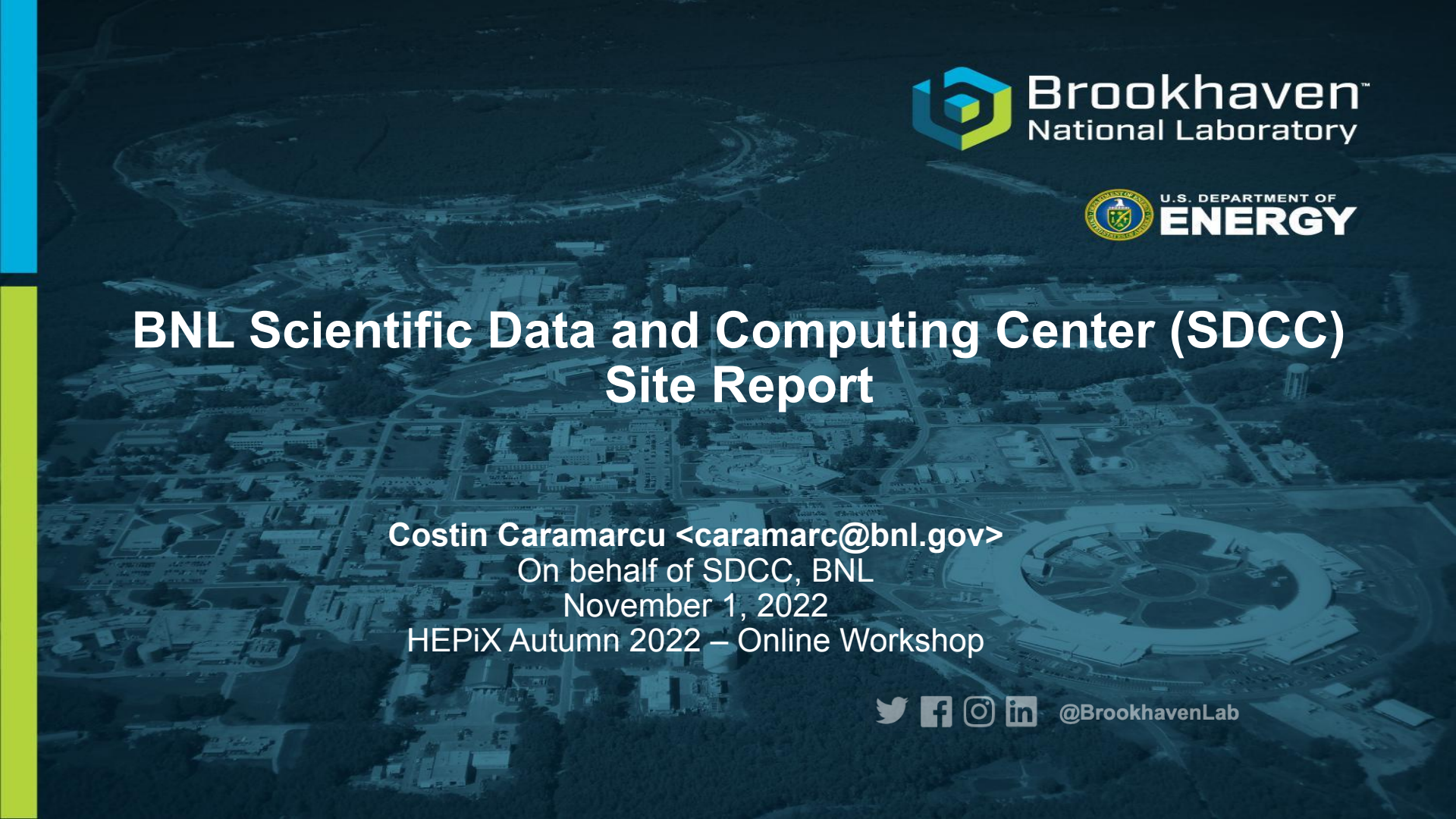


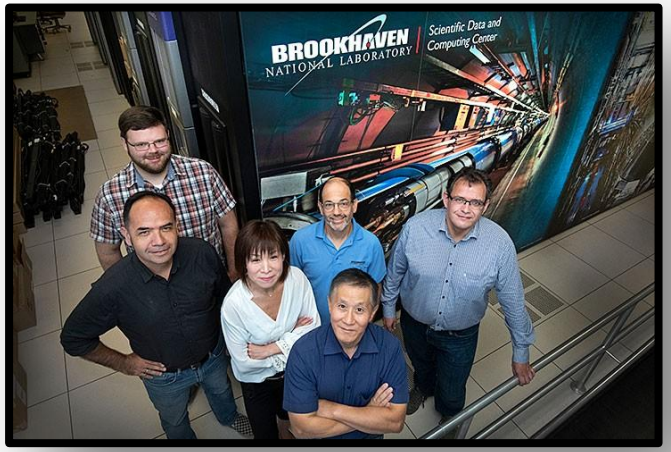
BNL Scientific Data and Computing Center (SDCC) Site Report

Costin Caramarcu <caramarc@bnl.gov>
On behalf of SDCC, BNL
November 1, 2022
HEPiX Autumn 2022 – Online Workshop

@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



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

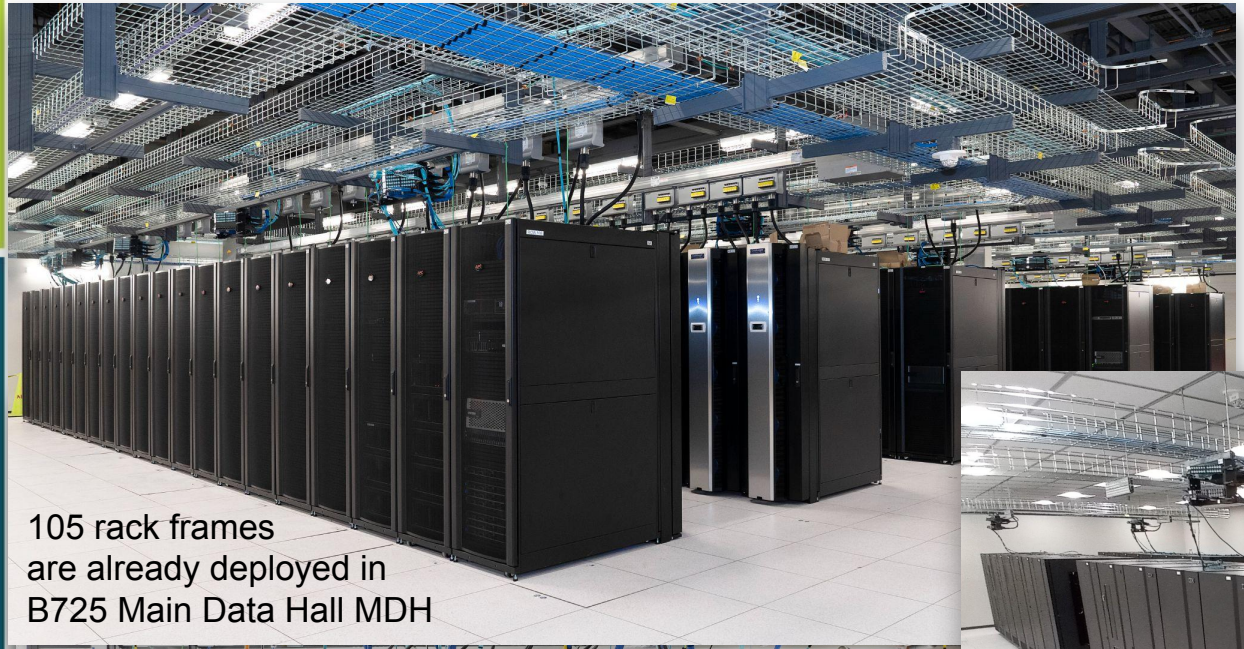
Scientific Data and Computing Center Overview

- 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 in NP and HEP
- 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 2023

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

New Data Center (Building 725) — FY22: 1st Complete Year 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 2022Q1-3 to the level of 54 racks populated with equipment in the B725 Main Data Hall (MDH)
  - The vast majority of equipment purchased by SDCC starting from 2021Q3 is being placed in the new data center
  - 18 existing CPU racks were moved from the old data center to the new one in 2022Q1
  - 7 new CPU racks and 19 high capacity JBOD racks were deployed directly in B725 in FY22
  - 20 more new CPU racks are expected to be added to B725 in 2022Q4-2023Q1
- The second diesel generator was added to the B725 diesel generator yard in FY22: more to be added in the years to come following the scaling of the IT load
- Two library rows in B725 Tape Room were populated with IBM TS4500 tape libraries to serve ATLAS and sPHENIX experiments in 2021Q4-2022Q1 (4 libraries, 128 tape drives in total)
- Preparations are being made for deploying new HPC clusters and associated storage systems in B725 MDH in 2023Q2-3
- 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



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

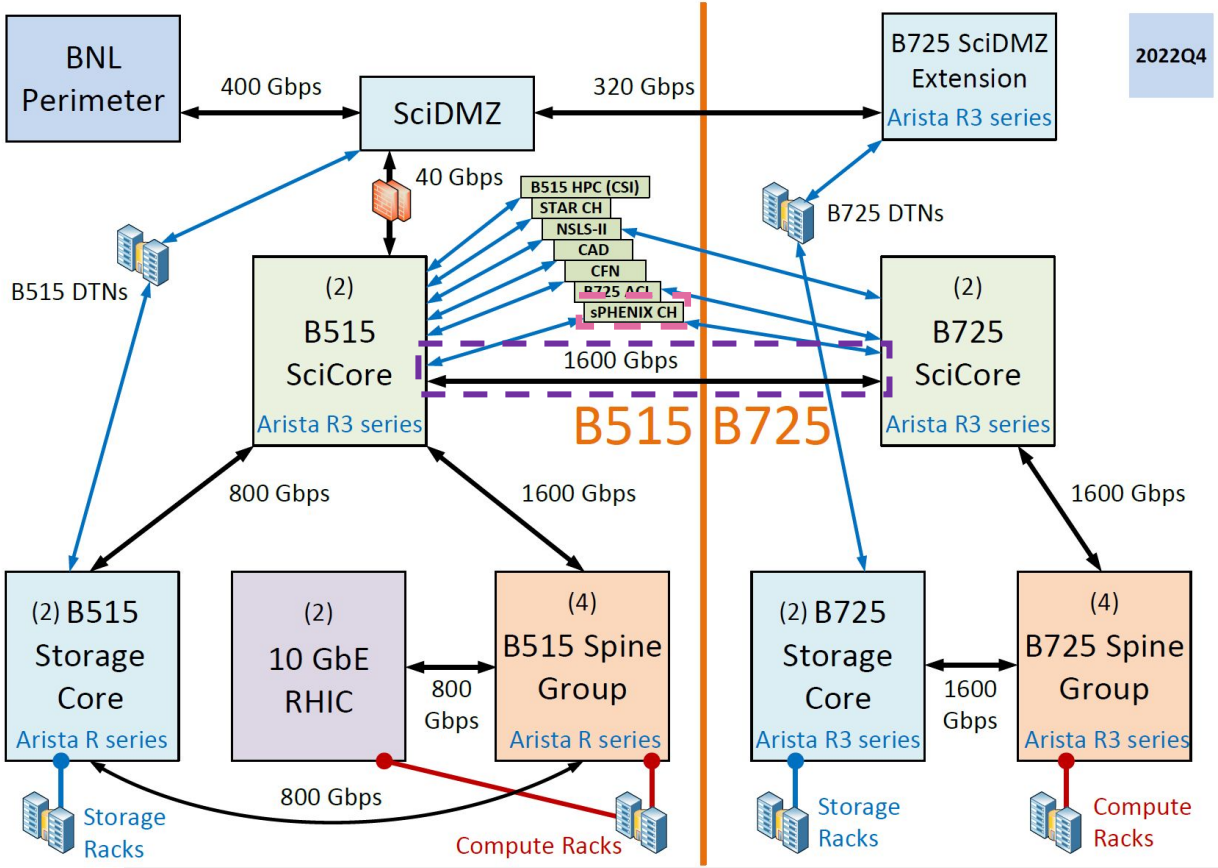
84 RDHx units deployed in B725 MDH, out of which 54 are on racks with equipment while 30 are deployed for the future growth

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

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

5

Network Systems of B515 and B725 Data Centers (as of Oct'22)



- 1.6 Tbps (LR) B515/B725 interbuilding link is functioning as expected
- sPHENIX Counting House (CH) uplinks to both B515 & B725 data centres are activated at 200 Gbps to B515 plus 400 Gbps to B725

Brookhaven National Laboratory

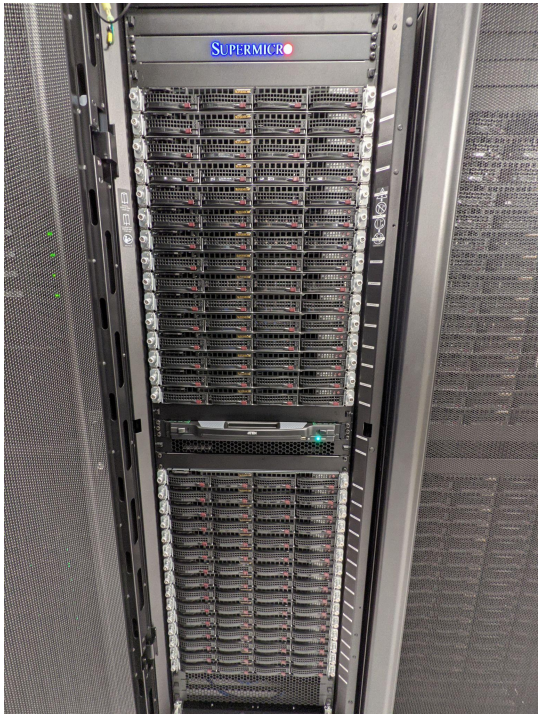
High Throughput Computing

- Providing our users with ~1,900 HTC nodes:
  - ~90,000 logical cores
  - ~1050 kHS06
  - Managed by HTCondor 9.0

- Purchased 648 Supermicro SYS-610C-TR nodes for ATLAS and the RHIC experiments (~62k logical cores total)
  - Expected delivery Jan/Feb 2023
  - 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

- All nodes running Scientific Linux (SL) 7
  - Testing/preparations for an OS upgrade to Rocky Linux 8/9 in progress

- Beginning preparations for Condor 10.0 update



2021 Supermicro SYS-6019U-TR4 Servers

Brookhaven National Laboratory

High Performance Computing

Currently supporting 5 HPC clusters

- **Institutional Cluster gen1 (IC)**
  - 216 HP XL190r Gen9 nodes with EDR IB
  - 108 nodes with 2x Nvidia K80
  - 108 nodes with 2x Nvidia P100
- **Skylake Cluster**
  - 64 Dell PowerEdge R640 nodes with EDR IB
- **KNL Cluster**
  - 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

NSLS2 HPC Cluster

Institutional Cluster gen2 (IC)
Order placed for the IC gen2 with the following specs:

- 2x Intel Xeon (Ice Lake)
- 512GB DDR4-3200 CPU nodes
- 1TB DDR4-3200 GPU nodes
- NDR200 InfiniBand interconnect (200Gbps per uplink)
- 4x Nvidia A100 80GB

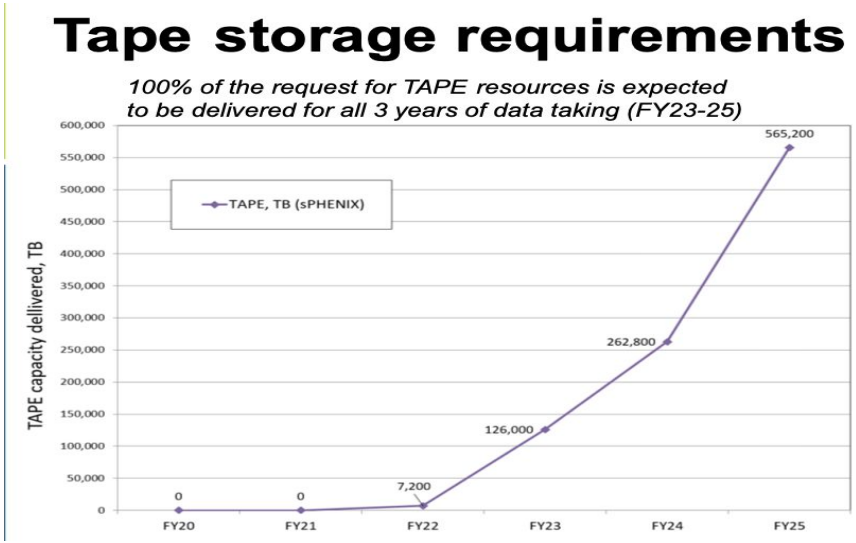
We are looking at a performance of 3x from current IC node: ~7.9 TF to ~25TF IC gen2 node.

Brookhaven National Laboratory

Tape System

- Currently ~200 PB of data in HPSS with ~70k tapes
- New HPSS hardware in the newly commissioned data center for sPhenix experiment
  - 10GB/sec data injection requirement
  - High performance/capacity disk cache (2.1PB, 330 HDDs)
  - Two new IBM TS4500 tape libraries
  - Total of 64 new LTO-9 tape drives in the TS4500 libraries

sPHENIX Tape Storage Space Requirements



Central Disk Storage

- Support various experiments
  - NSLS-II, sPHENIX, ATLAS, LQCD and STAR
  - Total capacity: 50 PB and 600 million files.
- Upgrade
  - Growing footprint for Lustre (2.12.8)
  - Added 25PB to sPHENIX Lustre. Excellent streaming sequential performance with aggregate throughput of 210 GB/s
  - Lustre running mostly stable.

Storage: dCache/XROOTD

dCache

Total ~70 PB in dCache

- ATLAS (v7.2.16)
  - Upgrade from 6.2 to 7.2 release
- Belle II (v7.2.19)
  - Upgrade from 6.2 to 7.2 release
- PHENIX (v5.2.9)
  - Mix of central and farm node storage
- DUNE (v8.2.2)

Please see Thursday storage track session, there is a presentation on BNL dCache services

XROOTD

~11 PB total storage for STAR

- Mix of central and farm node storage

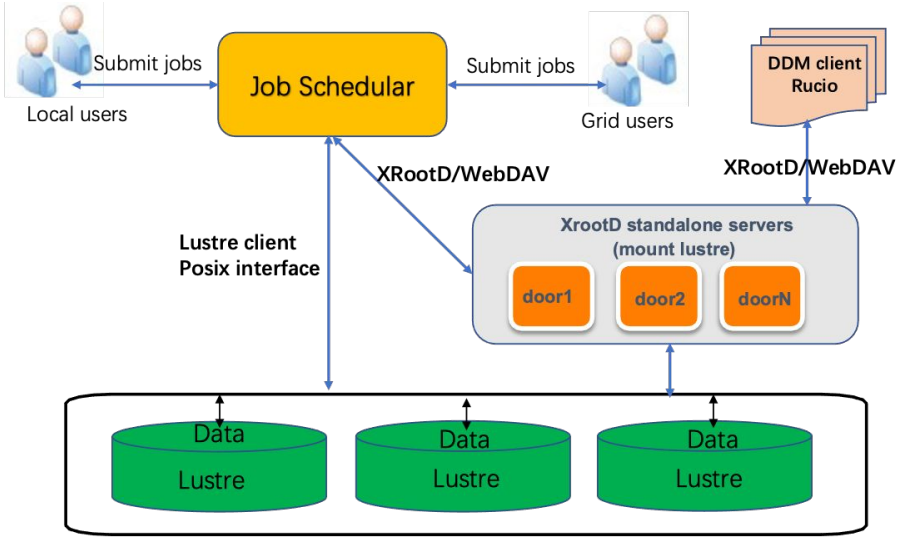
Activity of the Future Storage

Motivation

- The huge amount of data storage from sPHENIX (in the near-term) and ATLAS (medium-term, in LHC-HL run)
  - No Grid specific interface required
  - Many local/POSIX accesses
  - Hundreds of PB disk storage
  - Reliable and resilient interface to archive system (HPSS)

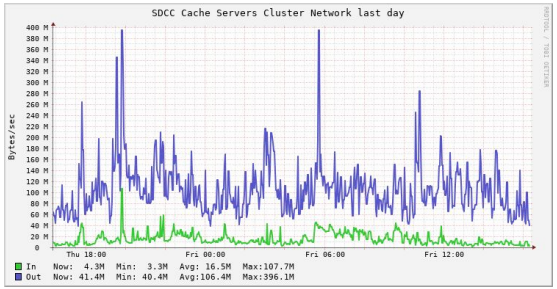
We are gaining experience with different (HPC) storage technologies

- LustreViaXrootdRW
  - Lustre read/write through the OSG Xrootd Standalone server
  - Provide grid XROOTD and WebDAV protocols above Lustre



CVMFS

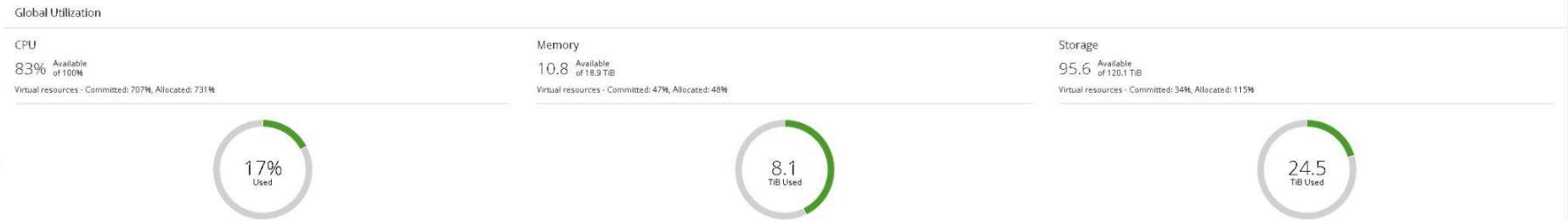
- All server & replica services running version 2.9.4 (latest)
- Server
  - Stratum Zero for 13 locally hosted repositories
  - Occupying 1.1 TB NAS disk
  - For facility and experiment use:
    - ASTRO, ATLAS, DayaBay, DUNE, EIC, OSG, PHENIX, SDCC, sPHENIX, STAR
      - Coming soon: sPHENIX calibration data publishing
- Replica
  - Stratum One for 112 replicated WLCG and OSG repositories
  - From 10 domains and four main sources (BNL, CERN, OSG, RAL)
  - Occupying 52 TB NAS disk
- Plans to virtualize, migrate all services to new platform, OS, storage, infrastructure in FY 2023



Brookhaven National Laboratory

Redhat Virtualization

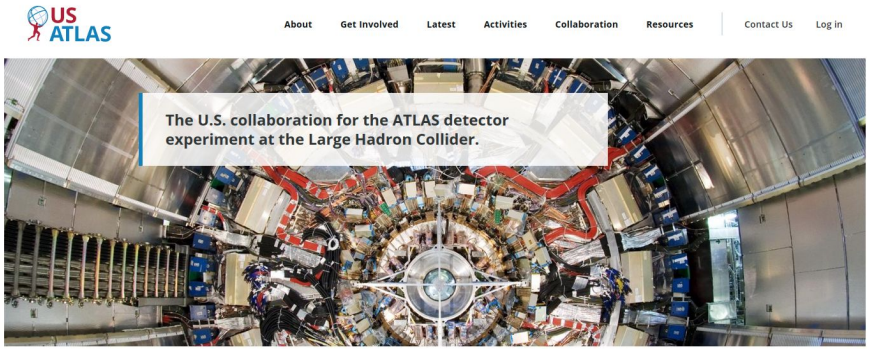
- Redhat Virtualization is end of life in 2024.
- Evaluating Openshift Virtualization and VMware.
  - Leaning towards VMware due to product maturity, features and pricing.
- Moving 600 VMs from RedHat Virtualization to VMware or Openshift will take time.
- Since RHEL7 is also EOL in 2024, we can rebuild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



SDCC USATLAS Site

US ATLAS Drupal site launched June 1st, 2022

- usatlas.org
- New site is created with Drupal content management system
- Built in conjunction with Data Art and US ATLAS team members
- MFA and CERN federated login
- Public and private pages for US ATLAS documentation



OKD Clusters at BNL

- Two production OKD clusters were brought online in 2022
  - **ATLAS cluster**
    - Primarily for Analysis Facility (AF) services that require k8s
      - ServiceX
      - REANA
      - Note that our analysis facility JupyterHub deployment does not require k8s and uses a modified batchspawner plugin to utilize our existing large batch (HTCondor/SLURM) farms
        - REANA is also capable of leveraging batch resources
  - **sPHENIX cluster**
    - Primarily for Panda service, and conditions database (CDB) deployment
      - Panda developers have created numerous helm charts
    - Example of a developer/user-deployed service in OKD
      - Collaboration between two groups at BNL
        - SDCC managing the OKD software/hardware
        - CDB and Panda deployments maintained/managed by the NPPS (Nuclear and Particle Physics Software) group at BNL, with SDCC support
- Separate clusters for now as the ATLAS AF services are currently in development

Brookhaven National Laboratory

OKD Cluster Details

- Each cluster running OKD 4.10, and is provisioned with:
  - 7 Dell R640 Servers
    - 3 HA control plane nodes, 4 worker nodes
      - Running Fedora CoreOS (FCOS) 35 deployed via OKD Installer-Provisioned Infrastructure (IPI)
        - CRI-O used as container runtime
    - Specs:
      - 2x Xeon Silver 4210 CPU @ 2.20 GHz
      - 128 GB RAM
      - 4x 25 Gbps NICs
      - 3x 480 GB SSDs
  - NetApp A250 Storage Appliance
    - 14 x 1.92 TB NVME drives (~26 TB raw)
    - ONTAP NetApp OS allows dynamic PV provisioning via Trident

TRIDENT

ATLAS OKD Cluster Hardware

Brookhaven National Laboratory

Thanks to the SDCC team for contributing to this presentation

Questions?