BNL Scientific Data and Computing Center (SDCC) Site Report

Chris Hollowell <hollowec@bnl.gov>

HEPiX Fall 2019 – Nikhef Amsterdam, Netherlands

BROOKHAVEN NATIONAL LABORATORY | Scientific Data and Computing Center

U.S. DEPARTMENT OF ENERGY

Scientific Data and Computing Center Overview

Located at Brookhaven National Laboratory (BNL) on Long Island, New York

- Multidisciplinary US Department of Energy laboratory with well established nuclear, high energy and photon science physics programs

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

- Tier0 computing center for the RHIC experiments
- STAR is currently the only RHIC experiment taking new data
- However, preparations for a new experiment at RHIC underway
  - sPHENIX - scheduled to start taking data in 2023
  - Designed to bridge and extend the useful lifetime of RHIC into a potential future eRHIC
- RHIC may transition to e-RHIC/EIC (pending site selection) in the late 2020s

SDCC/RACF is currently supporting sPHENIX's requests for computing resources (MC simulation) and collaborative tools (Invenio-based document repository, JIRA for activity tracking, mailing lists, etc.)

- We expect sPHENIX to request support across our entire service portfolio by 2021

SDCC Overview (Cont.)

US Tier1 Computing facility for the ATLAS experiment at the LHC
- Also one of two ATLAS shared analysis (Tier3) facilities in the US

US Belle II Tier1 Computing center

Also providing computing resources for various smaller/R&D experiments at BNL
- DUNE, EIC, LSST, etc.

Increasingly supporting the HPC needs of various groups beyond HEP/NP including the Center for Funtional Nanomaterials (CFN), National Synchrotron Light Source II (NSLS-II), Biology, Simons Foundation and the National Nuclear Data Center (NNDC)

Serving more than 2,000 users from >20 projects

SDCC Overview (Cont.)

Besides providing computing/storage resources for our user community, we've recently expanded our emphasis on developing and administrating new collaborative tools

- Invenio, Jupyter, BNL Box, Gitea, Mattermost, etc.

Currently employ 33 full time employees

SDCC Collaborative Tools Developers/Administrators

Hiring if you're potentially interested in joining SDCC

- Expect a number of positions to be posted in the near future
- Check https://jobs.bnl.gov for updates

New Datacenter (CFR) Update

Computing and storage requirements for US ATLAS T1 for HL-LHC, and sPHENIX at RHIC are too large to operate the equipment within our existing datacenter

- Cannot be accommodated from both a space and power/cooling perspective

Current datacenter – Building 515

- 3 rooms – BCF, Sigma-7 and CDCE
  - BCF was commissioned in the 1960s, with antiquated 1-foot raised floor, and aging central basement air handler units
- Air-cooled, without hot/cold isle containment
- 1.5 MW facility
- Mix of battery UPS and flywheel/diesel generator backup systems
- Can support up to ~13 kW racks

New datacenter (CFR – Core Facility Revitalization) expected to come online in Building 725 in February 2021

- 3.6 MW facility
  - Backup power provided by diesel generators (battery UPS for generator startup window)
- Liquid-cooled rear-door heat exchangers for racks
  - Supports up to ~30 kW per rack

New Datacenter (CFR) Update (Cont.)

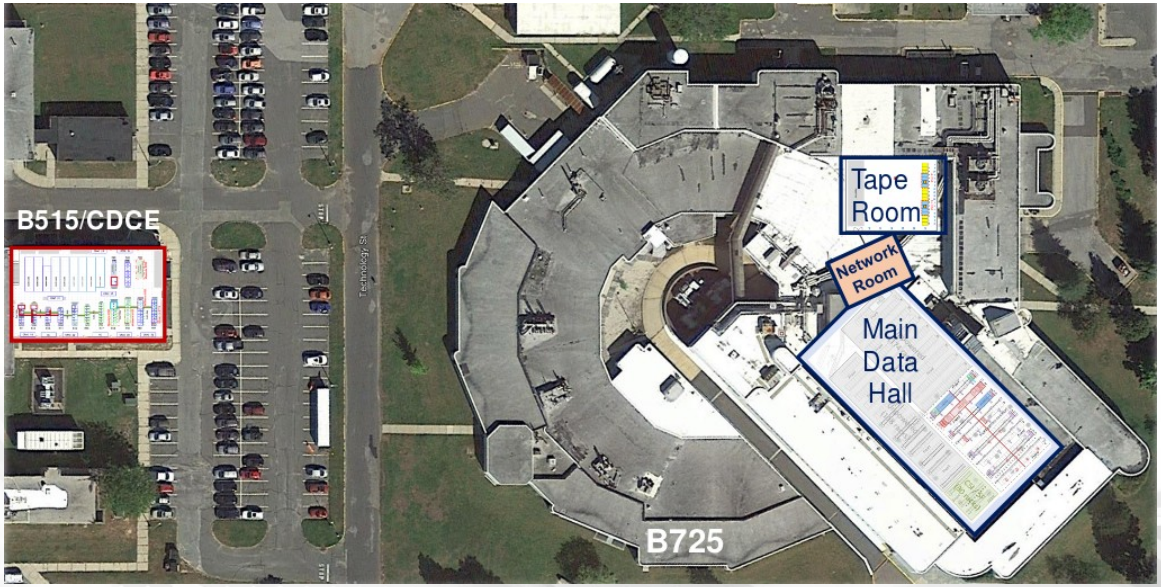
Construction began in May 2019

However, work stopped for several months over the summer due to contract issues

- These have been resolved
- Hopeful this doesn't delay the targeted 2/2021 delivery

The BCF portion of the old datacenter will be retired gradually via attrition

- Most existing equipment won't be moved to the new datacenter: it will be allowed to operate in place until end of life
- CDCE will remain operational for tape storage indefinitely



New SDCC Datacenter in Building 725

Network

Total facility aggregate Ethernet bandwidth at the endpoints:

- ~15 Tbps (unidirectional)
- Including 3.5 Tbps added since Spring 2019

The number of Arista 7020TR-48 ToR/leaf gigabit switches in production for HTC compute nodes has grown to over 70

- These are all connected to central spine group of 4 Arista 7280QR switches

FY19 EOL/EOS central switch equipment replacement campaign was completed successfully

- Only one legacy switch system remains in production for the entire facility
- Expected to be retired in FY20

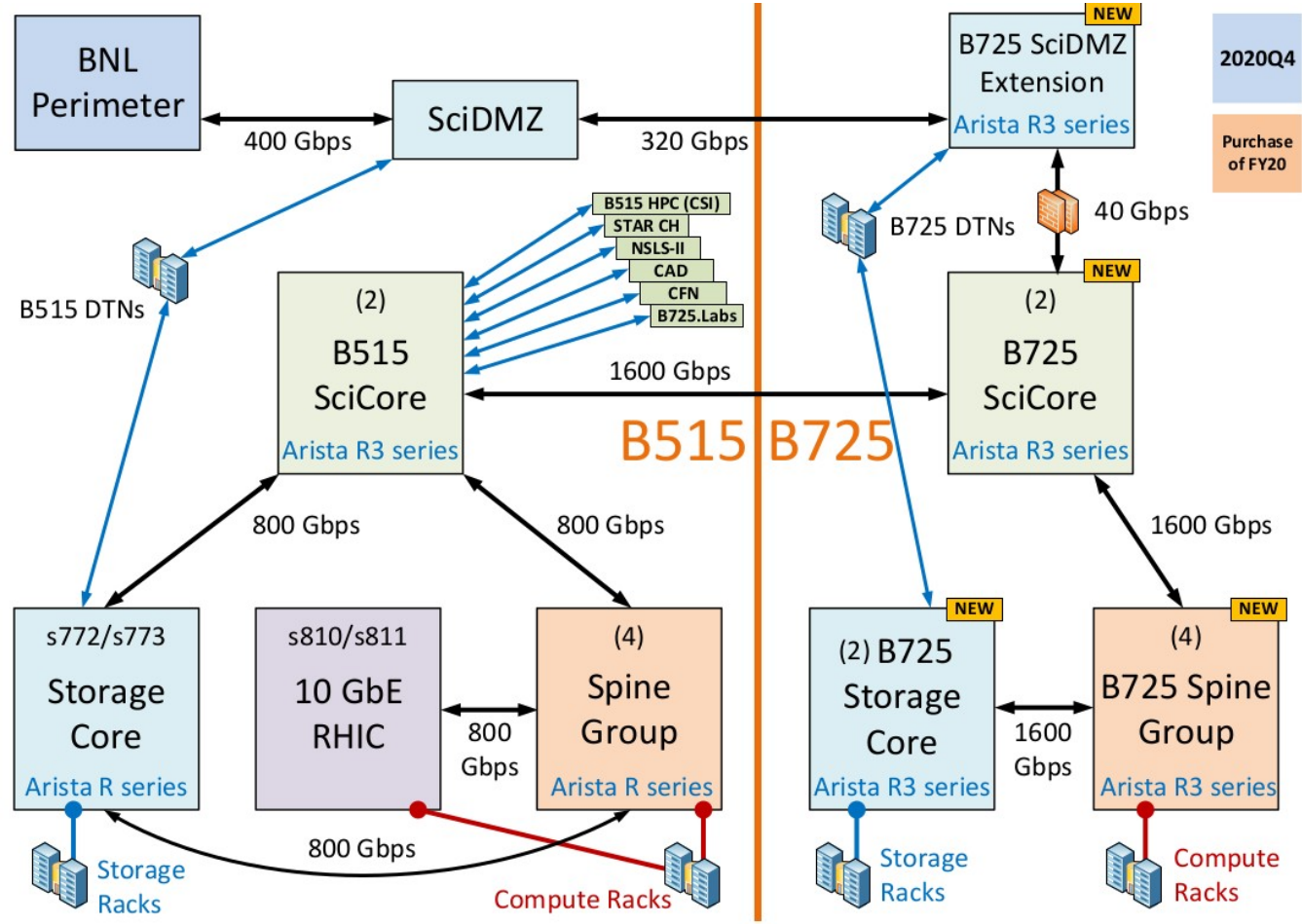
Optimization/cleanup of the facility's IP space is ongoing

- Attempting to move systems without external connectivity requirements to private IP addresses where possible

Need to connect our existing datacenter in building 515 to the new datacenter being built in building 725

- Construction project to connect the "last mile" of the inter-building link, avoiding the BCF area (to be dropped from the datacenter in FY23) is expected to commence in 2020Q1-2

Network (Cont.)



FY21 Planned Multi-Datacenter Network Architecture

High Throughput Computing

Providing our users with ~2,000 HTC nodes:
- ~65,000 logical cores
- ~790 kHS06

156 new Supermicro 1023US-TR4 servers brought online in September 2019
- Dual AMD EPYC 7351 CPUs @ 2.4 GHz
- 128 GB DDR4-2666 MHz RAM
- 4 x 4 TB 7200 RPM SATA6 drives
- 1U form factor
- 775 HS06/node = ~121 kHS06 total

All nodes running Scientific Linux 7 for some time
- SL6 Singularity containers provided to experiments which still require this OS

In the process of upgrading to HTCondor 8.8.3

Rack of New EPYC-based Supermicro 1023US-TR4 Servers

High Performance Computing

Institutional Cluster
216 HP XL190r Gen9 nodes
- 2 Intel Xeon E5-2695v4 CPUs @ 2.1 GHz
- 256 GB DDR4-2400 MHz RAM
- Dual Tesla K80 GPUs in 1/2 the systems
- Dual P100 GPUs in the other half
- EDR Infiniband

KNL Cluster
142 KOI S7200AP nodes
- Intel Xeon Phi 7230 CPU @ 1.3 GHz
  - 256 Logical cores total
- 192 GB DDR4-1200 MHz RAM
- EDR Infinband

New ML Cluster – provisioned 9/2019
5 HP Proliant XL270d Gen10 nodes
- 2 Intel Xeon Gold 6248 CPUs @ 2.5 GHz
- 768 GB DDR-2933 MHz RAM
- EDR Infiniband
- 8 V100 GPUs (per node)
  - Each system using ~4 kW of power!

ML Cluster

JupyterHub

Providing JupyterHub instances for several of our supported experiments
- Both on our HTC/HTCondor and HPC/Slurm resources

Recently brought an HTCondor batch-spawned instance online for NSLS-II
- Have link-based notebook sharing/copying between users functional

Ongoing question of kernel management
- Do we maintain, or experiments/groups ?
- Should we factor out a common set of useful defaults?

Slurm Spawner Form

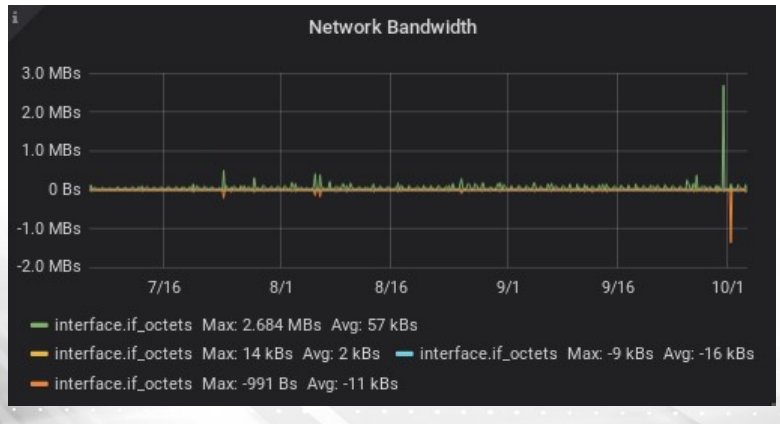
Developed a Slurm Spawner form which allows users to select Slurm Partition, GPU type, etc.

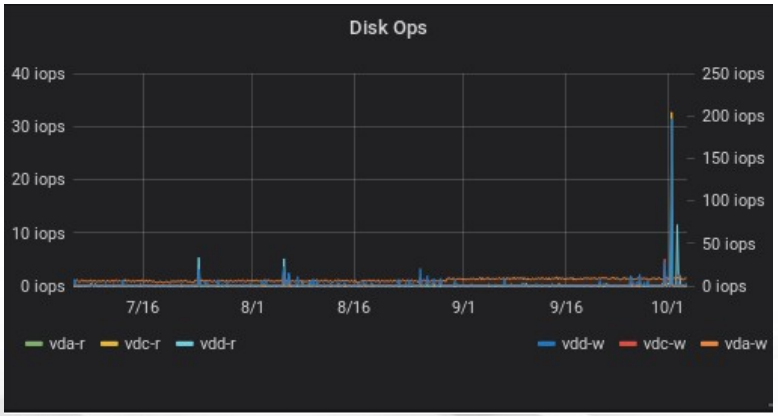
CVMFS

Servers running version 2.6.4

Stratum One continues to grow in size & utilization
- 26 TB of data in 81 replicated repositories

Stratum Zero in production for local experiments and groups
- 12 local repositories for BNL-based experiments
- Completed first re-publishing of local repository content to OSG for global distribution
  - sphenix.opensciencegrid.org
- As expected, tiered proxy/caching infrastructure working well
  - Little network/disk load on the Stratum Zero server





Stratum Zero Network and Disk Activity

Central Disk Storage

Currently have 7 GPFS filesystems
- Total of 14PB of raw storage and > 1 billion files

Running version 4.2.3-15 with nearly 3,000 GPFS clients

Isolated STAR GPFS metadata to dedicated storage nodes
- Improved interactive latency

GPFS contract with IBM expiring in early 2021
- Renewal price is very expensive
- Evaluating Lustre 2.12 as an alternative solution for central disk storage

NSLS-II dedicated 1PB Lustre filesystem will be provisioned next month

Tested Lustre Copytool which integrates with HPSS for transparent migration and automatic recall

dCache/XROOTD

dCache
- Managing over 50 PB of data total
  - ATLAS (v4.2)
  - BELLE-II (v4.2)
  - PHENIX (v3.2)
    - Mix of central and farm node storage
  - Simons (v4.2)

- QoS v5.1 testbed
  - BNL and FNAL pools

- Recently added 2 PB of JBOD storage to PHENIX pool

- Upgraded ATLAS GridFTP door hardware to new Dell R740 systems

XROOTD
- ~11 PB total storage for STAR
  - Mix of central and farm node storage
- Running version 4.7.1

Tape Storage - HPSS

~165 PB total data on tape managed by HPSS

Running HPSS v7.4.3.2 since Dec 2018
- Evaluating v7.5.1

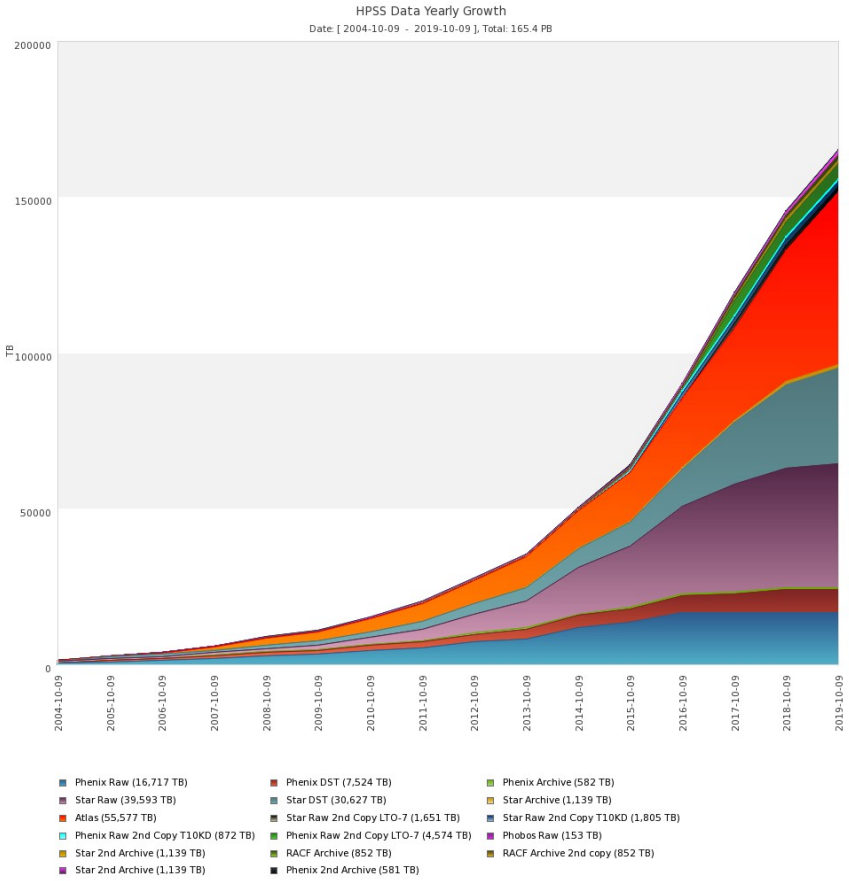
New data from Belle-II

Evaluating new library systems

LTO-8 deployment delayed due to unavailability of media

Both STAR and ATLAS still using LTO-7

PHENIX DST remains on LTO-6 due to lower volume of data



LTO-4 to LTO-7 migration in process to reclaim slots for new tape technologies

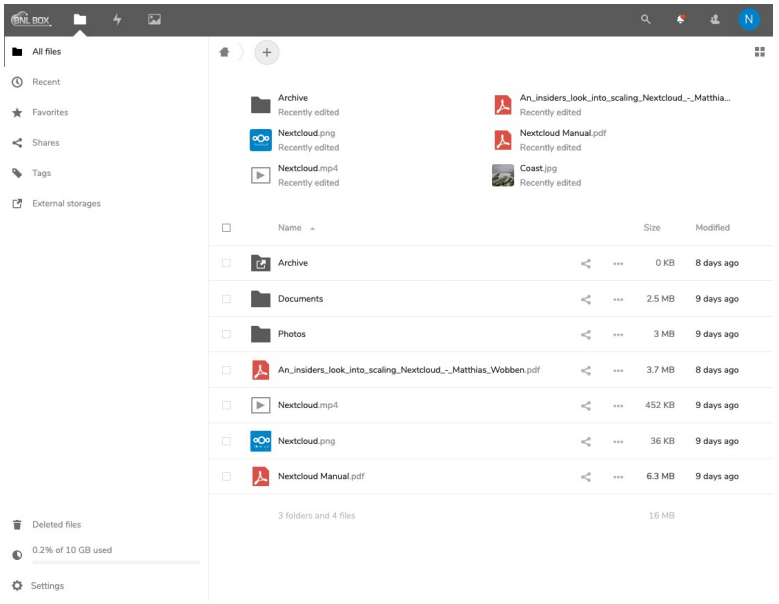
BNL Box

"Dropbox" style file syncing/sharing service based at SDCC

Easy access via browser, desktop and mobile clients

Open to all users with SDCC or BNL AD accounts (via Keycloak based SSO)

An integrated effort by many SDCC staff supporting back-end disk and tape storage, web and db services, AAI front-end, user interface, etc.



BNL Box UI

Completely redesigned using latest open source Nextcloud v16.0.4

Load-balanced, HA architecture

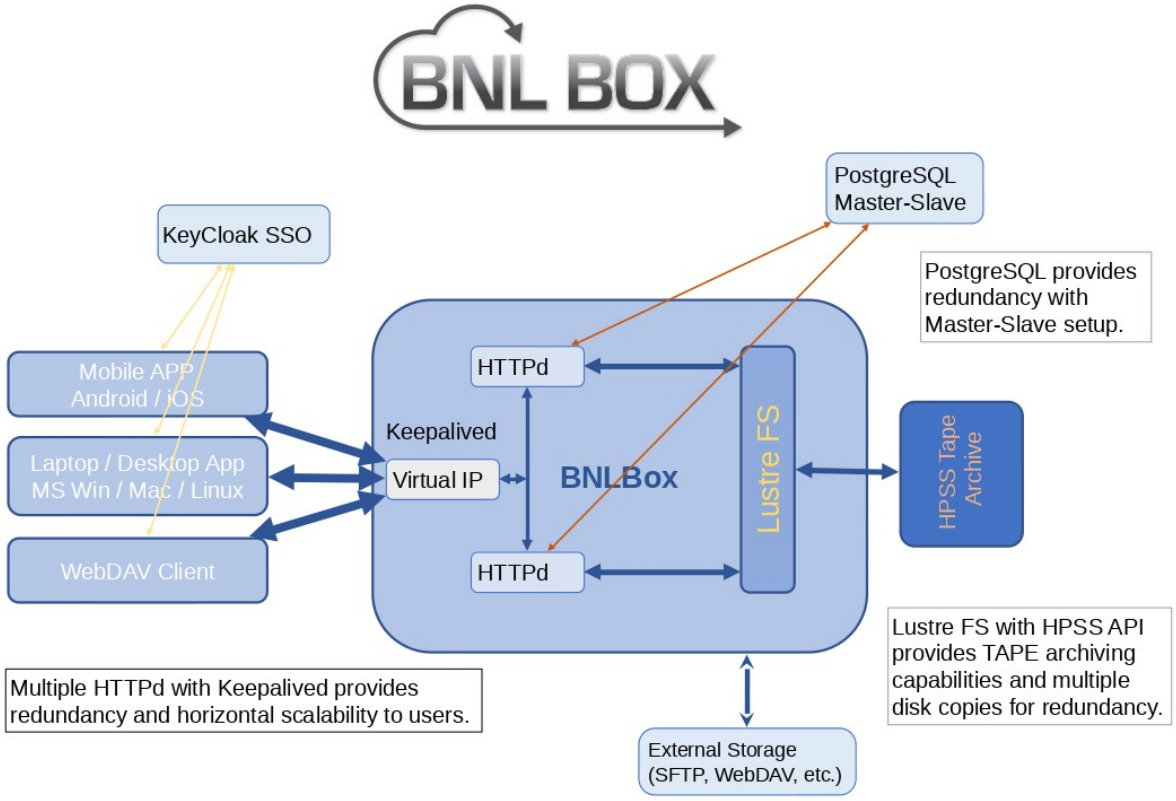
- Round-robin DNS + keepalived
- Shared config directory (Lustre), PSQL database (with standby), and memcache (redis)

User storage on Lustre (quota, TSM backup)

BNL Box (Cont.)

User “Archive” directory provided for offloading files to HPSS
- Mounted as “external” Lustre storage in Nextcloud
- Other “external storage” directory mounts under consideration

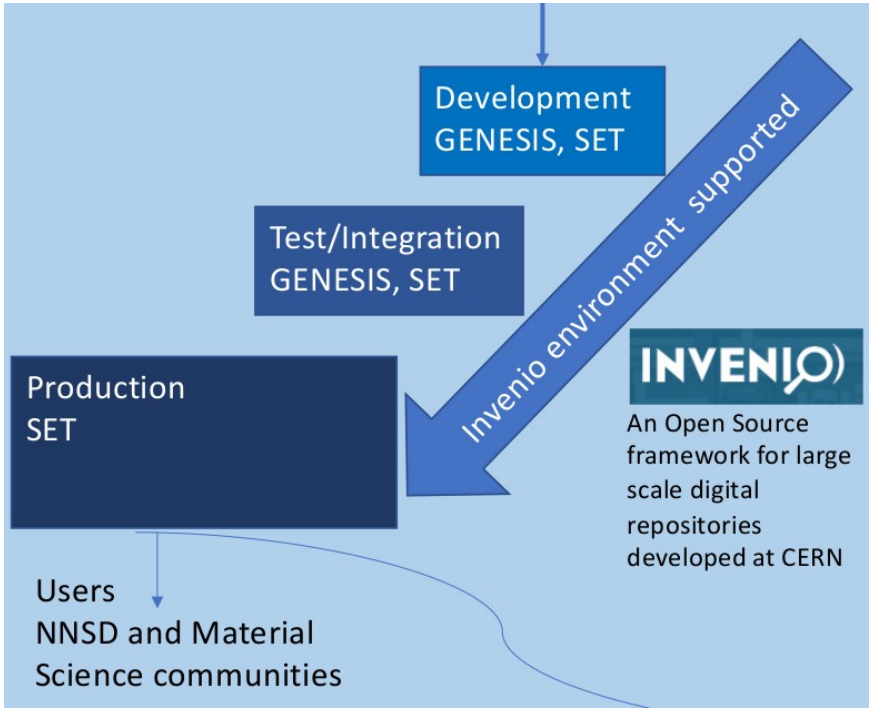
Currently migrating user data from previous Owncloud/Ceph-based BNL Box service



BNL Box Architecture

Invenio

Invenio V3 based custom applications
- SDCC supports custom applications based on INVENIO 3 for two different scientific communities
  - National Nuclear Security Administration (NNSD)
    - SET, Smuggling Detection and Deterrence Science and Engineering Team
  - Materials Science community
    - GENESIS Next-Generation Synthesis Center
  - Expect sPHENIX to utilize Invenio as well
- Maintaining both the hardware and Invenio framework



Invenio V3 Research Data Management (RDM)
- SDCC is working to build a research data management platform called InvenioRDM along with CERN and ten other multidisciplinary and commercial institutions

Single Sign-on and Federated Access

Deployed Keycloak as an SSO/Federated access solution

- FreeOTP MFA AuthN for interactive apps and services
- Allows use of BNL ITD ActiveDirectory accounts, SDCC IPA accounts, and federated ID via CILogon

Now transitioning new/existing websites to Keycloak integration

Considering the establishment of an SDCC standalone IDP to InCommon

- In addition to BNL's existing ITD InCommon IDP

Working on improvements to our AuthZ models, including testing a local COmanage instance

Mattermost

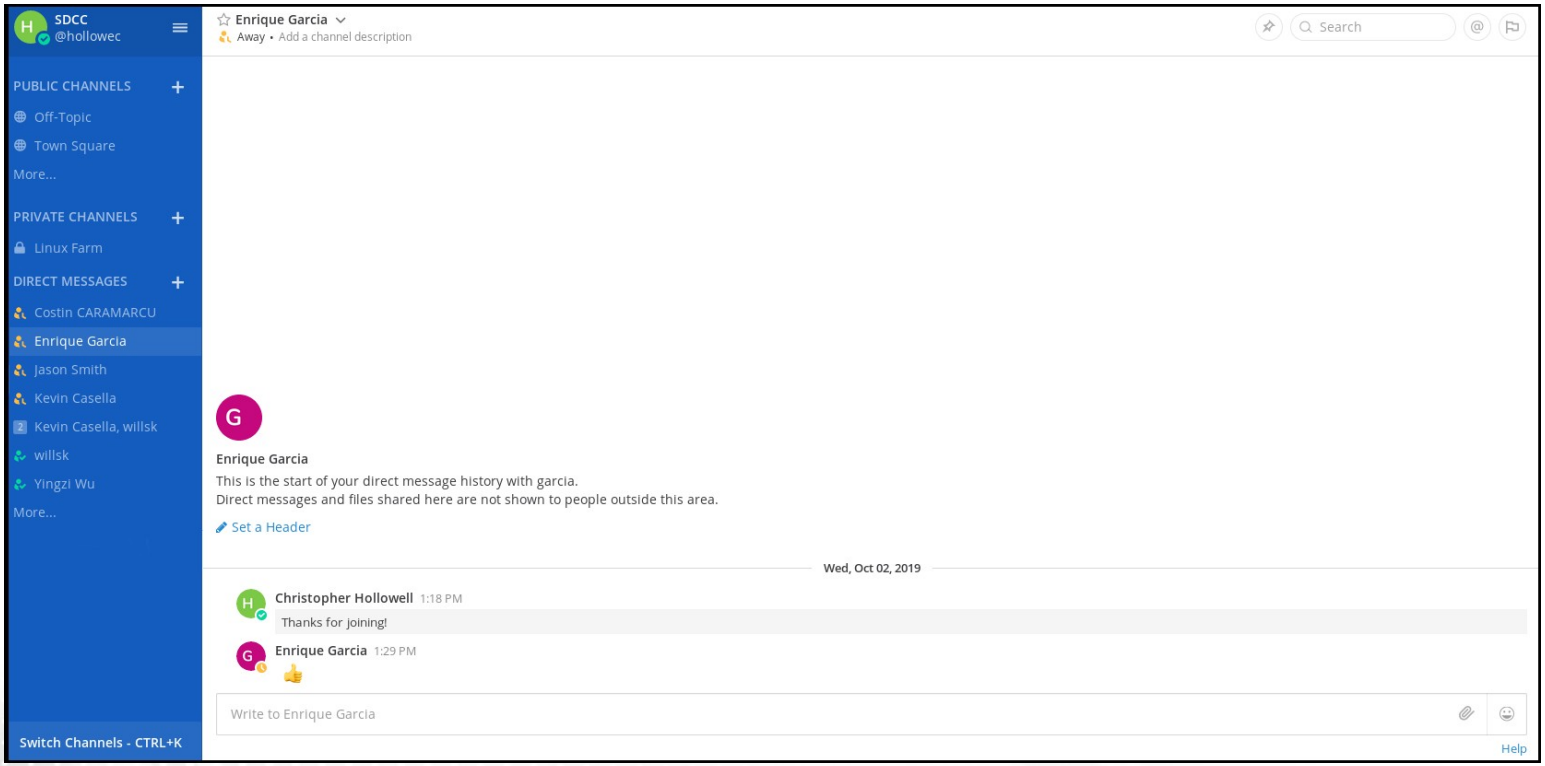
SDCC staff and user community interested in a facility-wide chat service
- After some testing, chose to implement Mattermost
  - Version 5.16.0

Features
- Locally hosting the service
- End-to-end encryption
- MySQL backend
- Web client
- Available mobile phone and desktop apps
- “Teams” concept/functionality
- Ability to create private channels
- Authentication via our Keycloak IDP
  - Users can invite others to join and create “external” accounts

Mattermost (Cont.)

System went into production in September 2019

Already being heavily utilized by SDCC staff, with growing adoption by our user community



Mattermost Web Client Interface

SDCC Talks @ HEPiX Fall 2019

HEP Workload Benchmarks: Design/Development
Chris Hollowell
Tuesday Oct. 15 @ 17:00

Federated ID/SSO at BNL's SDCC
Mizuki Karasawa
Thursday Oct. 17 @ 9:25

Integration & Optimization of BNL Storage management
Iris Wu
Thursday Oct. 17 @ 17:00

Questions?

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

Costin Caramarcu, Tim Chou, John De Stefano, Mizuki Karasawa, Carlos Gamboa, Hiro Ito, Tejas Rao, Ofer Rind, Jason Smith, Will Strecker-Kellogg, Tony Wong, Iris Wu, and Alex Zaytsev