BNL Scientific Data and Computing Center (SDCC) Site Report

Tejas Rao <raot@bnl.gov>

HEPiX Fall 2020 – Online Workshop

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

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

- ✓ Tier 0 computing center for the RHIC experiment.
- ✓ RHIC celebrating 20 years of success. RHIC run 20 ended in Sept 2020.
- ✓ US Tier1 Computing facility for the ATLAS experiment at the LHC
- ✓ Also one of two ATLAS shared analysis (Tier3) facilities in the US.
- ✓ 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 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.

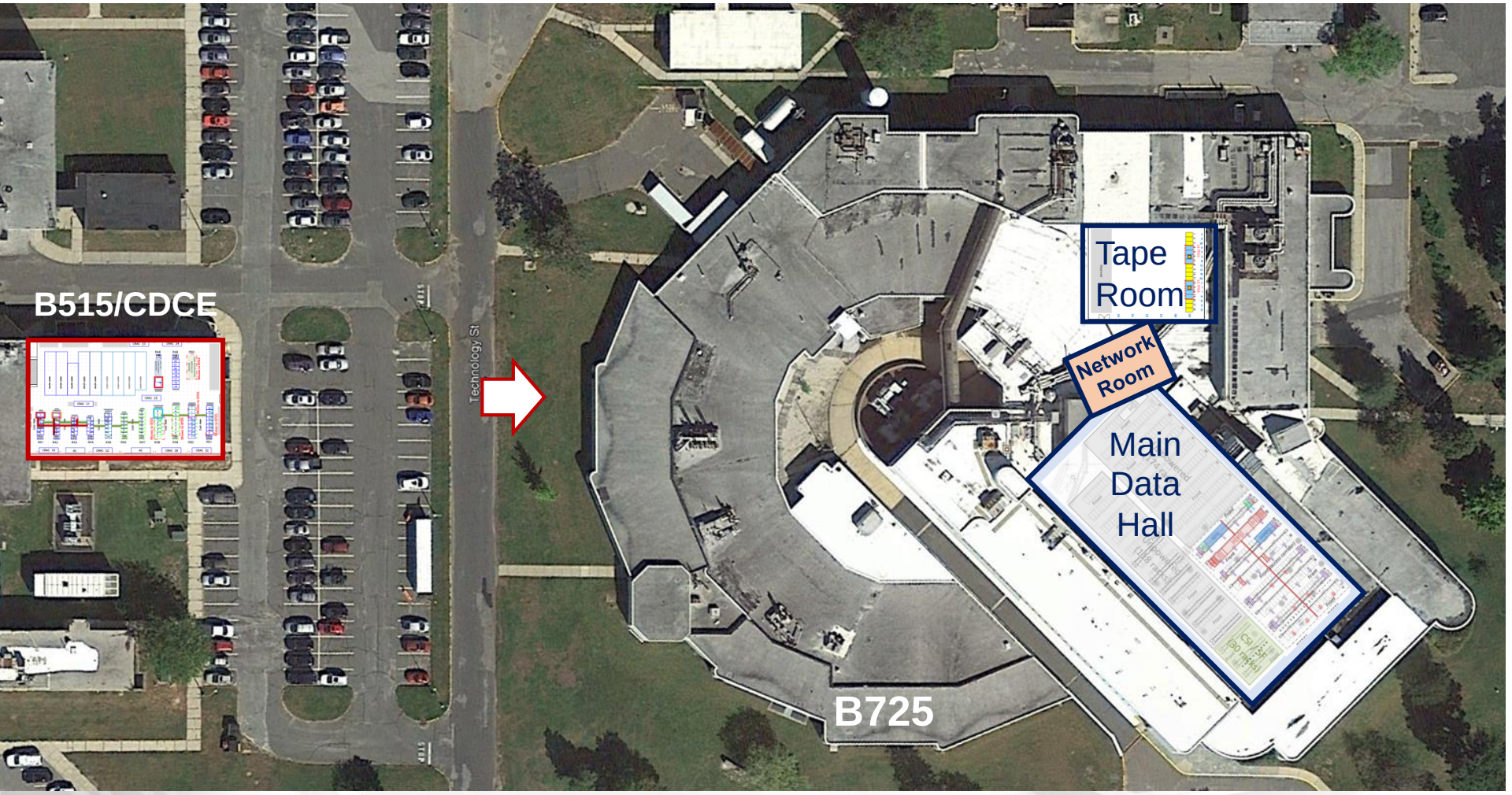
Serving more than 2,000 users from >20 projects

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, Discourse, Gitea, Mattermost, etc.

Currently employ 35 full time employees

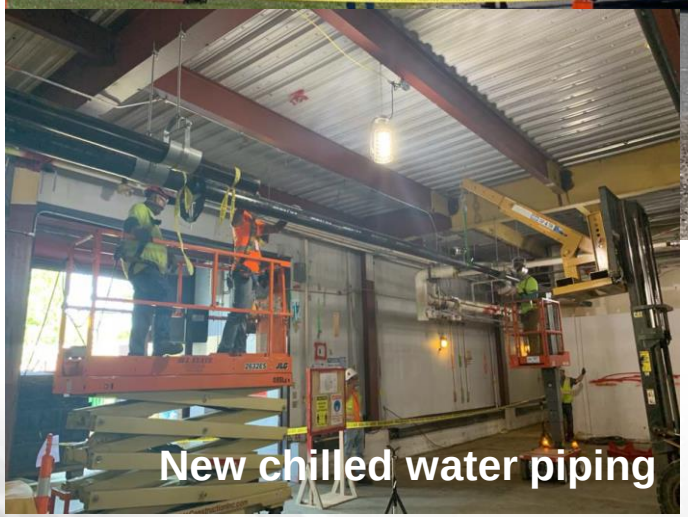
Hiring if you're potentially interested in joining SDCC
- Check https://jobs.bnl.gov for updates

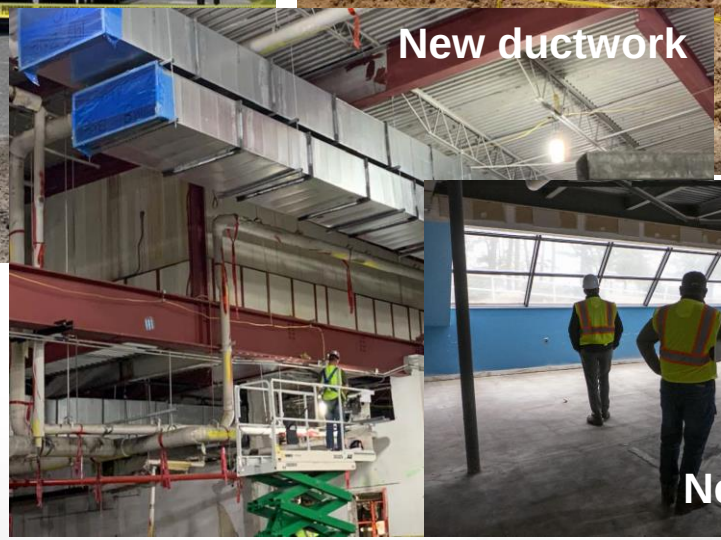
New datacenter is being designed & constructed for the SDCC Facility in B725 in FY19-21 period, with migration of most of the spinning disk storage and all of the compute capability to it from the existing B515 based datacenter to happen in FY21-23

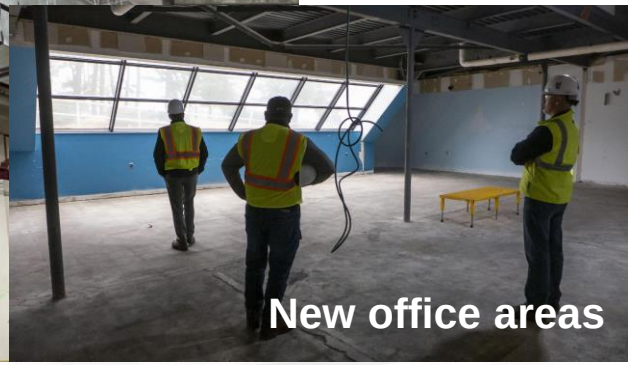


B725 Datacenter Construction

- Jul-Aug 2020: construction is going ahead after 3 months of delay in 2020Q2 due to COVID-19. The occupancy of B725 datacenter is expected to begin in Jun 2021 (early occupancy for network equipment deployment – starting from Apr 2021)





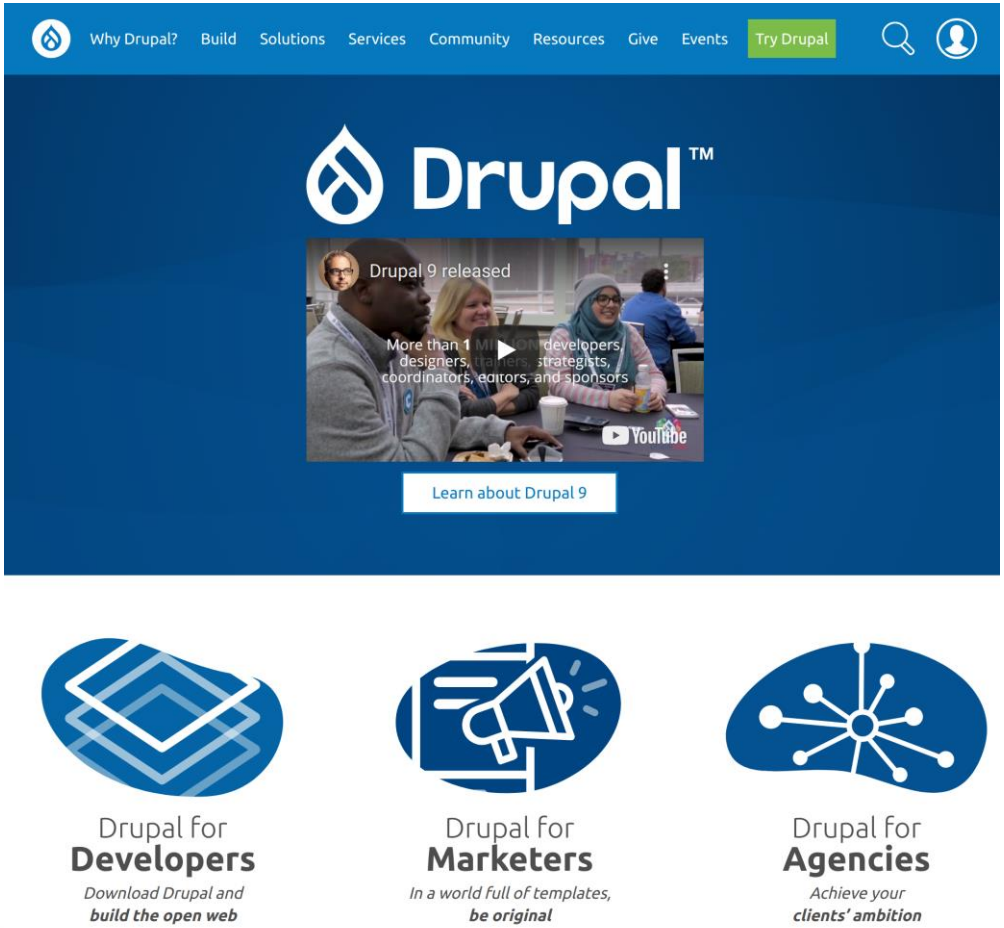


NSLS-II Activities

- NSLS-II is a user facility with 29 active beamlines and is expected to host more than 4000 users/year.
- Purchasing 30 node HPC cluster with EDR IB dedicated for NSLS-II.
- 4 PB usable Lustre storage with 2 X NetApp E5760 storage systems.
- New Ovirt virtualization cluster with 8 dedicated hypervisors able to host 100’s of virtual machines.
- Central Lustre storage will be connected directly to the beam lines.
- Dedicated Home directories for NSLS-II hosted on NetApp A300 all flash.
- Storage access will be via S3 API whenever possible and POSIX API will be used only for backward compatibility.

SDCC Drupal

- Quick deployment utilizing composer and prebuilt modules for pages, authentication, and themes
- Easily customizable from within the webui with further customization done by php development on backend
- Regularly updated with security patches with the ability to provide notifications via email to site administrators
- Updates are performed easily through composer with the added ability to keep modules at specific versions



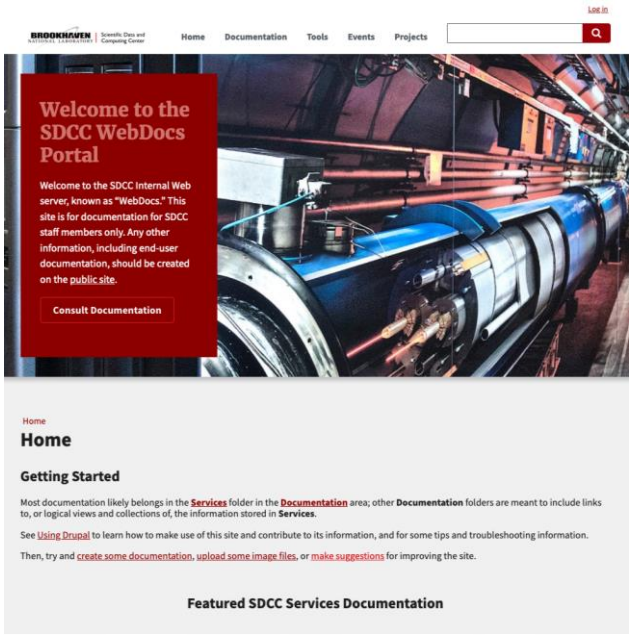
Initial Site Built with Drupal

SDCC WebDocs

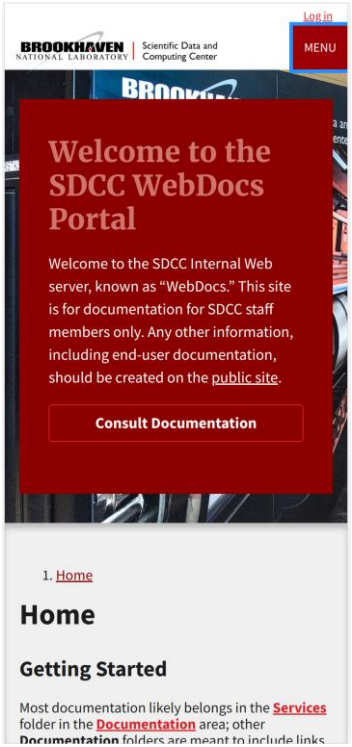
Internal documentation site available for users of SDCC computer resources.

Key features:

- Authentication easily managed through usage of Keycloak Drupal module
- Updated responsive web theme allowing for modernization and accessibility on multiple platforms
- User friendly WebUI for content creation and moderation



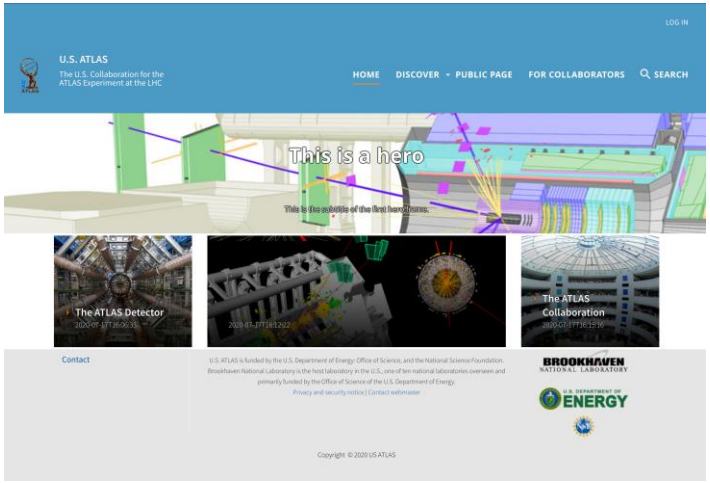
Site as seen on desktop browser



Site as seen on mobile

Customization

- SDCC custom deployment of USATLAS and sPHENIX sites



USATLAS Drupal instance

SPHENIX Drupal instance

Both sites were setup with the same deployment through puppet

BNL Box

Enterprise File Sync and Share Service (EFSS) integrated into the SDCC to provide flexible, easy-to-use, unified cloud storage for all BNL scientific users

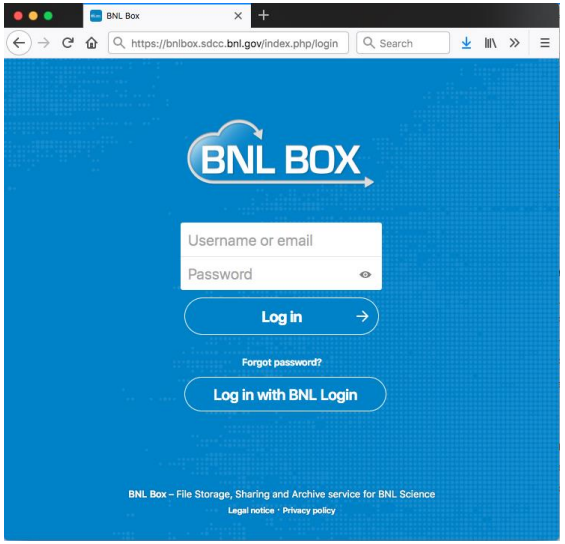
Released as a production service in Nov 2019, replacing the OwnCloud/Ceph prototype setup that had been running for the previous two years

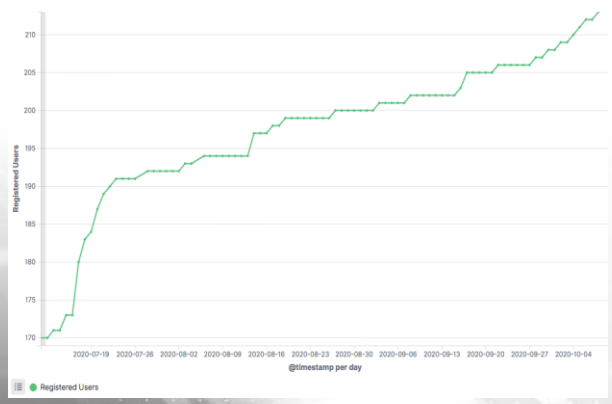
Based on Nextcloud 17 with a 1 PB Lustre FS

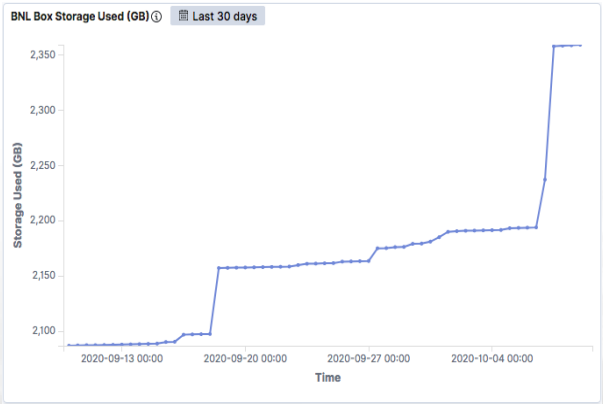
Currently 213 users, from across the BNL scientific landscape, storing 2+ TB and ~4M files

See presentation by Ofer Rind later today

https://bnlbox.sdcc.bnl.gov







ELK monitoring for SDCC services

BNL is using Elastic stack to monitor
- BNLBox (implementation of Nextcloud)
- Globus transfers

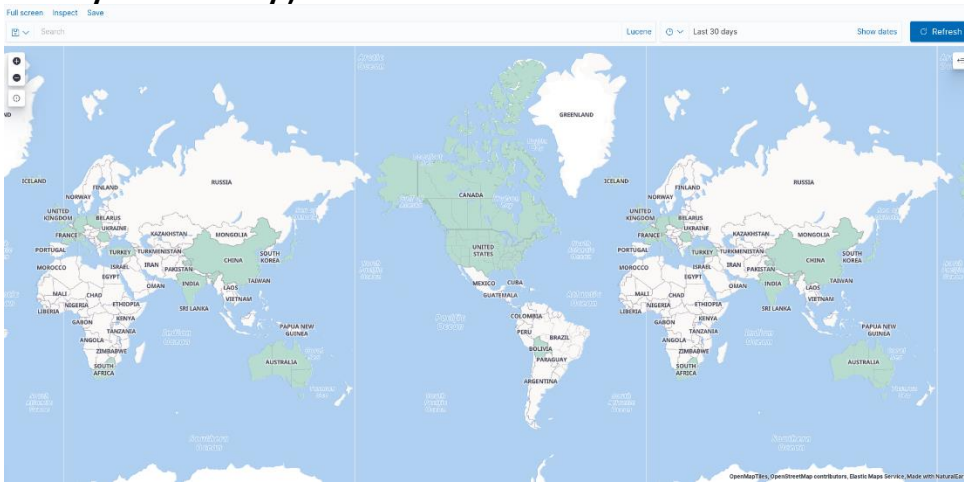
Filebeats is sending the application log and Apache log to logstash/elastic for ingestion

Metricbeats send system diagnostics of the BNLBox cluster and the elastic cluster itself
- (ex. cpu, memory, diskio, top 5 by cpu, top 5 by memory)

Some BNLBox metrics on Elastic
- current heavy users
- geo-location of user interactions

(logins, uploads, shares)
- user agent inspection

(what they're using to interact with BNLBox)



Some Globus metrics on Elastic
- geo-location of transfer destinations
- size of transfers in and out of SDCC Globus

Invenio based digital repositories

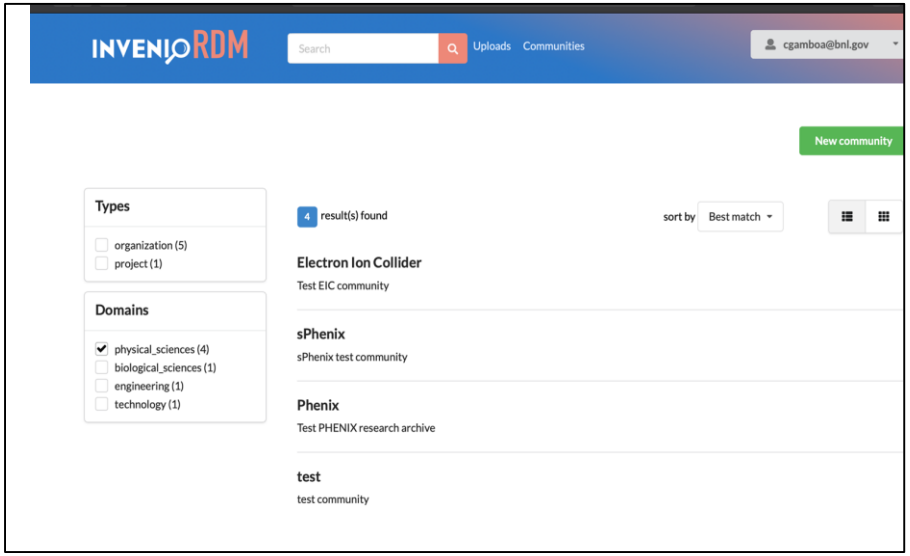
- **SET**: National Nuclear Security Administration repository integrated with DOE OneID federation for authentication
- **sPhenix digital repository:** Code in BETA release deployed in a production infrastructure
- **Electron Ion Collider (EIC):** New repository for EIC community based on Zenodo integrated with CILogon/COmanage federated authentication

invenioRDM project

BNL is contributing and it is a partner of the **invenioRDM** project:

Zenodo will be migrated to InvenioRDM once it is released

Testbed deployed using invenioRDM alpha (august 2020) release



For details please see Carlos Fernando Gamboa's talk on invenio based repositories at BNL presented later today.

High Throughput Computing

Providing our users with ~2,000 HTC nodes:
- ~75,000 logical cores
- ~855 kHS06

73 new Supermicro SYS-6019U-TR4 1U servers brought online in July 2020
- Dual Intel Xeon Cascade Lake 6252 CPUs @ 2.4 GHz (96 log. cores total)
- 12 x 16 GB (192 GB total) DDR4-2933 MHz RAM
- 4 x 1.8 TB SSDs
- 1U form factor
- 1140 HS06/node = ~83 kHS06 total

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

New Cascade Lake-based Supermicro 6019U-TR4 Servers

Upgraded to HTCondor 8.8.8
- Worked the Condor developers to resolve a group-quota issue affecting small groups

High Performance Computing

Currently supporting 4 HPC clusters

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 Omnipath 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

Purchase order out for new 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

Machine Learning Cluster

Containers

Updating to Singularity 3.6.3 on our HTC and HPC compute resources

Evaluating the use of the Podman container engine

- See Chris Hollowell's talk on "An Evaluation of Podman"
  October 15th 17:20 CEST.

Production k8s cluster deployed

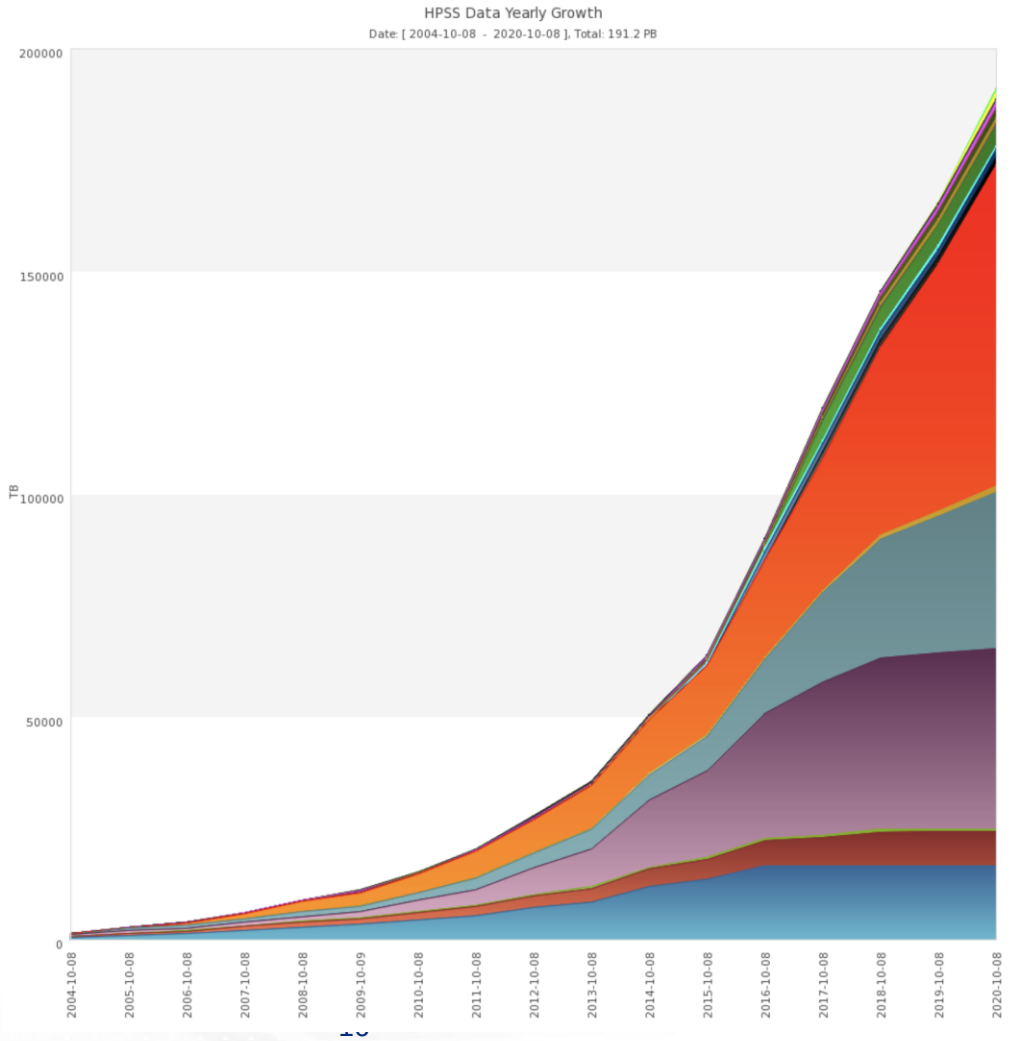
- Kubernetes v1.18
- 6 nodes with 2x10 Gbps interfaces
- For staff use

Evaluating Openshift/OKD as an alternative to k8s

- Provides an enhanced interface for deploying services/containers
- Implements some nice security features
- Could potentially be opened up to users for self-service

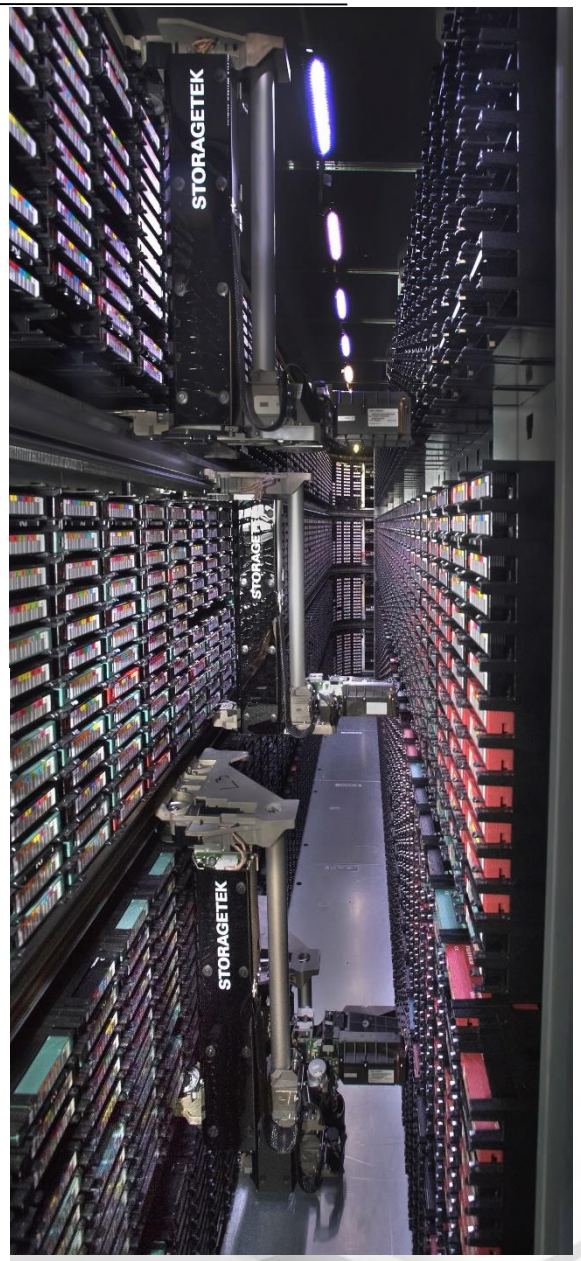
Tape Storage (HPSS)

- Running HPSS v7.4.3.2 since Dec 2018
- ~191.4 PiB accumulated data
- New data from Belle-II
- Evaluating HPSS 7.5.1
- Integrations with Lustre file system
- Upgrade of HPSS CORE server
- New IBM TS4500 tape library arrived



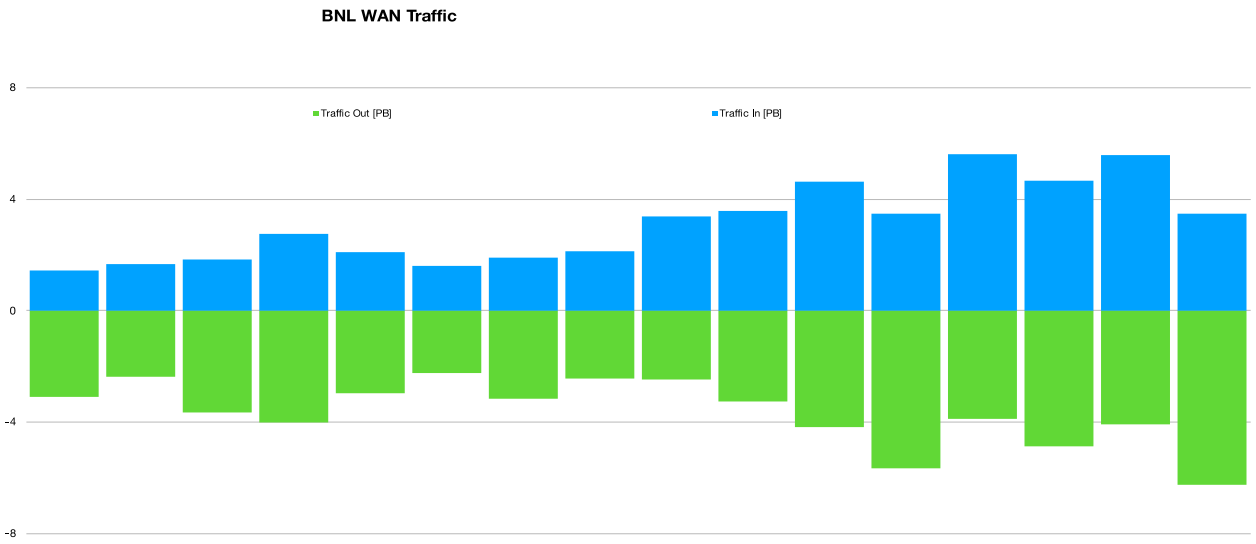
Tape Storage (HPSS)

- LOT8 deployment is in process
- Both STAR and ATLAS are still using LTO-7
- PHENIX DST remaining on LTO-6 due to lower volume
- Repacking LTO-4 to LTO-7 migration to reclaim cartridge slots for new tape technologies



Data Volume in 2019

- Data import : 20.0 PiB
- Data export : 27.4 PiB



Central Disk Storage

- Currently have 7 GPFS filesystems
  - Total of 14PB of raw storage and > 1.5 billion files
- Running version 4.2.3-15 with nearly 3,000 GPFS clients
- GPFS contract with IBM expiring in early 2021
  - Renewal price is very expensive
  - Using Lustre 2.12.5 on newer filesystems.

  - MinIO can export POSIX filesystems via S3 API.
  - Object storage has many advantages over POSIX filesystems.
  - Still in the early phases of evaluating S3 storage in favor of POSIX storage.

lustre®

dCache/XROOTD

dCache
Managing over 50 PB of data total.
Recent upgrades to v6.2 for several dCache systems, Version 6.2 supports QoS feature
Production dCache systems
- ATLAS (v6.2)
  - LAKE disk cache size increased to by 2.4PB in Sep 2020 for the total of 5PB.
  - 14 TB DMZ pools added for 3rd party HTTP/XROOTD transfer
- BELLE-II (v6.2)
  - DMZ pools added for 3rd party HTTP/XROOTD transfer
- PHENIX (v5.2)
- DUNE dCache (v5.2)
- Simons (v5.2)

QoS testbed (v6.2)
- BNL and FNAL pools

Sphenix dCache is coming

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

Running version 4.7.1

Single Sign-on and Federated Access

Joined InCommon federation platform in April 2020, enabled 2nd IdP for SDCC in BNL

Leverage CiLogon & CoManage for federated ID use cases for the projects of:

- Invenio / Zenodo instances for EIC, sPhenix, COVID-19 research
- Drupal CMS for USATLAS
- sPhenix DOMA Panda OSG Computing (alternative solution for VOMs)

Currently working on unified MFA solutions for future application authentication models.

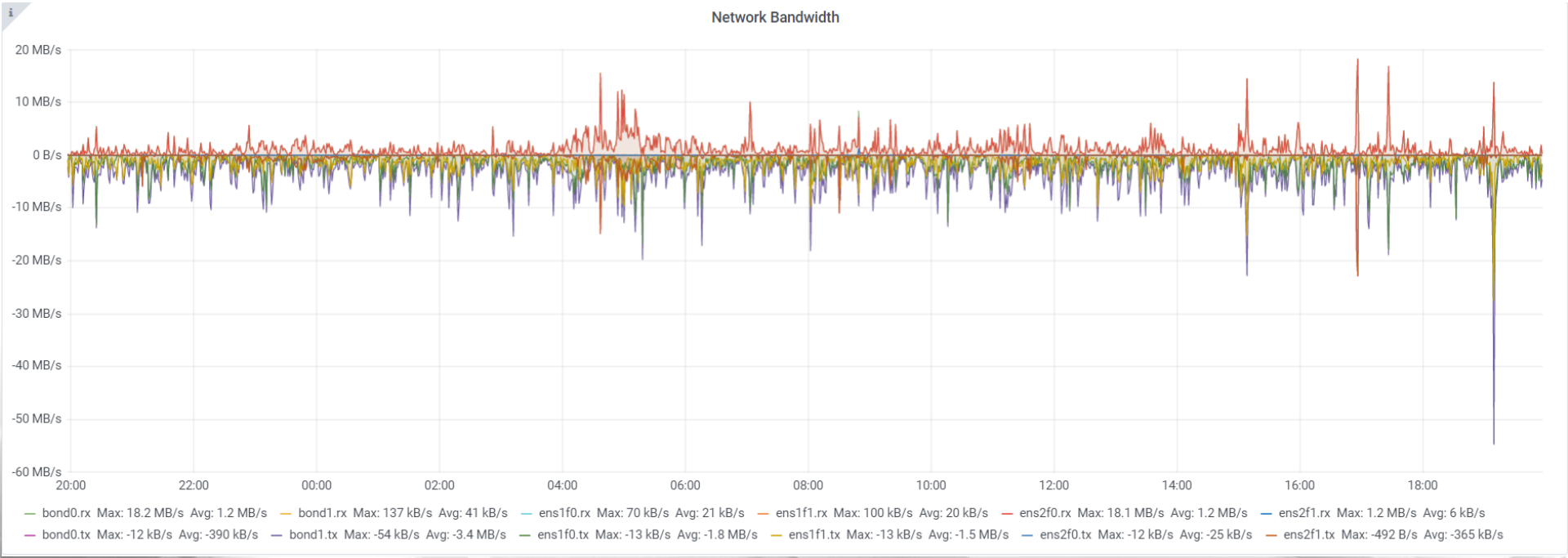
COmanage™

CVMFS

✓ All servers running version 2.7.4 (latest)

✓ Stratum One continues to grow in size & utilization
- 28 TB of data in 103 replicated repositories

✓ Stratum Zero in production
- 13 local repositories for BNL-based experiments and groups



SDCC Talks @ HEPiX Fall 2020

Invenio Based Digital Repositories at BNL
Carlos Fernando Gamboa - Monday Oct. 12 @ 18:40

A New CMS for a New Decade: Content Management at SDCC
Christian Lepore / Louis Pelosi - Monday Oct 12 @ 19:00

The BNLBox service at the SDCC
Ofer Rind - Monday Oct 12 @ 19:40

Federated Identity Management at BNL
Shigeki Misawa - Tuesday Oct 13th 18:45

HPSS migration to IBM tape technologies
Tim Chou - Wednesday Oct 14th 17:00

Watching File Storage and Transfers with Elastic Stack at BNL
Matthew Snyder - Wednesday Oct 14th 19:00

Cloud computing to support experiment online computing from the data center.
Shigeki Misawa - Thursday Oct 15th 17:00

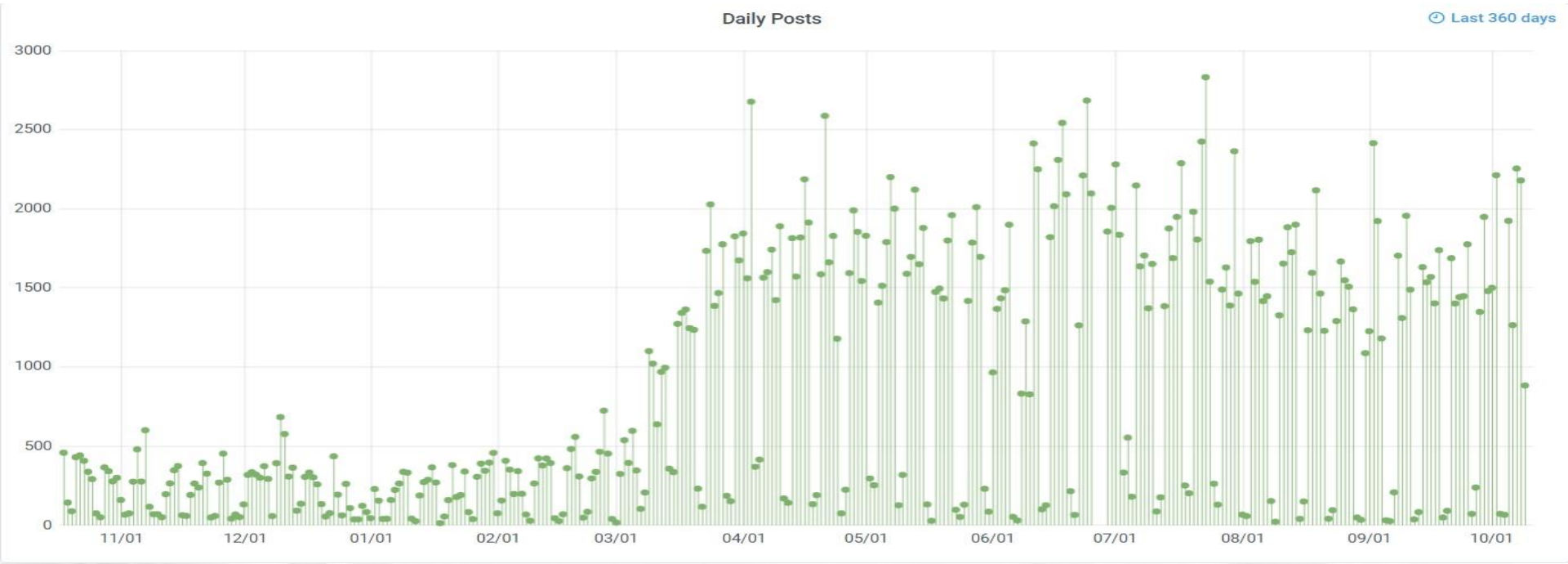
An Evaluation of Podman
Christopher Hollowell - Thursday Oct 15th 17:20

All times are in CEST (Paris timezone)

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, Matt Snyder, Louis Pelosi, Chris Hollowell, Christian Lepore, Jane Liu and Alex Zaytsev



Mattermost Chat usage after lockdown