dCache at BNL

Carlos Fernando Gamboa (on behalf of the storage group)
Brookhaven National Laboratory

Workshop on ATLAS Computing and Software Activities at BNL - Navigating Distributed Computing, Storage, Compute, and Beyond

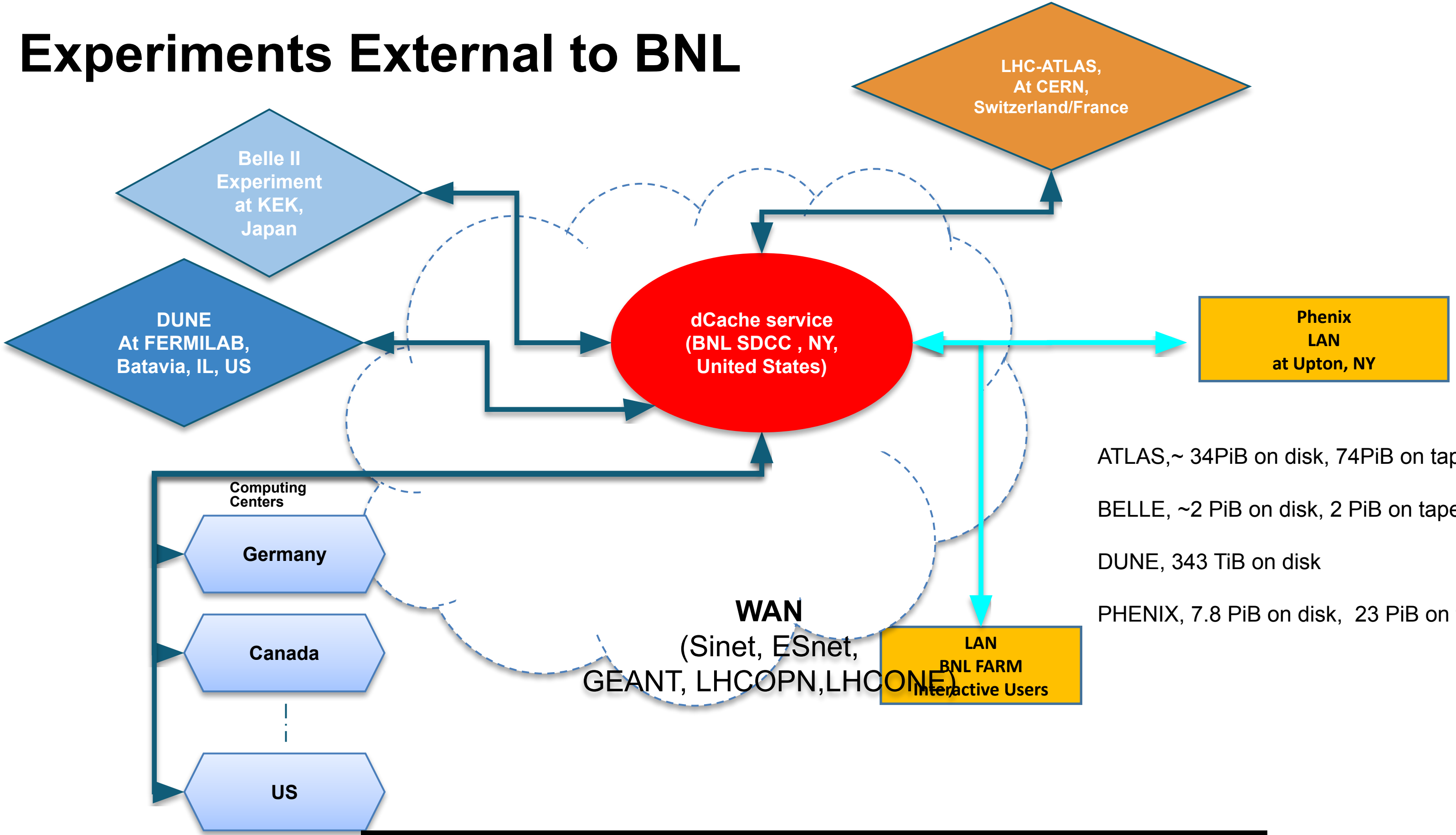
Outline

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

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, DUNE and Phenix SC store and manage 143PiBs (30% DISK) of data
  - Scientific Community data is produced outside BNL:
    - CERN (Switzerland/France),
    - KEK (Japan),
    - Fermilab(IL,US)
  - SC producing data at BNL
    - Phenix

Experiments External to BNL



ATLAS,~ 34PiB on disk, 74PiB on tape

BELLE, ~2 PiB on disk, 2 PiB on tape

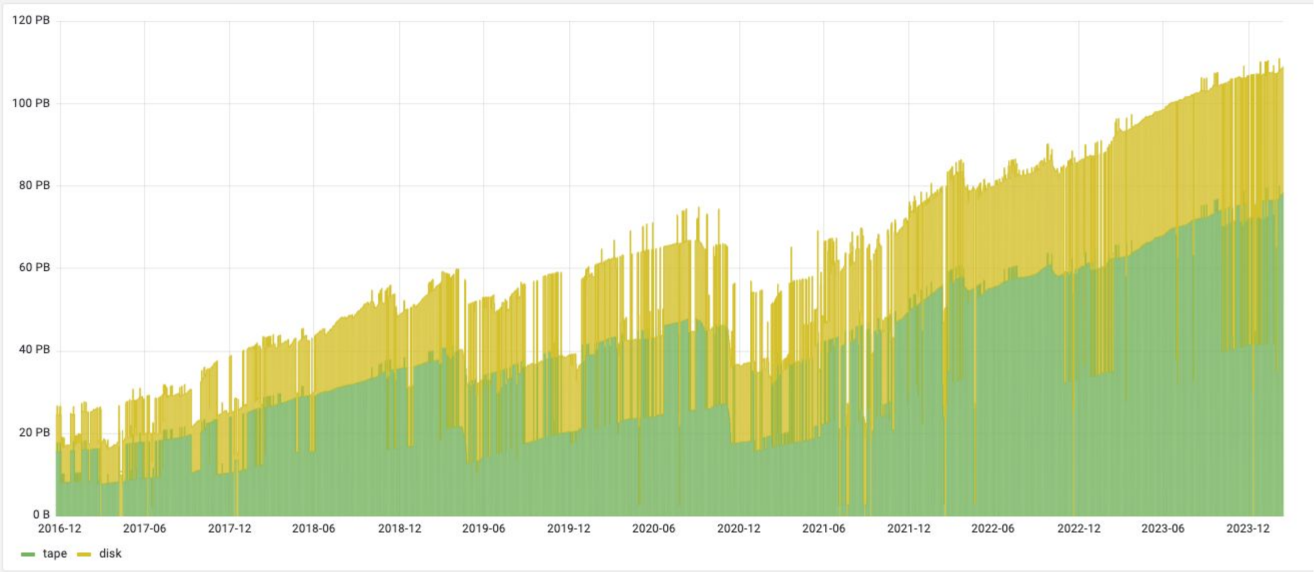
DUNE, 343 TiB on disk

PHENIX, 7.8 PiB on disk, 23 PiB on tape

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

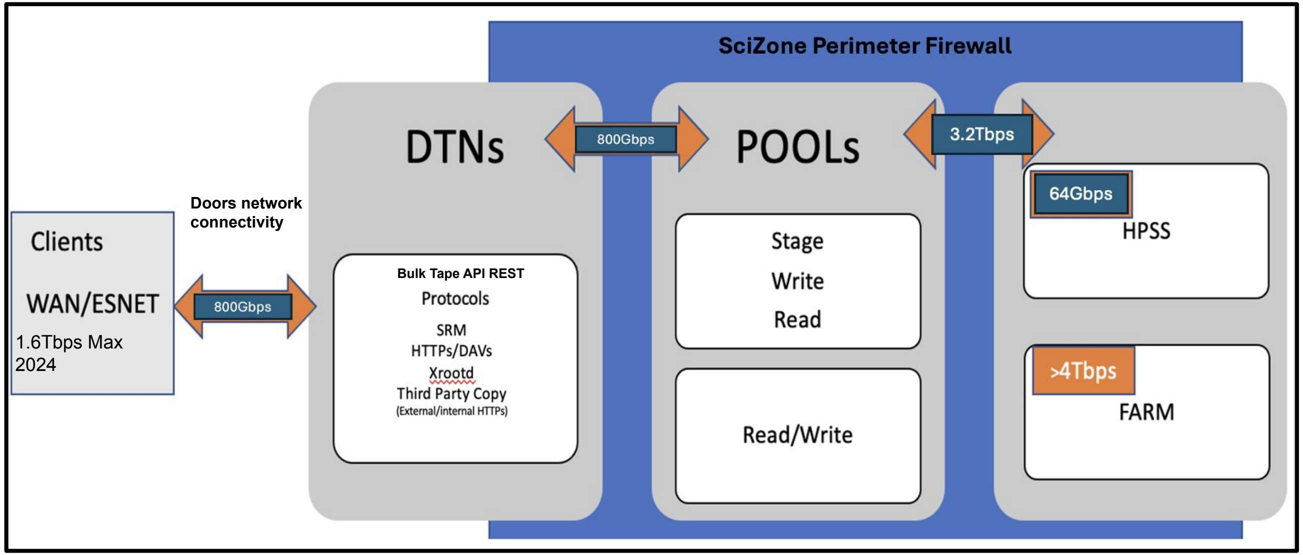
Evolution of Atlas SC storage

BNL provides more than 100PB for ATLAS



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)



HPSS architected to support up to 64Gbps sustained

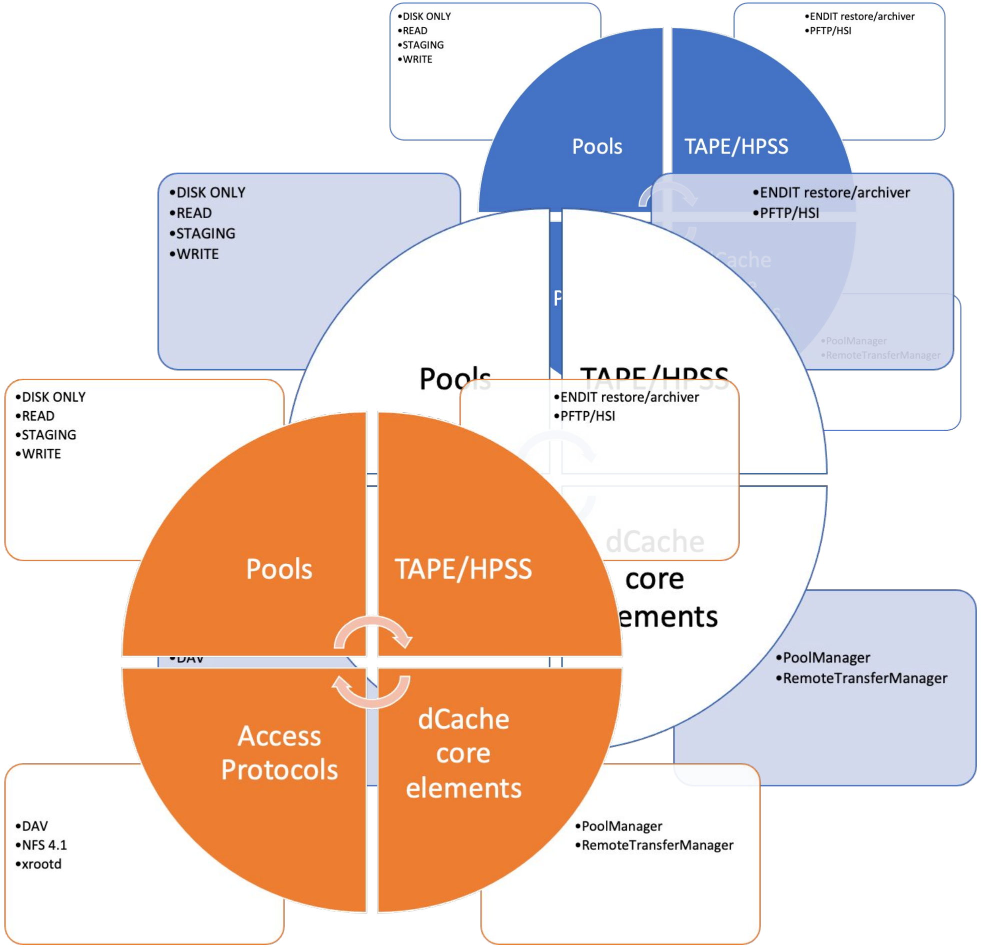
Comply with BNL cybersecurity policy disaggregation among external and internal resource accessibility

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

Brookhaven National Laboratory

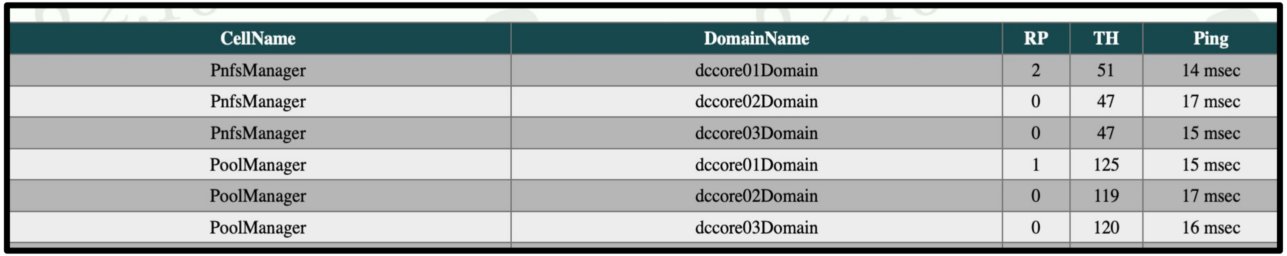
Towards an Improved dCache Operation

Areas of work:

- dCache SE multi-instance architecture
- Evolving dCache along with infrastructure
- Improving dCache data access workflows for client access
- Extending monitoring for dCache operations
- Adoption of ENDIT
    - (tomorrow’s session detailed overview)

Towards dCache SE multi-instance architecture

- Improving dCache SE components distribution
  - Reconfiguration in HA mode to "minimize single points of failures and enable rolling upgrades and, in some cases, horizontal scalability", cf. HA dCache Services
  -

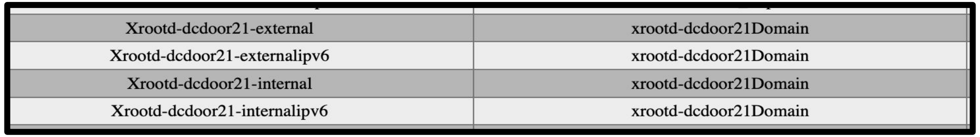


CellName	DomainName	RP	TH	Ping
PnfsManager	dccore01Domain	2	51	14 msec
PnfsManager	dccore02Domain	0	47	17 msec
PnfsManager	dccore03Domain	0	47	15 msec
PoolManager	dccore01Domain	1	125	15 msec
PoolManager	dccore02Domain	0	119	17 msec
PoolManager	dccore03Domain	0	120	16 msec

- Refactoring puppet code for dCache administration
  - SDCC puppet transition 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
    - Common puppet class to manage all experiments

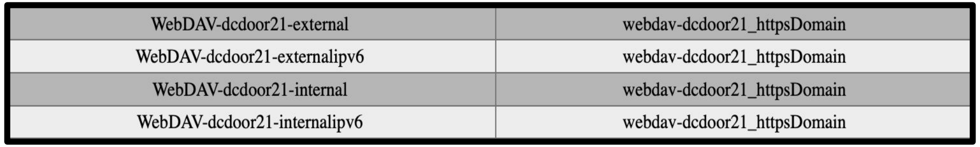
Improving dCache data access workflows for client access

- Tailor dCache data access workflows for LAN client access
    - Production workflows optimization for local resource access
        - BNL to BNL TPC
            - BNL to BNL TPC via p2p?
        - **Different scenarios being considered to improve DAVs-TPC will be discussed today**
- Xrootd external client access for direct write/read



Xrootd-dcdoor21-external	xrootd-dcdoor21Domain
Xrootd-dcdoor21-externalipv6	xrootd-dcdoor21Domain
Xrootd-dcdoor21-internal	xrootd-dcdoor21Domain
Xrootd-dcdoor21-internalipv6	xrootd-dcdoor21Domain

- Dual IPv4/IPv6 dCache application stack configuration



WebDAV-dcdoor21-external	webdav-dcdoor21_httpsDomain
WebDAV-dcdoor21-externalipv6	webdav-dcdoor21_httpsDomain
WebDAV-dcdoor21-internal	webdav-dcdoor21_httpsDomain
WebDAV-dcdoor21-internalipv6	webdav-dcdoor21_httpsDomain

Evolving dCache Along with Infrastructure

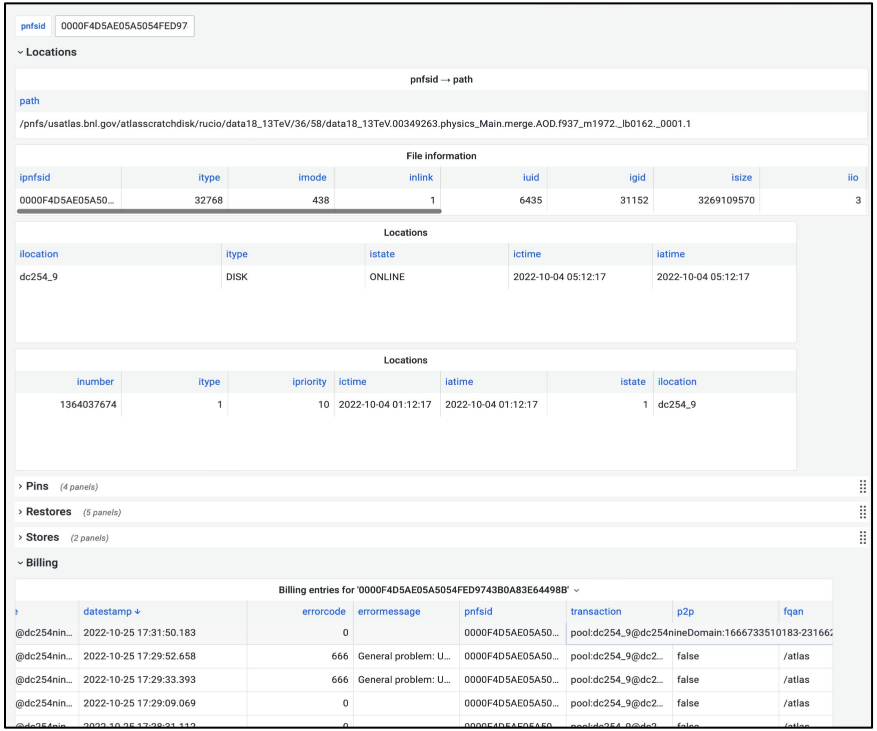
- SDCC puppet transition infrastructure evolving from puppet 3 to puppet 8
- RHEL 7 ~ 6 moths for end standard support, new hardware deployment on RHEL
- Adopting underlying ZFS to be underlying file system to host dCache data on pools

dCache instance	Number of VMs+Physical Hardware(PH)	OS RELEASE	dCache Version	Notes
ATLAS	87(95%PH)	RHEL 8.8 / Pools (7.9)	9.2.6+	24/56 pools servers to be decommissioned ~22 PiB data storage to be relocated (ongoing)
BELLE2	12(100%PH)	**RHEL 8.6 (Core services), Pools (7.8)**	8.2.26	Subject to a yearly schedule upgrades, mainly taking advantage of detector downtime
DUNE	12(33%PH)	RHEL 7.9	9.2.6	Legacy hardware in a resilient configuration 2 copy/file
Phenix	14 (100%PH)	RHEL 7.9	5.2.9	Recently moving to a centralized storage, decommissioning >400 pools on Farm nodes
Pre-production	12(20%PH)	RHEL 8/Pools (7.8)	9.2.14	WLCG REST API test endpoint Integrated with ATLAS DDM test infrastructure

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)

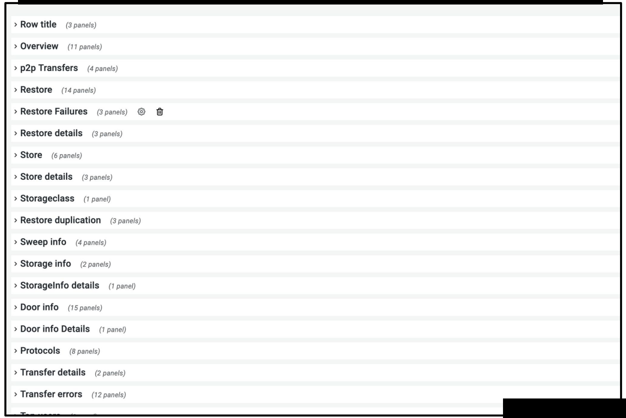
› Stores (2 panels)

⌄ Billing

Billing entries for '0000F4D5AE05A5054FED9743B0A83E64498B' ⌄

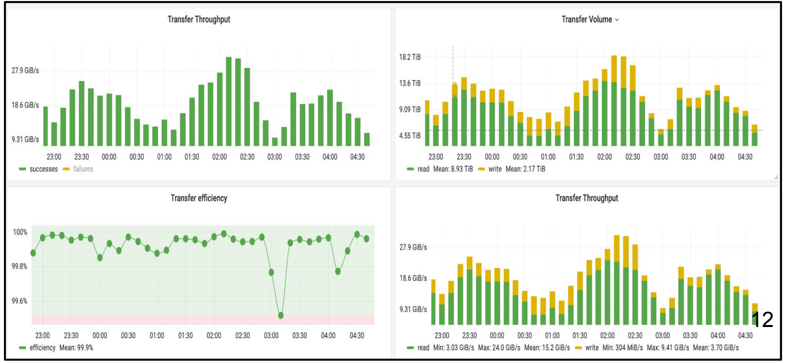
e	datestamp ↓	errorcode	errormessage	pnfsid	transaction	p2p	fqan
@dc254nin…	2022-10-25 17:31:50.183	0		0000F4D5AE05A50…	pool:dc254_9@dc254nineDomain:1666733510183-231662		
@dc254nin…	2022-10-25 17:29:52.658	666	General problem: U…	0000F4D5AE05A50…	pool:dc254_9@dc2…	false	/atlas
@dc254nin…	2022-10-25 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

Feature driven dashboards

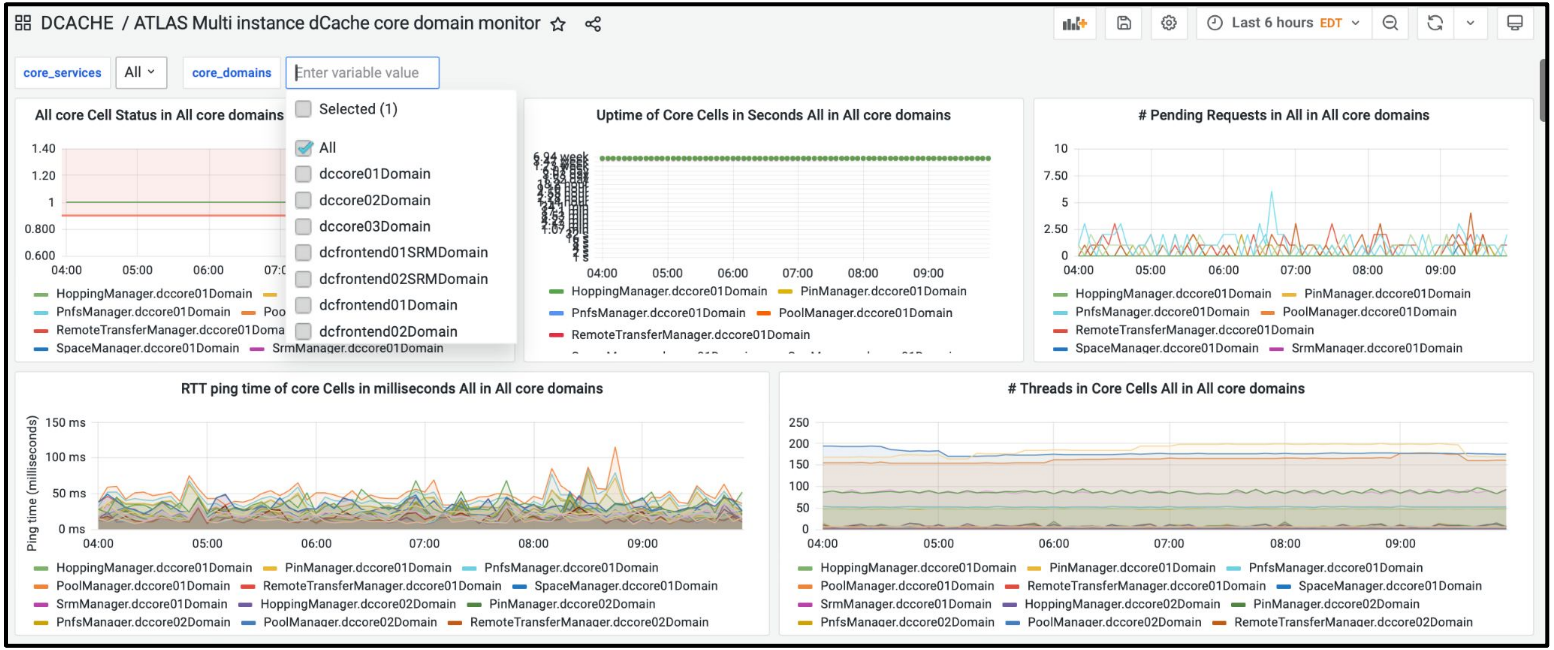


- › Row title (3 panels)
- › Overview (11 panels)
- › p2p Transfers (4 panels)
- › Restore (14 panels)
- › Restore Failures (3 panels)
- › Restore details (3 panels)
- › Store (6 panels)
- › Store details (3 panels)
- › Storageclass (1 panel)
- › Restore duplication (3 panels)
- › Sweep info (4 panels)
- › 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



Evolving monitoring to a multi-core architecture (on going)



DCACHE / ATLAS Multi instance dCache core domain monitor

core_services All

core_domains Enter variable value

□ Selected (1)

☑ All

□ dccore01Domain

□ dccore02Domain

□ dccore03Domain

□ dcfrontend01SRMDomain

□ dcfrontend02SRMDomain

□ dcfrontend01Domain

□ dcfrontend02Domain

Last 6 hours EDT

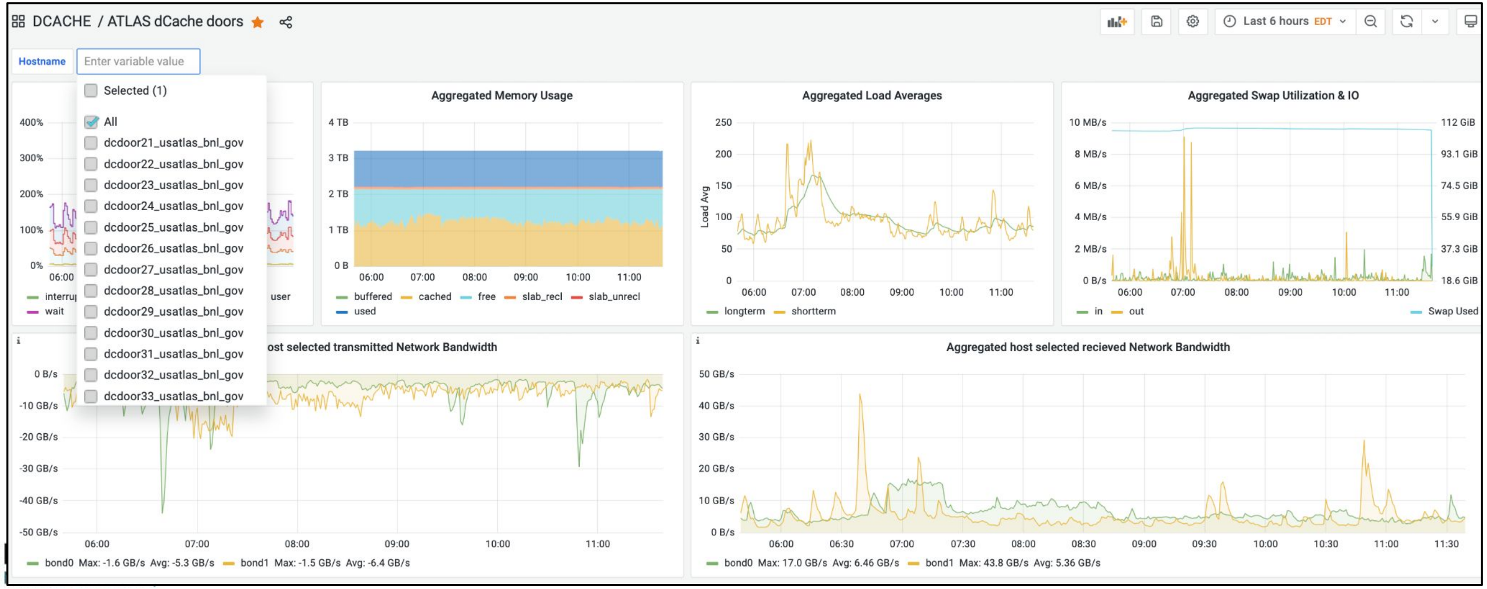
All core Cell Status in All core domains
1.40
1.20
1
0.800
0.600
04:00
05:00
06:00
07:0
HoppingManager.dccore01Domain
PnfsManager.dccore01Domain
Poo
RemoteTransferManager.dccore01Doma
SpaceManager.dccore01Domain
SrmManager.dccore01Domain
Uptime of Core Cells in Seconds All in All core domains
6.94 week
04:00
05:00
06:00
07:00
08:00
09:00
HoppingManager.dccore01Domain
PinManager.dccore01Domain
PnfsManager.dccore01Domain
PoolManager.dccore01Domain
RemoteTransferManager.dccore01Domain
# Pending Requests in All in All core domains
10
7.50
5
2.50
0
04:00
05:00
06:00
07:00
08:00
09:00
HoppingManager.dccore01Domain
PinManager.dccore01Domain
PnfsManager.dccore01Domain
PoolManager.dccore01Domain
RemoteTransferManager.dccore01Domain
SpaceManager.dccore01Domain
SrmManager.dccore01Domain
RTT ping time of core Cells in milliseconds All in All core domains
Ping time (milliseconds)
150 ms
100 ms
50 ms
0 ms
04:00
05:00
06:00
07:00
08:00
09:00
HoppingManager.dccore01Domain
PinManager.dccore01Domain
PnfsManager.dccore01Domain
PoolManager.dccore01Domain
RemoteTransferManager.dccore01Domain
SpaceManager.dccore01Domain
SrmManager.dccore01Domain
HoppingManager.dccore02Domain
PinManager.dccore02Domain
PnfsManager.dccore02Domain
PoolManager.dccore02Domain
RemoteTransferManager.dccore02Domain
# Threads in Core Cells All in All core domains
250
200
150
100
50
0
04:00
05:00
06:00
07:00
08:00
09:00
HoppingManager.dccore01Domain
PinManager.dccore01Domain
PnfsManager.dccore01Domain
PoolManager.dccore01Domain
RemoteTransferManager.dccore01Domain
SpaceManager.dccore01Domain
SrmManager.dccore01Domain
HoppingManager.dccore02Domain
PinManager.dccore02Domain
PnfsManager.dccore02Domain
PoolManager.dccore02Domain
RemoteTransferManager.dccore02Domain

Opportunistically uses BULK REST API to collect metrics

Monitoring Enhancement, Extending Grafana based metrics

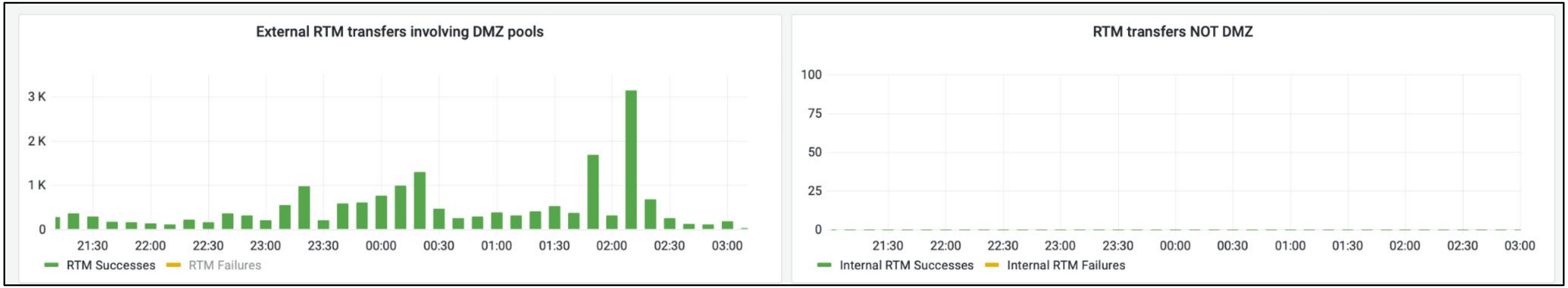
Rhel 7 based used Ganglia to monitor OS out of box aggregations

RHEL 8 based uses Grafana, custom based



Monitoring Used in ad-hoc Studies:

Allowed us to identify areas of improvement for dCache resource access



Internal (LAN) HTTP based TPC should not use DMZ pools resources

Future Work

- Consolidation of software stack on dCache migration to 9.2.X releases across instances

- Hardware refresh ATLAS/Belle2 pools
  - Hardware refresh cycle for pools→ RHEL 8.2 → Puppet 8
  - Refactorization puppet code for a multi instance dCache deployment
    - Belle, Dune…
  - Data migration
- Transition from JAVA 11 to JAVA 17

- Identify scenarios to improve TPC workflows

In Summary

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

Evolution of the dCache storage features adapted to SC

Thank you

Backup slides

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

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**

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 in tomorrow's session