Integration & Optimization of BNL Storage Management

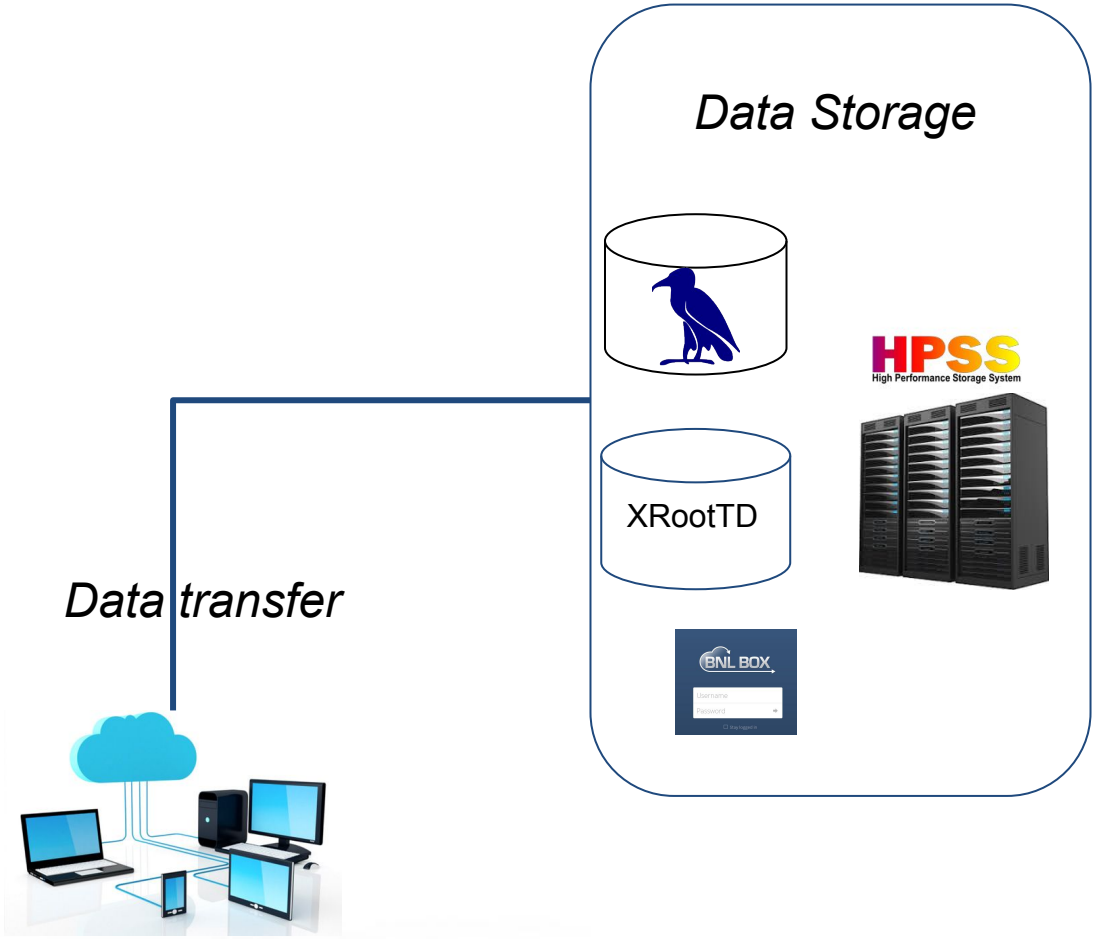
Iris Wu (BNL)

HEPiX, October 2019

Contributions: Guangwei Che, Tim Chou, Robert Hancock, Hironori Ito, Zhenping Liu, Ognian Novakov

Outline

- Data Storage
- Data Transfer
  - Protocols
  - Data carousel@BNL
- Data Access Pattern



Data Storage

ATLAS dCache	Belle II dCache	PHENIX dCache	STAR XRootD
ATLAS T1 site	Belle II T1 site	PHENIX T0 site	STAR T0 site
43PB	1.7PB	9.9PB(~3PB JBOD)	11PB(~3PB JBOD)

- JBOD
  - Relatively inexpensive
  - Relatively easier to scale
  - Easier Disaster Recovery

Data Storage (Cont.)

- PHENIX dCache
  - Hot data vs. Cold data
  - Farm work node vs. Central storage
    - dedication, stability & throughput

- US ATLAS/Belle II dCache
  - Secondary copy of disk file
  - Ceph technology (in the plan)
    - Erasure code
    - Re-use retired disk

Data Transfer - protocols

dCache interfaces to transfer data:

- Third-Party-Transfer (TPC)
  - Bulk data flows directly between endpoints
  - Two potential TPC protocols for the WLCG
    - Xrootd
    - HTTP/Webdav

Data Transfer-protocols (Cont.)

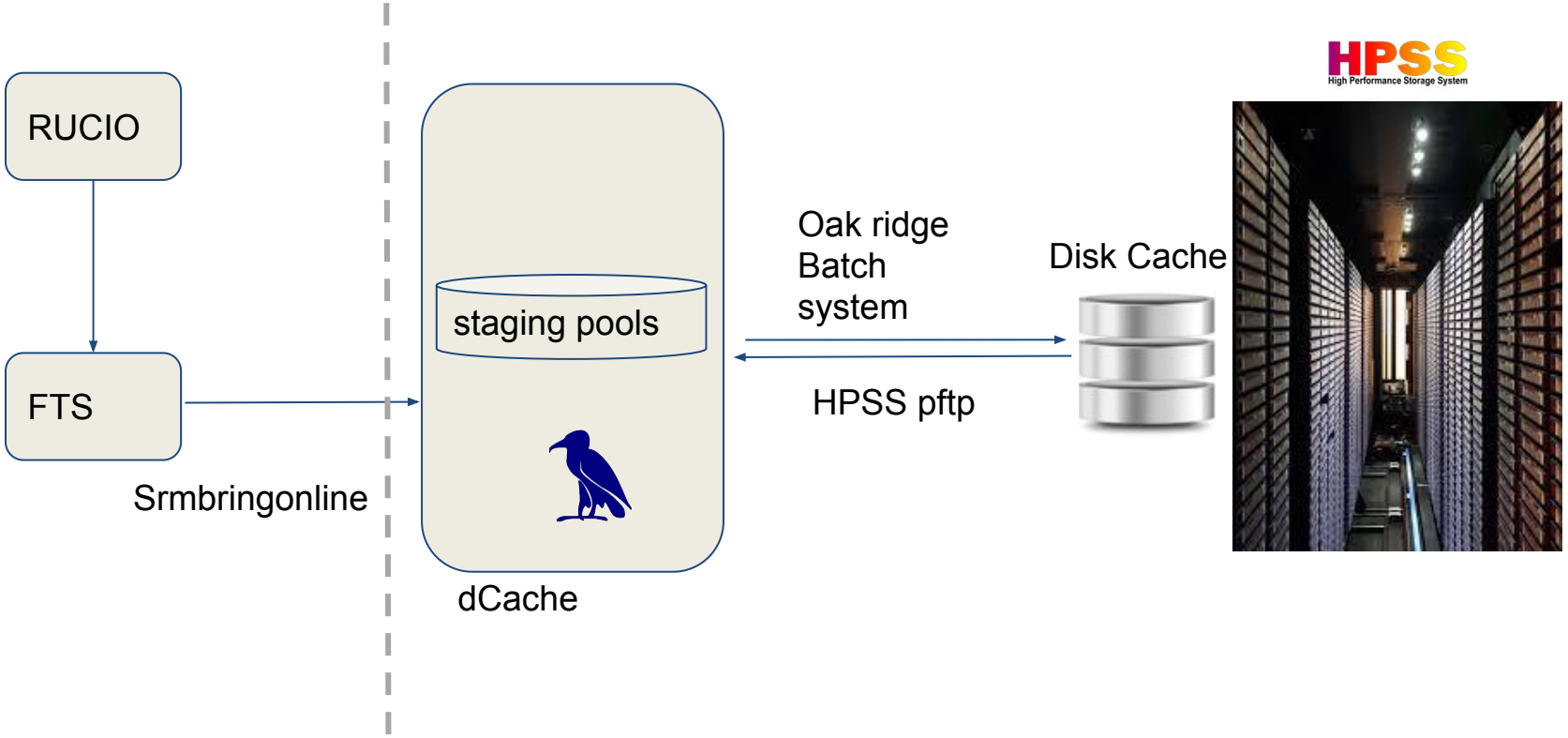
- US Atlas dCache
  - Stograte behind of firewall
  - Door outside of BNL firewall
    - SSD installed in door as hopping pool

- HTTP TPC in dCache
  - BNL Daily HTTP-TPC smoke test

HTTP-TPC Target:
https://dcdoor05.usatlas.bnl.gov:2881//pnfs/usatlas.bnl.gov/BNLT0D1/SAM/smoke-test-dcachetest

- Xrootd TPC in dCache
  - Hopping pool in door
  - New xrdcp client

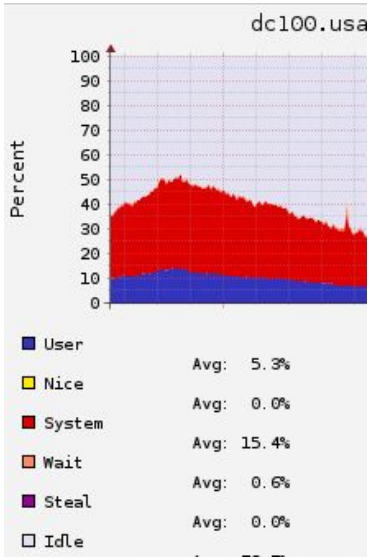
Data carousel@BNL overview



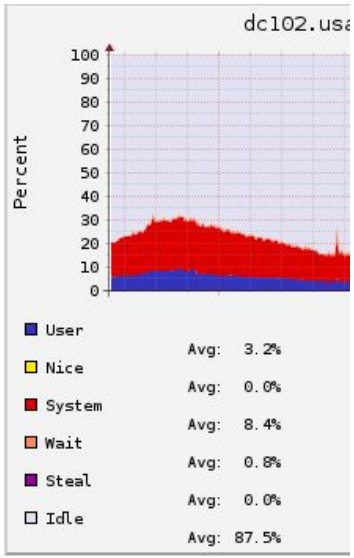
Increasing staging requests
less than 20,000------>20,000--------> 60,000---------->100,000

Data carousel@BNL-optimization

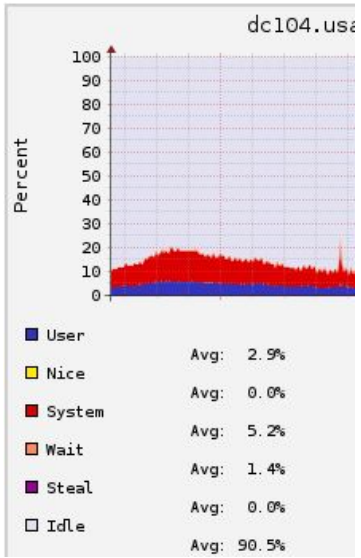
- dCache
  - Increase size of staging pools→ 2.8 PB
  - Decrease load of staging pools
    - Best dCache polling rate vs. CPU usage



2 minutes



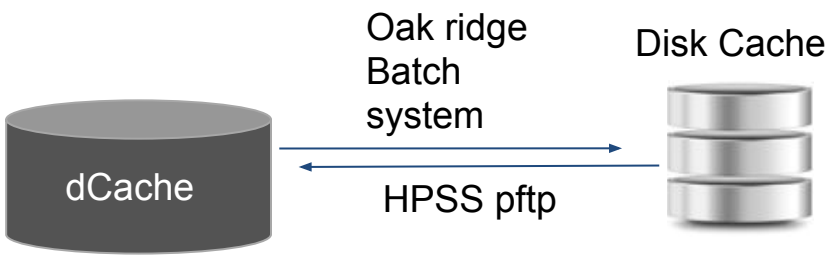
5 minutes



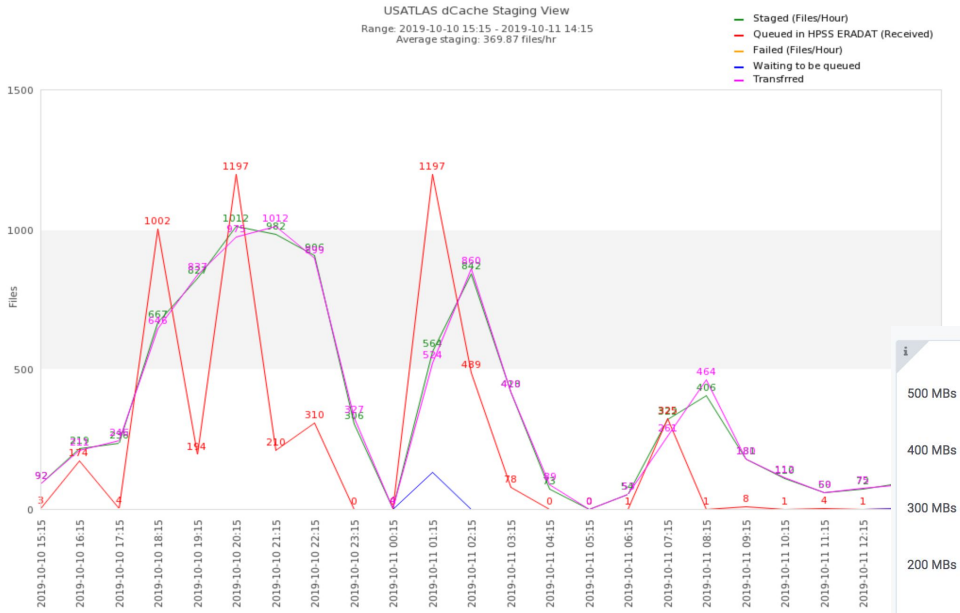
10 minutes

Data carousel@BNL-optimization(Cont.)

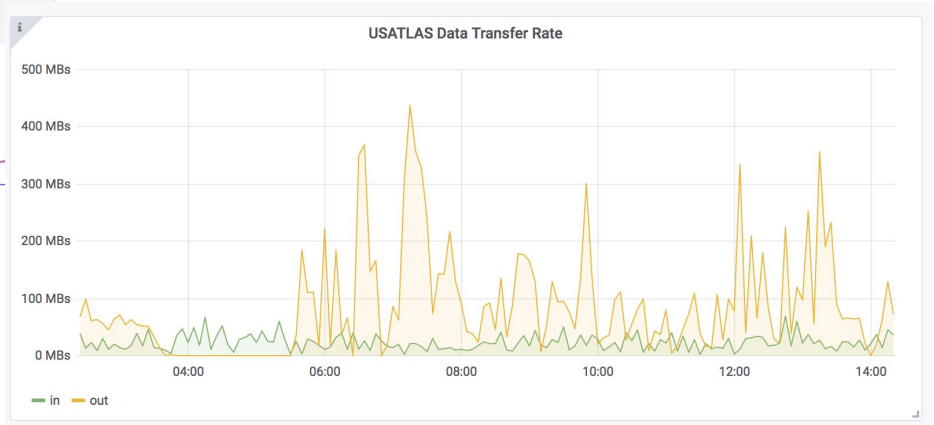
- Oak ridge BATCH system
  - Bulk submission
- HPSS
  - Check HPSS disk purge
  - Introduced load balancer in the frontend to distribute the workload to multiple pftp gateways in the back end



Data Carousel@BNL-monitoring

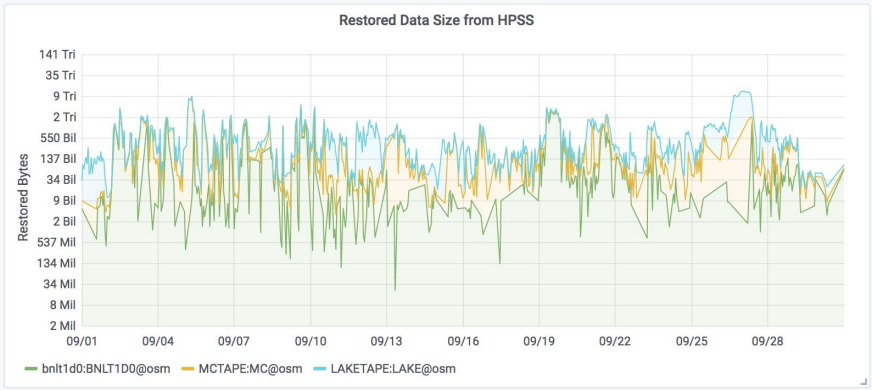


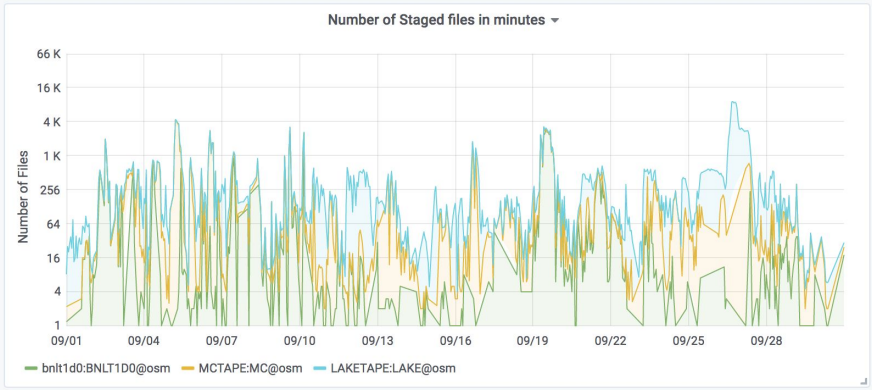
Monitor from HPSS

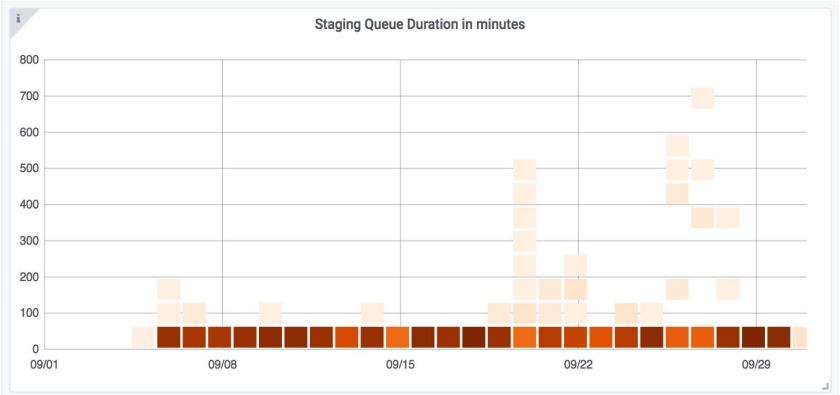


Data Carousel@BNL-monitoring(Cont.)

Monitor from dCache



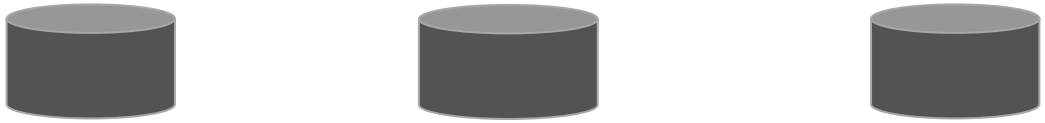




Data Access Pattern

- Access pattern
  - Never accessed
  - Data population

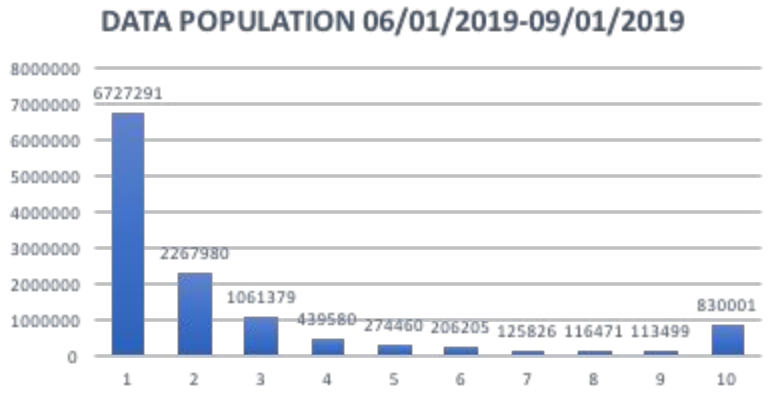
- Combine information of several dCache databases

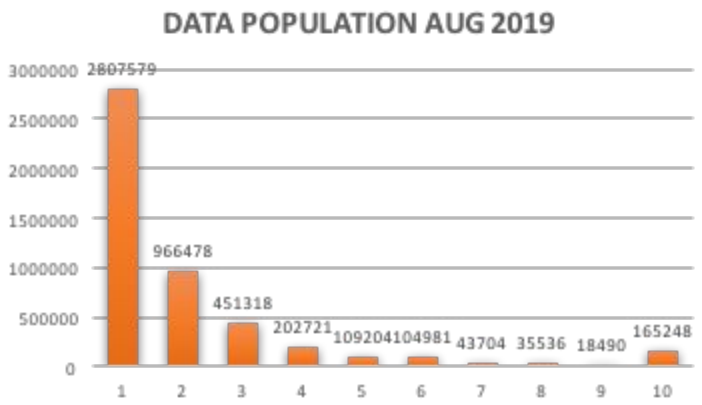


Chimera database | Billing database | Srm database

Data Access Pattern-population

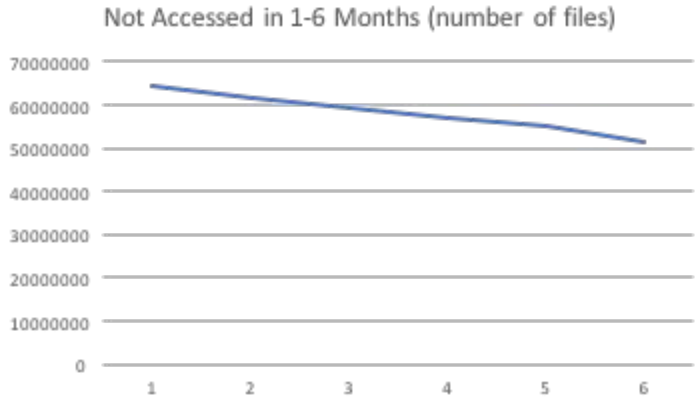
- Disk data population in ATLAS dCache
  - different time intervals





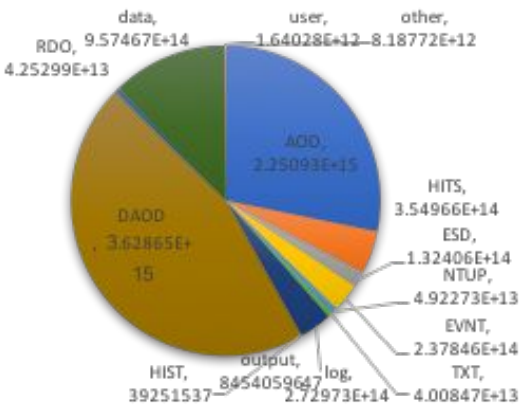
Data Access Pattern-unpopular(1/3)

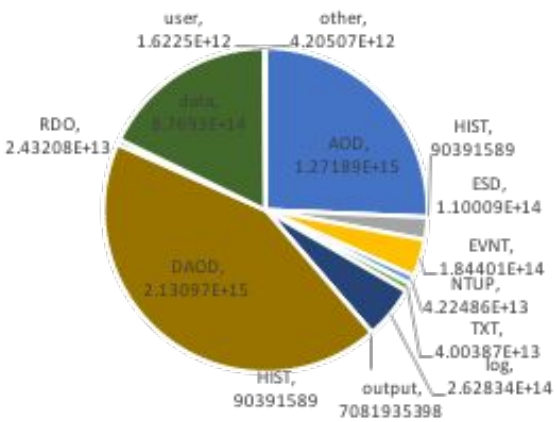
- Disk files in ATLAS dCache not accessed in last six months
  - Volume in bytes
  - Number of files

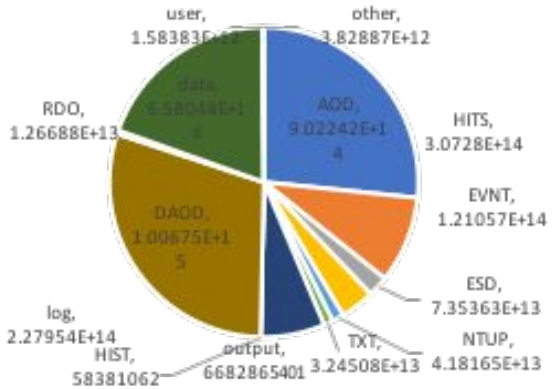


Data Access Pattern-unpopular(2/3)

- Disk files in ATLAS dCache not accessed in last six months
  - Volume in bytes







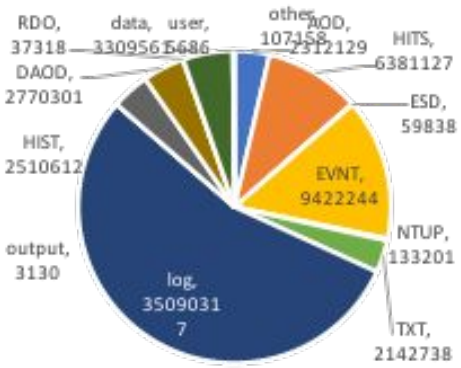
1 month

3 month

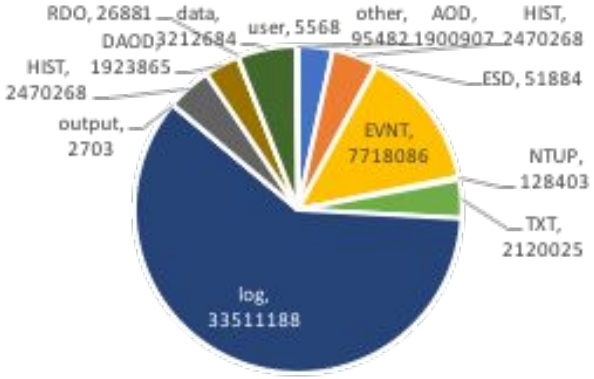
6 month

Data Access Pattern-unpopular(3/3)

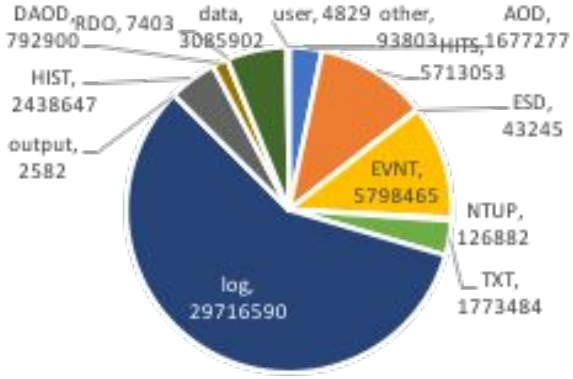
- Number of disk files in ATLAS dCache not accessed in last six months



1 month



3 month



6 month

Data Access Pattern- future work...

- Further development/improvement
  - Concentrate on certain Types
  - Not accessed in different time interval
  - Performance

- Support R&D projects


Thank You!