subMIT Overview

Josh Bendavid

Basic Computing Services (subMIT) Review

June 21, 2024

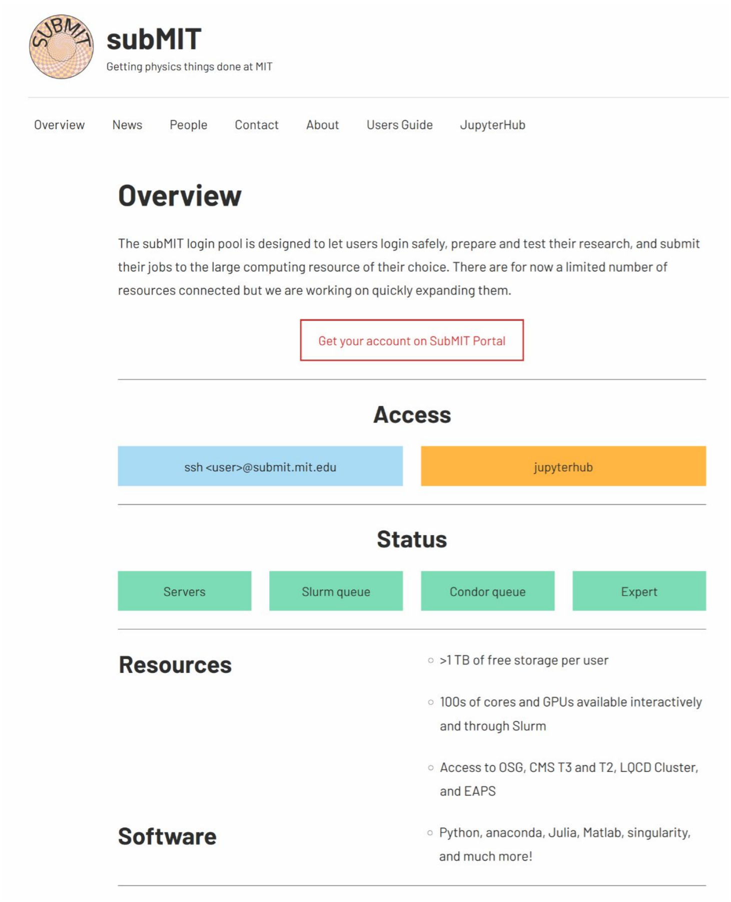
Introduction

- subMIT system provides an interactive login pool + scale-out to batch resources
  - Home directories
  - Convenient software environment (Alma Linux 9 native, docker/singularity images, conda)
  - SSH or Jupyterhub access
  - Local batch system with O(1000) cores, >50 GPU’s
  - Additional storage for software installation/development, large datasets
  - Convenient access to larger external resources (OSG, CMS Tier-2 and Tier-3, LQCD Cluster, EAPS)
- User support is a key feature of the system
  - Beyond basic troubleshooting
    - Help users make optimal use of the available resources
    - Expert advice on designing/improving workflows
    - Customize and evolve system configuration to accommodate user needs as appropriate

Introduction

- Storage and networking
  - 500TB of spinning disks
  - Local storage (1TB/user), 10’s of TB for larger group datasets
  - 40TB of ultra-fast NVME storage with room for future expansion
  - Fast networking: 100 Gbps ethernet
    - RoCE (RDMA over Converged Ethernet) has been partially tested/commissioned, should be possible for MPI applications
- System is located in B24 basement, with 100gbps uplink
- Additional resources recently integrated
  - More disk storage (100TB contributed from ABRACADABRA)
  - Integration of existing computing resources from research groups
  - Purchase of several large core count/high memory machines by research groups for additional computing resources and to support specialized workflows and/or R&D where large single node scaling is useful
    - Current “high density” template, Dual AMD EPYC 192 core/384 thread with 1.5TB of memory

Introduction: subMIT Website



subMIT

Getting physics things done at MIT

Overview

The subMIT login pool is designed to let users login safely, prepare and test their research, and submit their jobs to the large computing resource of their choice. There are for now a limited number of resources connected but we are working on quickly expanding them.

Access

Status

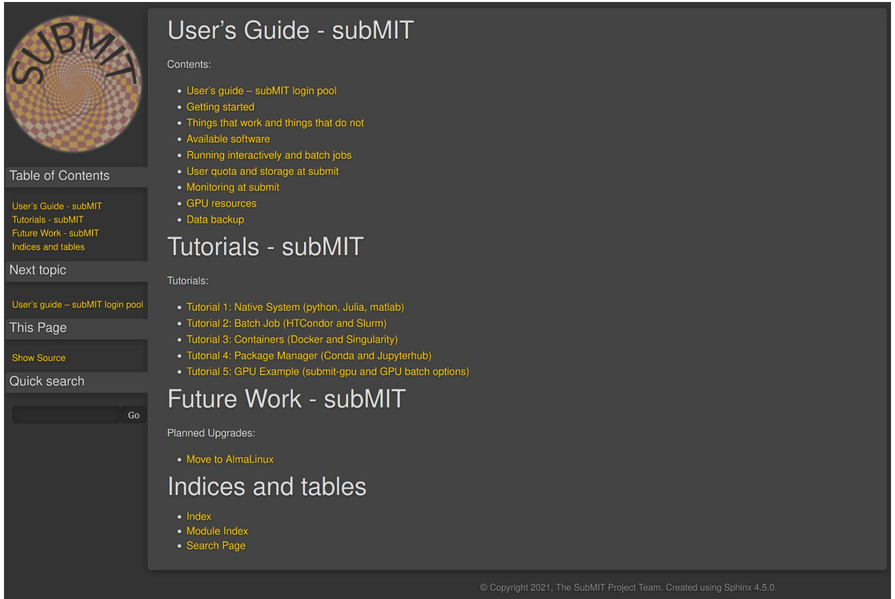
Resources

- >1 TB of free storage per user
- 100s of cores and GPUs available interactively and through Slurm
- Access to OSG, CMS T3 and T2, LQCD Cluster, and EAPS

Software

- Python, anaconda, Julia, Matlab, singularity, and much more!

- Website (with User's Guide/Instructions): https://submit.mit.edu/
  - Overview and general information
  - Direct JupyterHub access
  - User's Guide: https://submit.mit.edu/submit-users-guide/



User's Guide - subMIT

Contents:

- User's guide – subMIT login pool
- Getting started
- Things that work and things that do not
- Available software
- Running interactively and batch jobs
- User quota and storage at submit
- Monitoring at submit
- GPU resources
- Data backup

Tutorials - subMIT

Tutorials:

- Tutorial 1: Native System (python, Julia, matlab)
- Tutorial 2: Batch Job (HTCondor and Slurm)
- Tutorial 3: Containers (Docker and Singularity)
- Tutorial 4: Package Manager (Conda and Jupyterhub)
- Tutorial 5: GPU Example (submit-gpu and GPU batch options)

Future Work - subMIT

Planned Upgrades:

- Move to AlmaLinux

Indices and tables

- Index
- Module Index
- Search Page

© Copyright 2021, The SubMIT Project Team. Created using Sphinx 4.5.0.

Introduction: Project Organization

- Formally the project is organized as ***Basic Computing Services*** in the Physics Department
  - **Project Team:** Implementation/Operations/Maintenance of the system
  - **Users Group:** Contact point between the user community and the project team, forum for user feedback, requests, information flow to and from users
  - **Steering Committee:** Faculty oversight, funding, etc
  - See https://submit.mit.edu/?page_id=6

Users Group In Practice

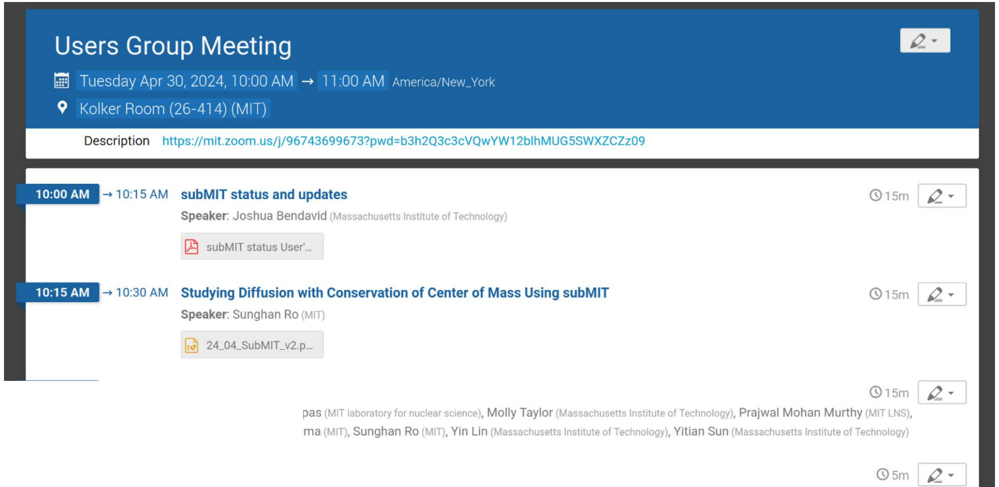
- Regular meetings (every few months)
  - Advertised and open to the broader community
  - Topical presentations from project team, Users Group representatives, or other users or community members
  - Forum for feedback and information flow between the user community and the project team
  - Regular timeslot: Tuesday 10:00-11:00 EST
  - Last meeting agenda: https://indico.mit.edu/event/1050/
- Users Group representatives
  - Identified representatives from research groups across the department
  - Attend the monthly meetings
  - Provide feedback from your groups/community
  - Distribute information/news from the project team

Users Group Representatives

- Users group has been formed (JB as coordinator)
- Current Users Group representative (associated faculty/group)
  - Yin Lin (Phiala Shanahan)
  - Siddharth Mishra-Sharma (Jesse Thaler)
  - Prajwal Mohan Murthy (Bob Redwine)
  - Kaliroë Pappas (LNS Neutrino/Dark Matter)
  - Amer Al-Hiyasat(Julien Tailleur)
  - Yitian Sun (Tracy Slatyer)
  - Molly Taylor (LNS Heavy Ion Group)

Users Group Meetings

- E.g. presentation from April Users Group meeting on Visualization of Geant Simulations on subMIT
- https://indico.mit.edu/event/1050/



Users Group Meeting

Tuesday Apr 30, 2024, 10:00 AM → 11:00 AM America/New_York

Kolker Room (26-414) (MIT)

Description https://mit.zoom.us/j/96743699673?pwd=b3h2Q3c3cVQwYW12blhMUG5SWXZCZz09

10:00 AM → 10:15 AM **subMIT status and updates** 15m

Speaker: Joshua Bendavid (Massachusetts Institute of Technology)

subMIT status User'...

10:15 AM → 10:30 AM **Studying Diffusion with Conservation of Center of Mass Using subMIT** 15m

Speaker: Sunghan Ro (MIT)

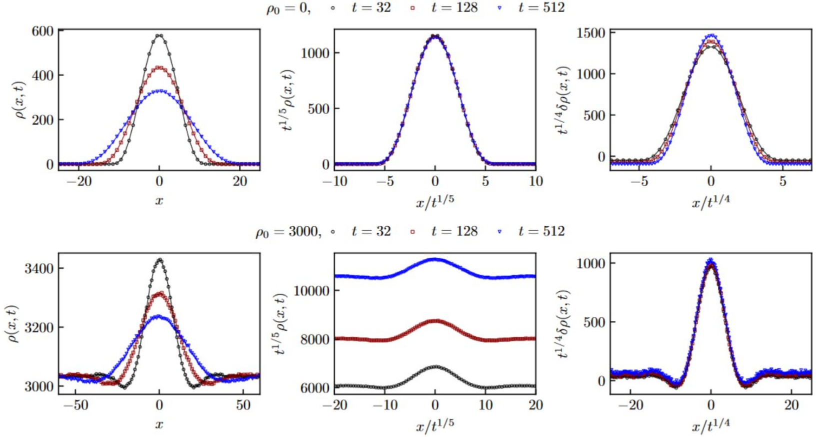
24_04_SubMIT_v2.p...

15m

pas (MIT laboratory for nuclear science), Molly Taylor (Massachusetts Institute of Technology), Prajwal Mohan Murthy (MIT LNS), ma (MIT), Sunghan Ro (MIT), Yin Lin (Massachusetts Institute of Technology), Yitian Sun (Massachusetts Institute of Technology)

5m

Crossover in finite sized systems

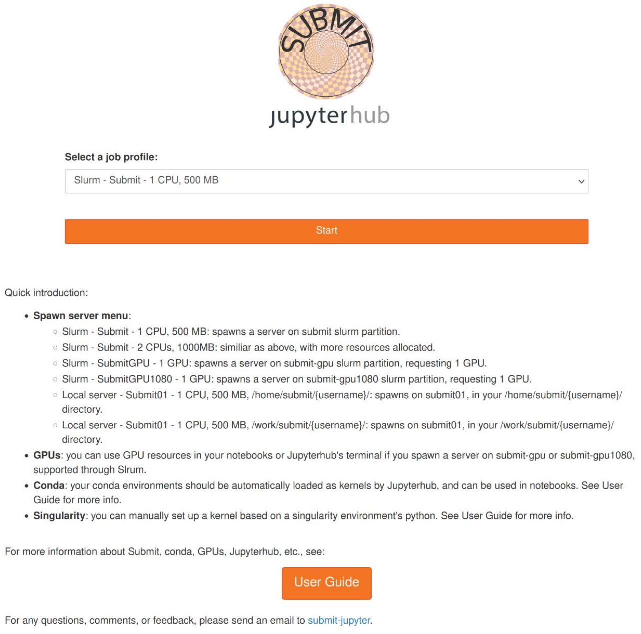


... thanks subMIT!

Storage breakdown

- Several different storage areas are available covering different use cases
  - /home/submit/<username>
    - Home directories (nfs server), redundant disk array with backups
    - 5GB quota
    - Use for software development and (small) critical data
  - /work/submit/<username>
    - Work directory (nfs server), no backups (but redundant disk array)
    - 50GB quota
    - Use for software installation (conda or docker/singularity images)
  - /data/submit/<username>
    - Large distributed disk system, no backups, but redundancy against disk failure (“erasure coding”)
    - 1TB user quota, larger quotas available in dedicated group directories
    - Store large datasets here
  - /scratch/submit/<username>
    - Fast NVMe SSD array
    - Commissioned by several groups for high performance data analysis
  - /cvmfs/
    - Read-only distributed storage for distributing software, singularity images, etc
    - Several CERN-related repositories are available
    - Local repository /cvmfs/cvmfs.cmsaf.mit.edu where additional software or data can be added if needed
- Flexible tiered storage system, can accommodate a wide range of user needs
- Larger datasets encouraged to use shared group space, but quotas can be increased when needed

Interactive Use: Terminal or JupyterHub



Select a job profile:

Slurm - Submit - 1 CPU, 500 MB

Start

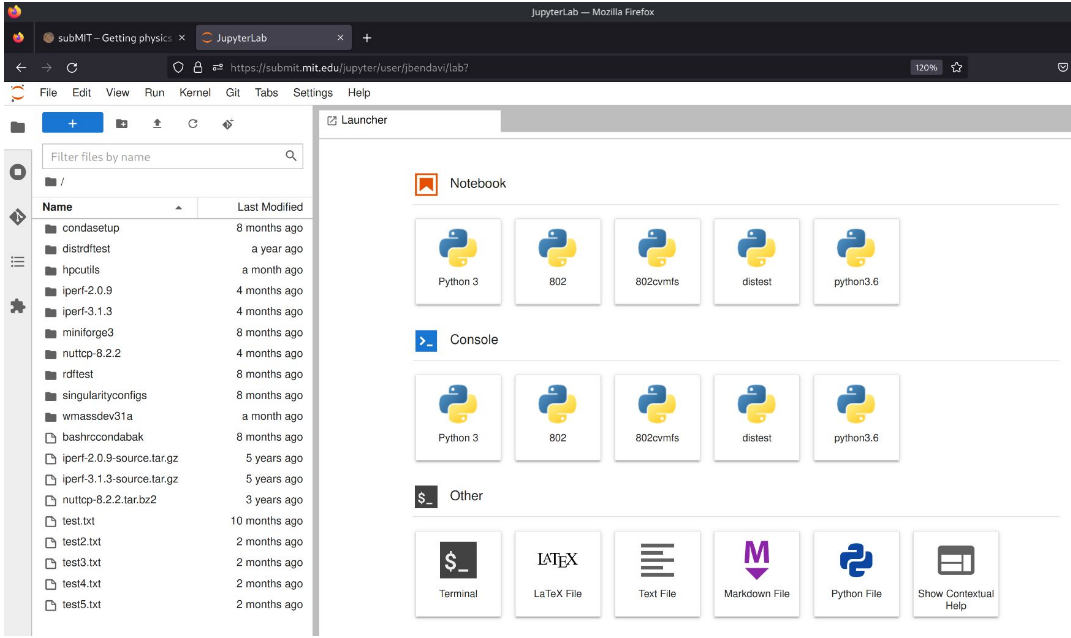
Quick introduction:

- **Spawn server menu**:
  - Slurm - Submit - 1 CPU, 500 MB: spawns a server on submit slurm partition.
  - Slurm - Submit - 2 CPUs, 1000MB: similiar as above, with more resources allocated.
  - Slurm - SubmitGPU - 1 GPU: spawns a server on submit-gpu slurm partition, requesting 1 GPU.
  - Slurm - SubmitGPU1080 - 1 GPU: spawns a server on submit-gpu1080 slurm partition, requesting 1 GPU.
  - Local server - Submit01 - 1 CPU, 500 MB, /home/submit/{username}/: spawns on submit01, in your /home/submit/{username}/ directory.
  - Local server - Submit01 - 1 CPU, 500 MB, /work/submit/{username}/: spawns on submit01, in your /work/submit/{username}/ directory.
- **GPUs**: you can use GPU resources in your notebooks or Jupyterhub's terminal if you spawn a server on submit-gpu or submit-gpu1080, supported through Slrum.
- **Conda**: your conda environments should be automatically loaded as kernels by Jupyterhub, and can be used in notebooks. See User Guide for more info.
- **Singularity**: you can manually set up a kernel based on a singularity environment's python. See User Guide for more info.

For more information about Submit, conda, GPUs, Jupyterhub, etc., see:

User Guide

For any questions, comments, or feedback, please send an email to submit-jupyter.



Name	Last Modified
condasetup	8 months ago
distrdftest	a year ago
hpcutils	a month ago
iperf-2.0.9	4 months ago
iperf-3.1.3	4 months ago
miniforge3	8 months ago
nuttcp-8.2.2	4 months ago
rdftest	8 months ago
singularityconfigs	8 months ago
wmassdev31a	a month ago
bashrccondabak	8 months ago
iperf-2.0.9-source.tar.gz	5 years ago
iperf-3.1.3-source.tar.gz	5 years ago
nuttcp-8.2.2.tar.bz2	3 years ago
test.txt	10 months ago
test2.txt	2 months ago
test3.txt	2 months ago
test4.txt	2 months ago
test5.txt	2 months ago

- Interactive Jupyter session available directly from website with touchstone authentication (subMIT account still required)
- SLURM is used to efficiently share resources between interactive and batch use
- Primary usage is research, but also being used for classroom exercises

Communication Channels

- User support mailing list: submit-help@mit.edu
- Experimental large language model application under development for interactive user support and to augment support ticket handling
  - Joint project with College of Computing, with dedicated funding
  - More discussion later + dedicated talk at LLM workshop on Friday https://indico.mit.edu/event/759/
- Slack workspace: https://mit-submit.slack.com
  - “help-desk” channel
- Monthly Users Group Meetings
  - Open for discussion
  - Open for user contributions: full set of Users Group representatives can be contacted at submit-usersgroup@mit.edu
- Annual subMIT workshop
  - February 2024 workshop: https://indico.mit.edu/event/956/
- In addition to direct interaction with the subMIT project team, users are encouraged to discuss with Users Group representative from their own group or “nearby” group

Linux Distribution Upgrade

- Previous CentOS 7 distribution reaches EOL for maintenance updates in June 2024
- Decision by Red Hat to reorganize CentOS project and releases disrupted the logical upgrade path from CentOS 7->8
- Decision was taken to upgrade from CentOS 7 to Alma Linux 9, considering:
  - Ease of transition
  - Support lifetime
  - Functionality
  - Direction being taken at other universities and labs (CERN, Fermilab, etc)
- Discussion included Users Group and broader community
- System has been fully upgraded, with user facing services (ssh, Jupyter, batch queues) switched to Alma 9 by default in April
- Dedicated documentation to ease user transition
  - https://submit.mit.edu/submit-users-guide/future/alma.html
- Well-supported and documented use of containers to keep older software environments available where needed

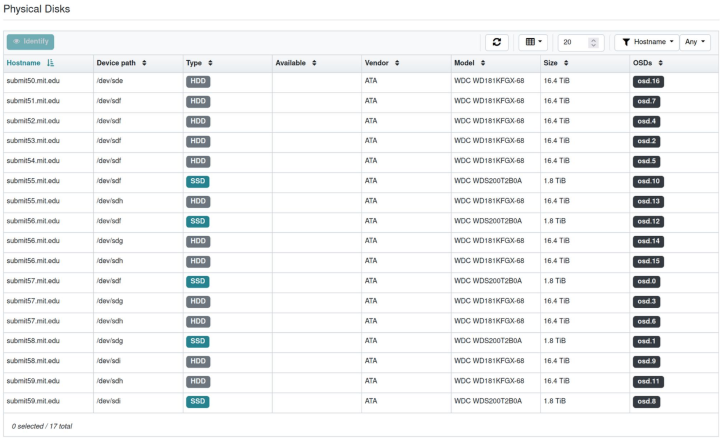
Mass Storage Upgrade

- Current mass storage system (/data/submit) is 500TB of spinning disks in a Gluster distributed filesystem
- **Users experience throughput bottlenecks for high performance analysis**
  - unable to effectively leverage the throughput of a large number of disks in parallel
- **Some other performance and operational issues related to user access patterns (system responsiveness/reliability problems for users)**
  - Large number of small files, large number of files in a single directory
  - Technical protections and user education to mitigate
- A number of limitations in maintenance and flexibility, plus suboptimal failure modes (files appear to have vanished when they are only temporarily offline)
- These issues also drive extra demand for scarce/expensive NvME storage (/scratch area)
- Migration in progress to higher performance CephFS

Mass Storage Upgrade

- Migration in progress to higher performance CephFS
- Test system was deployed using spare disks for initial testing/planning
- Storage servers hosting existing gluster system have been upgraded to Alma 9 as well
- Production Ceph system deployed in parallel on existing servers, 216TB of spinning disks from spares + disks borrowed from Tier 2 project
- Additional disks have been ordered (20x20TB) from funds allocated for storage upgrade hardware
- Migration plan:
  - Commissioning and performance tests being performed on 200TB system now
  - Additional disks to be incorporated into cluster
  - Migration of user data from gluster to CephFS
  - Decommissioning of gluster and incorporation of disks into Ceph
- Estimated complete migration by the end of July
- Net result will be a faster and more robust mass storage system, with almost 2x the existing capacity

Mass Storage Upgrade

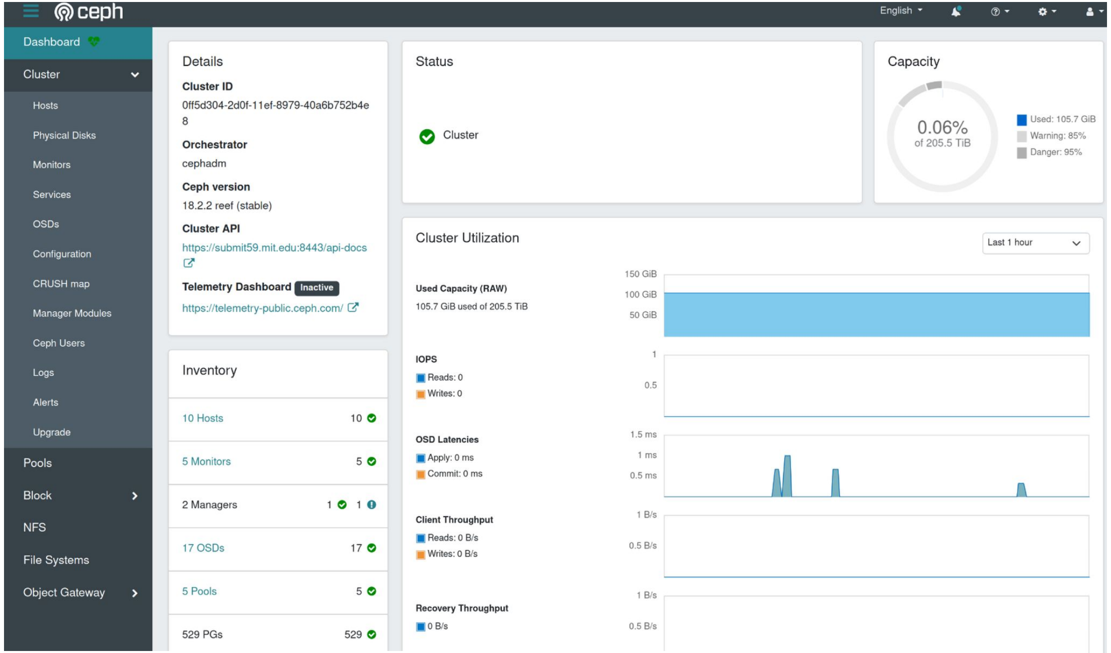


Physical Disks

Hostname	Device path	Type	Available	Vendor	Model	Size	OSDs
submit50.mit.edu	/dev/sde	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.16
submit51.mit.edu	/dev/sdf	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.7
submit52.mit.edu	/dev/sdf	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.4
submit53.mit.edu	/dev/sdf	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.2
submit54.mit.edu	/dev/sdf	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.5
submit55.mit.edu	/dev/sdf	SSD		ATA	WDC WDS200T2B0A	1.8 TiB	osd.10
submit55.mit.edu	/dev/sdh	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.13
submit56.mit.edu	/dev/sdf	SSD		ATA	WDC WDS200T2B0A	1.8 TiB	osd.12
submit56.mit.edu	/dev/sdg	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.14
submit56.mit.edu	/dev/sdh	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.15
submit57.mit.edu	/dev/sdf	SSD		ATA	WDC WDS200T2B0A	1.8 TiB	osd.0
submit57.mit.edu	/dev/sdg	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.3
submit57.mit.edu	/dev/sdh	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.6
submit58.mit.edu	/dev/sdg	SSD		ATA	WDC WDS200T2B0A	1.8 TiB	osd.1
submit58.mit.edu	/dev/sdi	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.9
submit59.mit.edu	/dev/sdh	HDD		ATA	WDC WD181KFGX-68	16.4 TiB	osd.11
submit59.mit.edu	/dev/sdi	SSD		ATA	WDC WDS200T2B0A	1.8 TiB	osd.8

0 selected / 17 total

- Monitoring and maintenance capabilities of new system already far superior



Details

Cluster ID
0ff5d304-2d0f-11ef-8979-40a6b752b4e8

Orchestrator
cephadm

Ceph version
18.2.2 reef (stable)

Cluster API
https://submit59.mit.edu:8443/api-docs

Telemetry Dashboard Inactive
https://telemetry-public.ceph.com/

Status

✓ Cluster

Capacity

0.06% of 205.5 TiB

Used: 105.7 GiB
Warning: 85%
Danger: 95%

Inventory

10 Hosts	10 ✓
5 Monitors	5 ✓
2 Managers	1 ✓ 1 ●
17 OSDs	17 ✓
5 Pools	5 ✓
529 PGs	529 ✓

Cluster Utilization

Last 1 hour

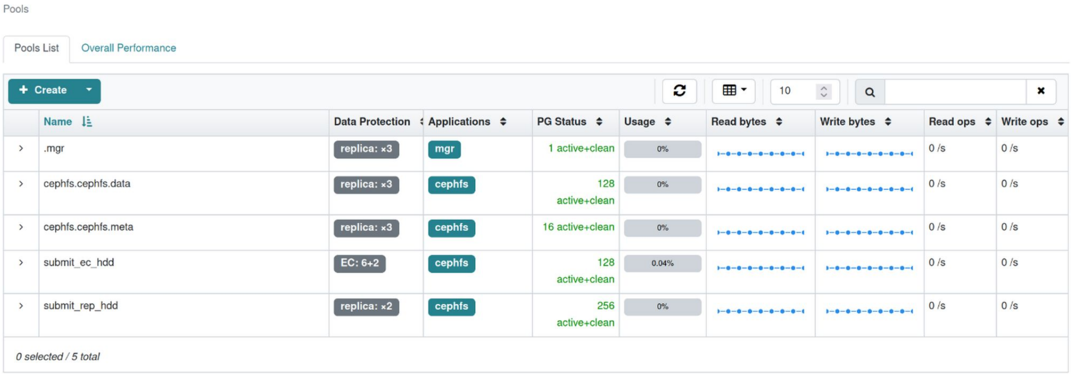
Used Capacity (RAW)
105.7 GiB used of 205.5 TiB

IOPS
Reads: 0
Writes: 0

OSD Latencies
Apply: 0 ms
Commit: 0 ms

Client Throughput
Reads: 0 B/s
Writes: 0 B/s

Recovery Throughput
0 B/s



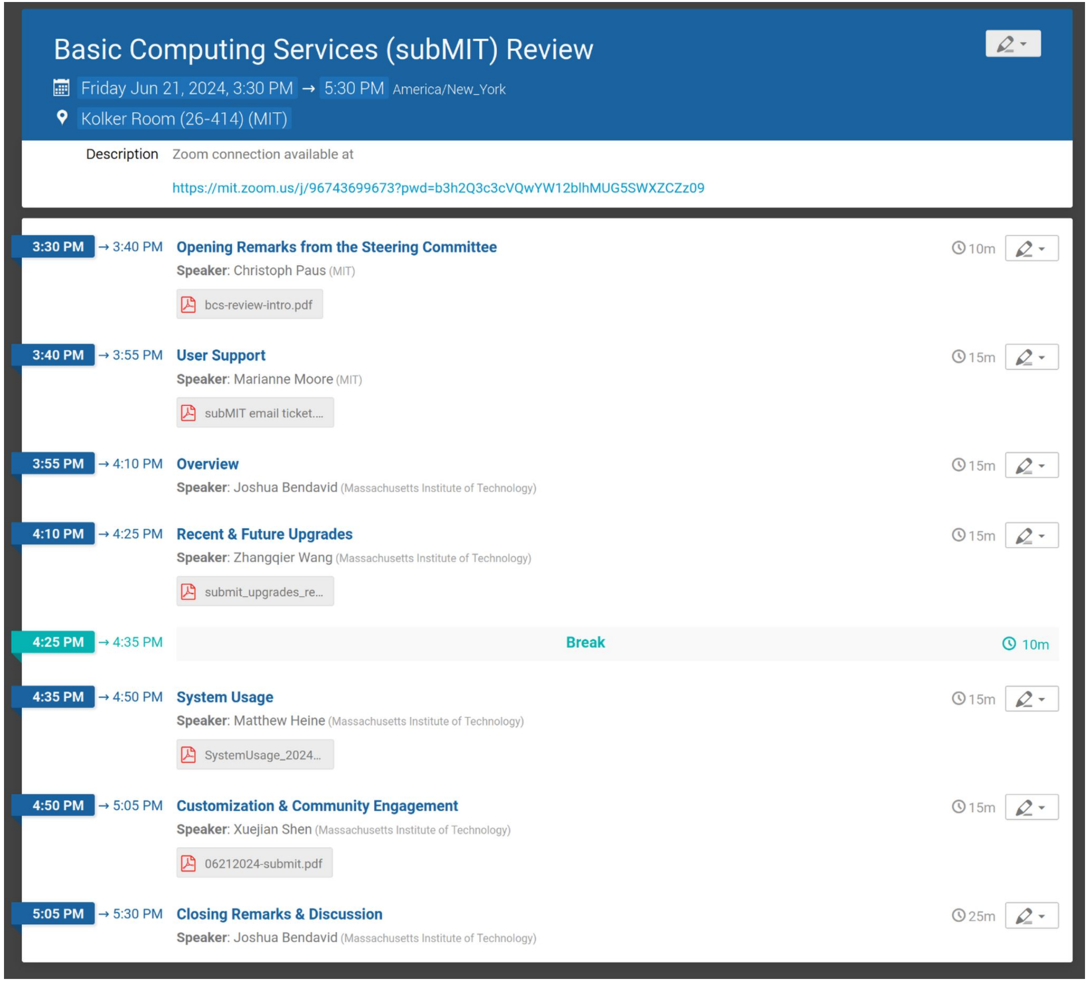
Pools

	Name	Data Protection	Applications	PG Status	Usage	Read bytes	Write bytes	Read ops	Write ops
›	.mgr	replica: ×3	mgr	1 active+clean	0%			0 /s	0 /s
›	cephfs.cephfs.data	replica: ×3	cephfs	128 active+clean	0%			0 /s	0 /s
›	cephfs.cephfs.meta	replica: ×3	cephfs	16 active+clean	0%			0 /s	0 /s
›	submit_ec_hdd	EC: 6+2	cephfs	128 active+clean	0.04%			0 /s	0 /s
›	submit_rep_hdd	replica: ×2	cephfs	256 active+clean	0%			0 /s	0 /s

0 selected / 5 total

Today’s Review

- Indico page with timetable and slides:
  - https://indico.mit.edu/event/1073/



Basic Computing Services (subMIT) Review

Friday Jun 21, 2024, 3:30 PM → 5:30 PM America/New_York

Kolker Room (26-414) (MIT)

Description Zoom connection available at

https://mit.zoom.us/j/96743699673?pwd=b3h2Q3c3cVQwYW12blhMUG5SWXZCZz09

3:30 PM → 3:40 PM **Opening Remarks from the Steering Committee** 10m
Speaker: Christoph Paus (MIT)
bcs-review-intro.pdf

3:40 PM → 3:55 PM **User Support** 15m
Speaker: Marianne Moore (MIT)
subMIT email ticket....

3:55 PM → 4:10 PM **Overview** 15m
Speaker: Joshua Bendavid (Massachusetts Institute of Technology)

4:10 PM → 4:25 PM **Recent & Future Upgrades** 15m
Speaker: Zhangqier Wang (Massachusetts Institute of Technology)
submit_upgrades_re...

4:25 PM → 4:35 PM **Break** 10m

4:35 PM → 4:50 PM **System Usage** 15m
Speaker: Matthew Heine (Massachusetts Institute of Technology)
SystemUsage_2024...

4:50 PM → 5:05 PM **Customization & Community Engagement** 15m
Speaker: Xuejian Shen (Massachusetts Institute of Technology)
06212024-submit.pdf

5:05 PM → 5:30 PM **Closing Remarks & Discussion** 25m
Speaker: Joshua Bendavid (Massachusetts Institute of Technology)