Data access interfaces to Data Lake

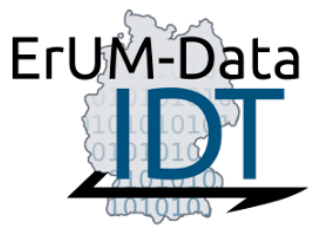
Current status

Nikolai Hartmann

on behalf of the DOMA/ACCESS group

LMU Munich

March 23, 2020

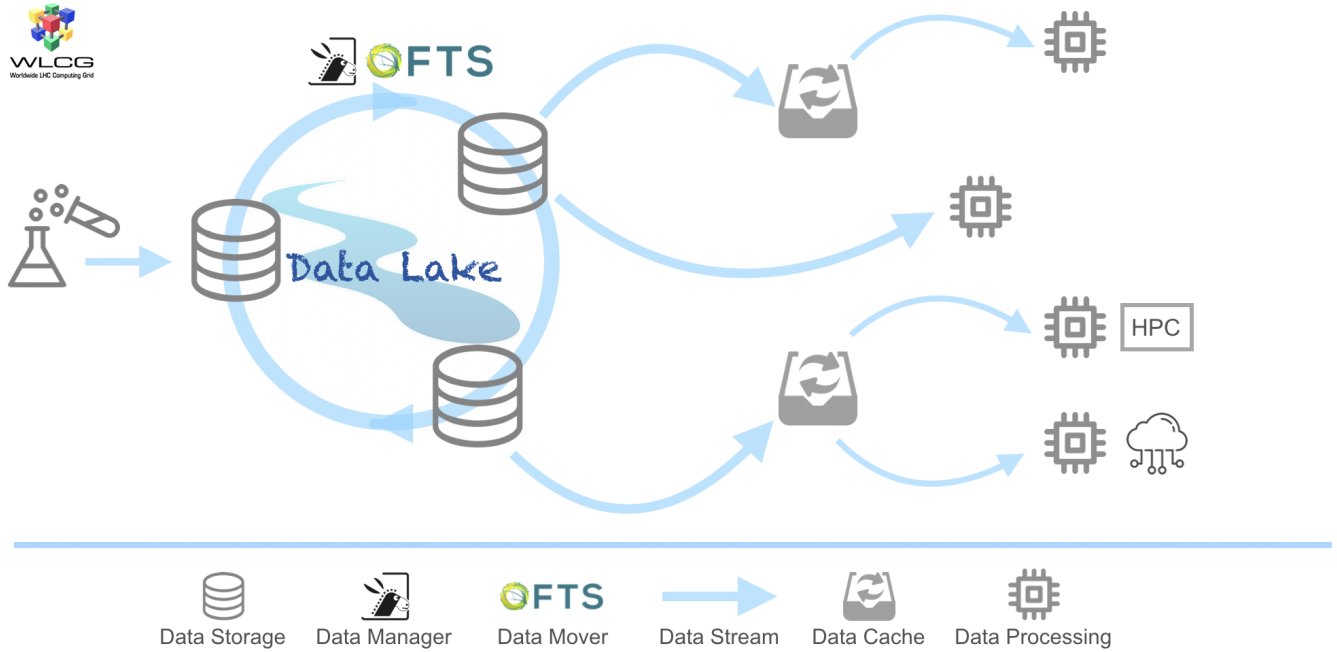


Scope of this presentation

- Future storage model mainly in the context of grid computing (rather than end-user analysis or analysis facilities)
- Focus on caching
- Focus on Xcache as caching technology
- Studies mainly from ATLAS and CMS

The model: Storage consolidation

Consolidate (managed) storage to fewer large storage centers



Datalakes, latency hiding and caching - Xavier Espinal (CERN)

- Less maintenance effort and cost, trade storage against network I/O $\rightarrow$ bandwidth expected to increase more than storage
- Allows sites to run computing only, integrate heterogeneous resources
- Caches can help to
  - Hide latency (read-ahead)
  - Reduce WAN traffic for reused files
  - Simplify storage (JBOD, no redundancy required)

Xcache

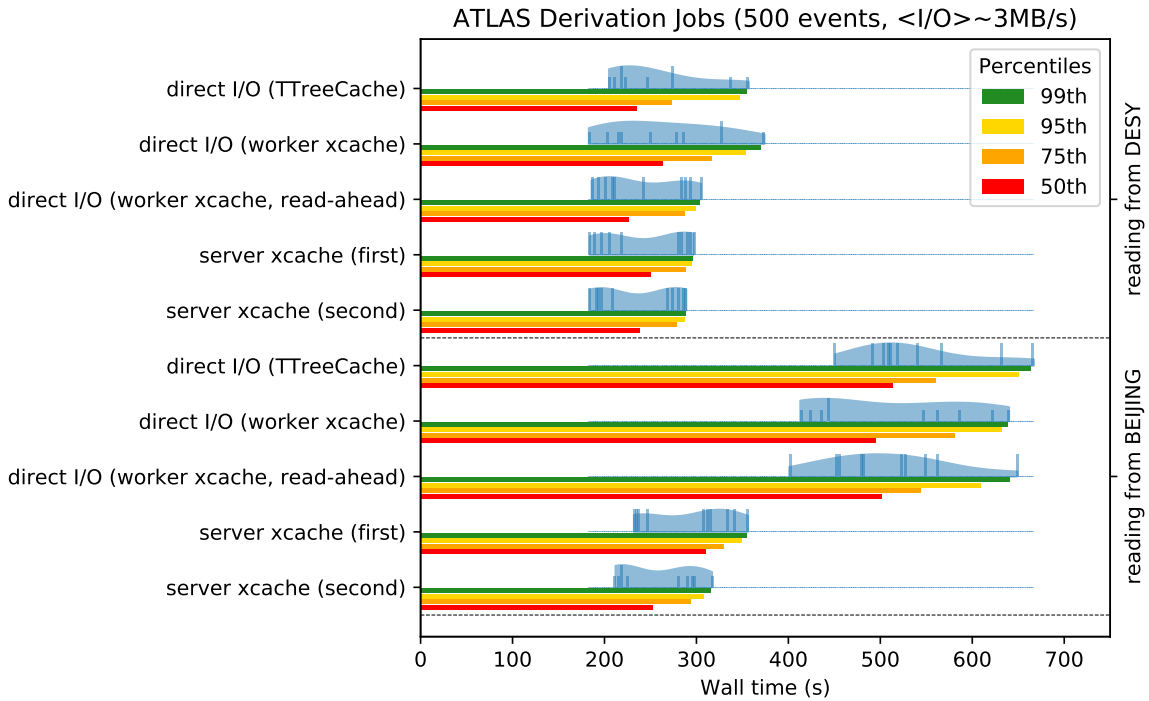
Currently most studied system at ATLAS and CMS:

- Proxy file cache implemented in the xrootd framework
- Supports partial caching in blocks
- Data forwarded as soon as it arrives
  $\rightarrow$ "Streaming cache"
- Supports read-ahead (prefetch)
- Prepend xcache url to get transfer proxied via xcache:

  ```
  TFile::Open("root:[xcache-server]:[port]//[xrootd-path]")
  ```
- Plugins internally relate file to global identifier (e.g. from rucio) to identify same file in different locations
- Can operate with individually mounted disks and distribute data for efficient access
- Caches can be combined to a cluster via a redirector

Latency hiding

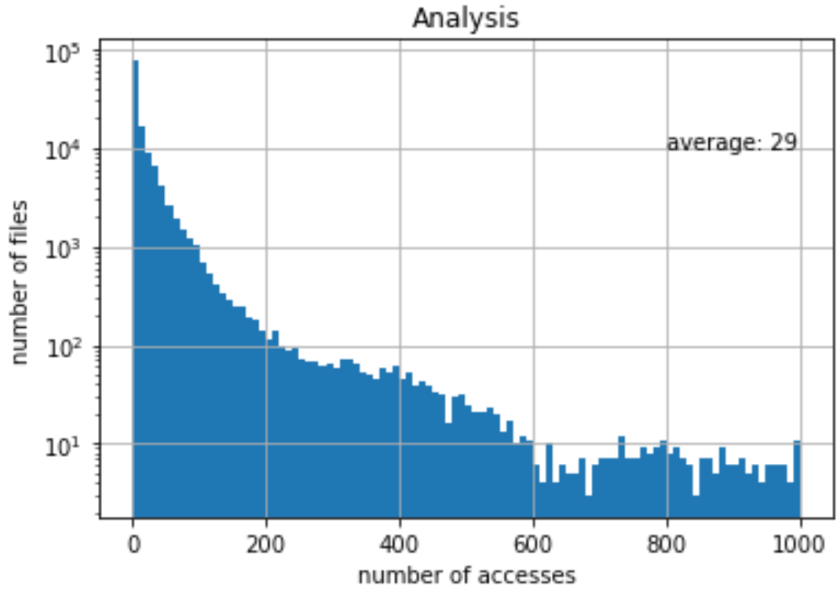
Test in Munich (idle server):

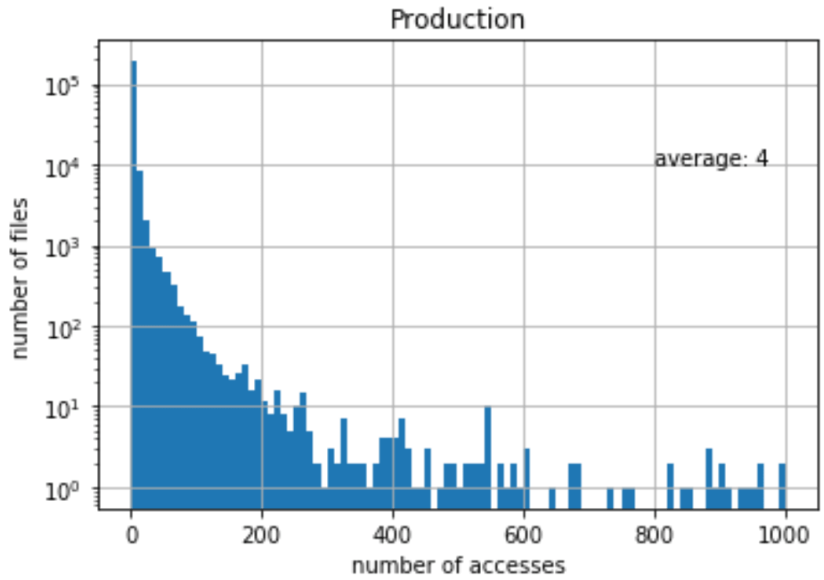


- Caching can hide latency and therefore provide similar performance for reading from far away sites as from close-by sites
- Possible on the worker node (TTreeCache, local Xcache, copy-to-scratch) or with server
- Server shows better performance
  $\rightarrow$ not fully understood yet, firewall/network setup might play a role

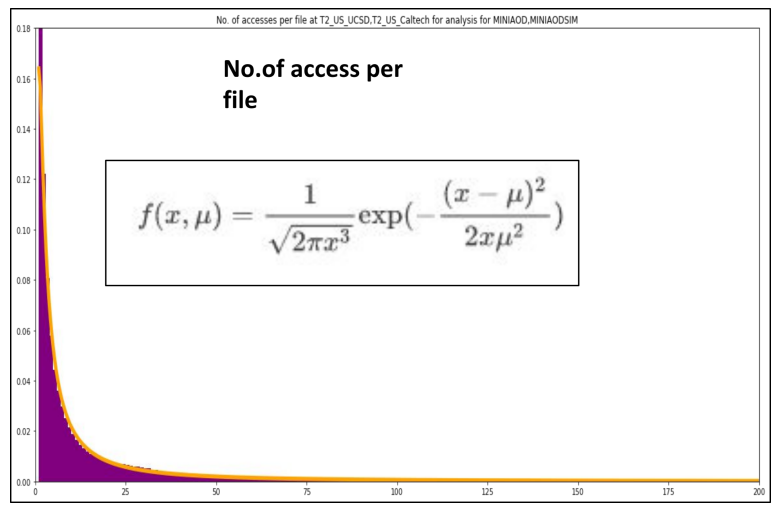
Data reuse

Study for CMS:





- Most data only accessed once
- Production data reused less often than analysis data
- Sometimes regular patterns seen (like inverse Gaussian)



Markus,Shreya@DOMA/ACCESS

https://indico.cern.ch/event/857868

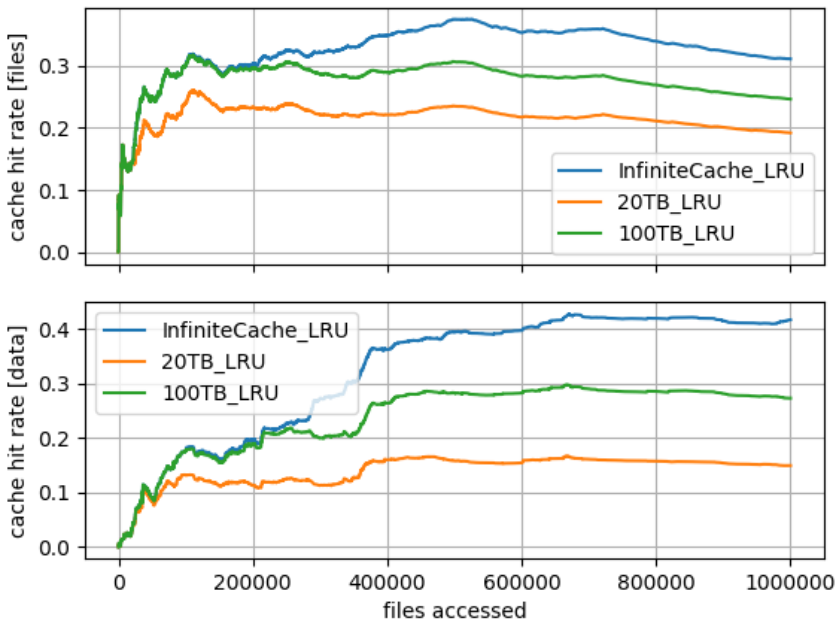
Small formats

Small analysis formats like CMS Mini/NanoAOD, ATLAS DAOD_PHYS(LITE) will play a big role

- Trend for less and smaller formats
  $\rightarrow$ More reuse
  $\rightarrow$ Ideal data for caching
- Already widely used at CMS
- Similar analysis model for ATLAS starting with Run3

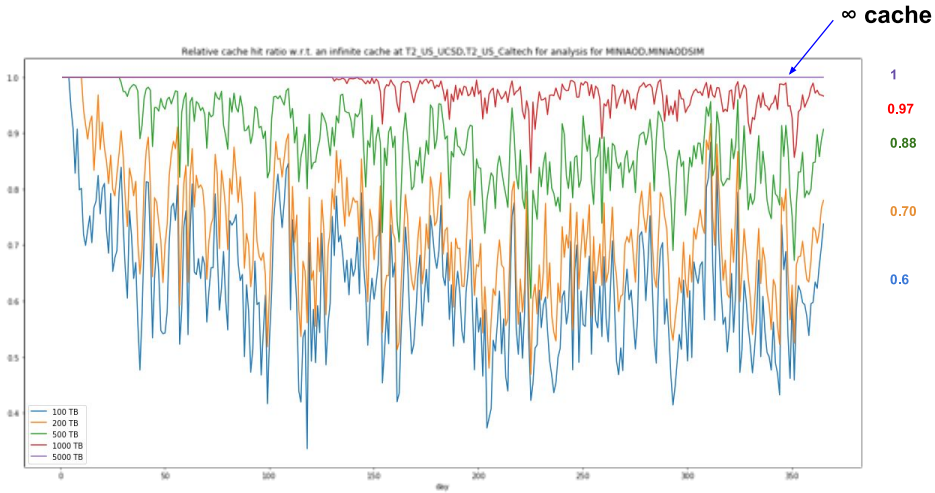
Cache size and hit rates

Hit rate vs cache size:
(ATLAS analysis + production mix)



(plot by Ilija)

Hit rate relative to ∞ cache
(CMS MiniAOD)

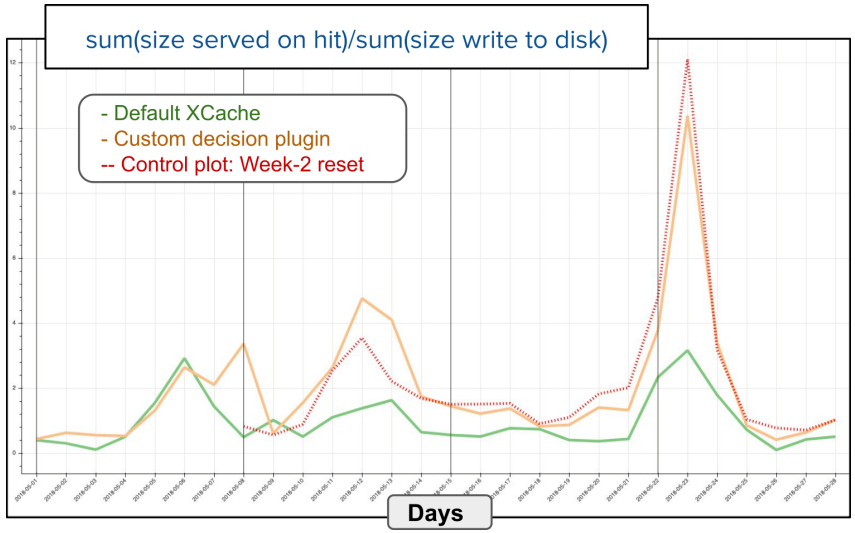


Shreya@DOMA/ACCESS
https://indico.cern.ch/event/834197

Cache hit rate saturates at a certain cache size (depending on site, jobs)
$\rightarrow$ could run with less storage than currently used

Cache (re-)placement policies

Study for CMS at INFN:

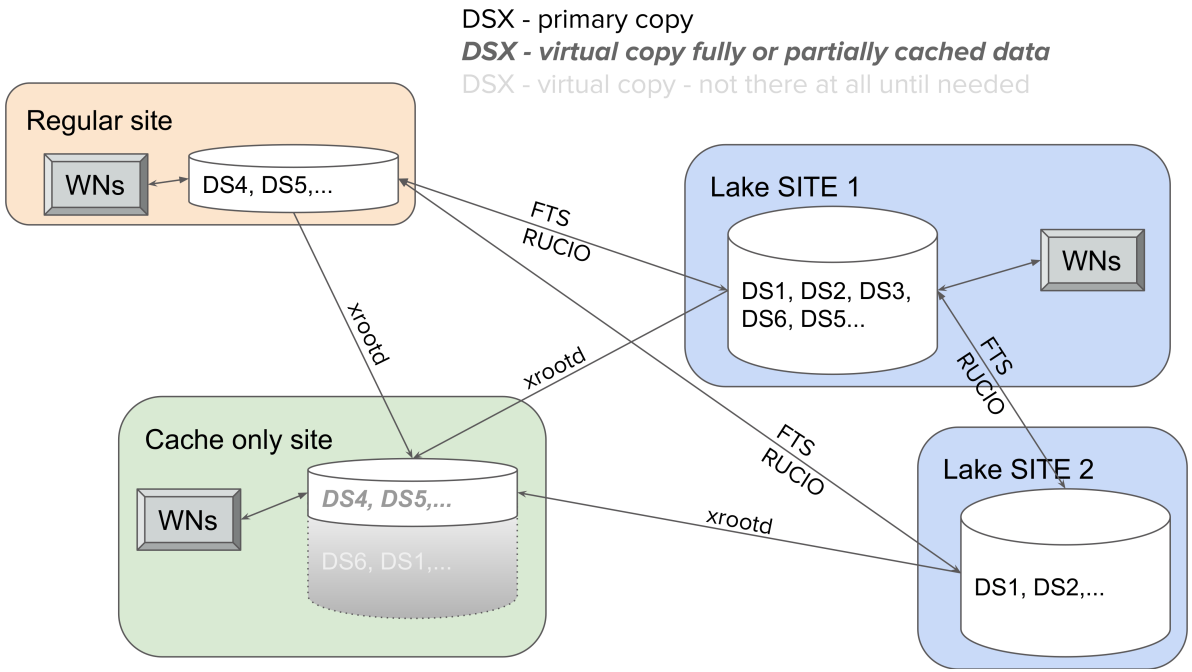


Daniele@CHEP2019, https://indi.to/yPvjG

Large potential for optimization - same hit rate for smaller cache?

- What data to consider at all for caching?
  → Default in Xcache: all
  → Decision algorithm can be added via plugin
- Which data to remove when cache full?
  → Default in Xcache: remove least recently used (check based on watermark or unconditionally purge cold files)
- Can we use ML to have dynamically adapting algorithm?
  → studies ongoing

Virtual placement

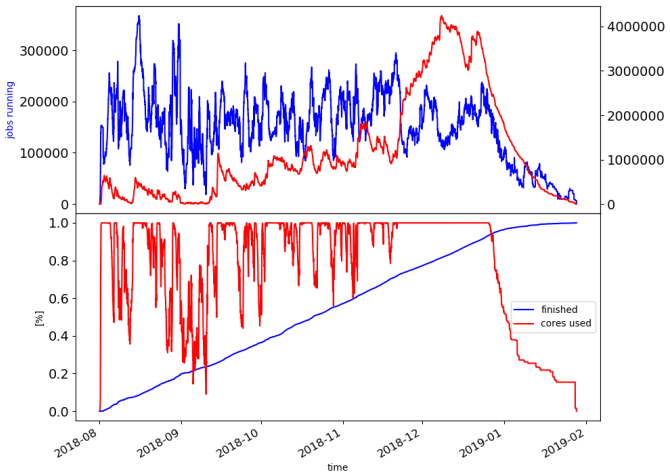


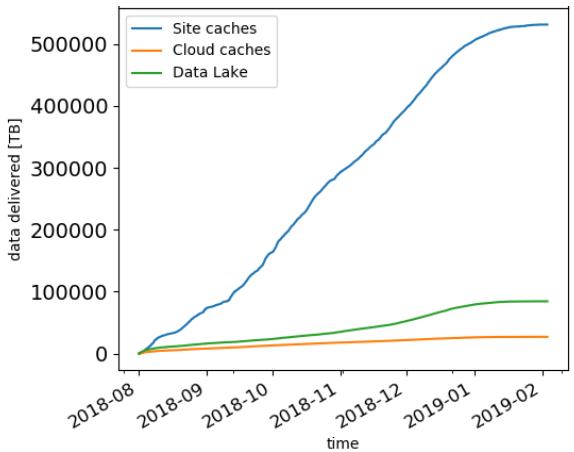
(plot by Ilija)

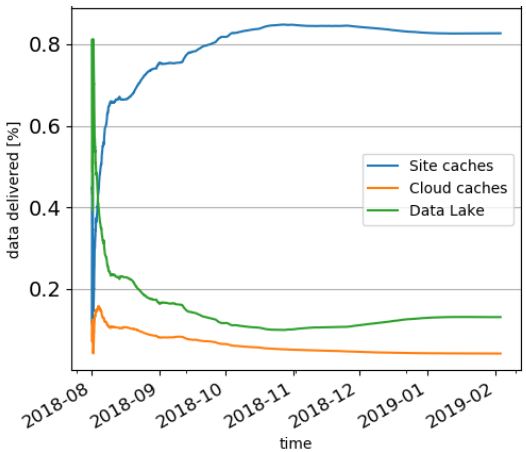
Virtually place datasets on cache only sites

- Compromise somewhere between managed storage and no management at all
- Ensures site gets jobs for same data again
- Can increase hit rates
- Currently in test phase at ATLAS

Virtual placement simulation







Ilija@DOMA/ACCESS
https://indico.cern.ch/event/769509/

Simulation for log data from current accesses:

- Time to completion similar to current situation
- 80% of data would be delivered by caches

Cost optimisation

If network and disk costs are known, can estimate the optimal cache size in terms of costs. Simulation with SoCal MiniAOD access data:

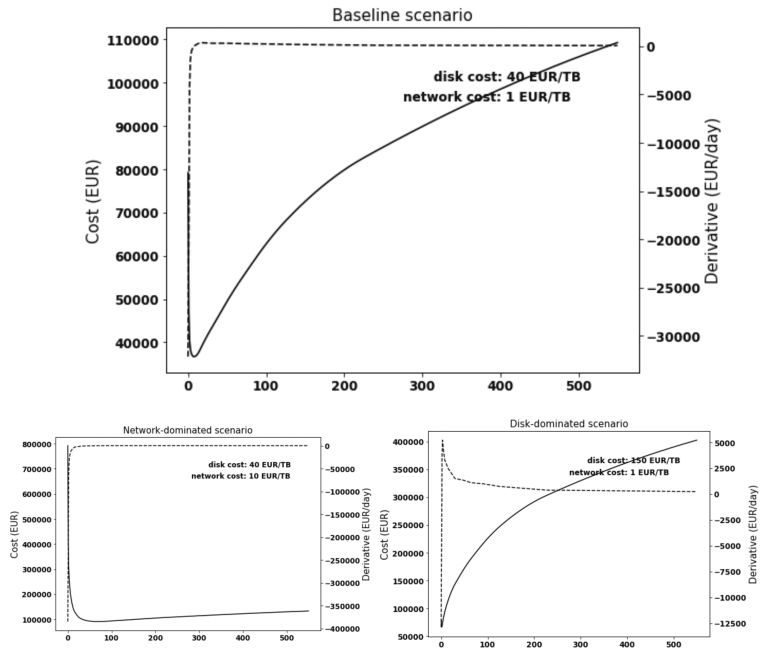
		Disk cost (EUR/TB)		
		40	100	150
Network cost (EUR/TB)	1	8	2	1
	10	68	36	26

Optimal file age (days)

		Disk cost (EUR/TB)		
		40	100	150
Network cost (EUR/TB)	1	522	252	175
	10	1262	980	870

Optimal cache size (TB)

These numbers are for 1.5 years



Andrea@DOMA/ACCESS

https://indico.cern.ch/event/866486/

Xcache in practice

Several deployments exist:

- SoCal: Common caching layer for Riverside, Caltech and San Diego for CMS MiniAOD
- Italy/INFN: Redirector at CNAF to caches across 3 CMS sites (Legnaro, CNAF, Bari)
- Caches managed via Slate at MWT2, AGLT2, BNL and PragueT2
- Munich/LMU: Xcache reading data for ATLAS production from neighbor site MPPMU
- UK: Birmingham reads ATLAS data from Manchester via an Xcache, other deployments in Cambridge and Edinburgh

Summary

- Consolidated storage ("WLCG-Data-Lake") simplifies storage organization
  $\rightarrow$ small sites can focus on CPU operation
- Caching helps to hide latencies and reduce WAN traffic
- Potential for optimization of cache (re-)placement policies
- Integration of caches into data/job management can help
  $\rightarrow$ Virtual Placement, rucio, currently tested
- Xcache provides a robust and flexible solution
  $\rightarrow$ several deployments exist
- Small formats like CMS Mini/Nano AOD, ATLAS DAOD_PHYS(lite) are ideal for caching
- More studies/evaluation needed for:
  - Xcache for analysis facilities
  - Xcache in front of HPC systems, commercial cloud

Questions for the Analysis experts

Many of the evaluations depend on how future analysis workflows will look like. We should know what we can't naively extrapolate from the current analysis model.

- How often will there be new versions of analysis formats? And how large are they?
- What fractions of files will be read?
  - Subset of events?
  - Subset of data per event?
- What will be the average time per event?
  (or in which range, will there be different classes in terms of Data intense/Compute intense?)
- Will the smallest formats (NanoAOD, DAOD_PHYSLITE) be processed on the grid or copied to local storage?