Fermilab Facilities Report

Ken Schumacher

USQCD All-Hands Collaboration Meeting

21 April 2022

Report number - **FERMILAB-SLIDES-22-022-SCD**

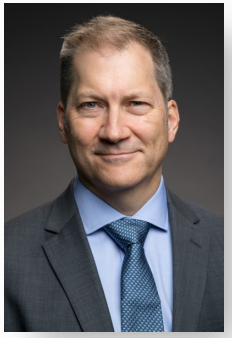
Introduction

1. Fermilab Scientific Computing Division
2. HPC Support Team and SNOW Tickets
3. Campus and Onsite Access Requests – Badge renewal
4. Fermilab's Lattice QCD Web pages
5. Fermilab's LQ1 Institutional cluster
6. Allocation Year 2021-22 (PY22) Usage
7. Storage Update (Disk and Tape)
8. SLURM Limits and Dispatch Priority
9. Scheduled Downtimes
10. Questions?

Fermilab Scientific Computing Division

- The Scientific Computing Division (SCD) provides software solutions and deploys and operates scientific computing facilities in support of the Fermilab program.
- SCD interacts with the experimental community to determine capacity, availability and capability requirements and to procure resources, as necessary.
- SCD engages in R&D activities required to maintain or advance capabilities necessary for the success of Fermilab's future physics program.
- Active Archive Facility (tape), CMS Tier-1 Center (40k cores), Data storage and handling, Data Centers, Fabric for Frontier Experiments (FIFE), General Purpose Grid, HEP Cloud and HPC.

Elizabeth Sexton-Kennedy
Chief Information Officer



James Amundson
Head of SCD

https://computing.fnal.gov/organization/
https://computing.fnal.gov/computing-facilities-and-middleware/

Fermilab

The HPC Support Team

- We have had considerable turnover and realignment

- Applications / Programming - - **Jim Simone** (Site Liaison), **Alexei Strelchenko** (Developer)
- Systems / Hardware - - SSI: Ed Simmonds (GL), **Tim Skirvin** (HPC Architect)
- Services / Support - - DCS: Ken Herner (GL), **Ken Schumacher** (Site Manager)
- Batch Management - - CSI: **Farrukh Kahn** (GL)
- Storage Services - - SSA: Robert Illingworth (GL), **Dan Szkola** (Lustre/ZFS)

User support and SNOW Tickets

- All emails to lqcd-admin@fnal.gov will automatically generate INC tickets.
- Ticket Types in SNOW (Service Now)
  - INC – Incidents are for when something is broken. Higher priority and tracked closely
  - RITM – Requests are for all other questions or assistance. Still tracked but less detailed reporting up to management or DOE
- We will convert INC to RITM when appropriate.
- Your email reply becomes part of the worklog of the ticket number in the subject

- Emailing lqcd-admin@fnal.gov is the best way to reach us and will elicit the quickest response. Please do not email any of our staff directly.
- Response time as well as time to resolve are tracked and reported
- Our goal is to resolve your issue within 3 business days.

https://computing.fnal.gov/lqcd/support/

Campus and Onsite Access Request Form

- Fermilab is starting to re-open for public access with various restrictions
- DOE has relaxed the maximum telecommute status, but on-site access is still limited with new controls in place
- Has required many changes to the staff and visitor authorization process

- Computing and Onsite Access Request form (for new and renewing Affiliates, Users)
  - Reason for access request – Choose "**Affiliate / Users**"
  - If you don't need on-site access, choose "**remote/logical ac**cess" (most of you)
  - If you will need on-site access, choose "**onsite with computing accounts access**"

We continue to update web pages

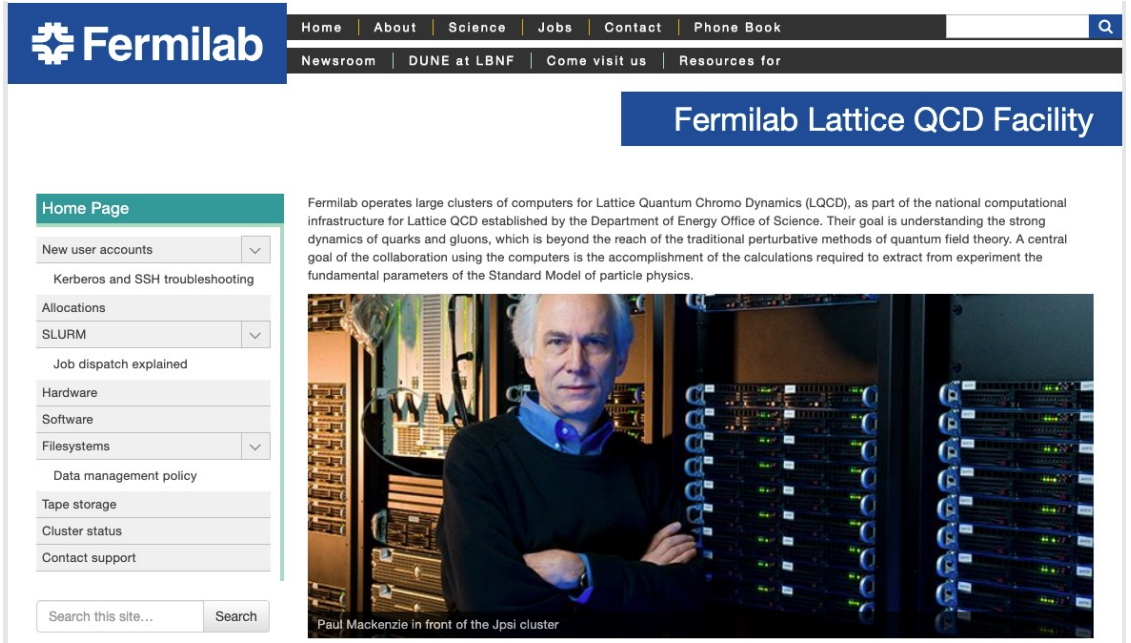
Our web address (WordPress):
https://computing.fnal.gov/lqcd/

- Brings the LQCD pages in line with Fermilab style guidelines
- Also easier for us to update and improve

It is our hope that you can find answers to all your questions on these pages.

We appreciate your feedback so that we can make these pages useful and easy to navigate.

Try the “search this site” function



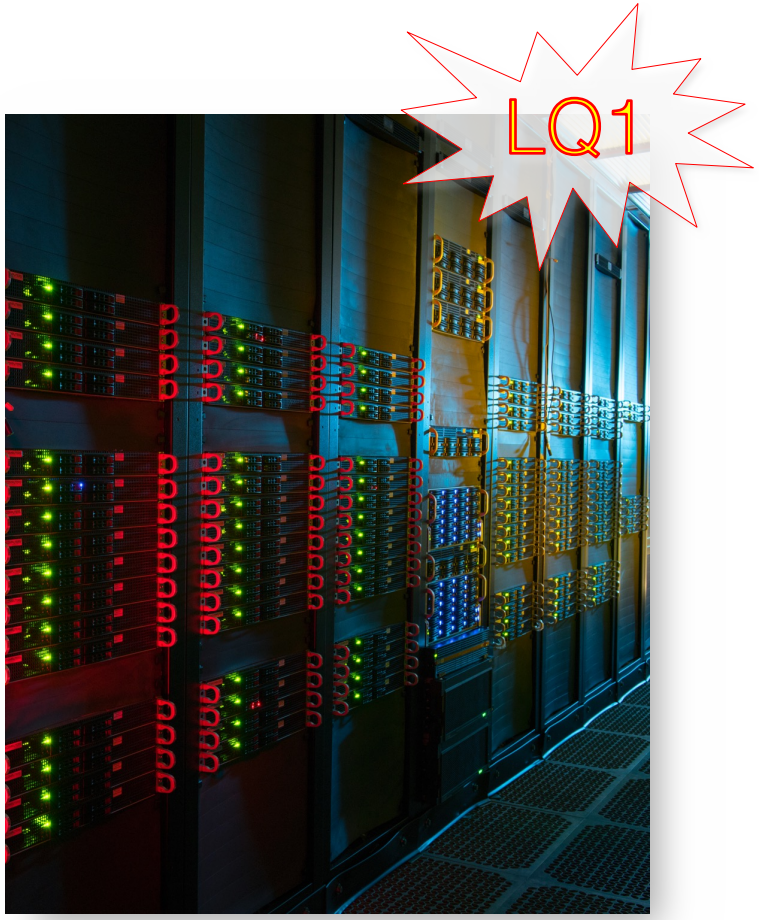
Fermilab's LQ1 Institutional cluster

- 181 nodes, 7,240 cores.
- Intel 6248 "*Cascade Lake*", 2:1 oversubscribed EDR Omni Path.
- 325 GFlops/node - - 59 TFlops total **

- Our /lustre1 filesystem is now 820 TB total. Current quotas add up to 742 TB
- Also watching quotas on /project (24 TB) and /home (1 TB) areas.

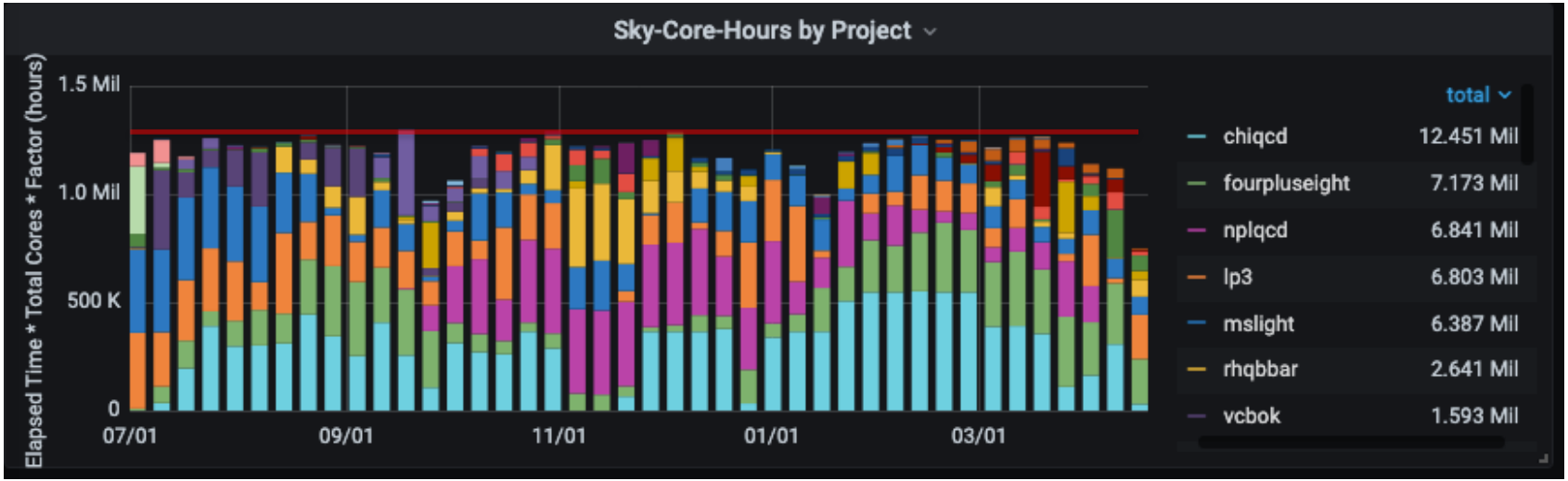
- Current Type A Allocations total 55.4 M sch (sky-core-hours)
- All Allocation Types total 57.94 M sch
- Delivered 2.9M Opportunistic sch so far, this PY

** not applying Amdahl's law.183 nodes,



Fermilab

All usage of the LQ1 cluster since July, 1, 2021



The few dips in fully loading the cluster came when there were not enough jobs in queue to keep the cluster fully loaded. We made temporary adjustments to max nodes per project to keep the entire cluster busy. I will continue to make this sort of adjustment.

Progress against allocations as of April 19, 2022

Project Name	Cluster	SPC Original Allocation (Sky-Core-Hours) ▲	Adjustments (Sky-Core-Hours)	SPC Adjusted Allocation (Sky-Core-Hours)	Project Used as of Jul 1, 2021 (Sky-Core-Hours)	Progress against Adjusted Allocation
lp3	FNAL-LQ1	14,000,000	-2,221,216	11,778,784	6,797,631	58%
chiqcd	FNAL-LQ1	9,000,000	3,617,595	12,617,595	12,453,510	99%
mslight	FNAL-LQ1	8,500,000	421,242	8,921,242	6,423,137	72%
fourpluseight	FNAL-LQ1	6,000,000	1,172,139	7,172,139	7,235,590	101%
nplqcd	FNAL-LQ1	4,500,000	2,328,054	6,828,054	6,840,582	100%
hadstruc	FNAL-LQ1	4,000,000	-3,814,060	185,940	218	0%
4fermi	FNAL-LQ1	3,500,000	-1,523,525	1,976,475	974,959	49%
rhqbbar	FNAL-LQ1	2,700,000	597,801	3,297,801	2,767,941	84%
ahisq	FNAL-LQ1	2,000,000	-298,458	1,701,542	1,162,534	68%
heavylight	FNAL-LQ1	1,200,000	-279,572	920,428	655,126	71%

Not shown: the hours used by 5 Type B (1.9M), 3 Type C (26K), and 7 opportunistic (3.2M) projects

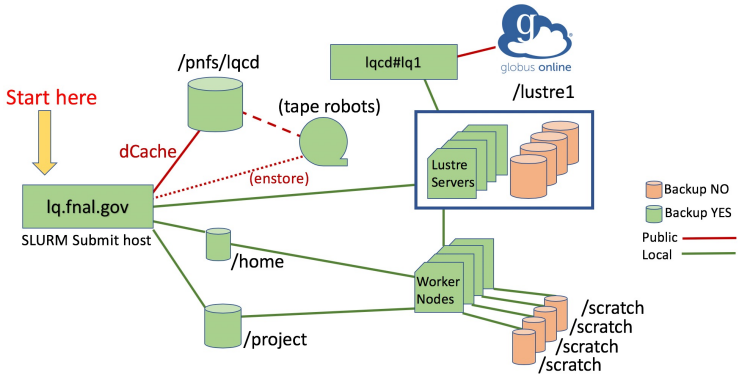
TOTAL	**FNAL-LQ1**	**57,795,000**	-	**57,795,000**	**50,441,702**	**81.5%**	**10,676,683**

These numbers from April 20 are at 80% through the program year

https://www.usqcd.org/fnal/clusterstatus/lq1/accounting.html

Storage (disk and tape)

- Please use **`lqio.fnal.gov`** for large IO transfers to Lustre or tape.
  - Has 100GigE interface to offsite networks.
- Globus endpoint lqcd#lq1 for /lustre1.
- For best job performance, use /scratch (890 GB)



Area	Description	Area	Description
/home	Home area. Visible on all cluster worker nodes via NFS. NOT suitable for configurations or propagators. SUITABLE as a "run" directory for light production or testing. Quota of 5 GB per user. Backups nightly.	/lustre1	Lustre storage. Visible on all cluster worker nodes. NOT suitable for large number of small files. SUITABLE for temporary storage (~month) of very large data files. Disk space usage monitored, and disk quotas enforced. NO backups.
/project	Space is allocated for all approved projects. Visible on all cluster worker nodes via NFS file-system. NOT suitable for fields e.g. configurations, quark propagators. SUITABLE for output logs, meson correlators and other small data files. Backups nightly.	/pnfs/lqcd	Tape storage. Visible on login head nodes only. NOT suitable for large number of small files, compression highly recommended. SUITABLE for permanent storage of parameter files and results. Must use special copy command: '`dccp`'

https://computing.fnal.gov/lqcd/filesystems/

Fermilab

Quotas for allocated storage in Lustre and /project

- We track /lustre1 usage and quota based on group ownership
- Periodically, we run scripts to correct group ownerships and sticky bits

- Since Lustre is designed for large file access, we have implemented limits on storage space and number of files

Low (>)	High (<=)	Number of files
0	5 TB	1.0 M
5 TB	50 TB	2.5 M
50 TB	100 TB	5.0 M
100 TB	150TB	7.5 M
150 TB		10.0 M

- We track /project area usage and quota based on ZFS project ids
- Since the number of files in ZFS is not an issue, there are no limits

https://www.usqcd.org/fnal/clusterstatus/lq1/diskusage.html

Data Management Planning

- Each project is responsible for their own data management plan.
- You need to keep copies of critical data somewhere safe. Lustre is not backed up.

- A project's storage allocation is tied to the program year. Think of it like a lease on an apartment.
  - At the end of the year, you either renew your lease, negotiate to a larger or smaller space, or you vacate and move to another location.
  - Clear out promptly please. We need to provide the space to the next "tenant".

- Short term tape storage will be separate tapes by program year vs Long term.
- We are working to provide new tape usage reporting.

https://computing.fnal.gov/lqcd/fermilab-data-management-guidelines-policies/

SLURM Limits and Dispatch Priority

- Maximum number of nodes per job and nodes per account is 64.
- Maximum number of jobs submitted per account is 125.

QoS Name	QoS Description	Priority	Group Resource Limit	Maximum Wall-time	Maximum Jobs Per User	Maximum Job Submits Per Account
admin	Reserved for Administrators	600				
test	Quick turnaround testing	500	cpu=80 (node=2)	00:30:00	1	3
normal	Normal QoS (default)	250		24:00:00		125
opp	Opportunistic	10		08:00:00		125

Job's requesting shorter wall-time limits are more likely to start sooner as Backfill jobs.

https://computing.fnal.gov/lqcd/slurm/ and https://computing.fnal.gov/lqcd/job-dispatch-explained/

Scheduled downtimes

- We plan for two downtimes per year for Kernel updates and other Security patches.
  - Early July during transition to new Allocations
  - Early January as the 6-month mark.

- When we need to do other scheduled maintenance, it will typically be on the third Wednesday of the month.

- We will continue to use the lqcd-users@fnal.gov mailing list to announce both scheduled downtimes and unscheduled outages.

- The scheduled maintenance window for Enstore and dCache is the third Wednesday of each month, beginning at 8am, normally 4-8 hours.

Questions ?

I know I just asked you to use the lqcd-admin@fnal.gov email address, but if you have questions about this presentation, you are welcome to contact me directly:

Ken Schumacher <kschu@fnal.gov>

The first baby bison of 2022 arrived Wednesday (April 13).

USQCD Fermilab Cluster Accounting (as of 2022-04-20)

Project Name	Cluster	SPC Original Allocation (Sky-Core-Hours) ▲	Adjustments (Sky-Core-Hours)	SPC Adjusted Allocation (Sky-Core-Hours)	Project Used as of Jul 1, 2021 (Sky-Core-Hours)	Progress against Adjusted Allocation	Remaining Allocation (Sky-Core-Hours)	30-day usage as of 04/20/2022	30-day burn rate as of 04/20/2022 (% of Alloc.)	Annual Pace YYYY-MM-DD
lp3	FNAL-LQ1	14,000,000	-2,221,216	11,778,784	6,797,631	58%	4,981,153	516,726	4.4%	2022-11-20
chiqcd	FNAL-LQ1	9,000,000	3,617,595	12,617,595	12,453,510	99%	164,085	725,199	5.7%	2022-04-23
mslight	FNAL-LQ1	8,500,000	421,242	8,921,242	6,423,137	72%	2,498,105	436,196	4.9%	2022-08-11
fourpluseight	FNAL-LQ1	6,000,000	1,172,139	7,172,139	7,235,590	101%	0	1,236,925	17.2%	2022-04-17
nplqcd	FNAL-LQ1	4,500,000	2,328,054	6,828,054	6,840,582	100%	0	521,763	7.6%	2022-04-19
hadstruc	FNAL-LQ1	4,000,000	-3,814,060	185,940	218	0%	185,722	-	0.0%	2705-09-24
4fermi	FNAL-LQ1	3,500,000	-1,523,525	1,976,475	974,959	49%	1,001,516	3,216	0.2%	2023-02-14
rhqbbar	FNAL-LQ1	2,700,000	597,801	3,297,801	2,767,941	84%	529,860	250,384	7.6%	2022-06-15
ahisq	FNAL-LQ1	2,000,000	-298,458	1,701,542	1,162,534	68%	539,008	345,823	20.3%	2022-09-02
heavylight	FNAL-LQ1	1,200,000	-279,572	920,428	655,126	71%	265,302	142,856	15.5%	2022-08-16
comphiggs	FNAL-LQ1	510,000	-	510,000	449,987	88%	60,013	103,073	20.2%	2022-05-29
qfenpct	FNAL-LQ1	500,000	-	500,000	174,013	35%	325,987	3,739	0.7%	2023-10-20
gluonpdf	FNAL-LQ1	500,000	-	500,000	570,143	114%	0	189,804	38.0%	2022-03-14
hadtensor	FNAL-LQ1	450,000	-	450,000	379,913	84%	70,087	4,368	1.0%	2022-06-13
lgncqcd	FNAL-LQ1	375,000	-	375,000	353,084	94%	21,916	134,640	35.9%	2022-05-08
vacuumdecay	FNAL-LQ1	20,000	-	20,000	5,535	28%	14,465	-	0.0%	2024-05-24
largenc	FNAL-LQ1	20,000	-	20,000	9,091	45%	10,909	-	0.0%	2023-04-06
hisqvec	FNAL-LQ1	20,000	-	20,000	11,445	57%	8,555	1,025	5.1%	2022-11-25
vcbok	FNAL-LQ1	-	-	-	1,593,347	-	-	-	-	-
safe	FNAL-LQ1	-	-	-	1,040,701	-	-	387,884	-	-
nptmd	FNAL-LQ1	-	-	-	4,072	-	-	480	-	-
nme	FNAL-LQ1	-	-	-	334,570	-	-	-	-	-
nedm	FNAL-LQ1	-	-	-	173,393	-	-	-	-	-
lqcdadmin	FNAL-LQ1	-	-	-	31,052	-	-	-	-	-
lattsusy	FNAL-LQ1	-	-	-	128	-	-	-	-	-
TOTAL	**FNAL-LQ1**	**57,795,000**	-	**57,795,000**	**50,441,702**	**81.5%**	**10,676,683**	**5,004,101**	-	-