Prioritization



Projected CPU hour consumption

Actual CPU hour consumption

# S3DF Software & Services

## Platform & Services



OnDemand provides an integrated, single access point for all of your HPC resources.

### Message of the Day

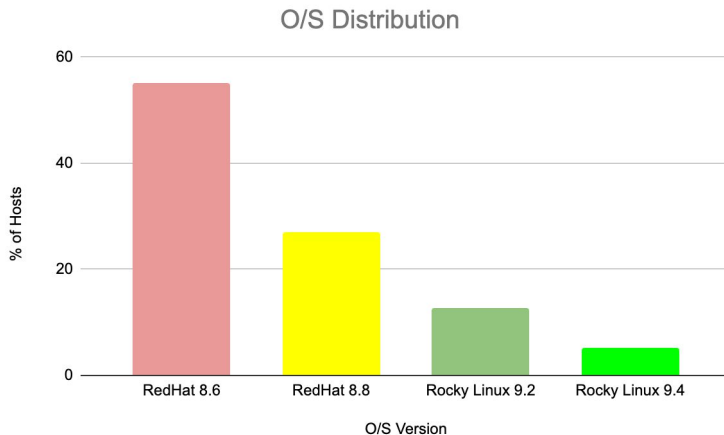
=====

This is a Federal computer system and is the property of the United States Government. It is for authorized use only. Users (authorized or unauthorized) have no explicit or implicit expectation of privacy.

By using this system you expressly consent to the terms and conditions in <https://www.slac.stanford.edu/comp/policy/use.html>

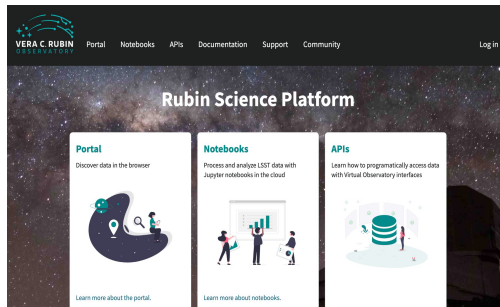
SDF Documentation: <https://s3df.slac.stanford.edu>  
 SDF Slack Channel: [https://slac.slack.com/app\\_redirect?channel=comp-sdf](https://slac.slack.com/app_redirect?channel=comp-sdf)

=====



### LCLS Controls & Data Systems - Web tools & Documentation catalog

| Applications for User Experiments   | Experiment/Match Management  | Computing System  | Engineering Databases  |
|---|--|---|--|
| <ul style="list-style-type: none"> <li>eLog (aka Data Manager)</li> <li>UED standalone eLog</li> <li>Analysis docs - LC131 (gsasl)</li> <li>Analysis docs - LC132 (gsasl)</li> <li>PCDS docs</li> <li>JupyterHub</li> </ul> | <ul style="list-style-type: none"> <li>Instrument operators</li> <li>Group management logs</li> <li>Questionnaire (Run 23)</li> <li>Questionnaire (Run 22)</li> <li>Questionnaire (Run 21)</li> <li>Questionnaire (Run 20)</li> <li>Questionnaire (Run 19)</li> <li>Questionnaire (Run 18)</li> <li>Questionnaire (Run 17)</li> <li>Questionnaire (Run 16)</li> <li>UED Questionnaire (Run 4)</li> </ul> | <ul style="list-style-type: none"> <li>News</li> <li>Current Outages</li> <li>Speed test</li> <li>Grafana Monitoring</li> <li>Grafana Monitoring (Dev)</li> <li>M/Monit Monitoring</li> <li>OpenSearch</li> <li>Controls Grafana</li> </ul> | <ul style="list-style-type: none"> <li>NucCAPTAR (Cable management)</li> <li>Machine Configuration Database</li> <li>EPICS Archive Viewer</li> </ul> |
| Development Support Tools & Documentation   |  |   |  |
| <ul style="list-style-type: none"> <li>Jenkins</li> </ul>   |  |   |  |



nomachine



https

# S3DF System Usage

## User Portal

- CoAct Portal allows users to see their usage & allows PIs to control allocation of resources
- Contributing members (“Facility”) can have multiple user groups using the resources they are entitled to - controlled as “Repos” with specific allocations
- Allows for “pre-emptible” jobs to enable effective utilization of the system

The screenshot displays the CoAct User Portal interface. At the top, there is a navigation bar with 'Coact' and tabs for 'Facilities', 'Repos', and 'Requests'. Below this, a list of facilities is shown, including LCLS, CryoEM, SUNCAT, Ruben, Neutrino, Fermi, MLI, SuperCDMS, FACET, EPPTTheory, DESC, KIPAC, SSR, RFAR, LDMX, HPS, AD, and SIMES. A secondary row lists EXO, ATLAS, CDS, SRS, FADERS, TOPAS, RP, Projects, and SCS. The main content area is divided into several sections:

- Details:** Shows 'Name: LCLS' and a 'Register new users' button.
- Czars:** Lists users: mshankar, wilko, thorsten, and snelson, with an 'Add/Remove Czars' button.
- Service Accounts:** Shows 'User Group: lcls-pcdsdata' and 'lcls-pcdsmgr'.
- Compute +:** A table showing resource usage for various clusters.
- Storage +:** A table showing storage class usage.

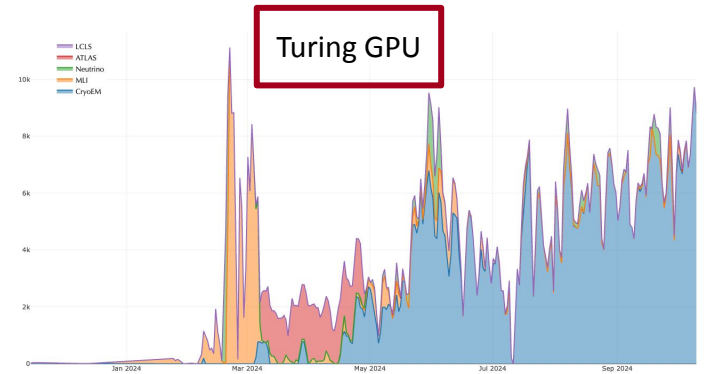
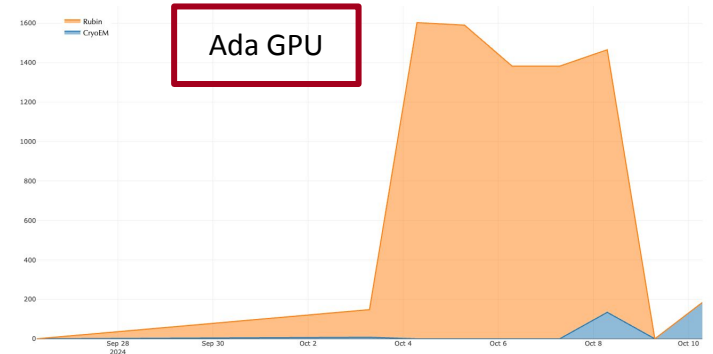
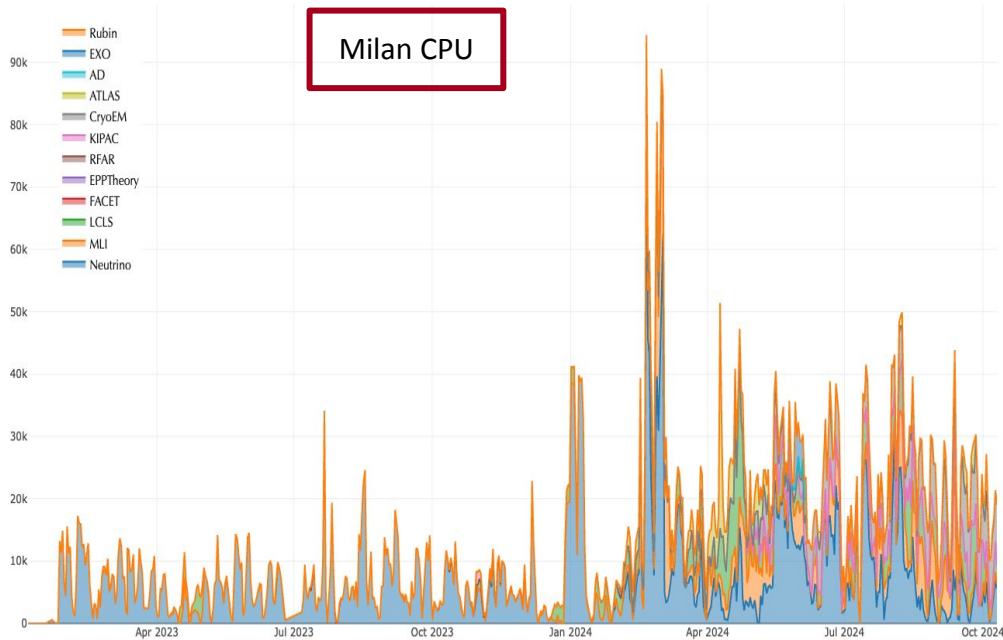
| Cluster                | Acquired nodes | Total allocated (%) | Past hour | % Used Past day | Past week |
|------------------------|----------------|---------------------|-----------|-----------------|-----------|
| <a href="#">ada</a>    | 0              | 3420                | 0.00%     | 0.00%           | 0.00%     |
| <a href="#">ampere</a> | 4              | 21510               | 25.20%    | 27.38%          | 26.51%    |
| <a href="#">milano</a> | 88             | 22250               | 3.84%     | 35.48%          | 10.85%    |
| <a href="#">roma</a>   | 22             | 21260               | 0.00%     | 0.82%           | 0.12%     |
| <a href="#">turing</a> | 1              | 15785               | 0.00%     | 0.00%           | 1.02%     |

| Storage Class              | Purpose                 | Acquired | In TB Allocated | Used |
|----------------------------|-------------------------|----------|-----------------|------|
| <a href="#">sdfdata</a>    | <a href="#">data</a>    | 9070.00  | 0.00            | 0.00 |
| <a href="#">sdfhome</a>    | <a href="#">group</a>   | 20.00    | 0.00            | 0.00 |
| <a href="#">sdfscratch</a> | <a href="#">scratch</a> | 150.00   | 0.00            | 0.00 |

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University U.S. DEPARTMENT OF ENERGY

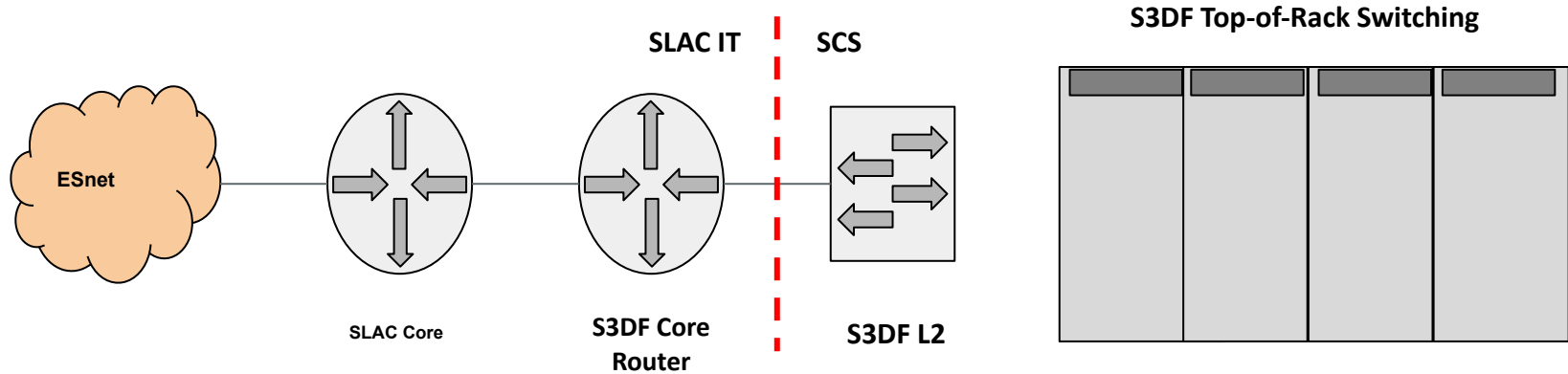
# S3DF System Usage





# S3DF System

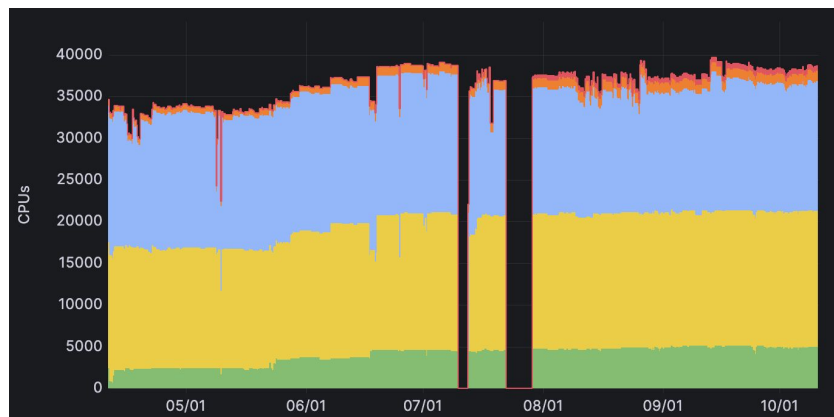
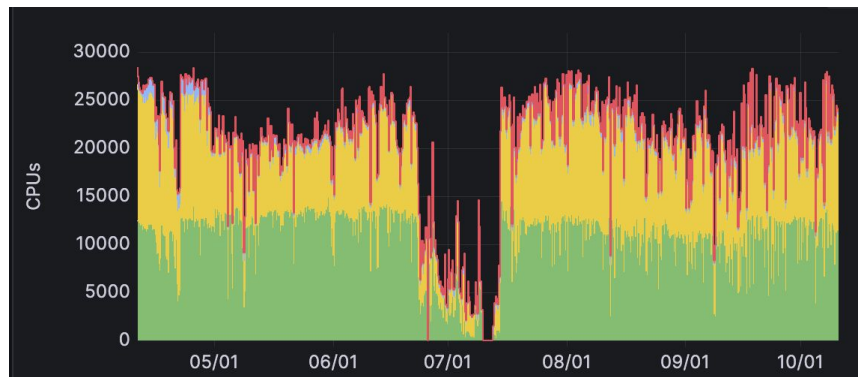
Network Infrastructure - split role with SLAC IT Networking Group



# S3DF Metrics

## How is the system performing?

- Over the last 8 months as we have moved into SRCF-2 system reliability has gradually improved
- Main issues stem from
  - Network reliability
  - Storage issues (also tied to Network)
  - Power issues (SLAC-wide)
- Power issues largely restricted to B50 hardware
  - SRCF has ridden through power issues
- We monitor all aspects of the system
  - Node health
  - Batch system usage
  - Services



# S3DF Roadmap

---

## Priorities for the next Year

- Prepare for CY2025 hardware purchases
  - AMD EPYC Milan is now 4 years old  
Start evaluating new CPU,GPU and storage
  - Continue build out of datacenter racks, PDUs and networking
- Network upgrades for improved reliability and performance
  - Deploy Arista Trident (leaf) and Jericho (spine) switches
  - Eliminate bottlenecks and inconsistent throughput
- Monitoring and Alerting
  - Refine our monitoring and logging tool stack: Loki, Grafana, Telegraf....
- Security
  - Partner with SLAC Cyber Team and science leadership on risk analysis
  - Key areas: Intrusion detection, Multi Factor Authentication
- Communication
  - Responsive incident management
  - Careful consideration of impactful system changes
  - Continual engagement with computing czars and users

# We Are Hiring!

---

<https://careers.slac.stanford.edu/>

- Linux Storage Engineer #6109
- Linux Networking and Security Specialist #6094
- Experimental Support Associate #5976
- Linux System Administrator #5894
- Scientific Database Administrator



SLAC

OUR SHARED COMMITMENT TO A

## RESPECTFUL WORKPLACE

Our values inspire us to the highest standards of conduct and foster the respectful workplace that is essential to everything we do.

Excellence  
Integrity  
Collaboration  
Creativity  
Respect

The advertisement features a man and a woman wearing headsets, working at computers. The man is on the left, looking at a monitor displaying code. The woman is on the right, looking at a laptop. A large, stylized graphic of a particle detector or accelerator structure is overlaid on the right side of the image. The SLAC logo is in the top right corner.