Laboratoire d'Annecy de Physique des Particules

DOMA Access & Caching report

S. Jézéquel

On behalf of the DOMA Access&Caching coordination

cnrs

UNIVERSITÉ SAVOIE MONT BLANC

Actions over last year

- During the first phase of the working group, our meetings = opportunity to collect and distribute information in our meetings
  - R&D activities on new components (mainly xcache)
    - Gain in functionalities
    - Reliability
    - Operational ‘burden’
  - Future computing models and file formats (nanoAOD/phys_lite)
  - Still interesting that ATLAS and CMS have different approaches
- Monitoring current data access
  - Understand current access pattern
  - Simulation of access pattern depending on new tool or data management policy
    - Network usage optimisation coming at same level as storage resources
- Strawman model on: Data Access on a Data Lake
  - Still at discussion level evaluating pros and cons
  - Progress will rely on outcome of component R&D and data format

Xcache status

- Most popular component these days
  - Pros : Caching, read-ahead, root aware (although http protocol should be available)
  - Gain :
    - Re-think and re-adapt storage in view of resource optimisation and less operational burden
    - Direct access to remote sites with caching for popular datasets
    - Different implementation and models being evaluated with real workflow (analysis and production):
      - US (ATLAS/CMS), UK (ATLAS), DE (ATLAS), IT (CMS)
      - New contributors coming (Spain, Russia)
- Development : Globally ready for production → only focus on consolidation
  - Long term reliability
  - Reliability under heavy load

- Deployment
  - Discussions on SLATE from security point of view ramping up (R. Gardner, R. Wartel)

Presentation over summer

‘CMS cache studies based on data lifetime’ (Shreya Krishnan. Open Lab)

Link

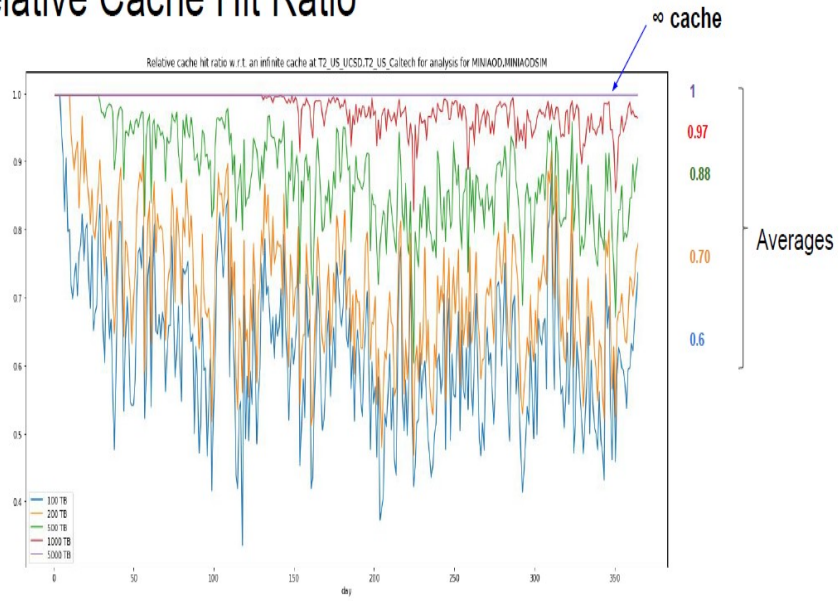
- Update at a future DOMA ACCESS meeting (24th September)

- Quantify trade-off between cache size and WAN traffic
  - Relative change in hit rate and data lifetime
  - Network activity

  as function of
  - Cache size
  - Proactive deletion policy (Delete file in cache if not used after N days)
- Extrapolation from CMSSW Popularity Data Set
  - CMS analysis jobs : Focus on MINIAOD (mostly reused format)
- Starting with SoCal (UCSD and CalTech)

Presentation over summer (2)

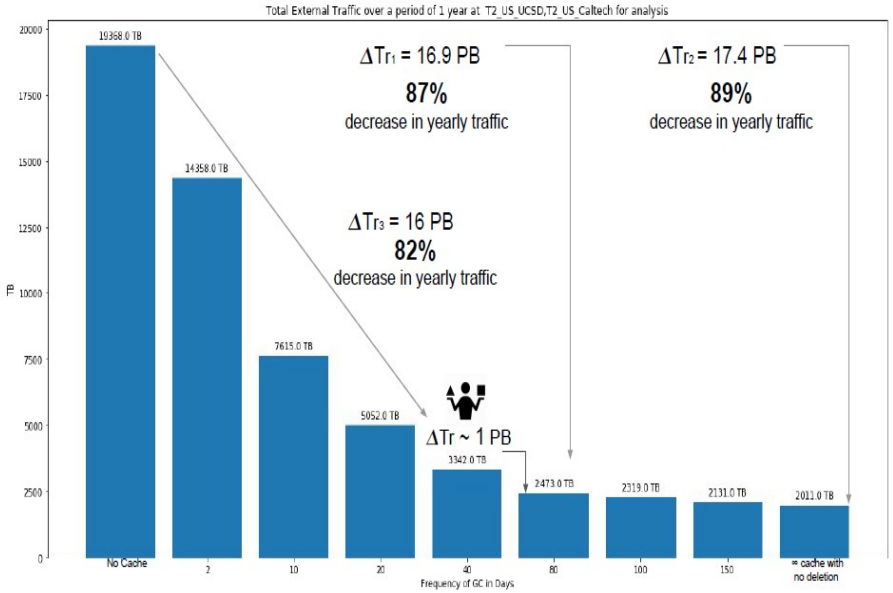
Relative Cache Hit Ratio



#Note: High water mark algorithm was used here.

Yearly External Traffic at SoCal

ΔTr = Difference in External Traffic



8

Preliminary conclusions

- ★ Caches using purging strategies where N ≥ 80 work a lot like ∞ cache with no deletion.
  - This means that a large proportion of files are not re-accessed after 80 days.
- ★ Results are more or less consistent with the current active deletion timeline of 85 days

CHEP

* 1 overview presentation on DOMA ACCESS + many others

  → Ensure to give a coherent picture would be usefull (to be started now)