Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

BNL Facility Report and Acquisition

Zhihua Dong (site manager)
Costin Caramarcu (site architect)
Chulwoo Jung (site architect)

USQCD All Hands Meeting - 4/21/2022

@BrookhavenLab

Scientific Data and Computing Center (SDCC)

SDCC was initially formed at BNL in the mid-1990s as the RHIC Computing Facility.

- 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
- Data center for US Belle II experiment
- BNL selected as the site for the upcoming major new facility Electron-ion Collider (EIC/eRHIC)
- sPHENIX - scheduled to start taking data in 2023

SDCC: The Scientific Data and Computing Center

ATLAS EXPERIMENT

Center for Functional Nanomaterials
Brookhaven National Laboratory

RBRC
RIKEN BNL Research Center

STAR ☆

SIMONS FOUNDATION

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

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

Brookhaven National Laboratory

SDCC Overview (Cont.)

- Support for various programs:
    - RHIC, LHC ATLAS, BER ARM, LSST, DUNE, LQCD,RIKEN, BES, Center for Functional Nanomaterials(CFN), National Synchrotron Light Source(NSLS) II, National Nuclear Data Center…

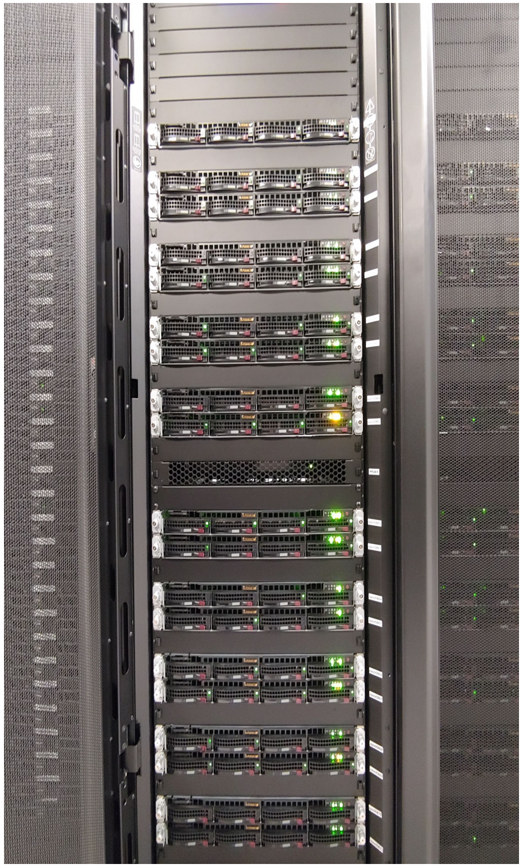
- Serving more than 2,000 users from > 20 projects

Currently we are 44 full time employees.

High Throughput Computing

Providing our users with ~1900 HTC nodes:

~90K logical cores
~1 MHS06



Latest online nodes (Feb 2022) :
210 Supermicro SYS-6019U-TR4
1U servers in 7 racks

- Dual Intel Xeon Cascade Lake 6252 CPUs @ 2.4 GHz (96 log. cores total)
- 12 x 32 GB (384 GB total) DDR4-2933 MHz RAM
- 4 x ~2 TB SSDs
- 1U form factor
- 1140 HS06/node = ~240 kHS06

All nodes running Scientific Linux (SL) 7

Have been evaluating RHEL8 and RHEL8-based distributions including Rocky Linux 8 due to SL discontinuation and CentOS 8 early EOL

Available to users via HTCondor batch system - upgraded to v9 (including token support)

Dedicated subset of per-experiment interactive VMs for login and job submission

High Performance Computing

Currently supporting 5 HPC clusters

- Allocations subject to approval

Institutional Cluster (IC)
216 HP XL190r Gen9 nodes with EDR IB
2x Intel Broadwell Xeon E5-2695v4 CPUs (36 cores total)
256 GB RAM (DDR-2400)
2x K80 or P100 GPUs

Skylake Cluster
64 Dell PowerEdge R640 nodes with EDR IB
2x Intel Skylake Xeon Gold 6150 CPUs (36 cores total)
192 GB RAM (DDR4-2666)

KNL Cluster
142 KOI S7200AP nodes with dual rail Omnipath Gen.1 interconnect
1x Intel Xeon Phi 7230 CPU (256 log. Cores total)
192 GB RAM (DDR4-1200)

ML Cluster
5 HP Proliant XL270d Gen10 nodes with EDR IB
2x Intel Xeon Gold 6248 CPUs (40 cores total)
768 GB RAM (DDR4-2933)
8x V100 GPUs NVlink

HPC cluster for NSLS2
30 1U nodes with EDR IB
2x Intel Cascade Lake Xeon Gold 6252 (48 cores total)
768 GB RAM (DDR4-2933)
12 of the hosts with 2x V100 GPUs

NSLS2 Cluster

Storage

- DISK
  - GPFS: 7 filesystems, 14PB
  - Lustre: 7PB
  - dCache: 65PB
  - XROOTD: 11PB

- HPSS:
  - ~216 PB accumulated data
  - 9 Oracle SL8500 and CSI TS4500 in old building.
  - Four new 8 frame TS4500 tape libraries and data movers

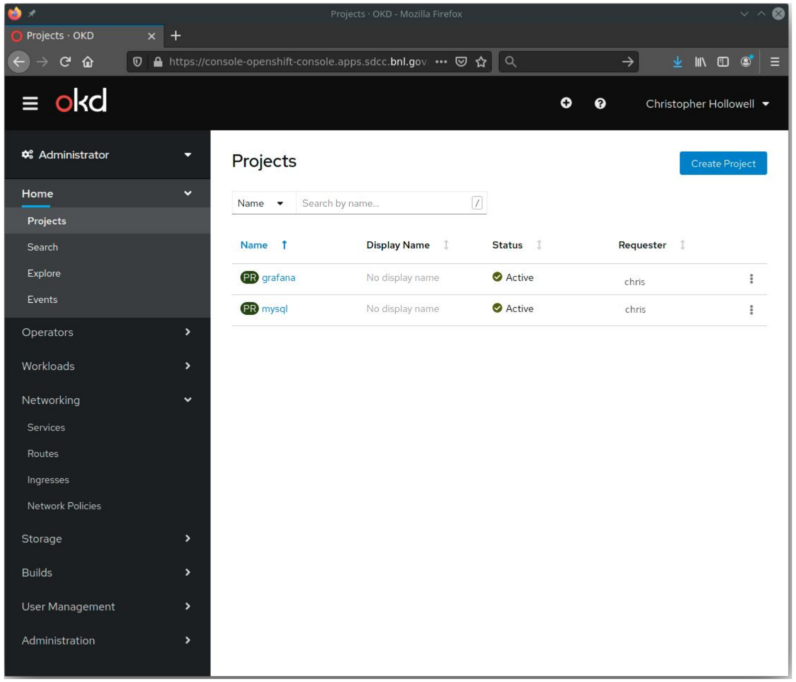
    2 in production (LTO8, ATLAS),

    2 currently in testing (LTO9, sPHENIX)

OKD @ SDCC

- **OKD provides a platform for container orchestration, similar to Kubernetes (k8s)**
  - Community-supported release of Openshift
  - Allows for simplified deployment of services via helm charts and Openshift templates
  - Contains numerous security enhancements out of the box vs k8s
    - Users are never root by default

- **Two OKD 4.10 clusters for sPhenix and ATLAS In production at SDCC**
  - Each have 7 nodes



Projects

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

Other tools and services

- Jupyter : https://jupyter.sdcc.bnl.gov

  Interface with HTC and HPC via HTCondor and Slurm

  A Federated instance for ATLAS is ready
- BNL Box : 253 users ~4TB since production 2 years ago
- MatterMost : https://chat.sdcc.bnl.gov

  Heavily used by our community and SDCC staff as major collaboration tool
- Invenio : For digital repositories
- Indico : Switching to Federated access.
- Gitea : git service , 2980 users, 307 repos
- Reana : Test Instance on staff K8s
- Discourse
- Cloud based service :

  Overleaf : Group license

  Jira license for NPP users – project and task management.

BNL Core Facility Revitalization (CFR) Project: Design & Construction of B725 Data Center

New Data Center (Building 725):

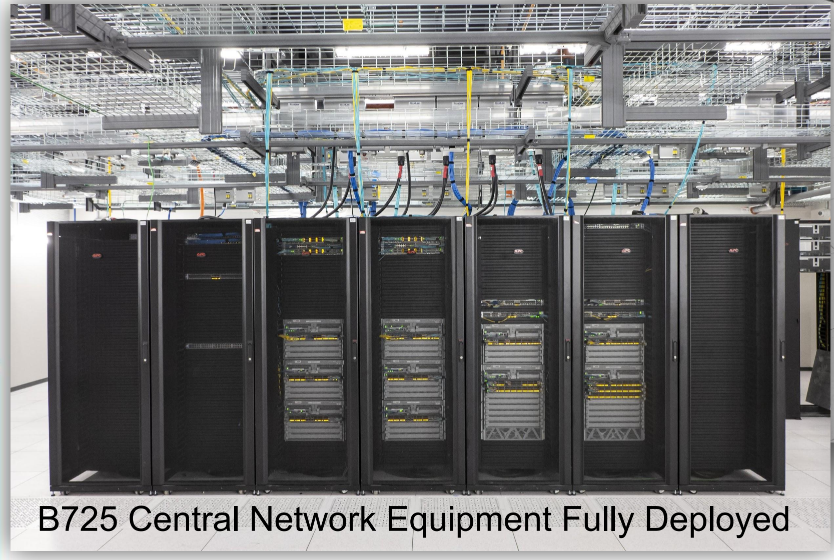
- A single large data hall for storage and compute resources (Main Data Hall):
  - Low density area (up to 20 kW per rack): 16 rows of maximum 20 racks each + one row of 16 racks.
  - High density area (up to 30 kW per rack): 14 rows of 10 racks each + one row of 8 racks.
  - Capable of supporting 478 rack in total with up to 9.6 MW of IT load combined.

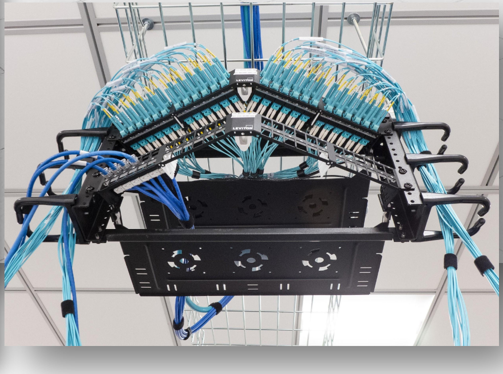
- Dedicated Network Room hosting all the central network equipment of the data center is operational since Aug 2021
- Dedicated Tape Room capable of supporting up to six 8-frame IBM TS4500 libraries

4x 8-frame IBM TS 4500 libraries installed in the B725 Tape Room

Installation of the first batch of RDHx units progressing in B725 MDH

One of the two storage / infrastructure rows deployed in B725 MDH in late FY21

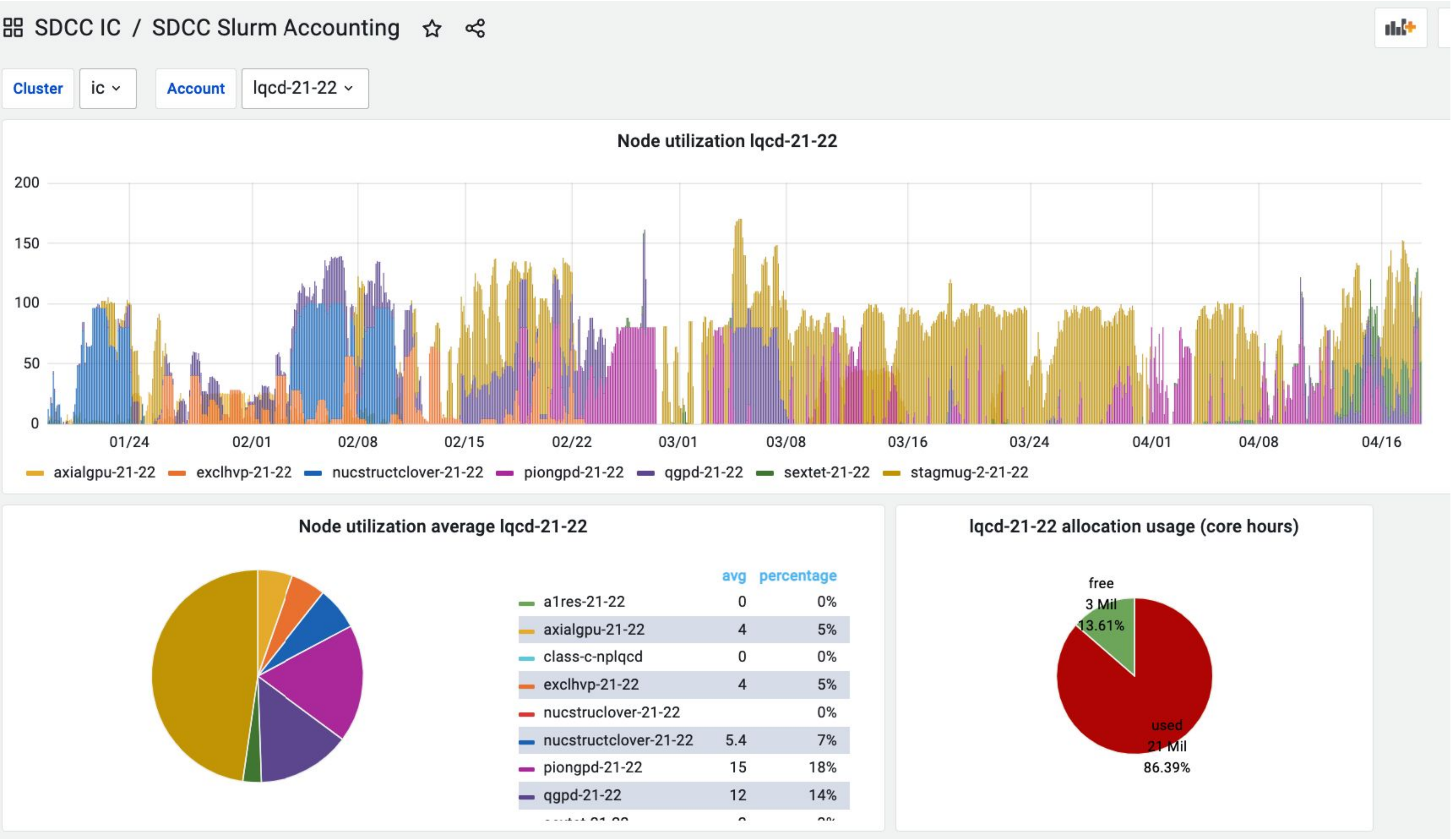

B725 Central Network Equipment Fully Deployed


Typical overhead mini-rack serving fiber & copper connectivity

Monitoring

- Several tools available
  - Graphical interface here (**authentication required**)
    - https://monitoring.sdcc.bnl.gov/pub/grafana/
  - Accounting information
    - https://monitoring.sdcc.bnl.gov/pub/allocation
    - LQCD only
      - https://monitoring.sdcc.bnl.gov/pub/allocation/lqcd.html
      - After loading module lqcd, Command line “lquota” for same information

Accounting https://monitoring.sdcc.bnl.gov/pub/grafana/



SDCC IC / SDCC Slurm Accounting

Cluster: ic

Account: lqcd-21-22

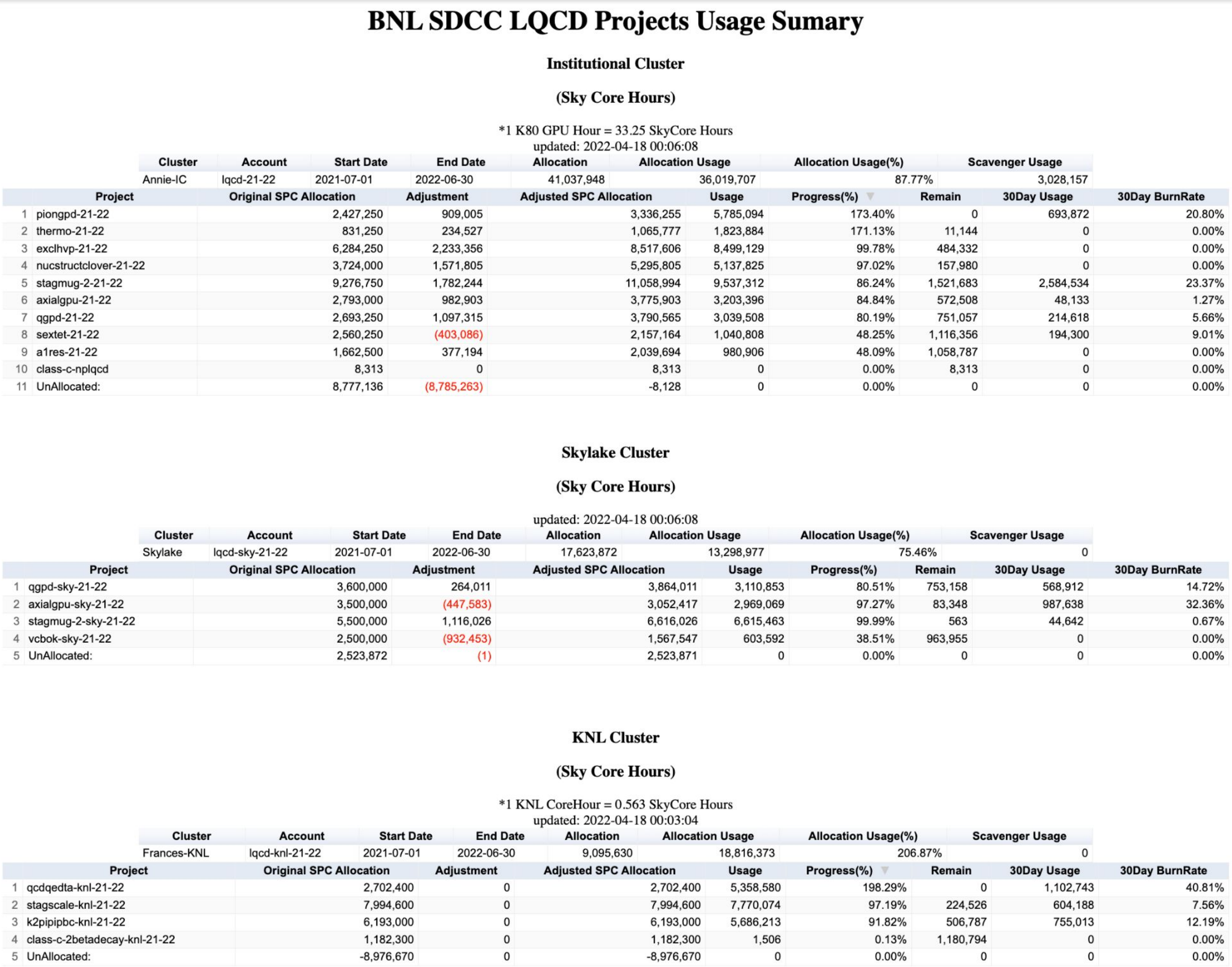
Node utilization lqcd-21-22

Node utilization average lqcd-21-22

	avg	percentage
a1res-21-22	0	0%
axialgpu-21-22	4	5%
class-c-nplqcd	0	0%
exclhvp-21-22	4	5%
nucstruclover-21-22		0%
nucstructclover-21-22	5.4	7%
piongpd-21-22	15	18%
qgpd-21-22	12	14%
sextet-21-22	0	0%

lqcd-21-22 allocation usage (core hours)

Accounting https://monitoring.sdcc.bnl.gov/pub/allocation/lqcd.html



BNL SDCC LQCD Projects Usage Sumary

Institutional Cluster

(Sky Core Hours)

*1 K80 GPU Hour = 33.25 SkyCore Hours
updated: 2022-04-18 00:06:08

Cluster	Account	Start Date	End Date	Allocation	Allocation Usage	Allocation Usage(%)	Scavenger Usage
Annie-IC	lqcd-21-22	2021-07-01	2022-06-30	41,037,948	36,019,707	87.77%	3,028,157

	Project	Original SPC Allocation	Adjustment	Adjusted SPC Allocation	Usage	Progress(%) ▼	Remain	30Day Usage	30Day BurnRate
1	piongpd-21-22	2,427,250	909,005	3,336,255	5,785,094	173.40%	0	693,872	20.80%
2	thermo-21-22	831,250	234,527	1,065,777	1,823,884	171.13%	11,144	0	0.00%
3	exclhvp-21-22	6,284,250	2,233,356	8,517,606	8,499,129	99.78%	484,332	0	0.00%
4	nucstructclover-21-22	3,724,000	1,571,805	5,295,805	5,137,825	97.02%	157,980	0	0.00%
5	stagmug-2-21-22	9,276,750	1,782,244	11,058,994	9,537,312	86.24%	1,521,683	2,584,534	23.37%
6	axialgpu-21-22	2,793,000	982,903	3,775,903	3,203,396	84.84%	572,508	48,133	1.27%
7	qgpd-21-22	2,693,250	1,097,315	3,790,565	3,039,508	80.19%	751,057	214,618	5.66%
8	sextet-21-22	2,560,250	(403,086)	2,157,164	1,040,808	48.25%	1,116,356	194,300	9.01%
9	a1res-21-22	1,662,500	377,194	2,039,694	980,906	48.09%	1,058,787	0	0.00%
10	class-c-nplqcd	8,313	0	8,313	0	0.00%	8,313	0	0.00%
11	UnAllocated:	8,777,136	(8,785,263)	-8,128	0	0.00%	0	0	0.00%

Skylake Cluster

(Sky Core Hours)

updated: 2022-04-18 00:06:08

Cluster	Account	Start Date	End Date	Allocation	Allocation Usage	Allocation Usage(%)	Scavenger Usage
Skylake	lqcd-sky-21-22	2021-07-01	2022-06-30	17,623,872	13,298,977	75.46%	0

	Project	Original SPC Allocation	Adjustment	Adjusted SPC Allocation	Usage	Progress(%)	Remain	30Day Usage	30Day BurnRate
1	qgpd-sky-21-22	3,600,000	264,011	3,864,011	3,110,853	80.51%	753,158	568,912	14.72%
2	axialgpu-sky-21-22	3,500,000	(447,583)	3,052,417	2,969,069	97.27%	83,348	987,638	32.36%
3	stagmug-2-sky-21-22	5,500,000	1,116,026	6,616,026	6,615,463	99.99%	563	44,642	0.67%
4	vcbok-sky-21-22	2,500,000	(932,453)	1,567,547	603,592	38.51%	963,955	0	0.00%
5	UnAllocated:	2,523,872	(1)	2,523,871	0	0.00%	0	0	0.00%

KNL Cluster

(Sky Core Hours)

*1 KNL CoreHour = 0.563 SkyCore Hours
updated: 2022-04-18 00:03:04

Cluster	Account	Start Date	End Date	Allocation	Allocation Usage	Allocation Usage(%)	Scavenger Usage
Frances-KNL	lqcd-knl-21-22	2021-07-01	2022-06-30	9,095,630	18,816,373	206.87%	0

	Project	Original SPC Allocation	Adjustment	Adjusted SPC Allocation	Usage	Progress(%) ▼	Remain	30Day Usage	30Day BurnRate
1	qcdqedta-knl-21-22	2,702,400	0	2,702,400	5,358,580	198.29%	0	1,102,743	40.81%
2	stagscale-knl-21-22	7,994,600	0	7,994,600	7,770,074	97.19%	224,526	604,188	7.56%
3	k2pipipbc-knl-21-22	6,193,000	0	6,193,000	5,686,213	91.82%	506,787	755,013	12.19%
4	class-c-2betadecay-knl-21-22	1,182,300	0	1,182,300	1,506	0.13%	1,180,794	0	0.00%
5	UnAllocated:	-8,976,670	0	-8,976,670	0	0.00%	0	0	0.00%

USQCD Access to SDCC Resources

- Current resources allocated 3 Clusters (7/1/2021-6/30/2022)
  - 617k node-hour allocation on CPU-GPU cluster used ~88%
  - 252k node-hour allocation on SKY cluster ~75%
  - 275k node-hour allocation on KNL cluster ~ 207%
  - 800 TB (increase from 600TB) of GPFS disk storage
  - Tape Storage:
    - BNL does not have “short” term tapes. We don’t delete tapes.
    - Total LQCD data on tape : ~3.4PB (since 1/2020)
    - Include Long Term Archive currently ~2.4 PB
- Usage policy
  - LQCD Jeopardy Policy (penalty/reward) apply at end of each month.
  - Opportunistic lower priority usage available for LQCD after sub-project allocation used up.

    We made adjustment of policy for IC, sub-project can use scavenger qos after finishing allocation. Usage does not count towards LQCD allocation !
  - Scavenger usage available when whole LQCD allocation finish and when cluster have available resource . (jobs subject to preemption)

User Support

- Facility website www.sdcc.bnl.gov .
  - New accounts
    - Instructions on website
    - Usually ~24 hours to process after verification
  - User support requests
    - SDCC policy is to respond within 3 business days. Majority is resolved within this period

- Bi-weekly meetings between facility staff and program/experimental Liaisons
  - Agenda on https://indico.bnl.gov/category/169/
  - Remote access via ZoomGov—Minutes of meeting posted for those who cannot join in person or remotely

New BNL Institutional Resources

April 21, 2022

Brookhaven National Laboratory

Background

- Computational Science Initiative (CSI) was formed in 2015 to consolidate and optimize BNL computing infrastructure
- The Institutional Cluster (IC) is a Lab-wide resource regulated by MOU's between CSI and PI's in the research community
  - IC helps to build computational science expertise at BNL and its partners
  - Is a ramp to Exascale computing systems
- IC arrived in 4 batches, first one (50% of cluster) was in Fall 2016. Warranty support due to expire in Fall 2021 —extended until Sep. 2023
- Augmented in subsequent years with KNL, SL (Skylake) and ML clusters
- Plan to purchase new IC after new data center is available (Autumn/Winter 2021)

Recent Activities

- An Advisory Panel was created to guide the acquisition process for the new BNL Institutional Cluster (IC). Panel members include SDCC, LQCD and other IC stakeholders
  - Panel charged with evaluating various configurations and making recommendations on composition of new IC
  - SDCC initiated the procurement process by obtaining budgetary quotes to set the scale of the procurement
- Hypothetical cluster (built with feedback from members of Advisory Panel)
  - ~45 dual cpu-only nodes
  - Dual cpu+gpu nodes (NVLINK-enabled)
    - ~70 nodes if 2-gpu configuration (1 evaluation node already setup for testing)
    - ~35 nodes if 4-gpu configuration
  - 1 PB of usable storage
  - NDR inter-connect network fabric

Members of Advisory Panel

Mark Hybertsen & Qin Wu – CFN

Doug Benjamin – Experimental HENP (ATLAS, DUNE and sPHENIX)

Peter Boyle – HEP Theory (Physics)

Stuart Campbell – NSLS-II

Robert Edwards & Jim Simone – LQCD

Chris Hollowell & Zhihua Dong – SDCC

Meifeng Lin & Adolfy Hoise – CSI

Andy Vogelmann – Environmental & Climate Sciences

Next Steps and other considerations

- Obtained program (LQCD, CFN, USATLAS, etc) commitments (Feb. 2022)
- Confirm configuration availability (Mar. 2022)
  - 4 A100 gpu’s with NVLINK only from Redstone model (Supermicro)
- Plan to break up procurements to:
  - Accommodate different ETA’s for various support components (IB switch fabric, ethernet switches, electrical power distribution units, storage, computing, etc)
  - Support components must be in place before new IC arrives
  - Concerns about COVID-related supply chain delays
- BNL, CSI & SDCC leadership are working on purchase plan
- Plan to run existing IC until Sep. 2023, and we expect availability of both IC’s to overlap, currently working on procurement plan including key deadlines that have to be met.

Questions ?