Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

BNL Scientific Data and Computing Center (SDCC) Site Report

Zhihua Dong
On behalf of SDCC, BNL

USQCD ALL Hands Meeting 2023 - 4/20/2023

@BrookhavenLab

SDCC: The Scientific Data and Computing Center

- Located at Brookhaven National Laboratory (BNL) on Long Island, New York
- SDCC was initially formed at BNL in the mid-1990s as the RHIC Computing Facility

STAR

ATLAS EXPERIMENT

Center for Functional Nanomaterials
Brookhaven National Laboratory

RBRC
RIKEN BNL Research

ARM
CLIMATE RESEARCH FACILITY

Belle II

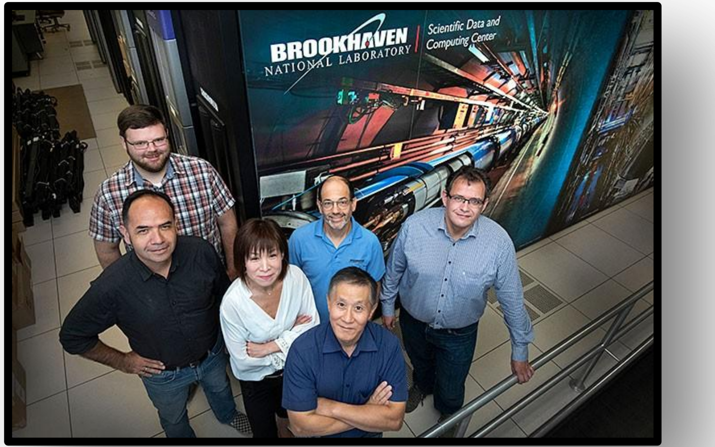
DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT

NNDC National Nuclear Data Center

USQCD US Lattice Quantum Chromodynamics

sPHENIX

PHENIX


BROOKHAVEN NATIONAL LABORATORY
Scientific Data and Computing Center

Brookhaven National Laboratory

Shared multi-program facility serving ~2,000 users from more than 20 projects

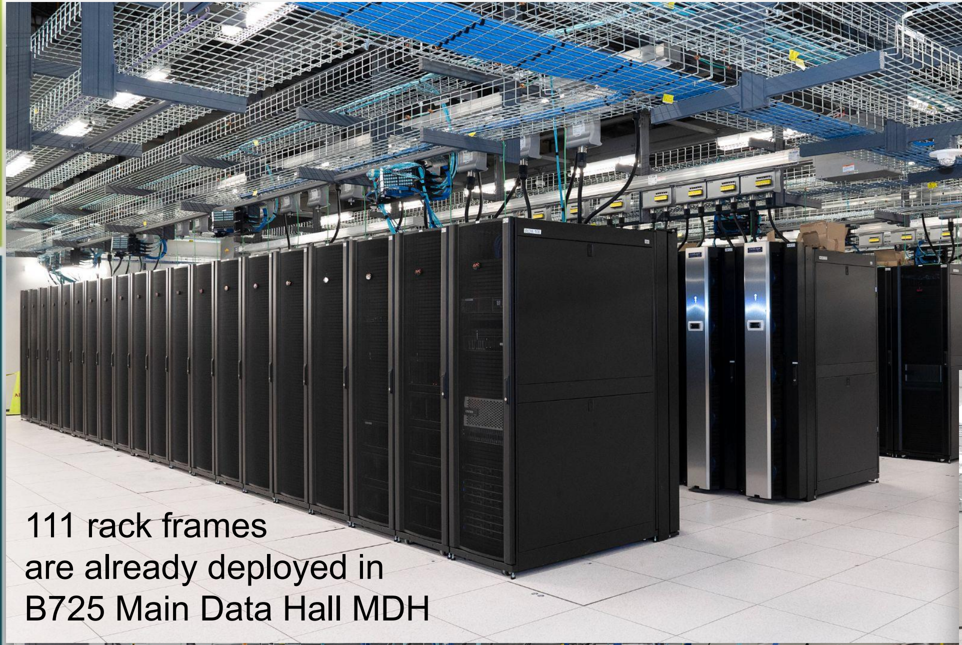
Scientific Data and Computing Center Overview

- Tier-0 computing center for the RHIC experiments
- US Tier-1 Computing facility for the ATLAS experiment at the LHC, also one of the ATLAS shared analysis (Tier-3) facilities in the US
- Computing facility for NSLS-II
- US Data center for Belle II experiment
- Providing computing and storage for proto-DUNE/DUNE along w/ FNAL serving data to all DUNE OSG sites
- Also providing computing resources for various smaller / R&D experiments .
- Serving more than 2,000 users from > 20 projects
- Developing and providing administrative/collaborative tools:
  - Invenio, Jupyter, BNL Box, Discourse, Gitea, Mattermost, etc.
- BNL was selected as the site for the upcoming major new facility Electron-Ion Collider (EIC/eRHIC)
- sPHENIX - scheduled to start taking data in May

BNL Core Facility Revitalization (CFR) Project: New Data Center

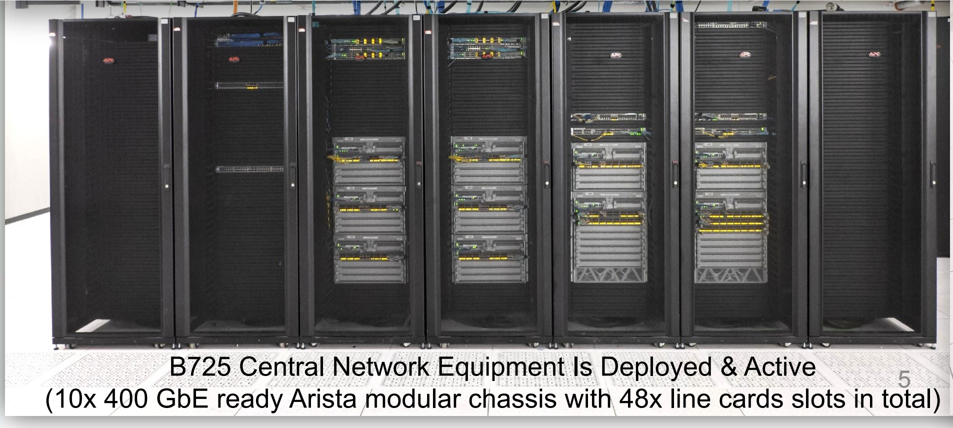
New Data Center (Building 725) — 2023Q1: 1.5 Years of Production Operations

- CFR project finished the design phase in the first half of 2019 and completed the construction phase by the end of FY21
- The occupancy of the B725 data center for production CPU and DISK resources for all programs started in 2021Q4 and ramped up in 2022Q1-2023Q1 to the level of 62 racks populated with equipment in the B725 Main Data Hall (MDH)
  - 10 more storage / infrastructure racks are in the process of being configured as of 2023Q1
  - 20 more new HTC CPU racks and 2 more HPC rack are expected to be added to B725 in 2023Q2
- Currently we have two diesel generators installed in B725 diesel generator yard providing covering up to 1.2 MW of total IT payload with N+1 redundancy
  - Two more diesel generators are planned to be added in FY24-25 to provide all IT payload in the B725 data center as it scales beyond 1.2 MW and 2.4 MW thresholds for combined IT payload deployed
- Two library rows in B725 Tape Room are populated with IBM TS4500 tape libraries to serve ATLAS and sPHENIX experiments (4 libraries, 128 tape drives in total).
  - One more library row is expected to be populated in FY24 (2 more IBM TS4500 sPHENIX libraries)
- The completion of the transition of the majority of CPU and DISK resources deployed in SDCC environment to the new B725 datacenter is still expected to be achieved by the end of FY23
  - The vast majority of equipment purchased by SDCC starting from 2021Q3 is being placed in the new data center in preparation for the retirement of the oldest areas of B515 datacenter by Sep 30, 2023



111 rack frames
are already deployed in
B725 Main Data Hall MDH

84 RDHx units deployed in B725 MDH, out of which 59 are on racks with equipment while 25 are deployed for the future growth



B725 Central Network Equipment Is Deployed & Active
(10x 400 GbE ready Arista modular chassis with 48x line cards slots in total)

4x 8-frame IBM TS4500
tape libraries are installed
in the B725 Tape Room

High Throughput Computing

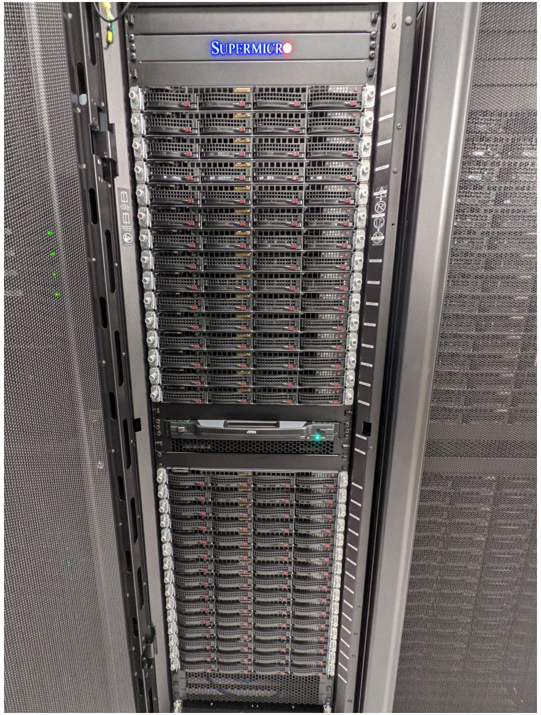
- Providing our users with ~1,900 HTC nodes:
  - ~90,000 logical cores
  - ~1050 kHS06
  - Managed by HTCondor

- Purchased 648 Supermicro SYS-610C-TR nodes for ATLAS and the RHIC experiments (~62k logical cores total)
  - 120 nodes just delivered , rest expect in May
  - Housed in 20 racks
  - System specs:
    - Dual Intel Ice Lake Xeon Gold 6336Y 24-core processors
    - 12x32 GB 3200 MHz ECC DDR4 RAM (384 GB total)
    - 4x2 TB SSD drives
    - 1U form factor
    - 10 Gbps NIC

- Will be purchasing some Supermicro ARM test nodes in April
  - With Ampere Altra CPUs

- All nodes running still running Scientific Linux (SL) 7
  - Preparations for an OS upgrade to Alma Linux in progress

- HTCondor 10.0 fully tested, and a rolling upgrade has begun



Supermicro SYS-6019U-TR4 Servers

High Performance Computing

Currently supporting 5 HPC clusters

- **Institutional Cluster gen1 (IC)   (Retire in Fall)**
  - 216 HP XL190r Gen9 nodes with EDR IB
  - 108 nodes with 2x Nvidia K80
  - 108 nodes with 2x Nvidia P100
- **Skylake Cluster   (Retire in Fall)**
  - 64 Dell PowerEdge R640 nodes with EDR IB
- **KNL Cluster (Retire in Fall)**
  - 142 KOI S7200AP nodes with dual rail Omnipath Gen.1 interconnect
- **ML Cluster**
  - 5 HP XL270d Gen10 nodes with EDR IB
  - Each node has 8x Nvidia V100
- **NSLS2 Cluster**
  - 32 Supermicro nodes with EDR IB
  - 13 nodes with 2x Nvidia V100

New IC Gen2 Cluster

Institutional Cluster gen2 (IC) phase 1

36 CPU only nodes + 12 nodes with GPU for CSI, CFN, delivered recently, testing now,
4 LQCD CPU only nodes on order
Specs: 2x Intel Xeon (Ice Lake)

- 512GB DDR4-3200 on CPU nodes
- 1TB DDR4-3200 on GPU nodes
- NDR200 InfiniBand interconnect (200Gbps per uplink)
- 4x Nvidia A100 80GB on GPU nodes

HPL performance from current IC gpu node: ~7.9 TF to ~60TF IC gen2 GPU node. More than **7x**

Brookhaven National Laboratory

Storage:

Disk

- **dCache:** ~74 PB
- **XROOTD:** ~11 PB
- **Lustre:** ~50 PB
- **GPFS:** ~9 PB + 3.4PB(raw) for IC Gen2 to be deployed soon
- **Home hosted S3 Storage for EIC : (now in house)**
  use native Object Storage(CEPH) & Federated ID access

Tape

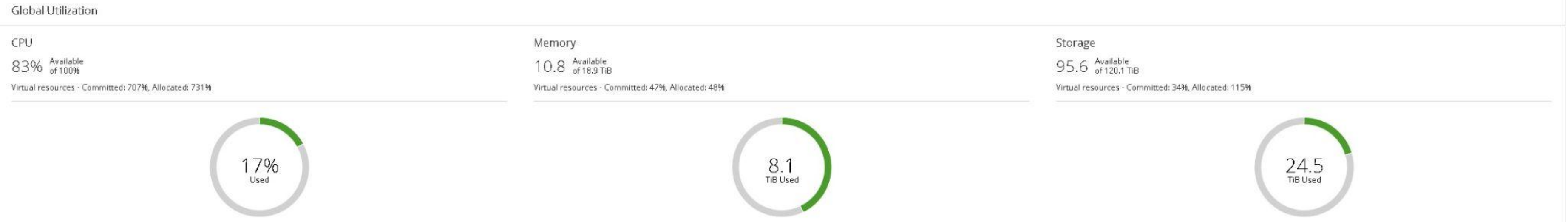
- Currently ~220 PB of data in HPSS with ~75k tapes
- 9 Oracle SL8500 and CSI TS4500 in old building.
- Four 8 frame TS4500 tape libraries and data movers

  2 for ATLAS (LTO8)

  2 for sPHENIX (LTO9)

Redhat Virtualization

- Redhat Virtualization is end of life in 2024.
- Evaluating Openshift Virtualization and VMware.
    - Leaning towards VMware due to product maturity, features and pricing.
- Moving 600 VMs from RedHat Virtualization to VMware or Openshift will take time.
- Since RHEL7 is also EOL in 2024, we can rebuild RHEL7 VMs as RHEL8 VMs in the new virtualization platform.



Global Utilization

CPU

83% Available of 100%

Virtual resources - Committed: 707%, Allocated: 731%

17% Used

Memory

10.8 Available of 18.9 TiB

Virtual resources - Committed: 47%, Allocated: 48%

8.1 TiB Used

Storage

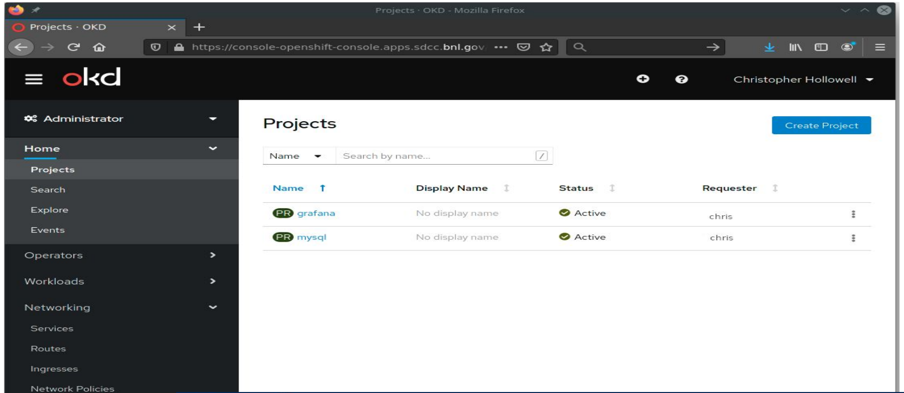
95.6 Available of 120.1 TiB

Virtual resources - Committed: 34%, Allocated: 115%

24.5 TiB Used

okd

- **Two OKD clusters (sPhenix and ATLAS) In production at SDCC**
  - Each have 7 nodes



Projects · OKD - Mozilla Firefox

Projects · OKD

https://console-openshift-console.apps.sdcc.bnl.gov

okd

Christopher Hollowell

Administrator

Home
- Projects
- Search
- Explore
- Events

Operators

Workloads

Networking
- Services
- Routes
- Ingresses
- Network Policies

Storage

Builds

User Management

Administration

Projects

Create Project

Name | Search by name...

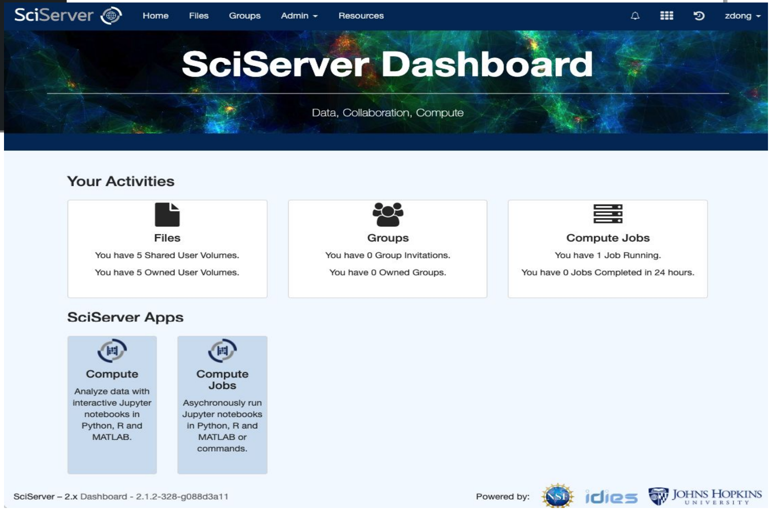
Name	Display Name	Status	Requester
PR grafana	No display name	Active	chris
PR mysql	No display name	Active	chris

SciServer

The Science Platform

A collaborative environment for server-side analysis with large datasets

- Test instance: available for test users
- For production : 4 Compute nodes in purchase: Each with 8xH100 SXM GPUs



SciServer

Home Files Groups Admin Resources zdong

SciServer Dashboard

Data, Collaboration, Compute

Your Activities

Files
You have 5 Shared User Volumes.
You have 5 Owned User Volumes.

Groups
You have 0 Group Invitations.
You have 0 Owned Groups.

Compute Jobs
You have 1 Job Running.
You have 0 Jobs Completed in 24 hours.

SciServer Apps

Compute
Analyze data with interactive Jupyter notebooks in Python, R and MATLAB.

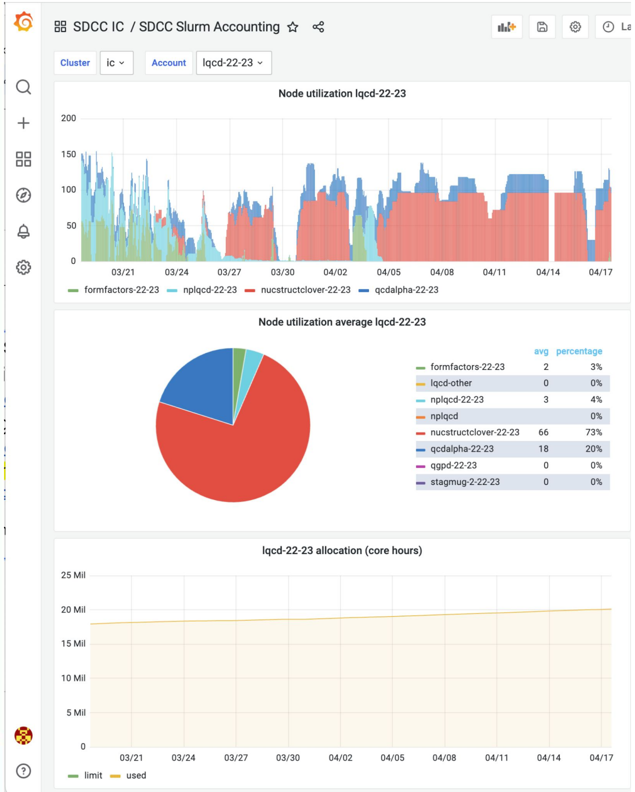
Compute Jobs
Asychronously run Jupyter notebooks in Python, R and MATLAB or commands.

SciServer – 2.x Dashboard - 2.1.2-328-g088d3a11

Powered by: NSF idies JOHNS HOPKINS UNIVERSITY

Brookhaven National Laboratory

Monitoring (some need authentication)

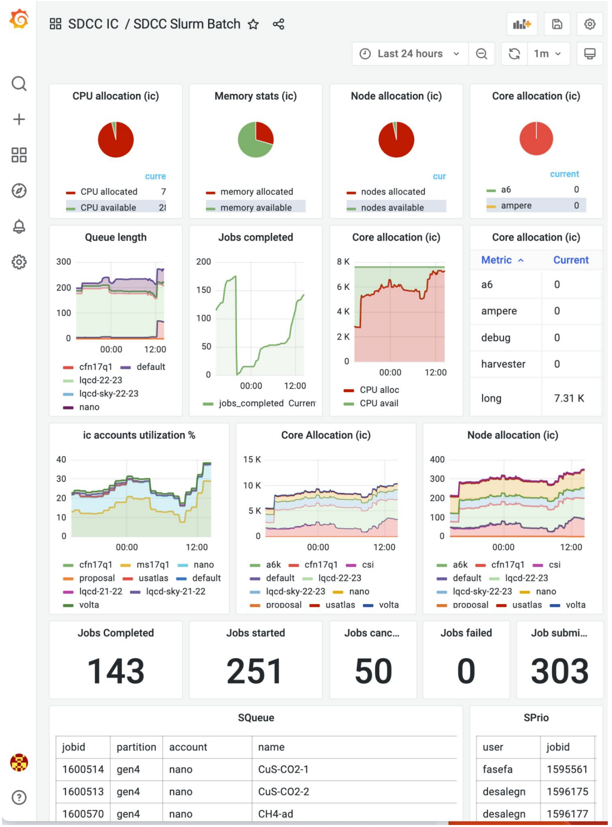


SDCC IC / SDCC Slurm Accounting

Cluster: ic · Account: lqcd-22-23

	avg	percentage
formfactors-22-23	2	3%
lqcd-other	0	0%
nplqcd-22-23	3	4%
nplqcd		0%
nucstructclover-22-23	66	73%
qcdalpha-22-23	18	20%
qgpd-22-23	0	0%
stagmug-2-22-23	0	0%

Brookhaven National Laboratory



SDCC IC / SDCC Slurm Batch

Last 24 hours · 1m

Core allocation (ic)

Metric ^	Current
a6	0
ampere	0
debug	0
harvester	0
long	7.31 K

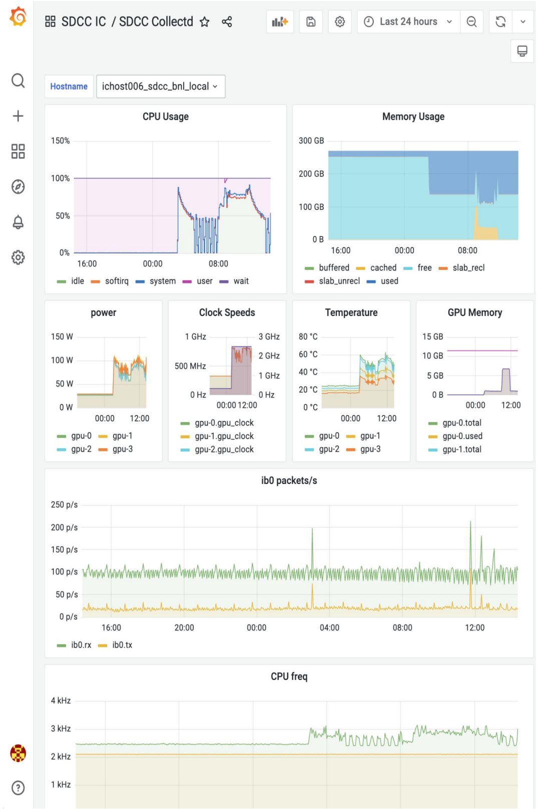
Jobs Completed	Jobs started	Jobs canc...	Jobs failed	Job submi...
143	251	50	0	303

SQueue

jobid	partition	account	name
1600514	gen4	nano	CuS-CO2-1
1600513	gen4	nano	CuS-CO2-2
1600570	gen4	nano	CH4-ad

SPrio

user	jobid
fasefa	1595561
desalegn	1596175
desalegn	1596177

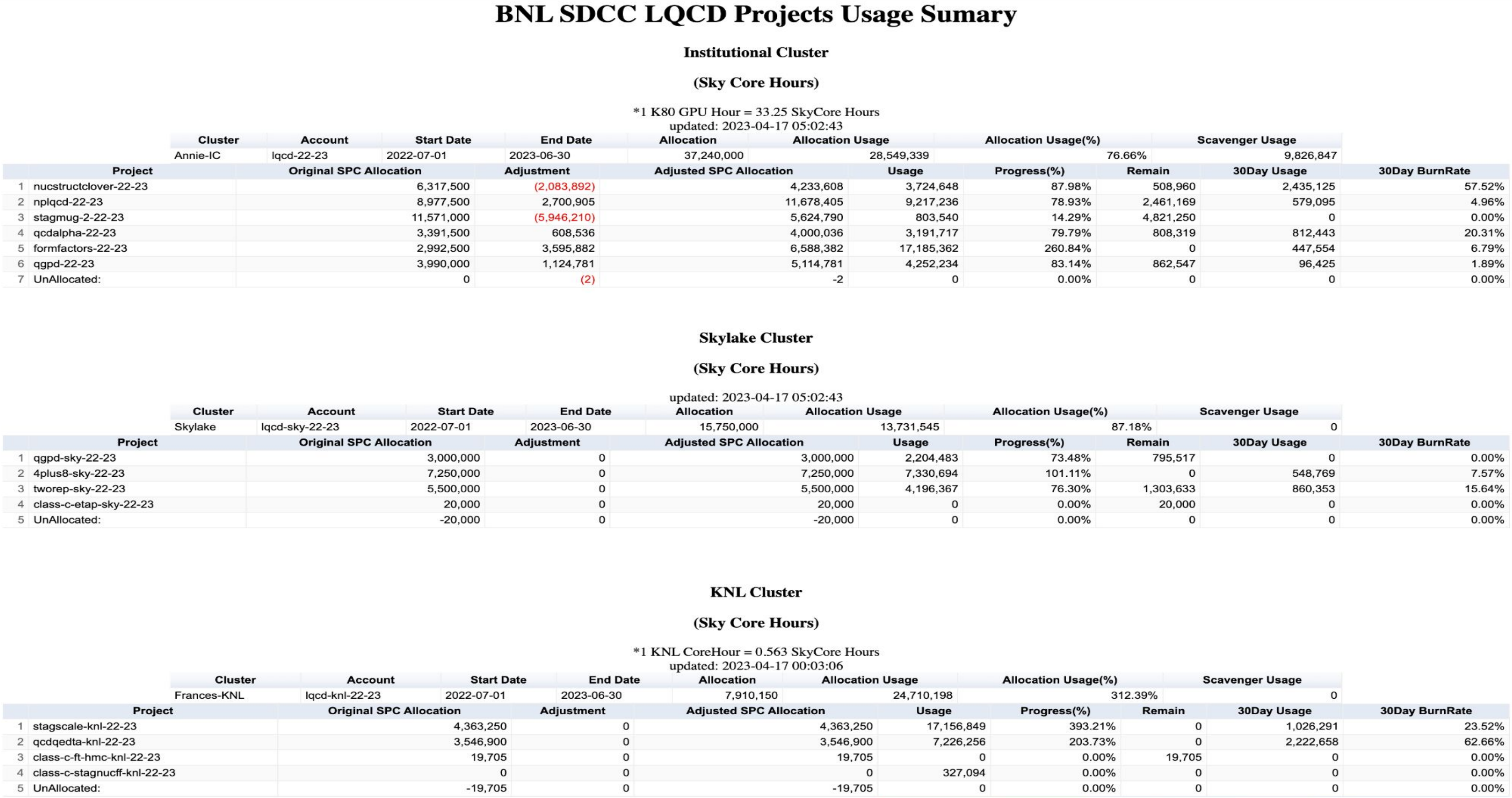


SDCC IC / SDCC Collectd

Last 24 hours

Hostname: ichost006_sdcc_bnl_local

Allocation Usage for LQCD projects



BNL SDCC LQCD Projects Usage Sumary

Institutional Cluster

(Sky Core Hours)

*1 K80 GPU Hour = 33.25 SkyCore Hours
updated: 2023-04-17 05:02:43

Cluster	Account	Start Date	End Date	Allocation	Allocation Usage	Allocation Usage(%)	Scavenger Usage
Annie-IC	lqcd-22-23	2022-07-01	2023-06-30	37,240,000	28,549,339	76.66%	9,826,847

	Project	Original SPC Allocation	Adjustment	Adjusted SPC Allocation	Usage	Progress(%)	Remain	30Day Usage	30Day BurnRate
1	nucstructclover-22-23	6,317,500	(2,083,892)	4,233,608	3,724,648	87.98%	508,960	2,435,125	57.52%
2	nplqcd-22-23	8,977,500	2,700,905	11,678,405	9,217,236	78.93%	2,461,169	579,095	4.96%
3	stagmug-2-22-23	11,571,000	(5,946,210)	5,624,790	803,540	14.29%	4,821,250	0	0.00%
4	qcdalpha-22-23	3,391,500	608,536	4,000,036	3,191,717	79.79%	808,319	812,443	20.31%
5	formfactors-22-23	2,992,500	3,595,882	6,588,382	17,185,362	260.84%	0	447,554	6.79%
6	qgpd-22-23	3,990,000	1,124,781	5,114,781	4,252,234	83.14%	862,547	96,425	1.89%
7	UnAllocated:	0	(2)	-2	0	0.00%	0	0	0.00%

Skylake Cluster

(Sky Core Hours)

updated: 2023-04-17 05:02:43

Cluster	Account	Start Date	End Date	Allocation	Allocation Usage	Allocation Usage(%)	Scavenger Usage
Skylake	lqcd-sky-22-23	2022-07-01	2023-06-30	15,750,000	13,731,545	87.18%	0

	Project	Original SPC Allocation	Adjustment	Adjusted SPC Allocation	Usage	Progress(%)	Remain	30Day Usage	30Day BurnRate
1	qgpd-sky-22-23	3,000,000	0	3,000,000	2,204,483	73.48%	795,517	0	0.00%
2	4plus8-sky-22-23	7,250,000	0	7,250,000	7,330,694	101.11%	0	548,769	7.57%
3	tworep-sky-22-23	5,500,000	0	5,500,000	4,196,367	76.30%	1,303,633	860,353	15.64%
4	class-c-etap-sky-22-23	20,000	0	20,000	0	0.00%	20,000	0	0.00%
5	UnAllocated:	-20,000	0	-20,000	0	0.00%	0	0	0.00%

KNL Cluster

(Sky Core Hours)

*1 KNL CoreHour = 0.563 SkyCore Hours
updated: 2023-04-17 00:03:06

Cluster	Account	Start Date	End Date	Allocation	Allocation Usage	Allocation Usage(%)	Scavenger Usage
Frances-KNL	lqcd-knl-22-23	2022-07-01	2023-06-30	7,910,150	24,710,198	312.39%	0

	Project	Original SPC Allocation	Adjustment	Adjusted SPC Allocation	Usage	Progress(%)	Remain	30Day Usage	30Day BurnRate
1	stagscale-knl-22-23	4,363,250	0	4,363,250	17,156,849	393.21%	0	1,026,291	23.52%
2	qcdqedta-knl-22-23	3,546,900	0	3,546,900	7,226,256	203.73%	0	2,222,658	62.66%
3	class-c-ft-hmc-knl-22-23	19,705	0	19,705	0	0.00%	19,705	0	0.00%
4	class-c-stagnucff-knl-22-23	0	0	0	327,094	0.00%	0	0	0.00%
5	UnAllocated:	-19,705	0	-19,705	0	0.00%	0	0	0.00%

Brookhaven™
National Laboratory

USQCD Access to SDCC Resources

- Current resources allocated 3 Clusters (7/1/2022-6/30/2023)
  - 657k node-hour allocation on CPU-GPU cluster used ~75% (88% if include scavenger)
  - 508k node-hour allocation on SKY cluster ~75%
  - 262k node-hour allocation on KNL cluster ~ 262%
  - 800 TB of GPFS disk storage
  - Tape Storage:
    - Total LQCD data on tape : ~4.4PB (since 1/2020)
    - Include Long Term Archive currently ~3.1 PB
- Next allocation Year , we only have 3 month for those clusters available

  Please avoid slow start which often happens during summer !

- Usage policy
  - LQCD Jeopardy Policy (penalty/reward) apply at end of each month.
  - Opportunistic lower priority and scavenger qos available for LQCD after sub-project allocation used up.

    Scavenger Usage does not count towards LQCD allocation ! Job subject to preemption.

User Support

- Facility website www.sdcc.bnl.gov .
  - New accounts
    - Instructions on website
    - Usually ~24 hours for SDCC to process after verification

      However BNL new user guest appointment could take some time. Always apply as early as possible.
  - User support requests (tickets)
    - SDCC policy is to respond within 3 business days. Majority is resolved within this period
  - !!! **Please do contact us** or submit ticket when could not find some information you need.

    You may miss the right pages from browsing SDCC web site.
- Bi-weekly meetings between facility staff and program/experimental Liaisons
  - Agenda on https://indico.bnl.gov/category/200/
  - Remote access via ZoomGov—Minutes of meeting posted for those who cannot join in person or remotely

Thanks to the following people at BNL for contributing to this presentation:

Costin Caramarcu, Tim Chou, Joe Frith, Vincent Garonne, Chris Hollowell, Jerome Lauret, Louis Pelosi, Alex Zaytsev, [...]

Questions?