Data Caching

Teng

26/10/2018

Outline

- **Ongoing activities**
- **XCache study**
- **Plan(s)**

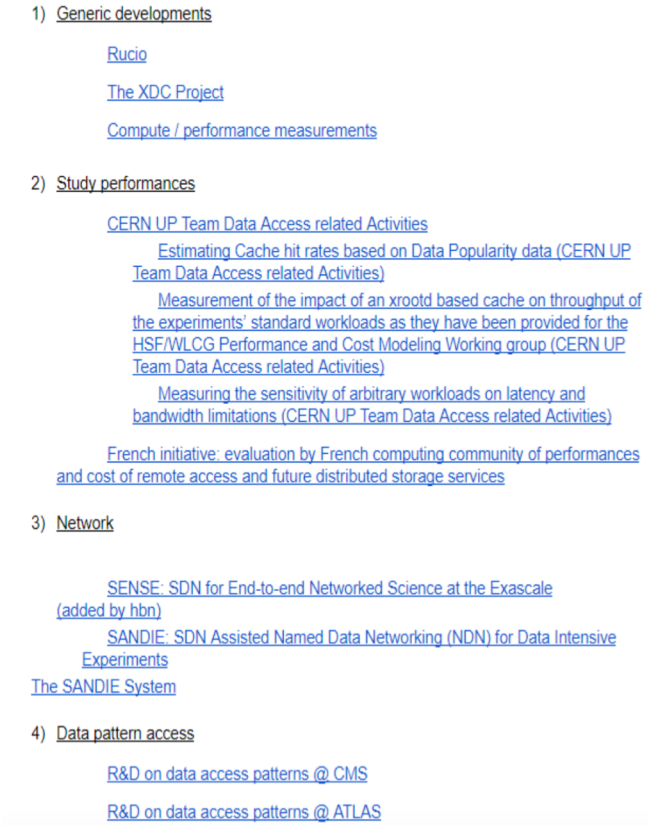
Ongoing activities

- **DOMA-Access sub-WG**
  - Gathering and coordinating works
  - To make advisories to communities/ sites

~ 15 projects on going

Deployment, performance measurement & study

Mainly ATLAS & (*CMS*) activates



1) Generic developments

Rucio

The XDC Project

Compute / performance measurements

2) Study performances

CERN UP Team Data Access related Activities

Estimating Cache hit rates based on Data Popularity data (CERN UP Team Data Access related Activities)

Measurement of the impact of an xrootd based cache on throughput of the experiments' standard workloads as they have been provided for the HSF/WLCG Performance and Cost Modeling Working group (CERN UP Team Data Access related Activities)

Measuring the sensitivity of arbitrary workloads on latency and bandwidth limitations (CERN UP Team Data Access related Activities)

French initiative: evaluation by French computing community of performances and cost of remote access and future distributed storage services

3) Network

SENSE: SDN for End-to-end Networked Science at the Exascale (added by hbn)

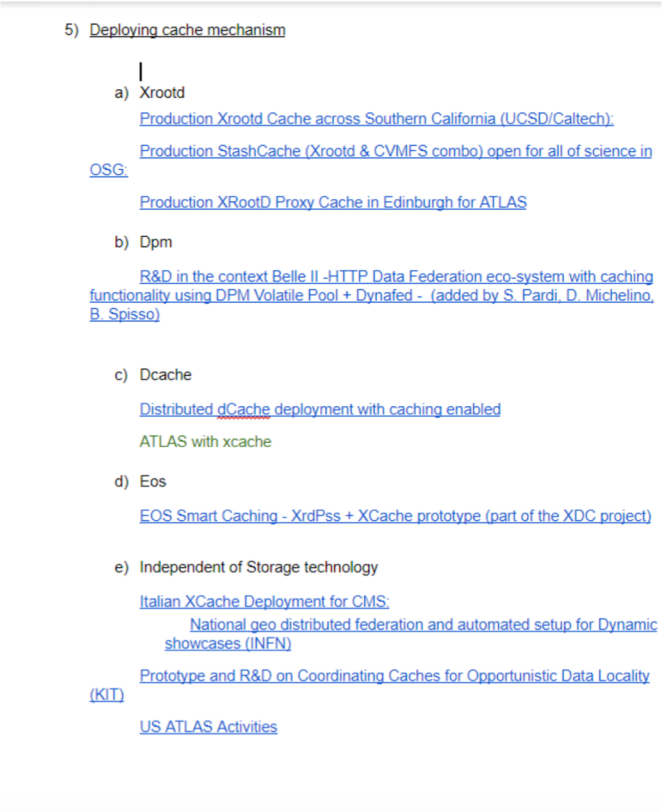
SANDIE: SDN Assisted Named Data Networking (NDN) for Data Intensive Experiments

The SANDIE System

4) Data pattern access

R&D on data access patterns @ CMS

R&D on data access patterns @ ATLAS



5) Deploying cache mechanism

a) Xrootd

Production Xrootd Cache across Southern California (UCSD/Caltech):

Production StashCache (Xrootd & CVMFS combo) open for all of science in OSG:

Production XRootD Proxy Cache in Edinburgh for ATLAS

b) Dpm

R&D in the context Belle II -HTTP Data Federation eco-system with caching functionality using DPM Volatile Pool + Dynafed - (added by S. Pardi, D. Michelino, B. Spisso)

c) Dcache

Distributed dCache deployment with caching enabled

ATLAS with xcache

d) Eos

EOS Smart Caching - XrdPss + XCache prototype (part of the XDC project)

e) Independent of Storage technology

Italian XCache Deployment for CMS:

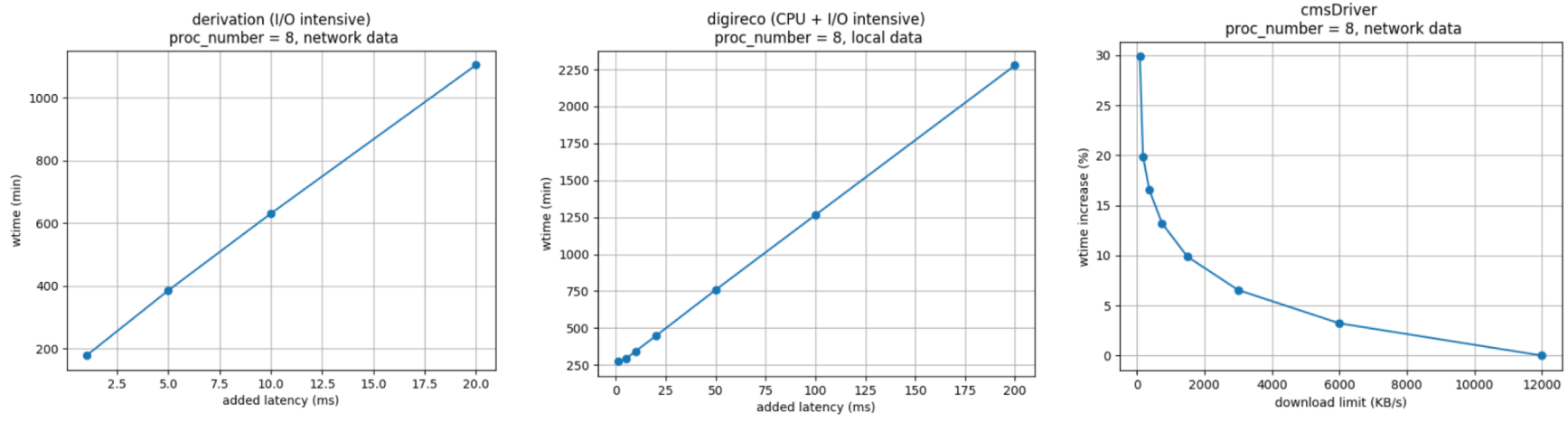
National geo distributed federation and automated setup for Dynamic showcases (INFN)

Prototype and R&D on Coordinating Caches for Opportunistic Data Locality (KIT)

US ATLAS Activities

Some highlights

- **Study of job sensitivities towards data access performance**



By adding latency/download limit/swap usage using cgroup

- **Cache simulation based on XElasticsearch**
  - All ATLAS data access requests are recorded
  - Some code to simulate behaviors of the cache
  - Theoretically you can simulate any sites within hours

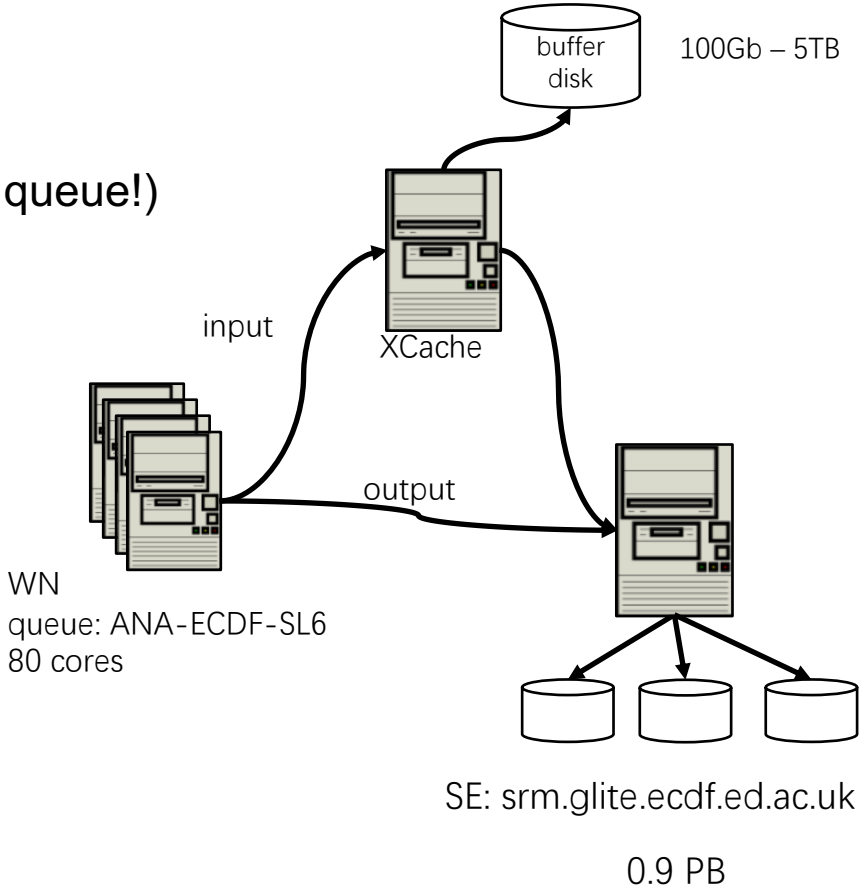
Some highlights

- **US plans of using cache**
  - Node local cache
  - Local storage speedup
  - Site with no storage
  - Full stack caching
- **Performance study of XCache/ Arc cache**
  - Arc cache
    - push model
    - opportunistic use

      82.8% naive cache-hit ratio for ATLAS workflow (after pre-warmed for some time)
  - XCache: next slides
- **More is coming in next weeks**

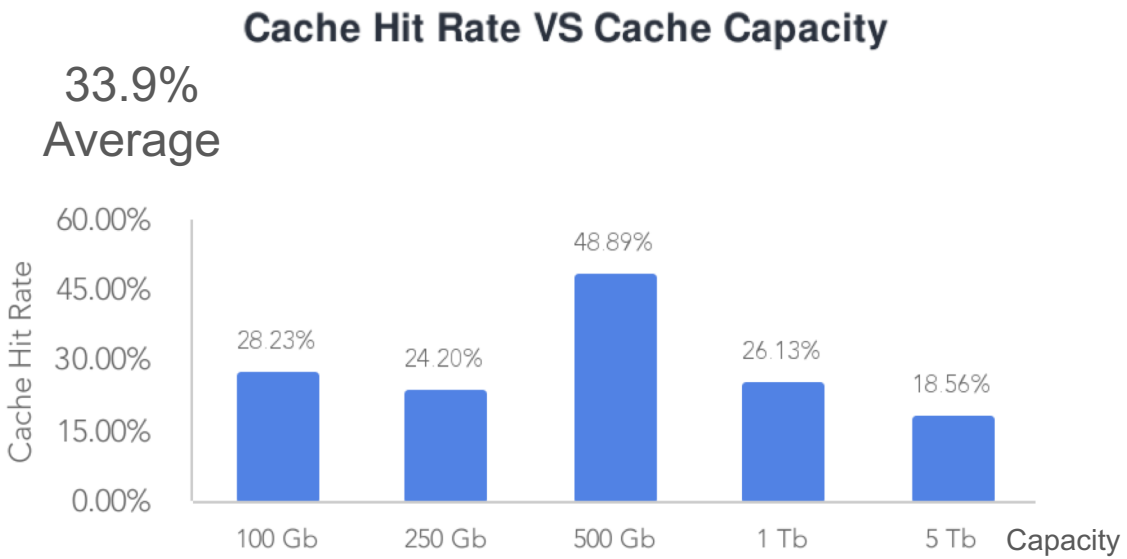
Study with XCache

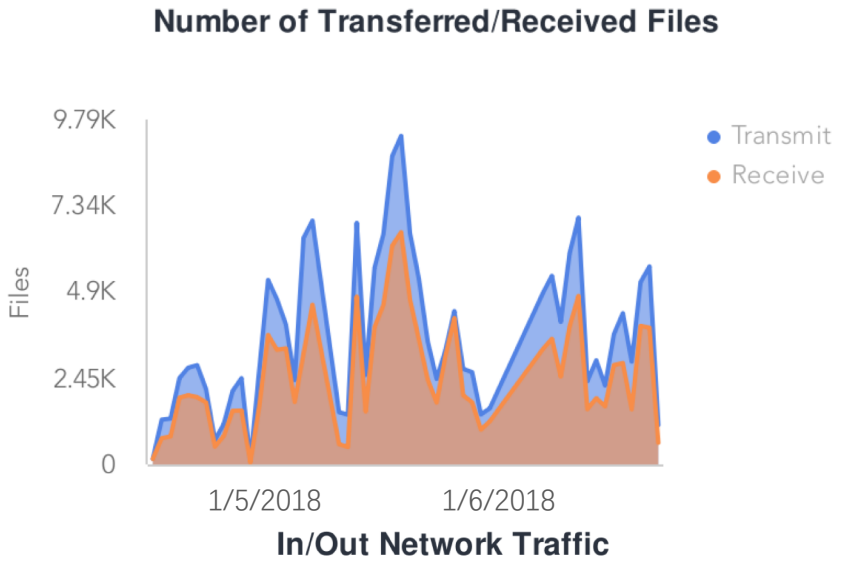
- Overview
  - Use an ATLAS analysis queue for testing
    - At very small scale (80 cores, we have a very small analysis queue!)
  - Simulating a CE attached to a remote SE (diskless site)
    - 0.9 Pb storage
  - Workflow
    - Input network traffic of WNs is redirected to XCache
    - Output network remains unchanged
    - Whole file mode is used
  - A XRootD client plugin is used to redirect the input url
    - root://srm.glite.ecdf.ed.ac.uk/file → root://xcache.url//root://srm.glite.ecdf.ed.ac.uk/file

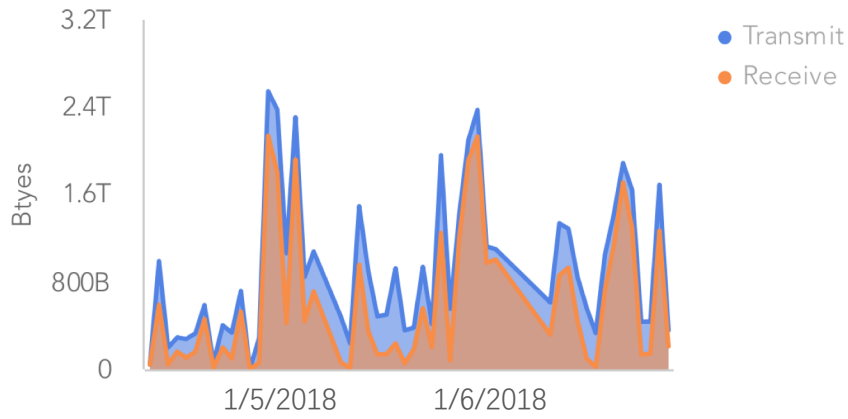


Study with XCache

- 4 months of data is taken to measure the cache performance
- Average cache hit rate is 33.9%.
- Different cache capacities are tested. Peak value reached ~50%. (Only for reference, since errors are high in production environment)





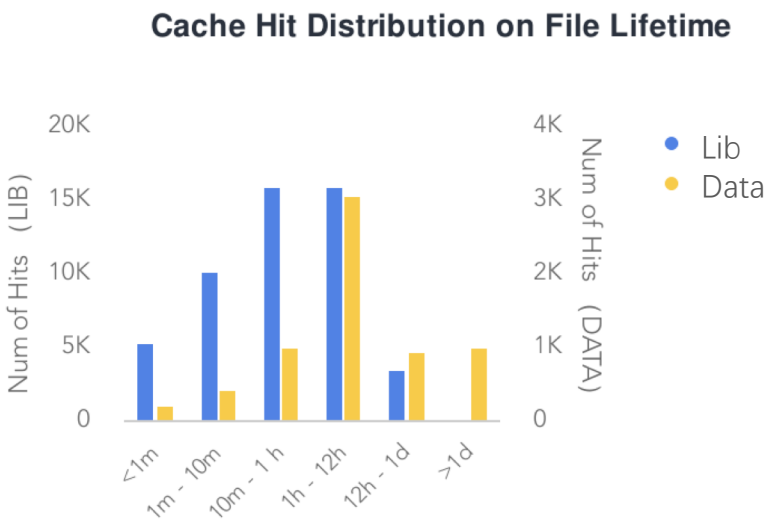


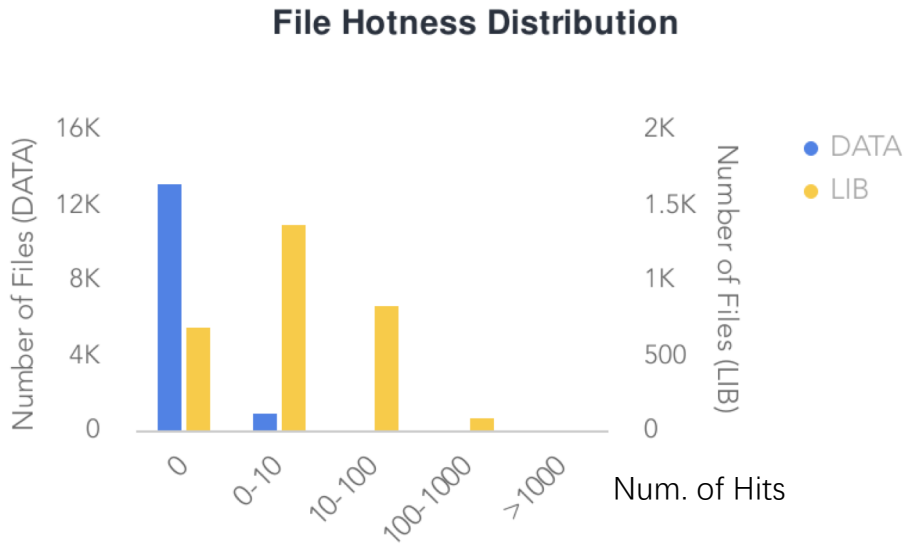
Study with XCache

4 kinds of files are cached:

- **input**: input data files (AOD, DAOD, …)
- **output**: user output
- **library**: user library files (dispatched by panda)
- **log**: job log files

type	portion in disk	hit contribution
Input	92.1%	70.6%
library	1.3%	29.1%
log	0.05%	0.27%
output	6.5%	~0



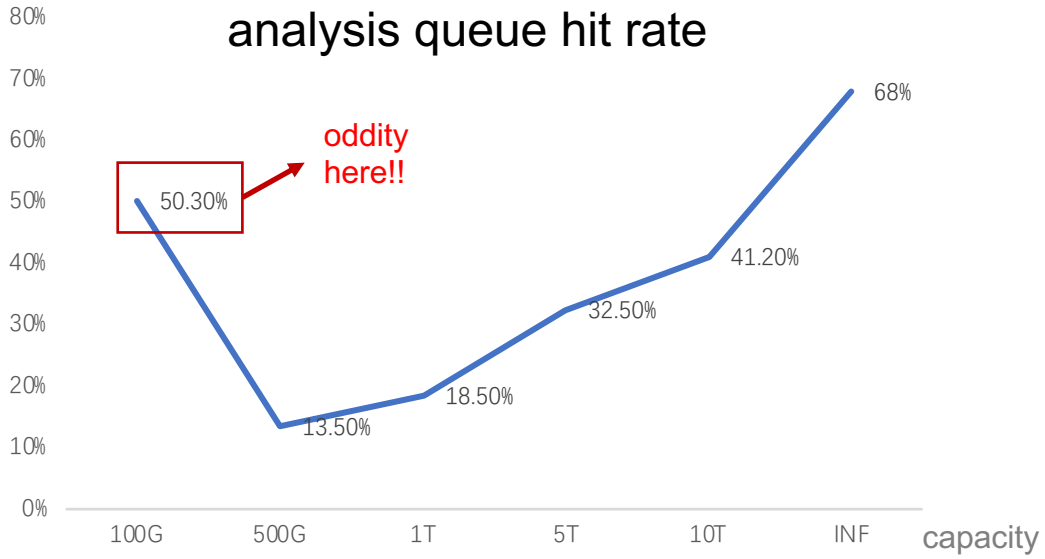


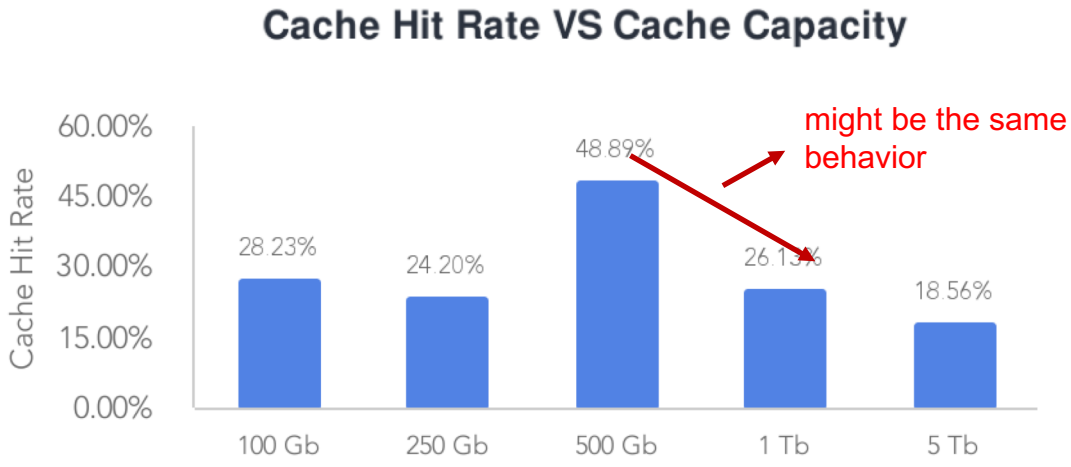
Summary

- Library files are extremely hot
- Most AOD input files are cold
  This definitely needs optimization
- Files are usually hot for the first 12 hours

Simulation @ ECDF

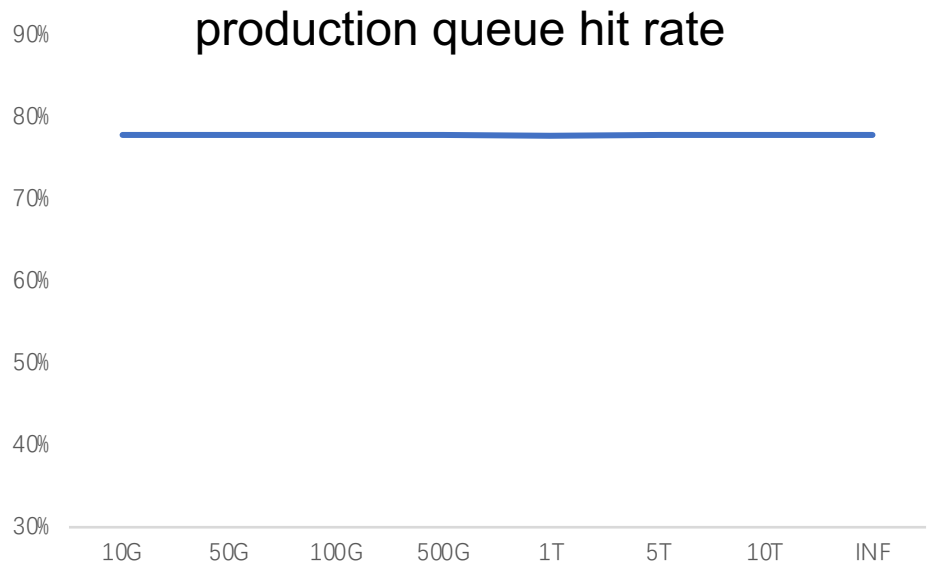
- Results of XCache simulation code
  - ECDF analysis queue (~100 cores) and production queue (~1k cores) are tested
  - 2~6 months of data, cache disk usage: 85%-95%, cleanup policy: largest access_time





Simulation @ ECDF

- Results of XCache simulation code
  - ECDF analysis queue (~100 cores) and production queue (~1k cores) are tested
  - 2~6 months of data, cache disk usage: 85%-95%, purge alg: access_time



Brief summary

- Cache hit rate varies greatly with capacity for analysis job
- Oddity with small capacity (might agree with real data)
- Cache hit rate for production jobs is higher and doesn't change with capacity
- This needs more investigation

Simulation @ ECDF

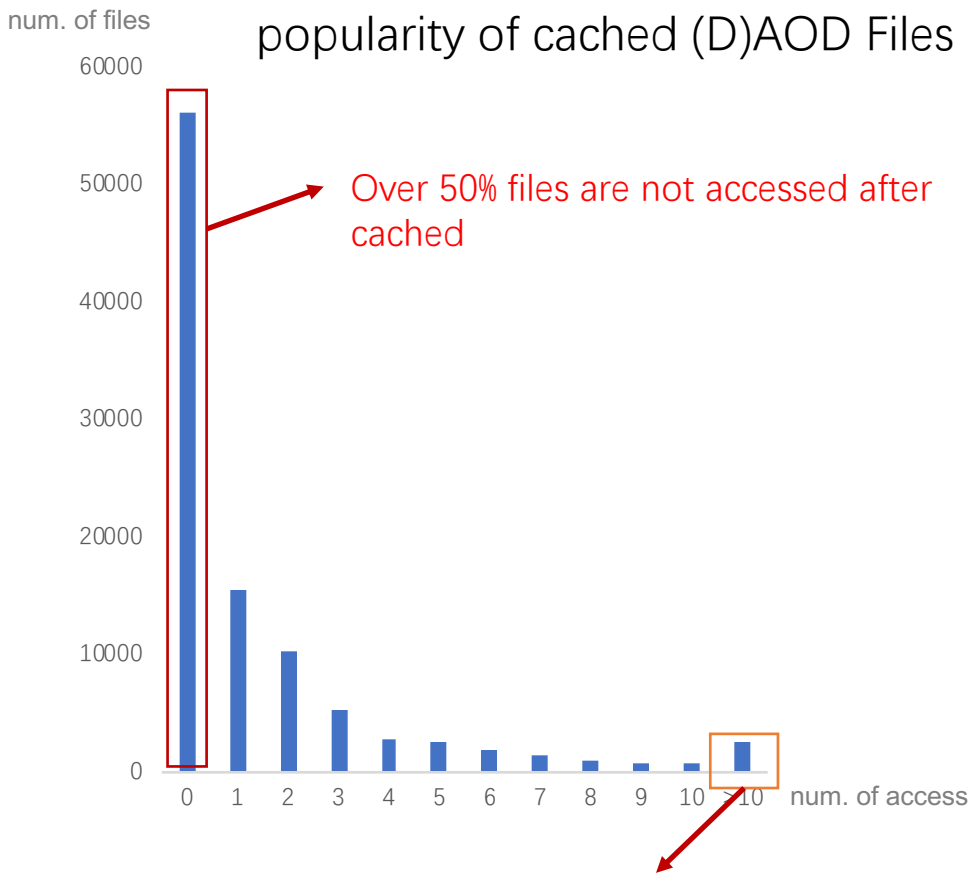
- Cached files
  - (D)AOD contributes most of the traffic and cache hit (optimization should focus on them)
  - Productions are easy

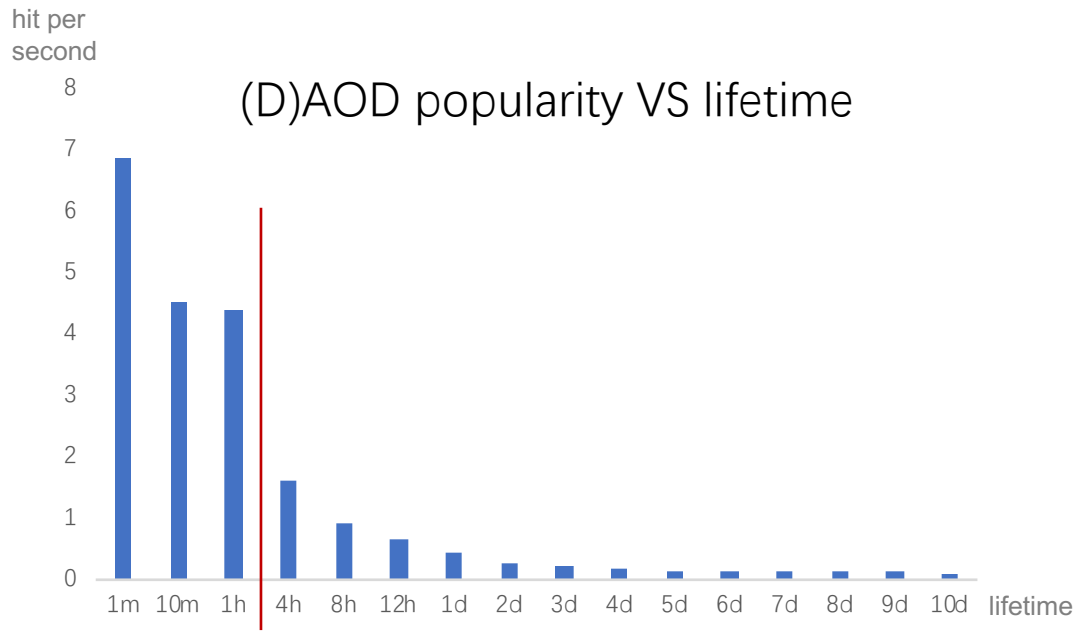
	Write into cache	Read from cache	Cache hit	Cache hit rate
AOD	110957	343629	232671	67.7%
library	173	5052	4879	96.6%
log	7.7	7.8	0.07	~0
output	1275	1371.7	96.6	7%

	Write into cache	Read from cache	Cache hit	Cache hit rate
AOD	4205.6	11490.9	7285.2	63.4%
DRAW_*	576	576	0	0
HIT	9.34	5047.8	5038.5	99.8%
TXT*	761.6	762.2	0.6	0
GEN*	1.39	448.6	447.2	99.7%
EVNT	2908.1	20060.4	17152.2	85.5%

* Gb as unit

Simulation @ ECDF





Plans

- Continue unfinished study to figure out oddities
- Look into optimization methods (desired to be VO/workflow agnostic)
- Simulate other GridPP sites
- Study partial file cache performance on ECDF analysis queue