Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

dCache at BNL

Carlos Fernando Gamboa on behalf of the BNL Storage / dCache Team
Brookhaven National Laboratory

17th International dCache workshop 2023, Berlin, Germany 2023

@BrookhavenLab

Outline

- Overview to dCache based storage services
- Toward an improved dCache service
- Challenges and future work

Brookhaven National Laboratory

Storage Services at BNL SDCC

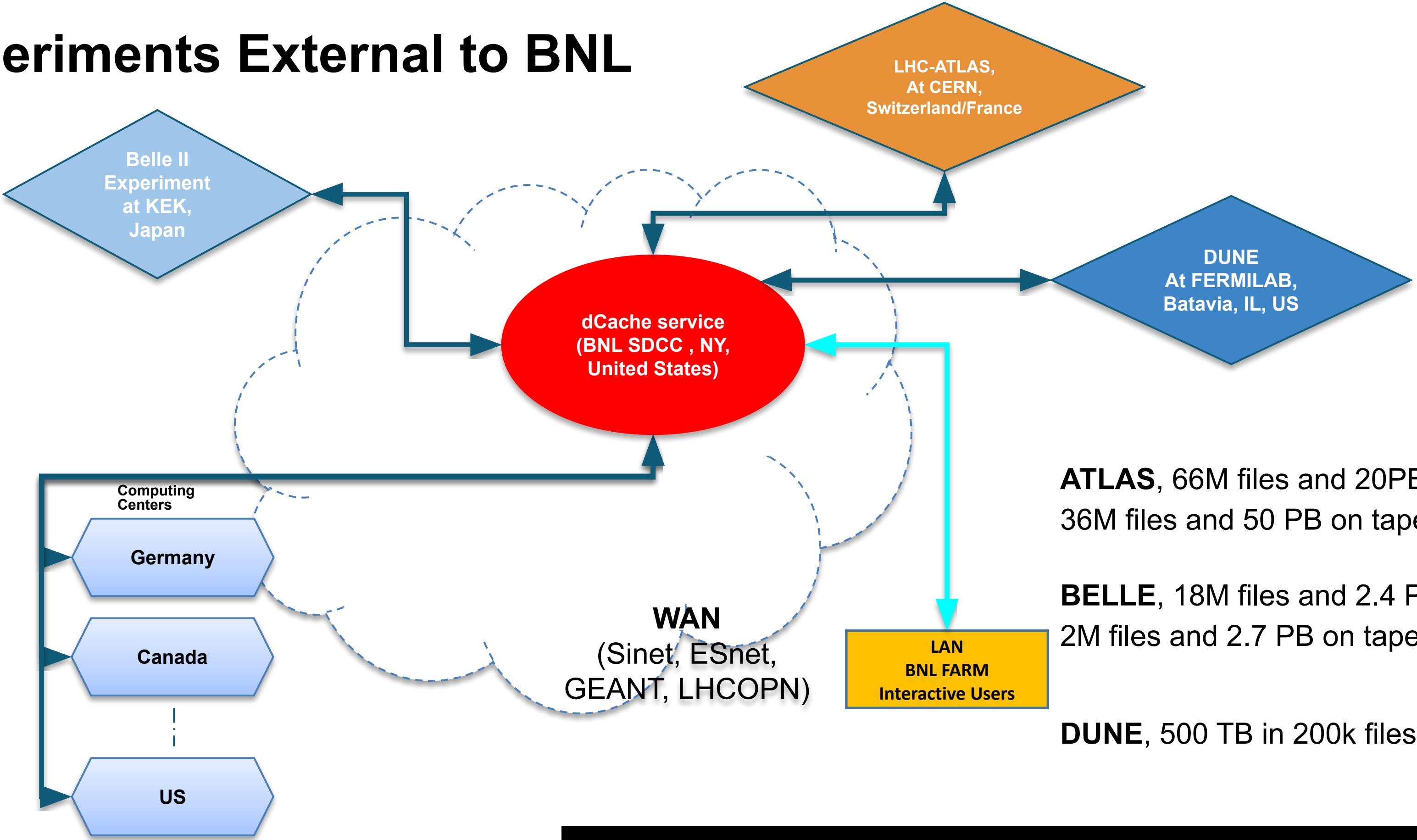
- BNL SDCC supports different storage services for a variety of Scientific Communities (SC) like NSLSII, Nuclear and High Energy Physics
- Diverse storage technologies are used to support the communities: dCache, Lustre and GPFS, please see past HEPIX 2023 BNL site report for specifics
- This talk will concentrate on **dCache storage** technology
    - dCache services for LHC-ATLAS, BELLE2 and DUNE SC store and manage 76PBs (30% DISK) of data distributed in around 122M files
    - Scientific Community data is produced outside BNL:
        - CERN (Switzerland/France),
        - KEK (Japan),
        - Fermilab(IL,US)

Who We Are

dCache Service Application and Operation Support:
Carlos Fernando Gamboa
Vincent Garonne
Qiulan Huang
Matt Snyder (newcomer)

Rob Hancock (Hardware)
Kevin Casela (Hardware)

Experiments External to BNL


LHC-ATLAS, At CERN, Switzerland/France
Belle II Experiment at KEK, Japan
DUNE At FERMILAB, Batavia, IL, US
dCache service (BNL SDCC , NY, United States)
Computing Centers
Germany
Canada
US
WAN (Sinet, ESnet, GEANT, LHCOPN)
LAN BNL FARM Interactive Users

ATLAS, 66M files and 20PB on disk, 36M files and 50 PB on tape

BELLE, 18M files and 2.4 PB on disk, 2M files and 2.7 PB on tape

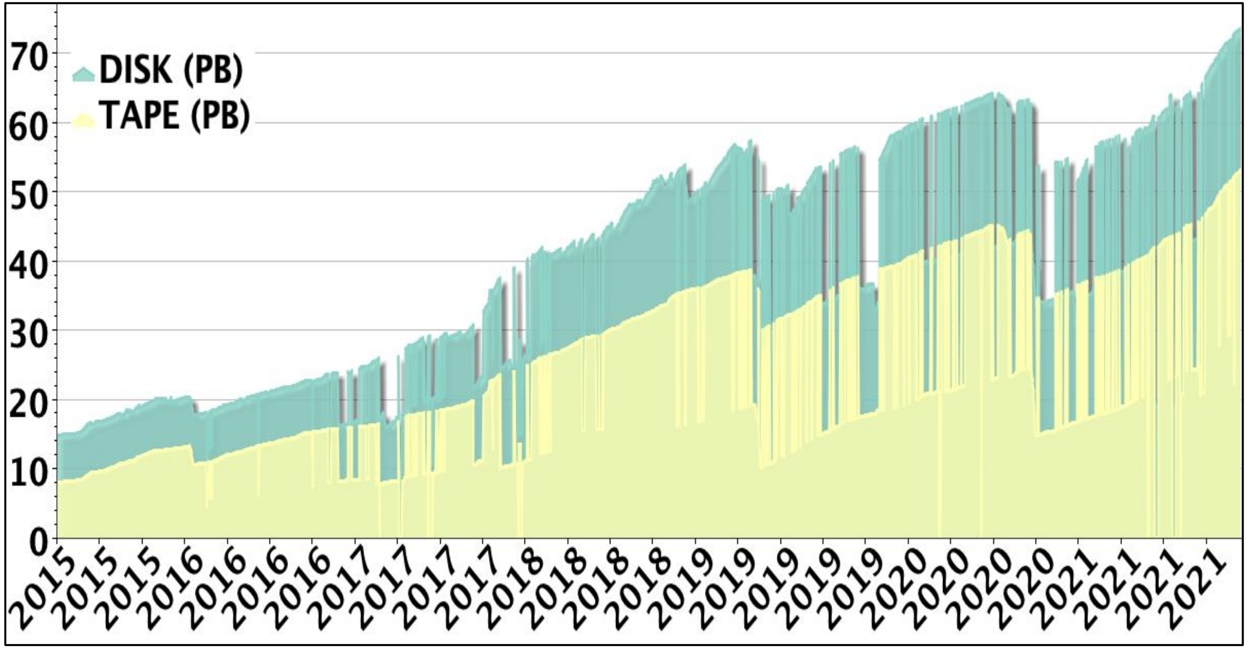
DUNE, 500 TB in 200k files on disk

ATLAS SC community driving the storage usage compared to other HEP SC supported at BNL

Evolution of Atlas SC storage

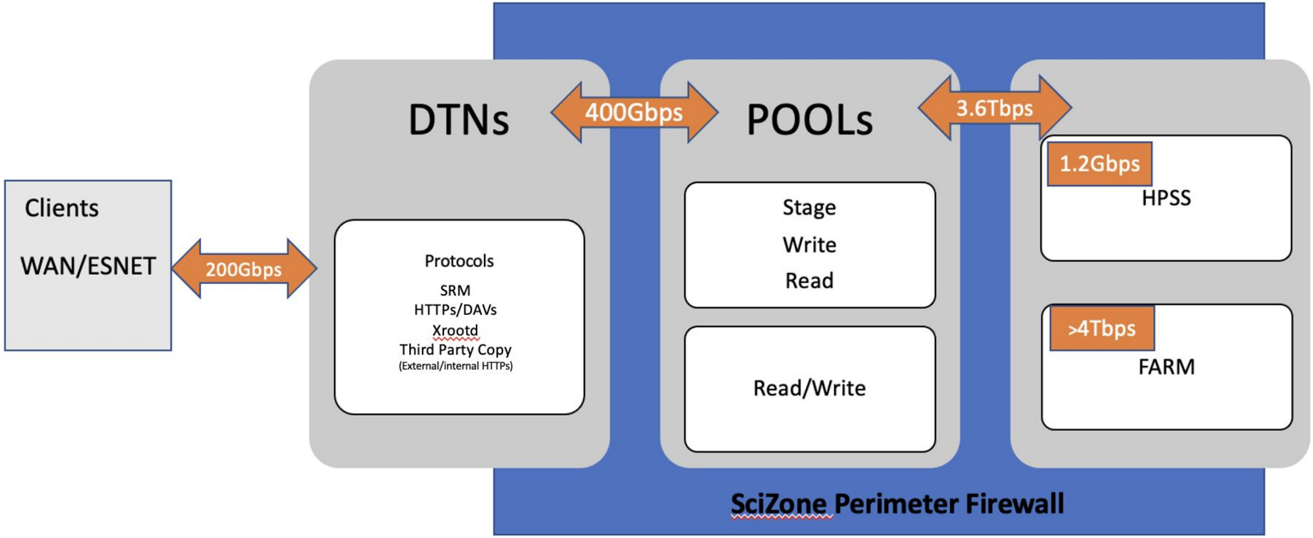
BNL provides more than 70PB of storage and hosts 100M files for ATLAS

We observe a factor of 7 in the past seven years for the total space.



The main challenge coming is HL-LHC and with the simple model of 3 to 4 order of magnitude increase in 10 years from now: 1B files, 700 PB, 300Mhz, 5-7PB/day

dCache General Layout (ATLAS)



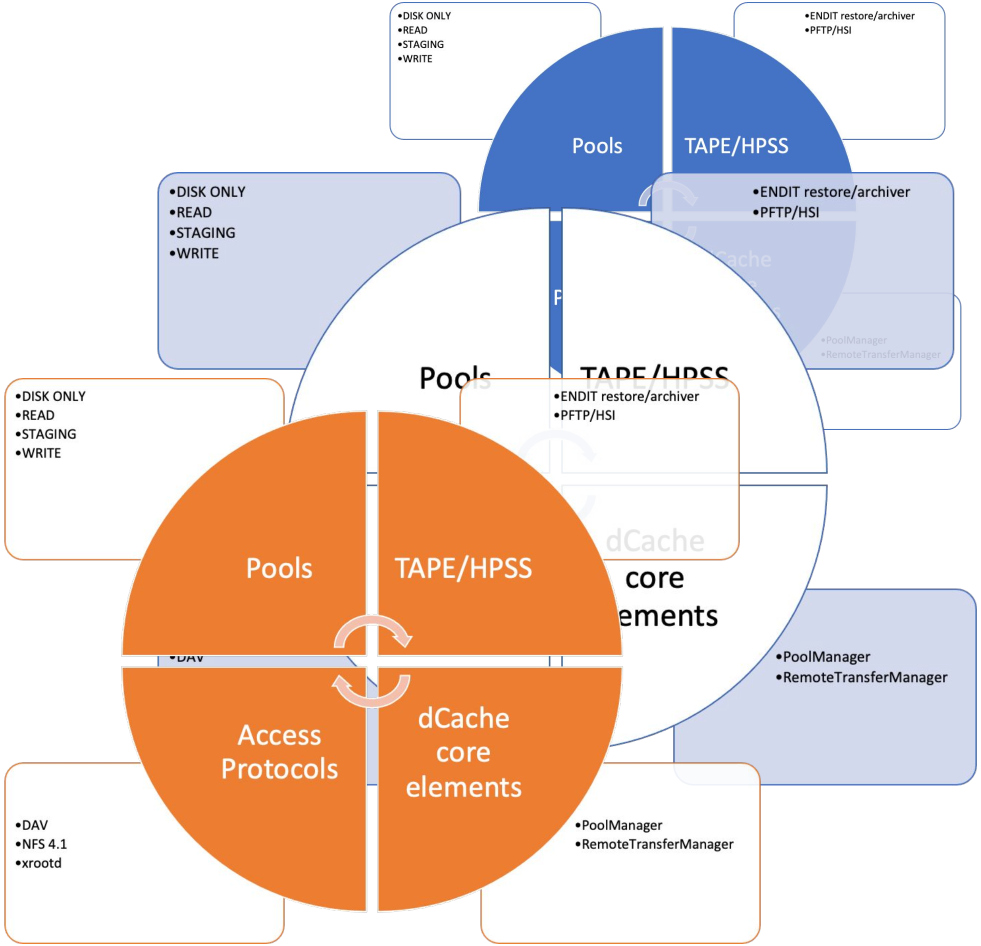
Comply with BNL cybersecurity policy disaggregation among external and internal resource accessibility

Reference deployment to be used as building block for other SC dCache.org

distributed storage for scientific data

dCache instances are isolated per SC

- SC diverge in their requirements
- Procurement and resource control
- Infrastructure supported on physical and virtual Machines



Storage technology flexible to support SC individual requirements

Towards an Improved dCache Operation

Areas of work:

- Enhancing software for interaction among dCache and TAPE HPSS systems
  - ENDIT archiver/retriever
- Improving dCache data access workflows for client access
  - Non firewalled Xrootd client access for write/read
  - DUAL IPv4/IPv6 dCache application stack configuration
- Extending monitoring for dCache operations
- Evolving dCache along with infrastructure

Improving Software to Interact with dCache and TAPE

ENDIT archiver/retriever

- Previous mechanisms used to instantiate restores from HPSS relied heavily on polling the dCache Poolmanager
- Stability of Poolmanager component at risk when > 100k concurrent requested restores
- Since ENDIT retriever adoption, **no more Poolmanager stability issues were observed**, more than 140k concurrent restore requests without any issue
- Successful adoption of ENDIT retriever permitted the extension of usability for writing interactions to HPSS
  - Allowed consolidate legacy software/code for writing to HPSS

Extended overview covered on this talk

Non Firewalled Xrootd Client Access for Write/Read

Standard xrootd client transfers involve pool redirections among client and dCache service

- Accessibility to clients outside BNL to pools is not permitted

Support for xrootd in proxy mode released on dCache 8.2.2

- Proactive functional test work along dCache Developers (Al Rossi et al.)
- First enabled on DUNE dCache to READ/WRITE via xrootd
- Later on successfully integrated on ATLAS dCache instance (8.2.15)
  - Xrootd standalone servers used to front dCache xrootd to provide xrootd external READ (ATLAS) decommissioned

ATLAS DUAL IPv4/IPv6 dCache Stack Configuration

Latest dCache upgrade (8.2.15) permitted to:

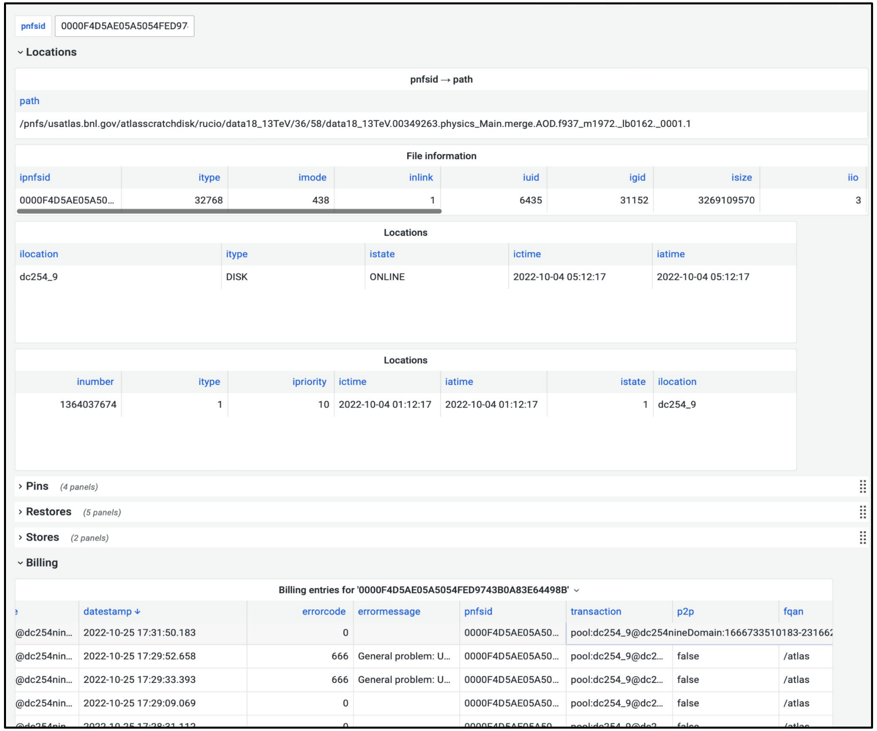
- Utilize dual-stack network infrastructure deployed on different components (doors, core, and pools)
- Configure the dCache stack to be able to support client requests on IPv6 and IPv4 in dual networks:
    - dCache doors configured to support different client accessibility
        - Clients internal to BNL LAN supporting only IPv4 or IPv6 (no proxy access)
        - Clients external to BNL proxied access for IPv6 and IPv4
    - **The use IPv6 when transferring data between two dual-stack machines for HTTP-TPC transfers**

Monitoring is key to help identify different data workflows

Monitoring Enhancement

Grafana based monitor using the dCache billing/chimera/srm databases to provide information use in operations

Allows aggregate information from different dCache events by entering the PNFSID (dCache file ID)



pnfsid 0000F4D5AE05A5054FED97

⌄ Locations

pnfsid → path

path
/pnfs/usatlas.bnl.gov/atlasscratchdisk/rucio/data18_13TeV/36/58/data18_13TeV.00349263.physics_Main.merge.AOD.f937_m1972._lb0162._0001.1

File information

ipnfsid	itype	imode	inlink	iuid	igid	isize	iio
0000F4D5AE05A50...	32768	438	1	6435	31152	3269109570	3

Locations

ilocation	itype	istate	ictime	iatime
dc254_9	DISK	ONLINE	2022-10-04 05:12:17	2022-10-04 05:12:17

Locations

inumber	itype	ipriority	ictime	iatime	istate	ilocation
1364037674	1	10	2022-10-04 01:12:17	2022-10-04 01:12:17	1	dc254_9

› Pins (4 panels)

› Restores (5 panels)

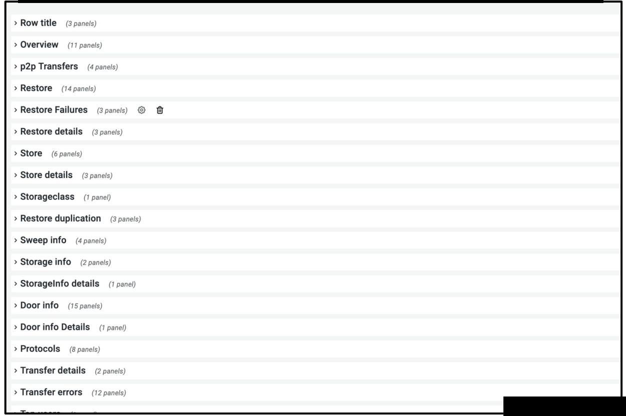
› Stores (2 panels)

⌄ Billing

Billing entries for '0000F4D5AE05A5054FED9743B0A83E64498B'

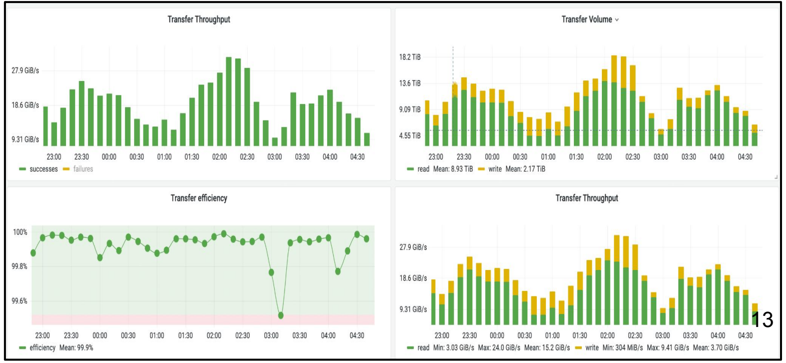
e	datestamp ↓	errorcode	errormessage	pnfsid	transaction	p2p	fqan
@dc254nin...	2022-10-25 17:31:50.183	0		0000F4D5AE05A50...	pool:dc254_9@dc254nineDomain:1666733510183-23166		
@dc254nin...	2022-10-25 17:29:52.658	666	General problem: U...	0000F4D5AE05A50...	pool:dc254_9@dc2...	false	/atlas
@dc254nin...	2022-10-25 17:29:33.393	666	General problem: U...	0000F4D5AE05A50...	pool:dc254_9@dc2...	false	/atlas
@dc254nin...	2022-10-25 17:29:09.069	0		0000F4D5AE05A50...	pool:dc254_9@dc2...	false	/atlas
@dc254nin...	2022-10-25 17:28:21.112	0		0000F4D5AE05A50...	pool:dc254_9@dc2...	false	/atlas

Feature driven dashboards



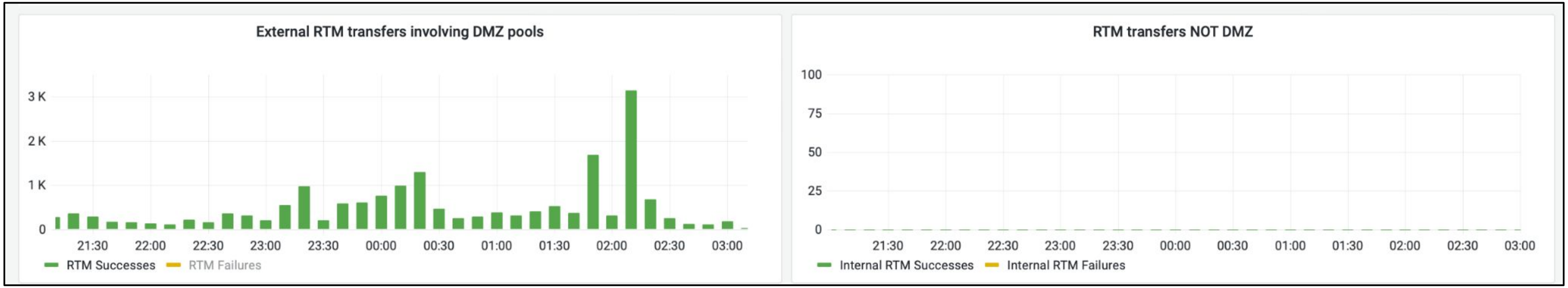
- › Row title (3 panels)
- › Overview (11 panels)
- › p2p Transfers (4 panels)
- › Restore (14 panels)
- › Restore Failures (3 panels)
- › Restore details (3 panels)
- › Store (6 panels)
- › Store details (3 panels)
- › Storageclass (1 panel)
- › Restore duplication (3 panels)
- › Sweep info (4 panels)
- › Storage info (2 panels)
- › StorageInfo details (1 panel)
- › Door info (15 panels)
- › Door info Details (1 panel)
- › Protocols (8 panels)
- › Transfer details (2 panels)
- › Transfer errors (12 panels)

Performance of dCache



Monitoring Used in ad-hoc Studies:

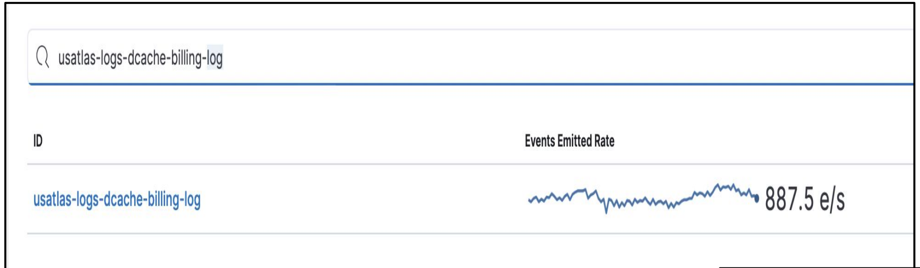
Allowed us to identify areas of improvement for dCache resource access



Ongoing work to optimize BNL to BNL HTTP-TPC resource data access

dCache and ELK Stack
Started to be Used in Operations

Filebeat / Logstash pipelines enabled for domain logs and billing logs

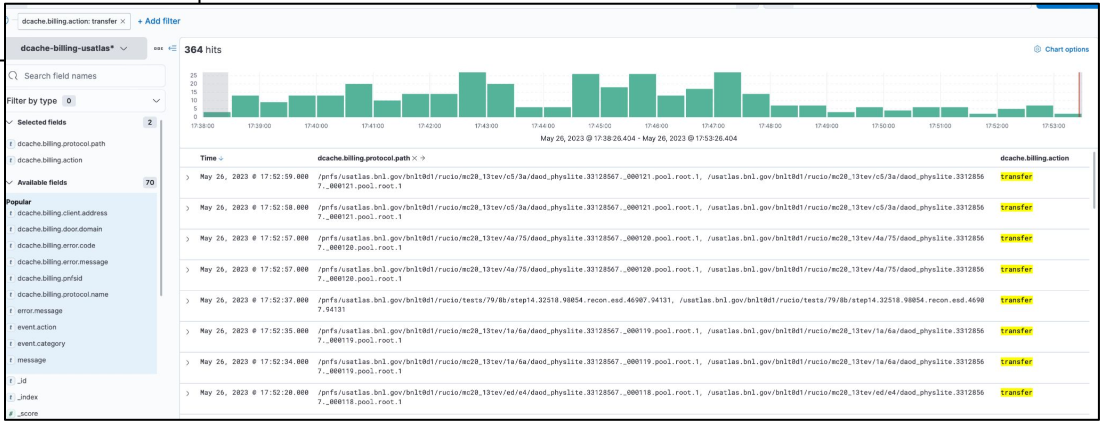


usatlas-logs-dcache-billing-log

ID	Events Emitted Rate
usatlas-logs-dcache-billing-log	887.5 e/s

Rows per page: 50

ELK use to mine the billing logs with arbitrary queries



Time	dcache.billing.protocol.path	dcache.billing.action
May 26, 2023 @ 17:52:59.000	/pnfs/usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/c5/3a/daod_physlite.33128567._000121.pool.root.1, /usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/c5/3a/daod_physlite.33128567._000121.pool.root.1	transfer
May 26, 2023 @ 17:52:58.000	/pnfs/usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/c5/3a/daod_physlite.33128567._000121.pool.root.1, /usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/c5/3a/daod_physlite.33128567._000121.pool.root.1	transfer
May 26, 2023 @ 17:52:57.000	/pnfs/usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/4a/75/daod_physlite.33128567._000120.pool.root.1, /usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/4a/75/daod_physlite.33128567._000120.pool.root.1	transfer
May 26, 2023 @ 17:52:57.000	/pnfs/usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/4a/75/daod_physlite.33128567._000120.pool.root.1, /usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/4a/75/daod_physlite.33128567._000120.pool.root.1	transfer
May 26, 2023 @ 17:52:37.000	/pnfs/usatlas.bnl.gov/bnlt0d1/rucio/tests/79/8b/step14.32518.98054.recon.esd.46907.94131, /usatlas.bnl.gov/bnlt0d1/rucio/tests/79/8b/step14.32518.98054.recon.esd.46907.94131	transfer
May 26, 2023 @ 17:52:35.000	/pnfs/usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/1a/6a/daod_physlite.33128567._000119.pool.root.1, /usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/1a/6a/daod_physlite.33128567._000119.pool.root.1	transfer
May 26, 2023 @ 17:52:34.000	/pnfs/usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/1a/6a/daod_physlite.33128567._000119.pool.root.1, /usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/1a/6a/daod_physlite.33128567._000119.pool.root.1	transfer
May 26, 2023 @ 17:52:20.000	/pnfs/usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/ed/e4/daod_physlite.33128567._000118.pool.root.1, /usatlas.bnl.gov/bnlt0d1/rucio/mc20_13tev/ed/e4/daod_physlite.33128567._000118.pool.root.1	transfer

Brookhaven National Laboratory

Evolving dCache Along with Infrastructure

- SDCC puppet new infrastructure evolving from puppet 3 to puppet 8
- dCache related puppet modules in principle ported to puppet 8
- New effort in refactoring dCache puppet classes for a multi-instance deployment
- RHEL 7 ~ 2 years for end standard support, new hardware deployment on RHEL 8

dCache instance	**Number of VMs+Physical Hardware(PH)**	**OS RELEASE**	**dCache Version**	Notes	Pools storage filesystem transition to ZFS from MDRAID
ATLAS	84(96%PH)	RHEL 7.8	8.2.15	Hardware for core services to be upgraded end 2023	20 servers
BELLE2	10(100%PH)	**RHEL 8.6 (Core services), Pools (7.8)**	7.2.19	Core services hardware recently refreshed (Aug/2022) Upgraded from 6.2.x	
DUNE	5(60%PH)	RHEL 7.8	8.2.2	Resilient manager recently commissioned	4 Servers
Pre-production	5(100%PH)	RHEL 8 Doors/Other (7.8)	8.2.18	WLCG REST API test endpoint Integrated with ATLAS DDM test infrastructure	1 server

Future Work

- Consolidation of software stack on dCache migration to 8.2.X releases across instances

- Migration of hardware ATLAS dCache to 8.2
  - Hardware refresh cycle→ RHEL 8.2 → Puppet 8
  - Refactorization puppet code for a multi instance dCache deployment
  - Migration of hardware opportunity to consolidate hardware into new datacenter

- Possibilities to enhance monitoring (ELK stack for Billing DB and components log events)

- Participation on HTTP REST Tape API testing

In Summary

BNL SDCC is supporting dCache based storage for a diverse of SC

Evolution of the dCache storage features adapted to SC

Priority work will concentrate on:

- Review/optimize TPC data flows for internal transfers (BNL to BNL)
- Improving the orchestration management of dCache software
- Consolidating dCache release levels and OS in different instances

Thank you / Danke schön

Brookhaven National Laboratory