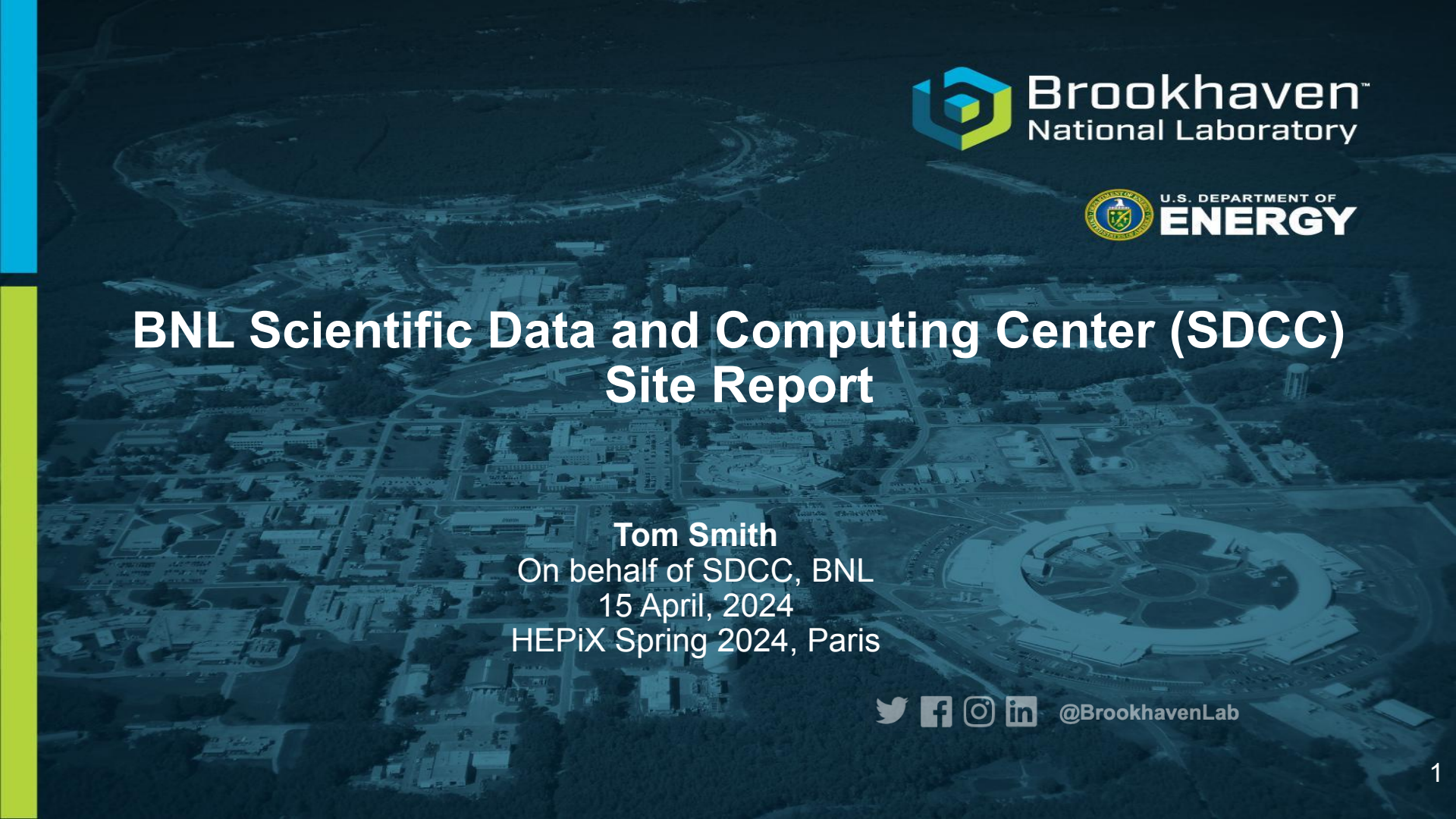


BNL Scientific Data and Computing Center (SDCC) Site Report

Tom Smith
On behalf of SDCC, BNL
15 April, 2024
HEPiX Spring 2024, Paris

@BrookhavenLab

Scientific Data and Computing Center Overview

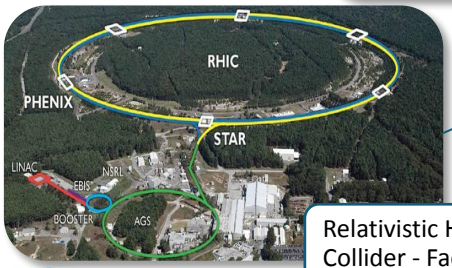
- Located at Brookhaven National Laboratory (BNL) on Long Island, New York
- Tier-0 computing center for the RHIC experiments
  - sPHENIX, STAR
  - BNL is host site for the future Electron-Ion Collider (EIC)
- US Tier-1 Computing facility for the ATLAS experiment at the LHC
  - Also one of the ATLAS 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

We are hiring!
Details at the end

2024Q1

National Synchrotron Light Source II, CryoEM

Center for Functional Nanomaterials



Relativistic Heavy Ion Collider - Facilities

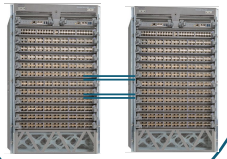
Brookhaven National Laboratory

0.8 Tbps

0.4 Tbps

0.7 Tbps

Science DMZ



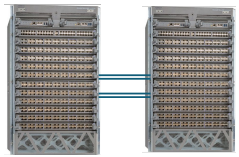
2x800 Gbps

0.2 Tbps

1.2 Tbps

DTNs
(52 systems)

High Throughput Science Network



400 Tbps capability
(60 Tbps deployed)

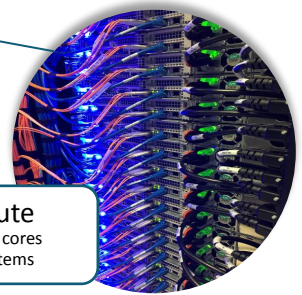
1.2 Tbps

12 Tbps

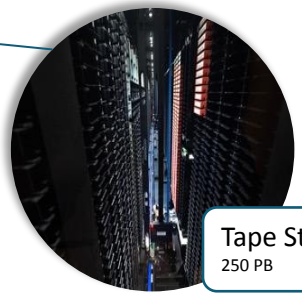
0.5 Tbps

18 Tbps

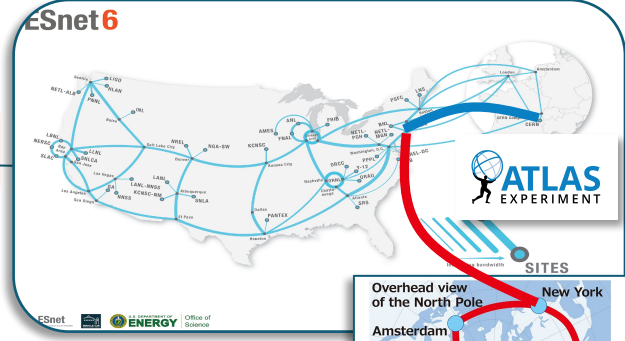
Disk Storage
125+ PB



Compute
140k CPU cores
5 HPC systems



Tape Storage
250 PB



CVMFS

All server & replica services running version 2.11.2 (latest)

- WN clients to be upgraded via migration to EL9 (Summer)

Server (Stratum Zero):

- Stratum Zero for 13 locally hosted repositories
  - Occupying **1.4** TB NAS disk
  - For facility and experiment use:
    - ASTRO, ATLAS, DayaBay, DUNE, EIC, OSG, PHENIX, SDCC, sPHENIX, STAR
    - Recently added sPHENIX calibration data publishing

Replica (Stratum One):

- Replication of 119 WLCG & OSG repositories
  - From 10 domains and four main sources (BNL, CERN, OSG, RAL)
  - Occupying **49 TB** NAS disk

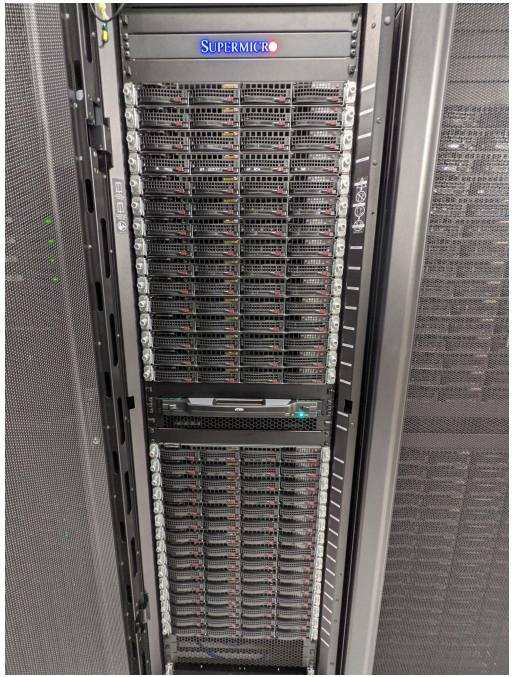
CVMFS [2]

Due to aging server HW, JBOD storage migrated to fully virtual environment

- RHEV VMs for server, replica, site caches & reverse proxies
- Server, replica storage migrated to NetApp A400 NAS
- Plans to move existing volumes on shared NetApp to dedicated CVMFS NAS

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

Evaluations for Linux Farm Procurement

- Supermicro Jumpstart Program Remote Testing:
  Dual-6448Y+ CPU: HEPscore **2219**; Peak Power **900** W
  (Sapphire Rapids)
- Supermicro Lab Remote Testing:
  Dual-6530 CPU: HEPscore **2295**; Peak Power **805** W
  Dual-6548Y+ CPU: HEPscore **2425**; Peak Power **728** W
  (Emerald Rapids)
- On-prem Evaluation Systems ETA soon™:
  Dual-6548Y+ CPU, Dual-6538Y+ CPU, Dual-6448Y+ CPU

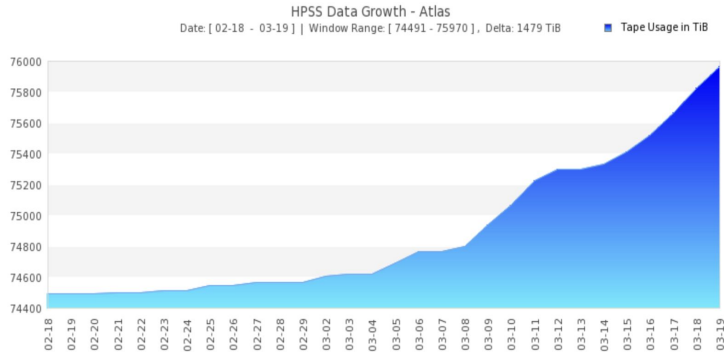
Alma 9 readiness

- Transition from **SL7** to **Alma Linux 9** coming soon™
  - Targeting upgrade of BNL ATLAS T1 condor pool first, then BNL shared condor pool
  - Porting and testing SL7 puppet v3 code to support Alma 9 and puppet v7+
  - IPv6
    - Some SL7 nodes have been migrated to dual stack for testing
    - For most of the hosts, IPv6 will be added at the time of Alma 9 rebuild

- Upgrade will be in rolling fashion: SL7 nodes rebuilt into Alma 9 in batches
  - Use of new provisioning infrastructure (foreman/puppet)

TAPE @ BNL

- 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**
- ATLAS staged **28.7PB** (7,898,544 files) in 2023, Injected **11.5PB** (5,633,412 files)
- sPhenix injected **11.6PB** in 2023
  - Two new IBM **TS4500** libraries installed (**16,000** slots)
  - **64** LTO9 drives
- Star injected **5.5PB** in 2023
  - 5 new LTO8 drives install (18 LTO8 drives total)
- Belle2 injected **0.3TB** (500 files) and staged **708.2TB** (740,826 files) in 2023
- Tape resources installed for EIC and QCD (data injection not yet started)
- Approx. **8,000** LTO5 tapes repacked to LTO8, **7,500** library slots freed up



Brookhaven National Laboratory

Lustre

- sPHENIX Lustre expansion
  - Before Expansion: 55 OSS (Object Storage Server), ZFS, Lustre 2.15.2, RHEL8.7
  - **After Expansion**: **+23** OSS, **~69.1** PiB total storage, ZFS, Lustre 2.15.2, RHEL8.7.
  - **Ongoing**: old/new OST (Object Storage Target) rebalancing for optimized performance.
- NSLS2 Lustre expansion
  - The new **6 OSSs** (~3.7PB) has been added
  - All OSS are configured with **HA**(pacemaker+ cornsync+fence) to ensure the high availability
- New SciServer Lustre Deployment
  - **Setup**: 1 MDS (2TB, ZFS, Lustre 2.15.4, AlmaLinux 8.9) + 6 OSS (5.5 PB total, ZFS, Lustre 2.15.4, AlmaLinux 8.9).
  - **Features**: dRAID for ZFS, DoM (Data on MDT) for small files.
- Generic Lustre Puppet Module Development
  - Monitoring (Barreleye, Lustre_exporter, Node_exporter, LOKI)
  - Others (OSS, MDS, Firewall rules, admin scripts)
- Whamcloud support for sPHENIX Lustre

Major dCache activities

- Towards dCache SE multi-instance architecture
  - Reconfiguration in **HA** mode to "minimize single points of failures and enable rolling upgrades and, in some cases, horizontal scalability", cf. HA dCache Services
- Refactoring puppet code for dCache administration
  - SDCC puppet transition infrastructure evolving from **puppet 3** to **puppet 8**
  - dCache related puppet modules ported to **puppet 8**
  - New effort in refactoring dCache puppet classes for a multi-instance deployment
    - Common puppet class to manage all experiments
- ATLAS Hardware lifecycle
  - Ongoing data migration from 24 hosts supporting ~22PiB total space presented to the dCache. This space is presented in 576 pools.
- Data Challenge 24 preparation, with stress testing and tuning, conducted with a dedicated FTS instance

InvenioRDM @ SDCC

- sPHENIX InvenioRDM:
  - New updates to improve user experience
    - Customized theming
    - WIP to develop automated author list
- EPIC InvenioRDM:
  - New instance developed for the EPIC group
  - Based on the sPHENIX deployment
  - Complete with a customized theme
- Continued development of both container deployment and native services deployment

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

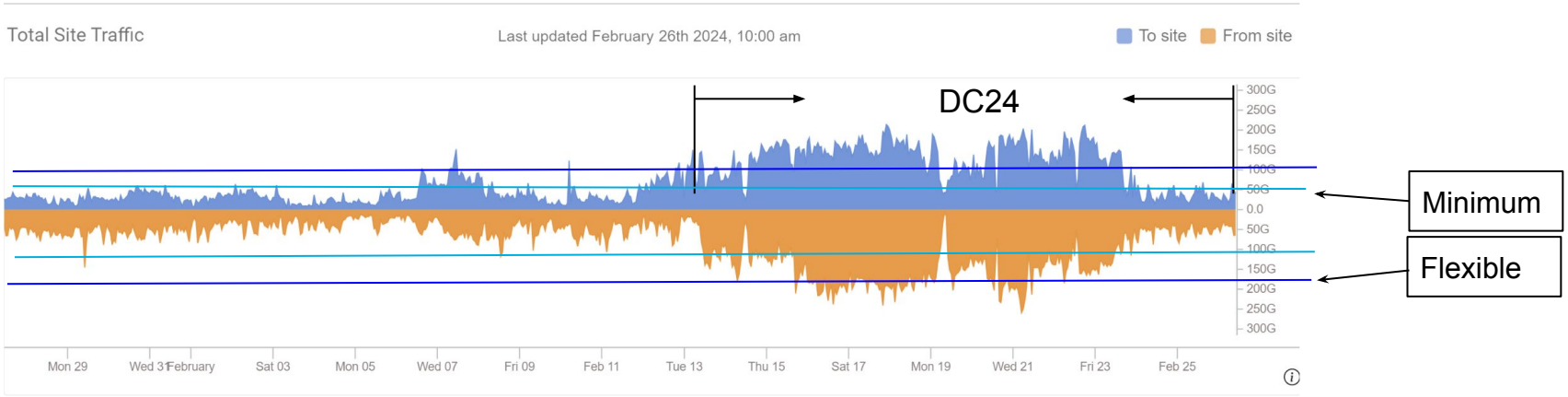
COmanage Registry

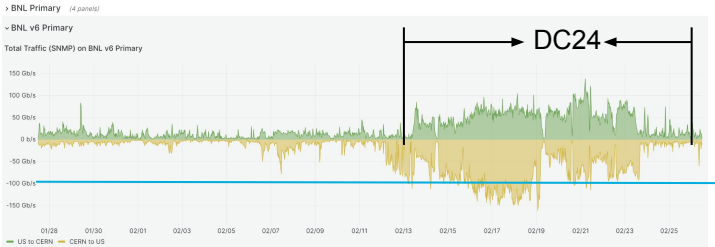
- Aggregate multiple identities so BNL services see only one.
- 354 identities currently aggregated to 307 unique users
- 5 production OIDC clients serving 26 unique service instances
- Service authorization can be controlled by active IDP and group membership
- Services are being added/converted as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

WLCG Data Challenge 24 (DC24)

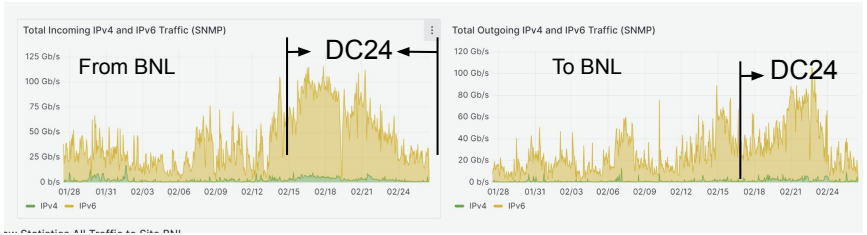
- BNL was involved with three experiments: ATLAS, Belle, and DUNE
- BNL met all of the peak transfer target goals
- > 200 Gb/s in and out reached
  - During joint USATLAS-USCMS pre DC24 tests between BNL and Univ. of Chicago and Univ of Michigan — utilizing H. Ito's testing suite
  - Discovered that we can not saturate the BNL network pipes
- Results of DC24 are being analyzed to identify bottlenecks in our infrastructure

DC24: BNL WAN Network Load

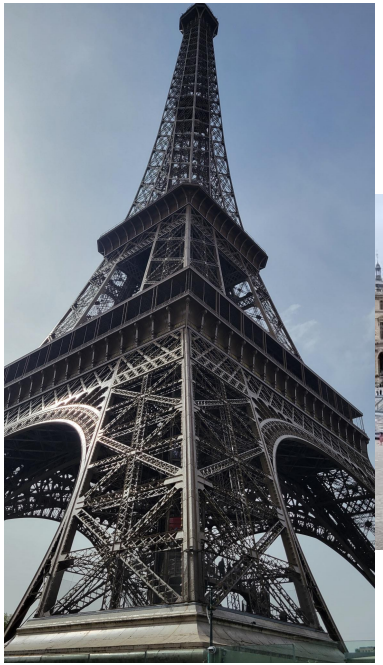


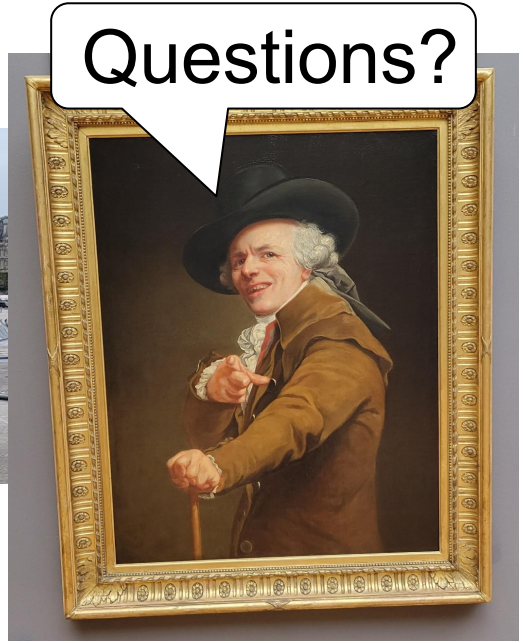


LHCOPN (T0 and T1s)



LHCONE





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.

Brookhaven National Laboratory

SDCC at BNL is hiring!

- First posting
  - https://jobs.bnl.gov/job/upton/senior-technology-engineer-technology-architect/3437/58652677120
- Second posting
  - https://jobs.bnl.gov/job/upton/technology-engineer-advanced-technology-engineer/3437/58652677296
- 3rd opening to be posted soon!
- For more details, please contact Ofer (rind@bnl.gov) or Tony (tony@bnl.gov)

Extra slides

ATLAS DC24 Transfer Target Rate

ATLAS DC24 transfer rates

(final version: 20240207)

red rate above the bandwidth available to ATLAS **red** rate > 80% of bandwidth available to ATLAS

grey text color: sites not participating in DC24 (only yearly ingress/egress averages)

Deletion rates are calculated from injected ingress DC24 bandwidth assuming 3GB average filesize

Table: DC24 (src)			Site WAN (Gb/s)		DC24 minimal scenario				DC24 flexible scenario				FTS active inbound / outbound
			Total	Usable by	T0	Total Gb/s & bandwidth		Space [TB/24h]	T0	Total Gb/s & bandwidth		Space [TB/24h]	
Site	Tier	Cloud	(Gb/s)	ATLAS	Export	∑ ingress	∑ egress	(deletions/hour)	Export	∑ ingress	∑ egress	(deletions/hour)	
CERN-PROD	T0	CERN	2100	891	257.0	23.4	282.5	246 (0)	257.0	88.9	392.8	937 (9825)	454 / 2037
T0 summary					257.0	23.4	282.5		257.0	88.9	392.8		
BNL-ATLAS	T1	US	400	400	60.0	88.9	67.1	938 (10719)	60.0	119.8	124.9	1263 (15342)	719 / 851
FZK-LCG2	T1	DE	400	144	32.0	58.7	35.6	619 (6637)	32.0	92.9	65.5	980 (11768)	473 / 410
IN2P3-CC	T1	FR	400	177	38.0	62.9	43.0	663 (7248)	38.0	93.6	77.8	986 (11849)	543 / 429
INFN-T1	T1	IT	240	62	23.0	38.2	26.0	402 (4470)	23.0	61.2	46.1	645 (7920)	230 / 209
NDGF-T1	T1	ND	200	149	15.0	16.6	23.3	175 (0)	15.0	95.6	33.7	1008 (11856)	593 / 106
SARA-MATRIX	T1	NL	400	238	15.0	32.6	16.5	343 (3575)	15.0	60.1	30.2	634 (7708)	164 / 139
pic	T1	ES	200	85	11.0	18.0	12.5	190 (2097)	11.0	29.1	20.9	306 (3750)	141 / 150
RAL-LCG2	T1	UK	400	177	38.0	67.7	40.3	714 (7177)	38.0	92.8	81.0	978 (10936)	1595 / 663
TRIUMF-LCG2	T1	CA	100	100	25.0	40.1	27.8	423 (4726)	25.0	60.0	50.9	632 (7704)	322 / 434
T1 summary					257.0	423.8	292.0		257.0	705.0	530.9		

Minimal: modified hierarchical model T0 → T1 ⟷ T1 → T2
Flexible: mesh of transfers T0 ⟷ T1 ⟷ T1 ⟷ T2 ⟷ T2 ⟷ T0