Overview of RCIC Resources. Some New Things. And BACKUP your STUFF

Philip Papadopoulos, Ph.D

ppapadop@uci.edu

https://rcic.uci.edu

- Build and maintain scalable computing and storage resources for researchers
- Work directly with researchers (grad students, faculty, post-docs, …) to define the computing environment
- “Clusters R Us” – computing and storage clusters. We work in midscale (10000 cores) computing and storage (5-10PB).
  - The next scale up (100K cores) and 50-100PB is handled better at national resource centers

RCIC Faculty Oversight

Executive Committee – Chair Filipp Furche, Professor, Dept. of Chemistry

- Help with strategic guidance and direction
- Approval chain for large purchases (> $100K) and high-level policy
- Meet approximately semi-annually (next meeting: 10/4/2023)

- Advisory Committee
  - About 30 researchers from disciplines across UCI
  - Key feedback on what RCIC does right and wrong. They are not shy about expressing their views.

Formation of RCIC was the result of the UCI Cyberinfrastructure Vision 2016

Key Resources @ RCIC

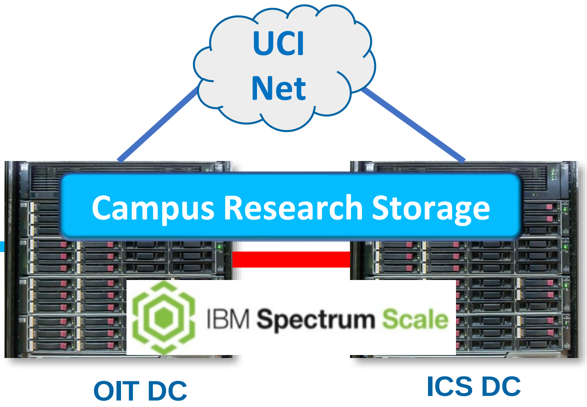
HPC3

- ~9600 Cores/224 Hosts
- 108 Nvidia GPUS(52 v100s, 48 A30s, 8 A100s) EDR (100Gbps) Infiniband
- 10GbE Ethernet
- Minimum
  - 4GB memory/core
  - AVX2 instruction set (Epyc/Intel CPUs)

Seven Parallel File Systems

DFS3b, DFS4,DFS5, …,DFS9

- 7.75 PB usable storage
- ~6GB/sec bandwidth/System
- Regular Backups



CRSP – Campus Research Storage Pool

- 1.1 PB usable storage
- Available anywhere on UCI Network
- Dual Copy of All Data
- Snapshots
- Highly available
- Regular Backup
- 87% Full

Driving Principles

FREE STUFF

- Every <u>Faculty</u> member has <u>no-cost access</u> to significant resources
  - Cost to go beyond baseline is based on the cost of hardware only
- Position resources to be significant - *but not a replacement* for national scale resources (like SDSC, NCSA, TACC, NCAR, …)
- Software environments need to be consistent and well-managed
  - RCIC spends significant effort spent to build/maintain domain-specific environments
  - Not possible to "cover the waterfront"
  - We build 1000s of individual software components. ~1 year cadence for updates to R, Python, Tensorflow, MATLAB, Conda, etc.
- Data integrity/availability are critical to success

HPC3 - Goals

1. Enables users to have access to a larger compute/analysis system than they could reasonably afford "on their own"
2. Enables access to specialized nodes (large memory, 64bit GPU)
3. Fosters a growing community across UCI to utilize scalable computing (HPC and HTC)* for their scientific research program and teaching
4. Provides a well-managed software environment that forms the basis of a *reproducible* and more secure research environment

** HPC – High-Performance Computing*
HTC – High-Throughput Computing

What does HPC3 look like?

- HPC3 is 100s of servers, 1000s of compute cores

- 9632 x86 Cores
- 224 servers
- 4 different brands of hardware
- Grows every year

This is a "Sea" of physical resources.

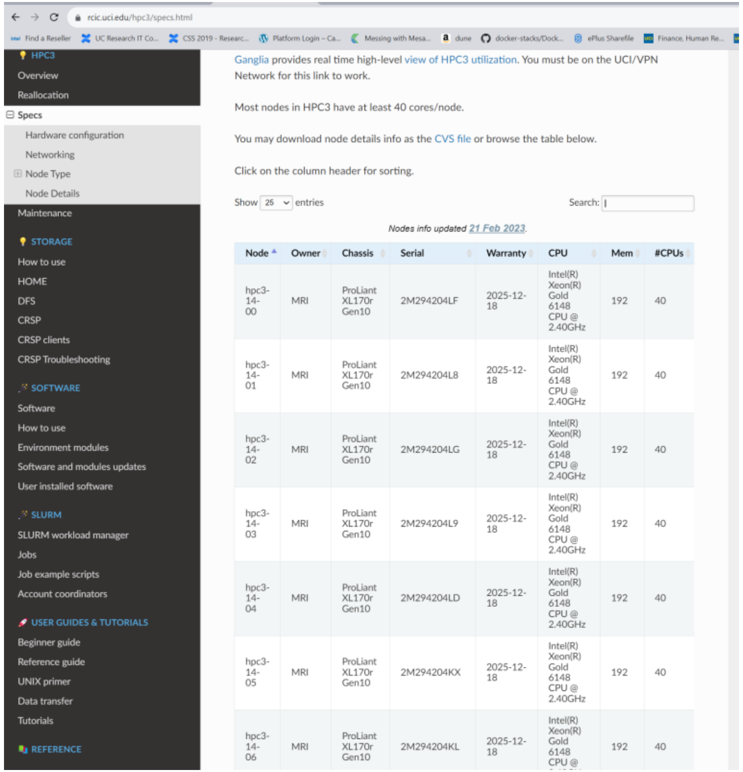
Interconnected by high-speed networking.

- And Many Petabytes (1 PB = 1000 TB = 1,000,000 GB) of Storage

- Seven Parallel File Systems (BeeGFS)
- One home area file system
- ~900 hard drives
- Every compute/GPU node has two local drives

Detailed Nodes Specs

- Searchable/Filtered online view
    - Owner
    - Warranty Date
    - Cores/node
    - Memory/Node
- Can query Slurm for more specifics



Ganglia provides real time high-level view of HPC3 utilization. You must be on the UCI/VPN Network for this link to work.

Most nodes in HPC3 have at least 40 cores/node.

You may download node details info as the CVS file or browse the table below.

Click on the column header for sorting.

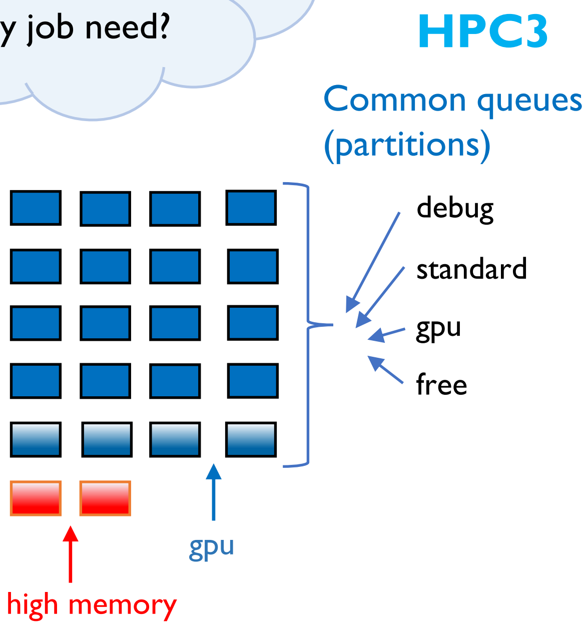
Show 25 entries Search:

Nodes info updated 21 Feb 2023.

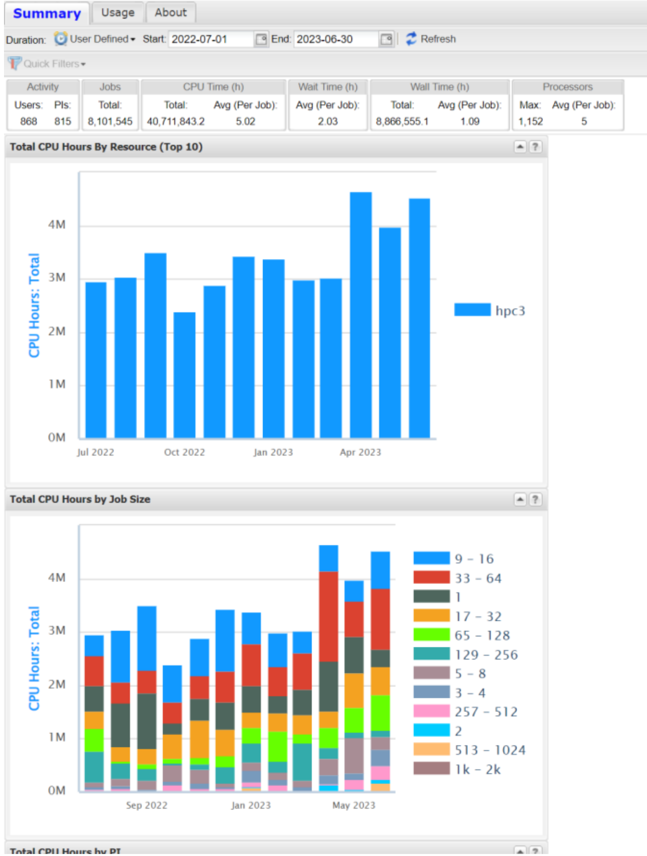
Node	Owner	Chassis	Serial	Warranty	CPU	Mem	#CPUs
hpc3-14-00	MRI	ProLiant XL170r Gen10	2M294204LF	2025-12-18	Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz	192	40
hpc3-14-01	MRI	ProLiant XL170r Gen10	2M294204L8	2025-12-18	Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz	192	40
hpc3-14-02	MRI	ProLiant XL170r Gen10	2M294204LG	2025-12-18	Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz	192	40
hpc3-14-03	MRI	ProLiant XL170r Gen10	2M294204L9	2025-12-18	Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz	192	40
hpc3-14-04	MRI	ProLiant XL170r Gen10	2M294204LD	2025-12-18	Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz	192	40
hpc3-14-05	MRI	ProLiant XL170r Gen10	2M294204KX	2025-12-18	Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz	192	40
hpc3-14-06	MRI	ProLiant XL170r Gen10	2M294204KL	2025-12-18	Intel(R) Xeon(R) Gold 6148 CPU @	192	40

Submitting Jobs: Queues on HPC3

- Submit a job to the desired queue, asking for resources (CPUs, memory) you need
- You will share the node with other (unrelated) jobs, but linux cgroups are used to reserve memory, cpu for your job
- There are some maximums on queues to prevent resource starvation (most never see these)



A year of usage: July 2022 – June 2023



Duration: User Defined Start: 2022-07-01 End: 2023-06-30

Activity		Jobs	CPU Time (h)		Wait Time (h)	Wall Time (h)		Processors	
Users:	PIs:	Total:	Total:	Avg (Per Job):	Avg (Per Job):	Total:	Avg (Per Job):	Max:	Avg (Per Job):
868	815	8,101,545	40,711,843.2	5.02	2.03	8,866,555.1	1.09	1,152	5

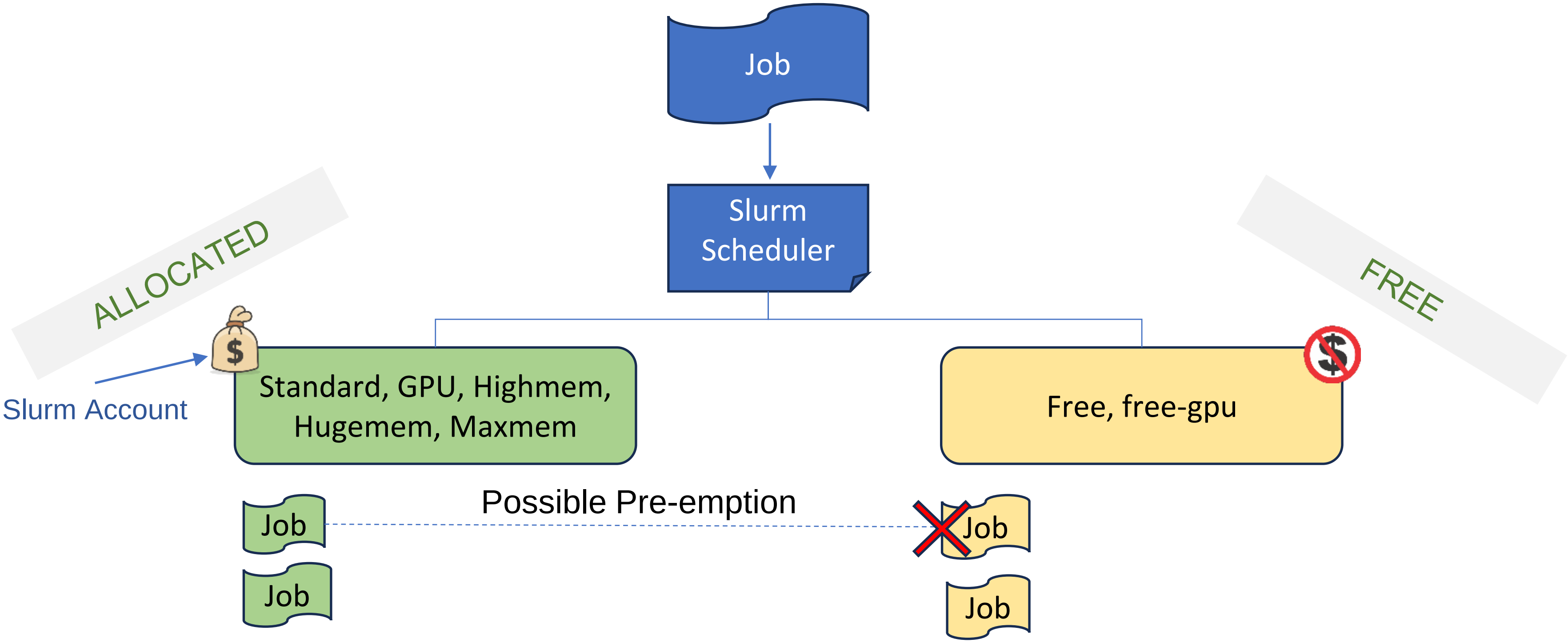
Total CPU Hours By Resource (Top 10)

Total CPU Hours by Job Size

Total CPU Hours by PI

- **868 unique users** (~ 15% of all faculty AND Grad students) ran **8.1M jobs**
- 40.7M CPU hours = 4650 core-years
- Max job size – 1152 cores
- Most impacted Month: April 2023
  - 4.65M CPU hours = 6460 core-months
- Users will see start to see wait times when instantaneous consumption is above 90%
- **→ There are, however, a *significant* # of unclaimed (and now lost) cycles**

Two Job Types on HPC3: Allocated and Free



Job

Slurm Scheduler

ALLOCATED

Slurm Account

Standard, GPU, Highmem, Hugemem, Maxmem

FREE

Free, free-gpu

Job

Job

Possible Pre-emption

Job

Job

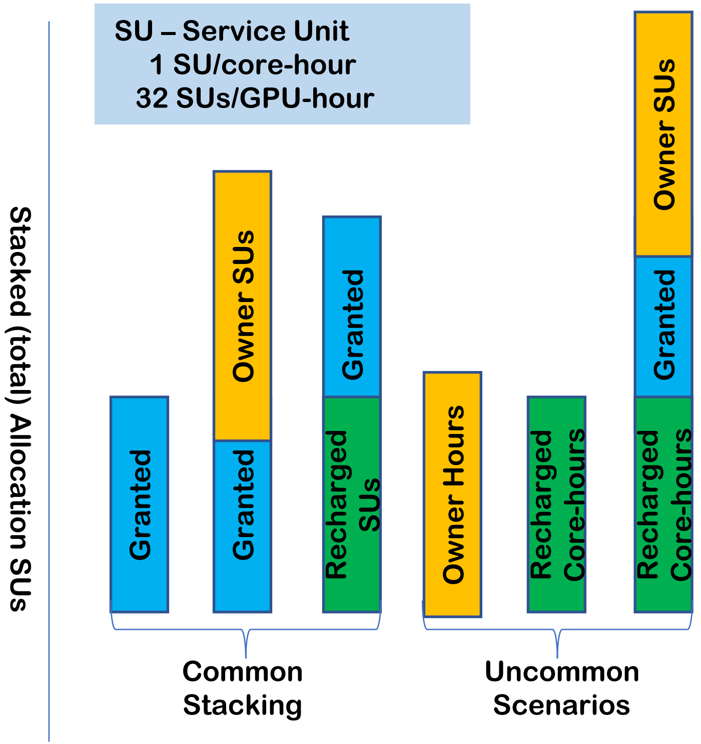
Resources charged to a Slurm Account
Jobs cannot be pre-empted

No cost.
Jobs can be **pre-empted (killed)** at any time

How Do you Get "Money" into your Slurm Account

- SLURM Account monetary unit = SU (Service Unit). Charges: 1/core-hour, 32/GPU-hour
- Account OWNER (Lab accounts) can grant access to any other HPC3 user
- No Cost Allocation for Faculty
  - CPU:We start you out at 100000 SUs/6 months. Refill based upon actual use. If no use, reduces to minimum of 12500 SUs/6 months
  - GPU: Can request 66K SUs/6 months. Refill based upon actual use. If NO use, account removed
  - https://rcic.uci.edu/hpc3/reallocations.html#reallocation
- Purchase Prepaid SUs
  - $.01/SU. > $5K = 20% discount.
  - Recharge via memorandum. We don't "post-bill hours"
- Buy Hardware
  - Must coordinate with RCIC prior to purchase. We'll assist with quotes from standard configs
  - SUs credited/year = 95% * (# cores + 32 * #GPUs)* 8760 hours/year
  - Credits computed every 6 months. Unused credits from previous 6 months are lost.
  - Cost of 48-core node, if every owner-credit is utilized:
    - 95% * 48 * 8760 * 6 years = 2.4M core hours. $13K/2.4M hours = $.0054/SU

Your allocation is "Stacking" of different types of SUs

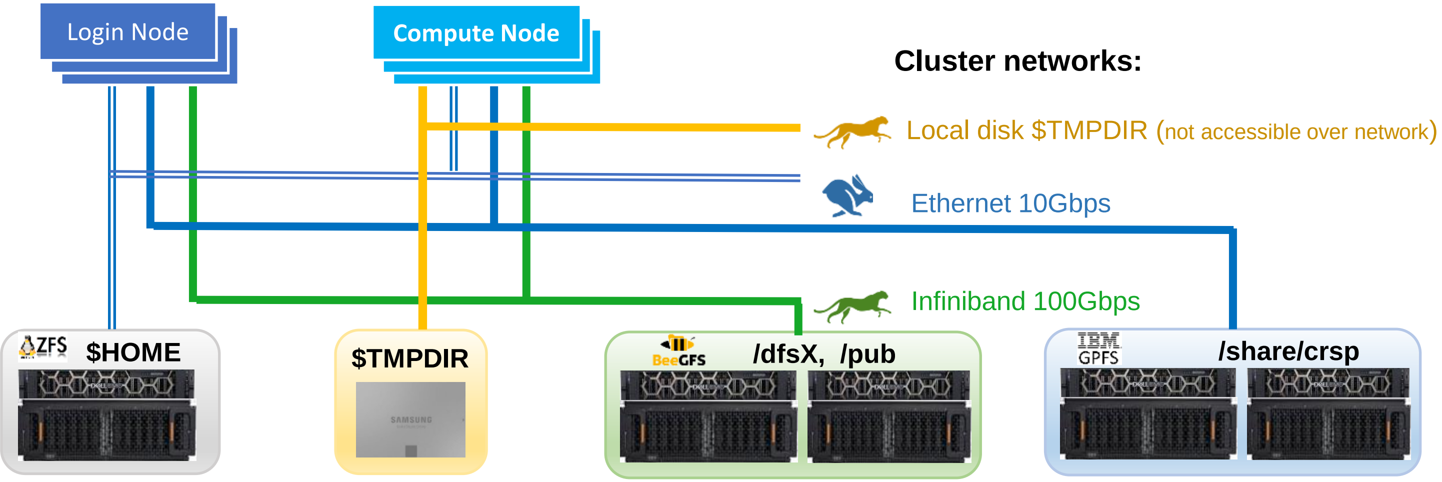


- Every 6 months, we recompute your Allocation
- If you have unused SUs from the previous 6 months, they are lost (no rollover)
    - Recharged SUs: 18-month lifetime
- If you don't use enough of your granted hours, they will be reduced in the next cycle

Details:

https://rcic.uci.edu/hpc3/allocations.html

Storage: Connectivity, File System architecture, and physical hardware all contribute to performance.



Home

DO NOT USE for data intensive batch jobs

1. Source code, binaries
2. Small (order of Mbs) data files
3. Convenience mount over NFS

Scratch

Use as LOCAL scratch storage for batch Jobs : many small files or make frequent small reads/writes

1. **Fastest performance**, data is removed when job completes
2. **@ Job Start:** explicitly copy input files from DFS/CRSP to $TMPDIR
3. **@ Job End:** explicitly copy output files from $TMPDIR

Parallel

Use for batch jobs

1. Source code, binaries
2. Best for processing Medium/Large data files (order of 100s Mbs/Gbs)
3. Most common place for data used in batch jobs
4. **Not optimal/advised for many small files**

Campus Storage

Use sometimes for batch Jobs

1. Source code, binaries
2. Best for processing Medium/Large data files (order of 100s Mbs/Gbs)
3. Convenience mount over NFS
4. **Usually better to use DFS or local $TMPDIR storage**

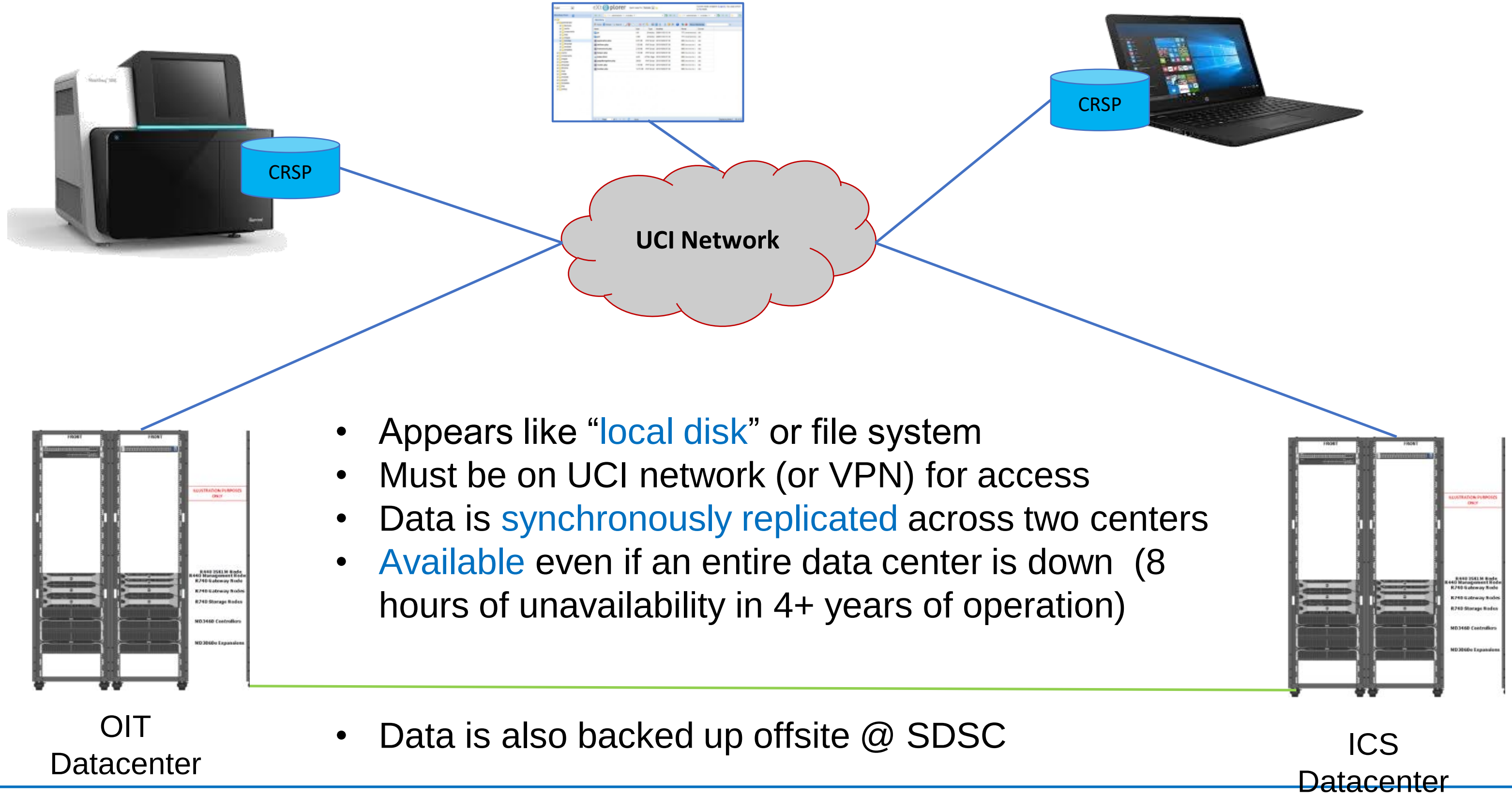
Storage Considerations

- Storage is shared among all users. **The nature of networked-storage makes it possible for a *single user to render* a file system *unusable* for all.**
- **User's responsibility**
  - **Understand how their code(s) interact with their data/storage.**
  - **Choose the appropriate file system**
  - **THINK! What do you think your code does if 1000 copies are running at the same time and accessing the same folder?**
- **Repeated access to LOTS of small files (< 256KB each) are problematic for everything except FLASH.**
- **Parallel file systems (DFS/CRSP) are ideal for big files (> 1MB). They are TERRIBLE for tiny files (< 64K)**
- **So, "Q: Why are there different file systems on HPC3? Flash would be so cool everywhere." A: Money**

Fastscratch - Soon on a cluster near you

- Adding another Storage Capability to the mix
- What is it?
  - 100TB of All NVMe (Flash) storage configured
  - An RDMA-based NFS file server
  - NO BACKUP! NO SNAPSHOTS. Delete that file and it's G-O-N-E, baby.
- Model of Use:
  - A user can request an allocation lasting not more than 4 weeks.
  - One allocation/user. Deleted if not utilized. At the end of an allocation, all data is deleted.
  - Target: lots of jobs that need to the same set of input files that
    - Either Don't fit on local flash drives
    - Cost (time) too much to copy the data in/out every job
- This is "Cache" and it can go away at any time.
- WHY? Abusive Bioinformatics and Genomics Codes
  - Some Older code attempts to treat a file system as a "database" with many small files, repeated access.
  - Local flash is the ONLY file system that can weather this kind of abuse (use it when warranted)
  - RDMA NFS is cluster-wide (shareable among nodes/jobs). Much better than DFS/CRSP for this use case

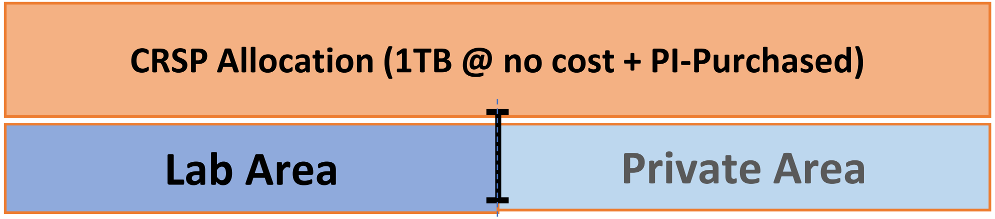
Campus Storage: CRSP High-Level Overview



- Appears like “local disk” or file system
- Must be on UCI network (or VPN) for access
- Data is synchronously replicated across two centers
- Available even if an entire data center is down (8 hours of unavailability in 4+ years of operation)

- Data is also backed up offsite @ SDSC

CRSP Storage – Lab-centric



CRSP Allocation (1TB @ no cost + PI-Purchased)

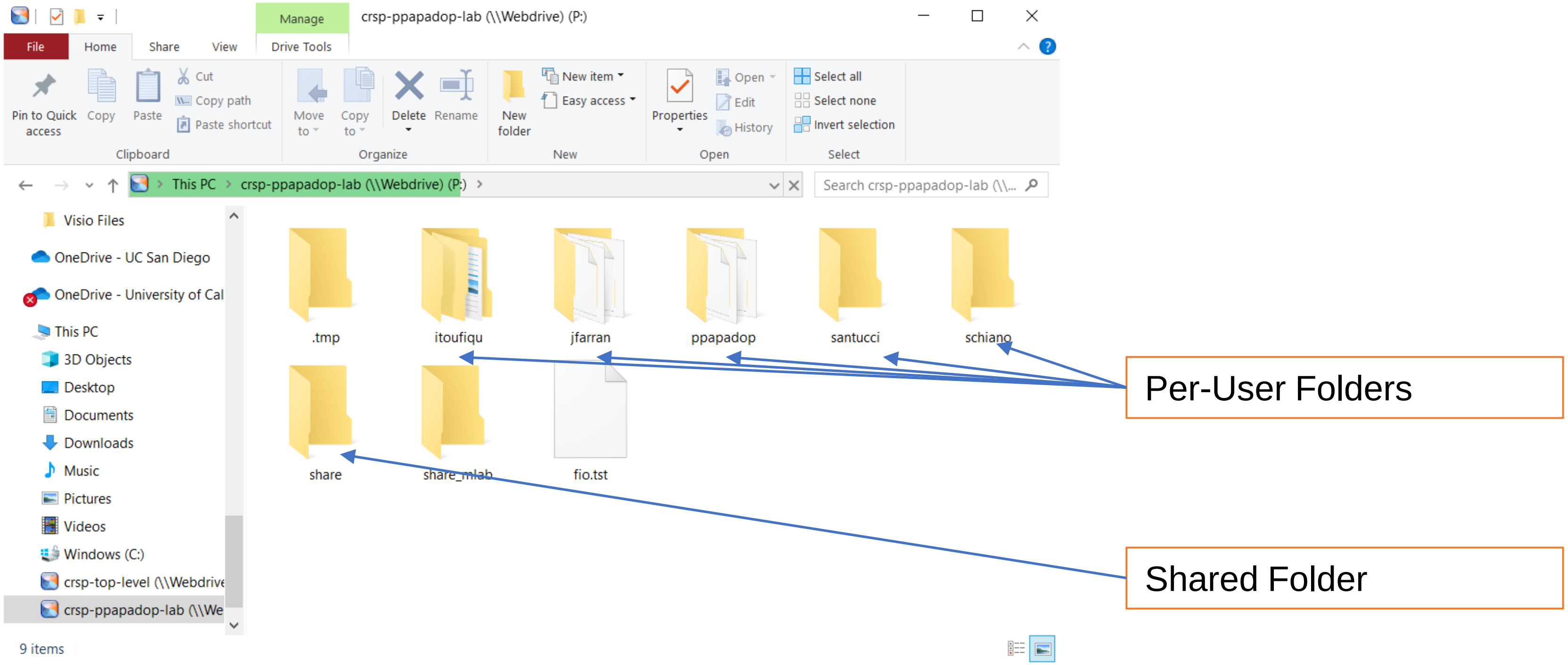
Lab Area

Private Area

- Behaves “like a disk”
- PI grants explicit access to others
- Lab-wide Share Folder
- Individual user folders
- PI has access to all files
- When someone separates from UCI, data remains. Ownership assigned to the Lab Owner

- Not intended for sharing with others
- If you want to share folders, they need to be in a different area on CRSP

A Sample Lab - ppapadop



crsp-ppapadop-lab (\\Webdrive) (P:)

This PC > crsp-ppapadop-lab (\\Webdrive) (P:) >

.tmp
itoufiqu
jfarran
ppapadop
santucci
schiano
share
share_mlab
fio.tst

9 items

Per-User Folders

Shared Folder

CRSP2 – In procurement

- CRSP is 85-88% utilized and is almost 5 years old
- Since January 2023 have been working on a replacement system via standard request for proposal (RFP)
- Update (yesterday): RFP failed (vendors were exceptionally greedy and we don’t have the deep pockets)
- New tactic – update all CRSP hardware and double capacity (~2.2PB). Building an RFQ (request for quote) for replacement hardware
- Reality – CRSP2 cannot be online before end of 2023 – working to recover from failed RFP – goal: end of Q1 2024.

Storage – You can lease storage from RCIC

- We focus on delivering reliable storage via two different classes
- CRSP – Campus Research Storage Pool – available anywhere on campus.
- DFS – Cluster-local, high-performance parallel file system

Feature	CRSP	Cluster-local (DFS)
Availability	Highly-available. Anywhere on campus, including HPC3	Infrequent downtimes occur. Cluster only
Cost	\$60/TB/Year[1]	\$100/TB/5 Years
Snapshots	Yes	No
Backups	Daily	Daily
dbGAP (P3)	Yes	Soon

[1] We expect this do go down somewhat with CRSP2.

High-level View costs-

FREE STUFF

No-Cost Allocations

Role	HPC3 Core Hours	GPU Hours	Home Area Storage	DFS Storage	CRSP Storage
Faculty	200K hours/year[1]	By Request ~4K hours/year[1]	50GB	1TB in Pub	1 TB
Student	1000 hours	---	50GB	1 TB in Pub	---

An Expansion Option: Core/GPU Recharge (vs. AWS UC Costs)

	HPC3 Core Hours	GPU Hours	Home Area Storage	DFS Storage	CRSP Storage
Faculty	$.008/core hour	$0.28/GPU hour	Not expandable	$100/TB/5 years	$60/TB/year
AWS Equivalent[3]	C5n.large $.029	P3.2xlarge $0.84	---	---	S3[2] Standard $145/TB/year

[1] Exact amounts dependent on # requests/available hardware

[2] Comparison difficult - S3 has higher durability, CRSP has no networking fee.

[3] modeled on three-year reserved with UC discounts, on-demand is twice as expensive

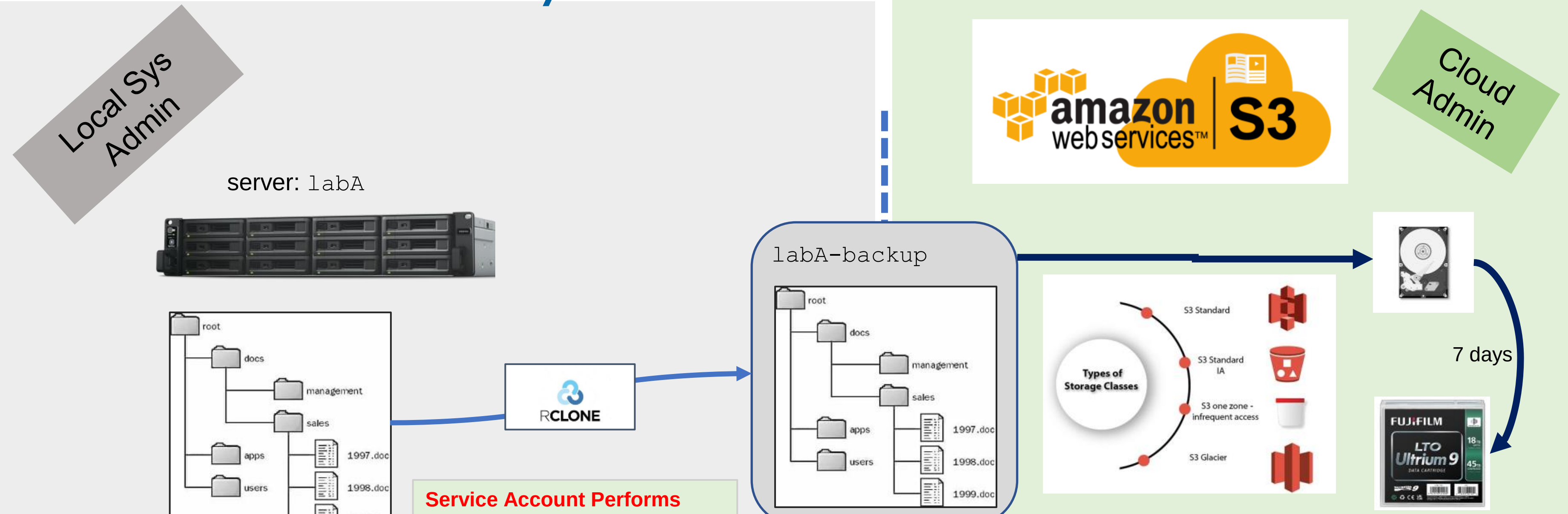
Backup of your STUFF (data)

- WE do.
  - CRSP – two copies on site + snapshots. Nightly backup to SDSC
  - DFS – Nightly backup to SDSC or SBAK (pub)
  - Home – Nightly backup to SBAK + Snapshots
- Desktops and Laptops (only for unmanaged systems in COHS)
  - Crashplan. Free. https://www.oit.uci.edu/services/research/crashplan/
  - 4 systems/user. Available for researchers (faculty, grad students, postdocs, research staff, undergrads in a research lab)
  - No limit on data but practical limit is 2-3TB
  - IF you have a use for it, PLEASE USE IT.
- Lab-based storage
  - Next Slide

Backing up Servers (10s to 1000s of TB)

- MANY Synology-based NAS (Network-attached Storage) in COHS
  - Backup to Google Drive MUST cease no later than March 2024 because unlimited storage is ending: https://www.oit.uci.edu/services/communication-collaboration/google/
  - Backup to OneDrive is not an option (They will be adopting similar limits that Google placed on all .edu)
- Backup to AWS
  - Currently funded by the Provost/Vice Chancellor ITD (Andriola)
  - Custom configuration, Some Software, on-boarding by RCIC.
  - Beta (Now) → Production.
  - Evaluated (6 months) some "commercial" software
    - EITHER didn't scale
    - OR "stupidly" expensive
  - Our cost target is ~ $35/TB/Year ($35K/PB, $350K/10PB)
    - Storage ~ $25/TB/Year (Glacier Tape-based)
    - Sync costs ~ $10/TB/Year

Basic Backup Overview: Utilize open source `rclone` for data synchronization to AWS S3



Service Account Performs Backup

1. Can only read/write backup the CLI (no AWS console access)
2. Deletion is not immediate

Scheduled Synchronization keeps Cloud copy up-to-date

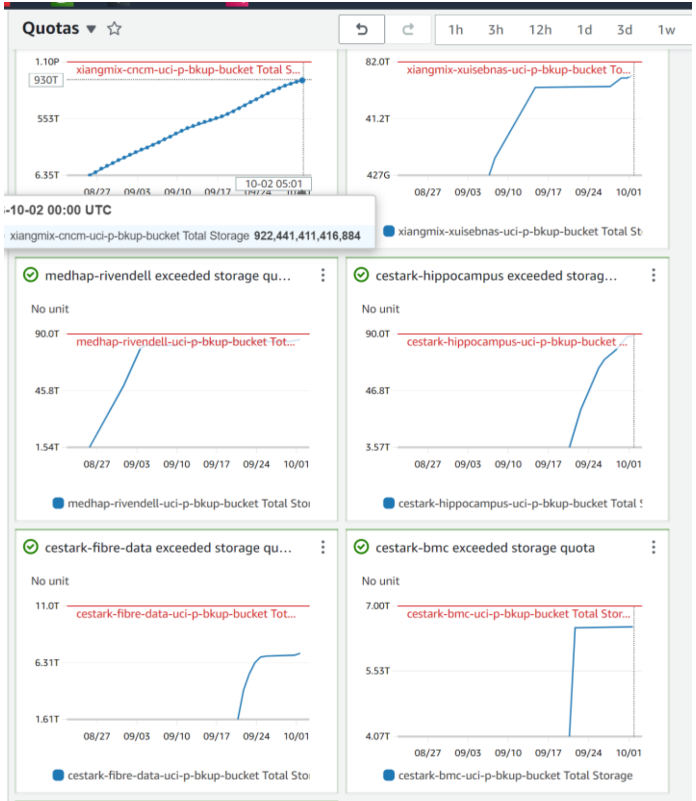
Security Separation
Service (Client) != Cloud Admin

Cloud Admin Creates AWS Environment

1. Defines Storage and Service Account
2. Role/policy limits what Service Account can Do
3. Lifecycle policies defined
   - Disk to Tape (after 7 days)
   - Recycle Bin: empties 90 days after deletion

Current State
RCS3 - https://github.com/RCIC-UCI-Public/rcs3

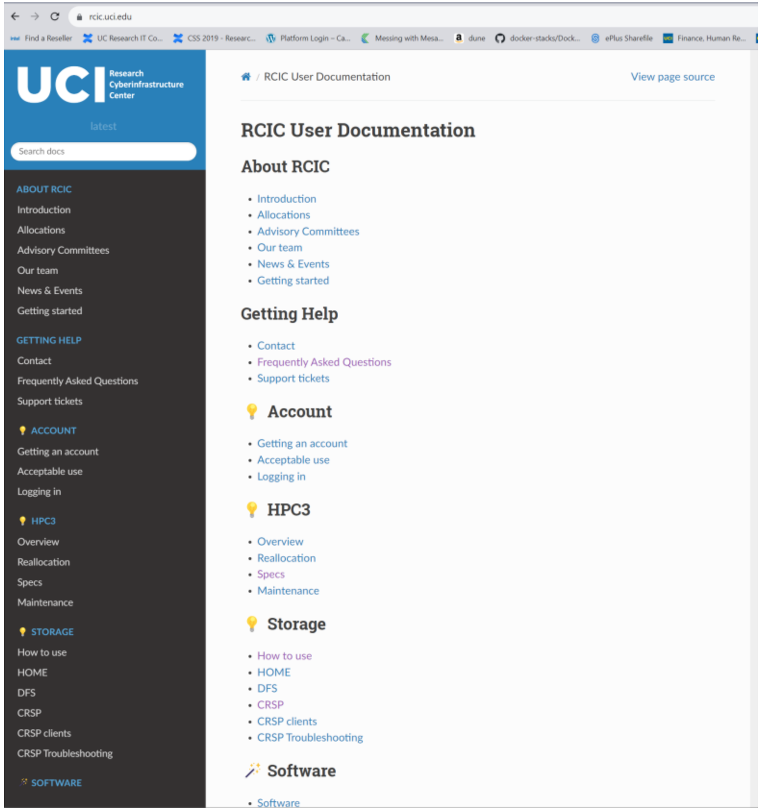
- 6 servers in Beta
  - Synology (Intel-based)
  - Linux (RedHat, Ubuntu)
  - > 1PB in aggregate
  - 30-60 minutes to onboard a new server (done over Zoom)
- RCIC only provisions the cloud storage, policies, permissions
- Local Admin installs/configures software
- Can restore, but RCIC must be involved.
  - Working on making this local-admin driven


Quotas
1h
3h
12h
1d
3d
1w
1.10P
930T
553T
6.35T
xiangmix-cncm-uci-p-bkup-bucket Total S...
10-02 05:01
08/27
09/03
09/10
09/17
09/24
10/01
82.0T
41.2T
427G
xiangmix-xuisebnas-uci-p-bkup-bucket To...
xiangmix-xuisebnas-uci-p-bkup-bucket Total St
medhap-rivendell exceeded storage qu...
No unit
90.0T
45.8T
1.54T
medhap-rivendell-uci-p-bkup-bucket Tot...
medhap-rivendell-uci-p-bkup-bucket Total Stor
cestark-hippocampus exceeded storag...
No unit
90.0T
46.8T
3.57T
cestark-hippocampus-uci-p-bkup-bucket ...
cestark-hippocampus-uci-p-bkup-bucket Total S
cestark-fibre-data exceeded storage qu...
No unit
11.0T
6.31T
1.61T
cestark-fibre-data-uci-p-bkup-bucket Tot...
cestark-fibre-data-uci-p-bkup-bucket Total Stor
cestark-bmc exceeded storage quota
No unit
7.00T
5.53T
4.07T
cestark-bmc-uci-p-bkup-bucket Total Stor...
cestark-bmc-uci-p-bkup-bucket Total Storage

-10-02 00:00 UTC
xiangmix-cncm-uci-p-bkup-bucket Total Storage 922,441,411,416,884

NEW! IMPROVED! BETTER! FASTER! CHEAPER (Our new website: https://rcic.uci.edu)

- We have spent a lot of time creating a site with good information
- Searchable
- Better Navigation than old site
- LOOK here FIRST!



RCIC User Documentation

About RCIC

- Introduction
- Allocations
- Advisory Committees
- Our team
- News & Events
- Getting started

Getting Help

- Contact
- Frequently Asked Questions
- Support tickets

Account

- Getting an account
- Acceptable use
- Logging in

HPC3

- Overview
- Reallocation
- Specs
- Maintenance

Storage

- How to use
- HOME
- DFS
- CRSP
- CRSP clients
- CRSP Troubleshooting

Software

- Software