BNL Scientific Data and Computing Center (SDCC) Facility Report

Zhihua Dong
On behalf of SDCC, BNL

USQCD ALL Hands Meeting 2024 - 4/18/2024

@BrookhavenLab

Scientific Data and Computing Center Overview

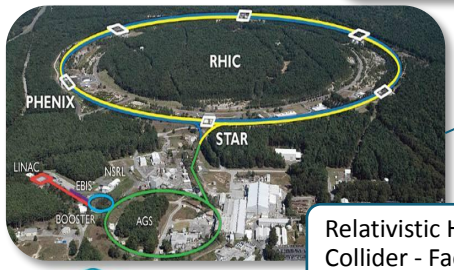
- Located at Brookhaven National Laboratory (BNL) on Long Island, New York
- Tier-0 computing center for the RHIC experiments
  - sPHENIX, STAR
  - BNL is host site for the future Electron-Ion Collider (EIC)
- US Tier-1 Computing facility for the ATLAS experiment at the LHC
  - Also one of the ATLAS shared analysis (Tier-3) facilities in the US
- RAW Data Center and Prompt Calibration Center for Belle II at KEK
- Computing facility for NSLS-II and CFN
- Providing computing and storage for proto-DUNE/DUNE along w/ FNAL serving data to all DUNE OSG sites
- Providing computing resources for a number of smaller experiments in NP and HEP
- Serving more than 2,000 users from >20 projects

Brookhaven National Laboratory

2024Q1

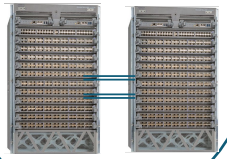
National Synchrotron Light Source II, CryoEM

Center for Functional Nanomaterials

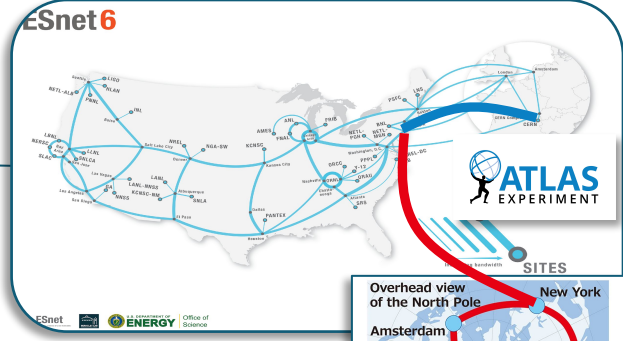


Relativistic Heavy Ion Collider - Facilities

Science DMZ



2x800 Gbps



0.2 Tbps

1.2 Tbps

DTNs (52 systems)

1.2 Tbps

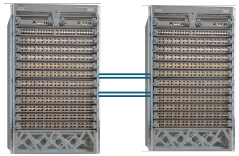
Disk Storage
125+ PB

0.8 Tbps

0.4 Tbps

0.7 Tbps

High Throughput Science Network

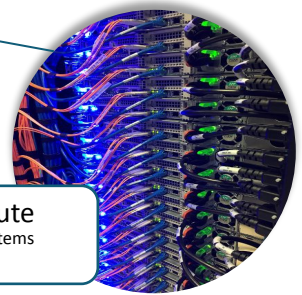


400 Tbps capability
(60 Tbps deployed)

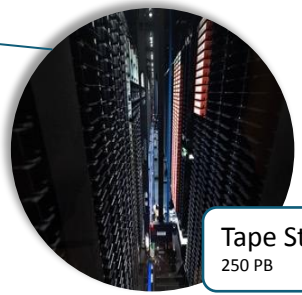
12 Tbps

0.5 Tbps

18 Tbps



Compute
3 HPC systems

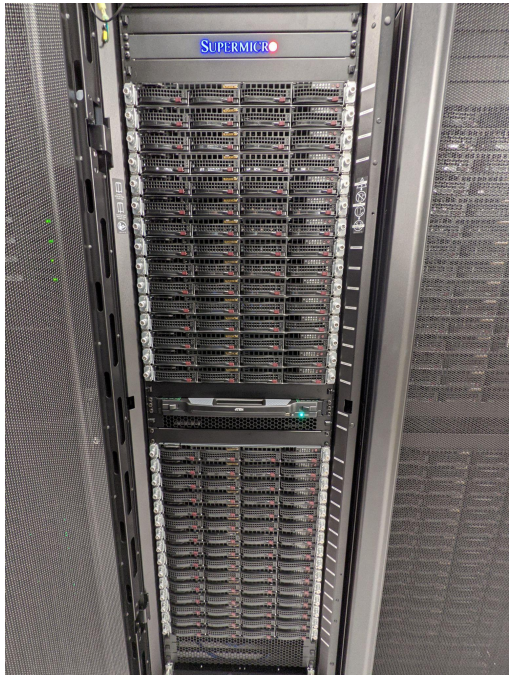


Tape Storage
250 PB

Brookhaven National Laboratory

High Throughput Computing @ SDCC

- Providing our users with ~2,300 HTC nodes:
  - ~140,000 logical cores
  - Managed by **HTCondor**
- **HTCondor 23.0** testing in progress
  - Test cluster has been created
    - central manager, submit, CE, worker nodes
  - Testing/altering current configs for Alma 9 / HTC23
- Provisioning and orchestration overhaul for the Linux Farm
  - Replacing dated custom build infrastructure with **Foreman**
  - Simplify the lifecycle management of nodes
- sPHENIX experiment at RHIC is a very high priority at BNL
  - ~68,000 logical cores (~880k HS23) currently available–nearly ~50% of total available HTC node count at the SDCC
  - Baseline plan will add ~46,000 cores (~620k HS23) in 2024



Supermicro SYS-6019U-TR4 Servers

High Performance Computing

Currently supporting 3 HPC clusters
(After retirement of KNL,IC,SKY on 10/1/2023)

- **Institutional Cluster gen2**
  - 39 CPU only nodes
  - 12 nodes with GPUs
  - 1TB memory on GPU nodes
  - 512GB on CPU nodes
  - 4x Nvidia A100 80GB SXM on GPU nodes
  - NDR200 Infiniband interconnect (200Gbps per link)

- **NSLS2 Cluster**
  - 32 Supermicro nodes with EDR IB
  - 13 nodes with 2x Nvidia V100

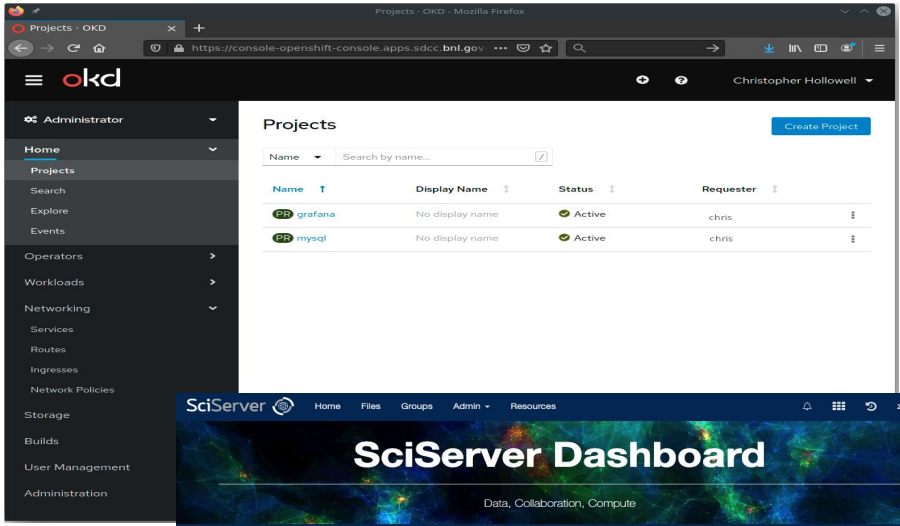
- **CFN Remix**
  - 54 2xNvidiaP100 GPU nodes with EDR IB

IC Gen2 Cluster

Brookhaven National Laboratory

3 Clusters

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



SciServer

The Science Platform

A collaborative environment for server-side analysis with large datasets

45 GPU Compute Nodes (in OKD and K8s)
4 of them with 8xH100 SXM GPUs
5 of them with 8xV100 SXM GPUs
6PB of underline Lustre Storage

Brookhaven National Laboratory

Storage:

Disk

- **dCache:** ~74 PB
- **XROOTD:** ~11 PB
- **Lustre:** ~79 PB
- **GPFS:** ~12 PB
- **Home hosted S3 Storage for EIC : (now in house)**
  use native Object Storage(CEPH) & Federated ID access

Tape

- Archive data size **257.69 PB** (239,125,036 files)
- Data movers: 25 servers
- Tape libraries: **14**
  - Oracle SL8500: 9
  - IBM TS4500: 5
  - Active tape volumes: **75,493**

COmanage Registry

- Aggregate multiple identities so BNL services see only one.
- 354 identities currently aggregated to 307 unique users
- 5 production OIDC clients serving 26 unique service instances
- Service authorization can be controlled by active IDP and group membership
- Services are being added/converted as time allows.

Web @ SDCC

- All SDCC managed Drupal deployments have been integrated with our COmanage instance
  - Allowing users the ability to collaborate across multiple institutions in one place
- Deployment of Hugo/Gitea based documentation site for internal use (in progress)
  - Static site generator using Go and Markdown

USQCD Access to SDCC Resources

Compute Allocations/usage:

From 7/1/2023 - 9/30/2023 (before IC, SKY, KNL Retired on 10/1/2023)

- 181k node-hour allocation on IC cluster used ~102% + ~14% scavenger)
- 117k node-hour allocation on SKY cluster ~94%
- 61k node-hour allocation on KNL cluster ~ 284%

After 10/1/2023 (We have an empty window for LQCD allocations)

LQCD only have 4 CPU nodes in IC Gen2,

Mainly for data access and few Class C projects.

In the process of procuring a new cluster hopefully online by end of FY24.

Storage Allocation/usage:

- 800 TB of GPFS disk storage ~470TB currently in use
- Tape Storage:
  - Total LQCD data on tape : ~4.7PB (since 1/2020)
  - Include Long Term Archive currently ~3.4 PB

Lattice QCD new equipment procurement for BNL FY24

BNL Lattice QCD Cluster Requirements Committee

Formed on Jan 2024

Report delivered Apr 2024

Members:

- Peter Boyle (chair) (BNL)
- Zhihua Dong (BNL)
- Josephine Fazio (FNAL)
- Chulwoo Jung (BNL)
- Imran Latif (BNL)
- Meifeng Lin (BNL)
- Alan Prosser (FNAL)
- James Simone (FNAL)
- Amitoj Singh (JLAB)

Brookhaven National Laboratory

Key Requirements Summary

Choices of node types based on benchmark

Type	Vendor	Platform	Performance	GPU count	HDR-200 NICs	Memory
CPU	Intel	Xeon+DDR	0.9		1	≥1TB
CPU	Intel	Xeon+HBM	1.3		1	≥1TB
CPU	AMD	EPYC Genoa	1.1		1	≥1TB
CPU	Nvidia	Grace/ARM	-		1	≥1TB
GPU	Nvidia	H100 x4	19	4	2	≥2TB
GPU	Nvidia	H100 x8	36	8	4	≥2TB
GPU	AMD	MI300X x4	19	4	2	≥2TB
GPU	AMD	MI300X x8	36	8	4	≥2TB

CPUs of core count 32 is recommended
Increasing cpu core count to more than 32 does not further increase performance

Requirements:

- **Cluster global host memory must > 20TB**
- **Must maximize geometric mean of global host memory (TB) and total Performance**
- **8TB local SSD for GPU nodes**
- **Constrained by Site :**
  - **power/cooling capacity**
  - **Additional network switch/ports**
  - **Work required to host**
  - **…**

BNL Lattice QCD Cluster Procurement Committee

Members:

- Tony Wong (chair) (BNL)
- Zhihua Dong (BNL)
- Chulwoo Jung (BNL)

Procurement process is starting
according rules set by requirement committee

Additional procurement considerations

- Supply chain delays (6+ months for Infiniband and some gpu choices) will affect cluster timeline for deployment
  - Is timely availability of new LQCD resources a selection criterium?
- Complexity of criteria for selecting “best” offer affects timeline of procurement process
  - Criteria must be approved by BNL procurement dept., as a precautionary measure (in case is is audited by DOE)
- Submittal of a requisition is contingent on funding availability, as per BNL procurement rules

Thanks to the great team at SDCC for contributing to this presentation:
Tony W. Ofer R. Costin C. Kevin C. Doug B. Tim C. Jane L. Hiro I. John D. Carlos G. Mark L. Vincent G. Robert H.

Questions?