Activités « DOMA-acess »
Content Delivery & Caching

L. Duflot

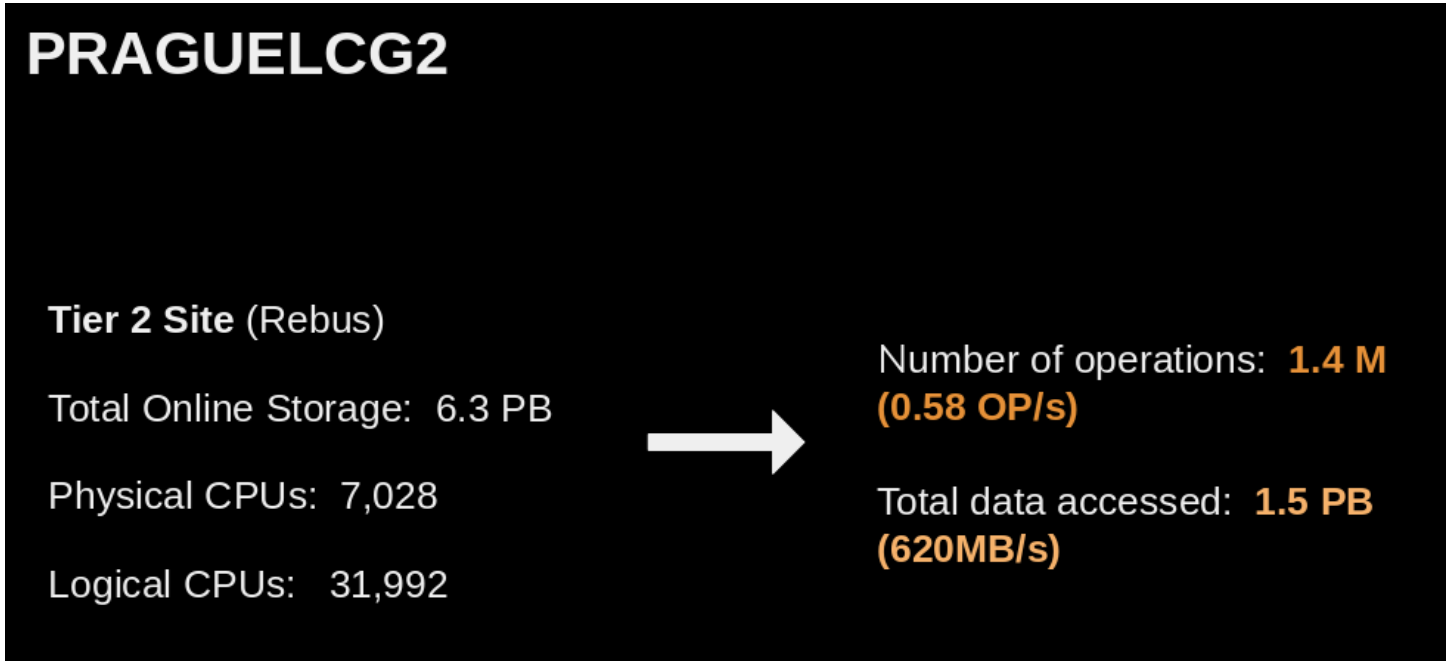
Introduction

- Ceci n'est pas un résumé mais une selection des sujets qui m'ont semblé les plus prometteurs pour WLCG (surtout HL-LHC)
- Ce groupe se concentre sur les moyens d'accès aux données mais inévitablement a un certain recouvrement avec les autres groupes DOMA : DOMA-access twiki - mandat en backup
  - Il a recensé les activités en cours DOMA-access links
  - Ces activités ont commencé à être présentées plus en détails dans les 5 dernières réunions INDICO DOMA
    - Sur les activités de caching pour commence avec pour but
      - Optimisation accès distant
      - Réduire la charge sur SE local
      - Viabilité pour un site SE-less ou cluster local
  - Coordinateurs : F. Wuerthwein (U. California), I Vukotic (U. Chicago), S. Jezequel (LAPP)

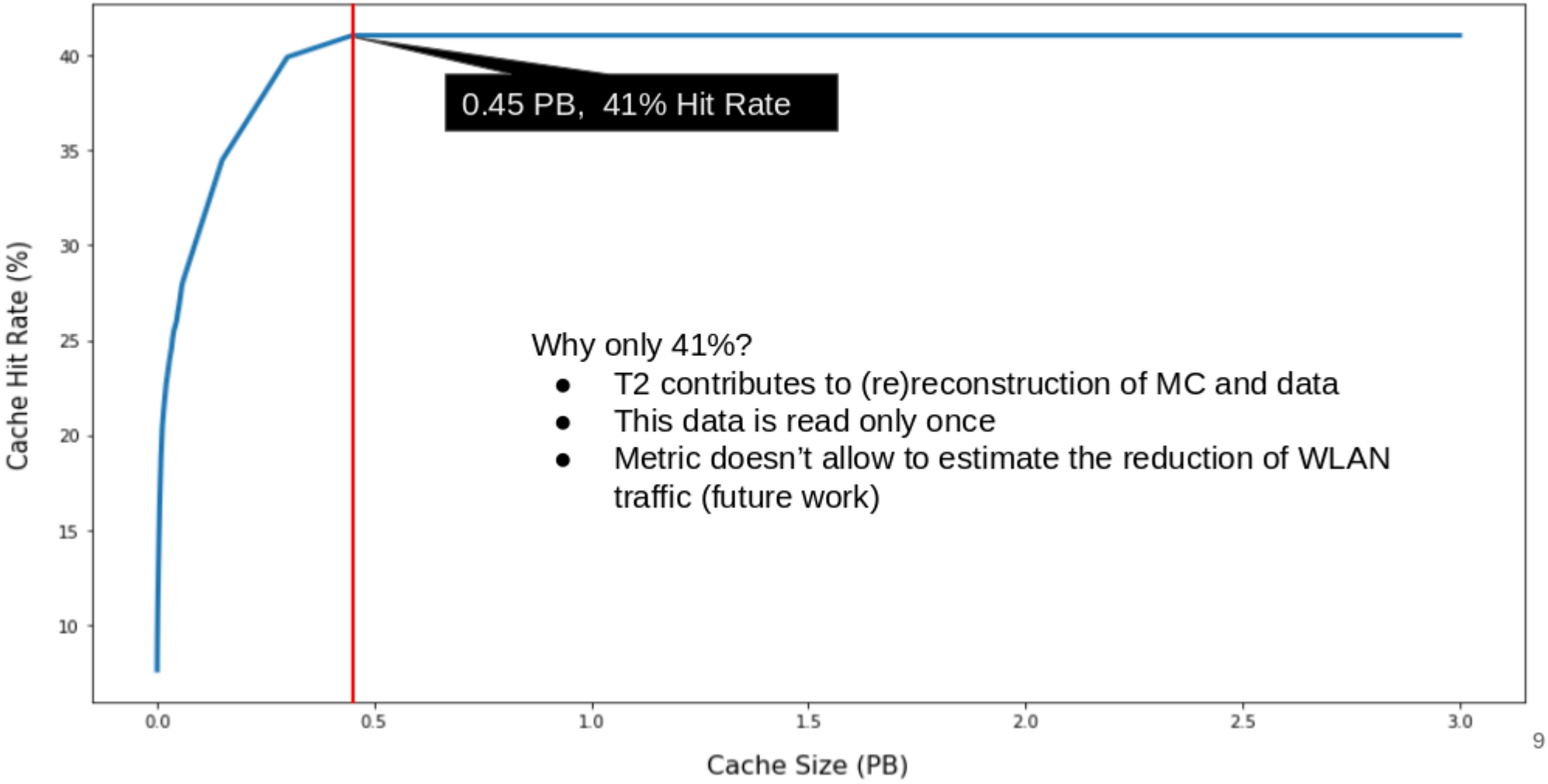
Simulation de cache

@CERN avec étudiant d'été

- Utilisation des logs du site T2 Prague sur un mois pour simuler l'utilité de cache
  - Fait de la production et de l'analyse
  - Accède essentiellement aux données locales
- Estimation du « Hit rate » : hit/ (hit+miss)



Hit Rate vs Cache Size for PRAGUELCG2



Why only 41%?

- T2 contributes to (re)reconstruction of MC and data
- This data is read only once
- Metric doesn't allow to estimate the reduction of WLAN traffic (future work)

9

Simulation de cache avec les logs ATLAS

XCache simulation

Pilot reports all file accesses to Rucio which stores info in HDFS. This gets indexed in ES at UC. Use ML platform to simulate cache at any site at any period of time.

All simulations:

- Full file caching
- Low water mark 85%
- High water mark 95%

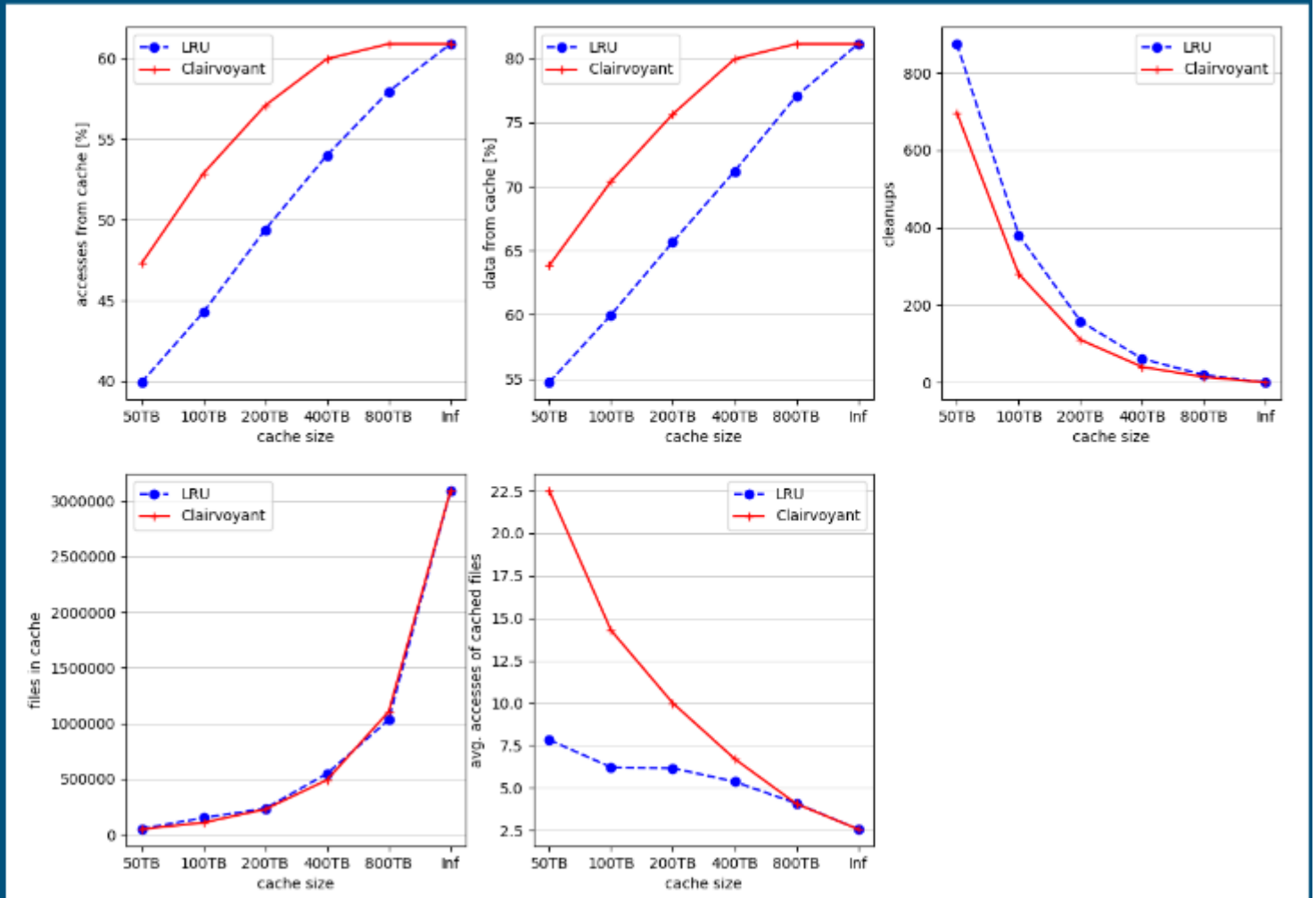
- **LRU** - least recently used cached files are expunged first
- **Clairvoyant** - first expunges files that have the next access furthest in the future. Impossible in practice but gives measure of what is theoretically best possible cache.

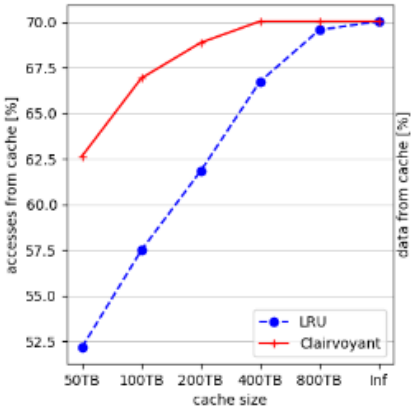
All the input files

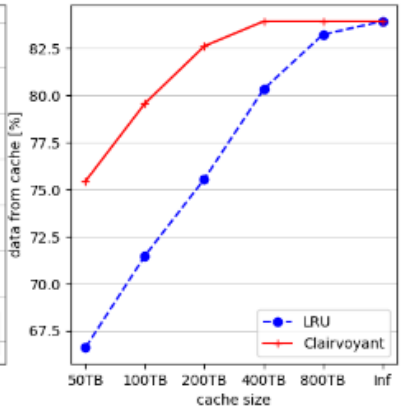
In total MWT2 jobs accessed 9.5 PB of data in August.

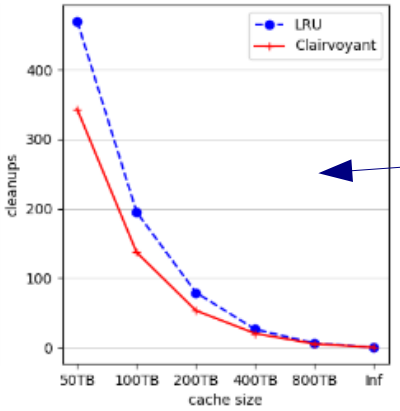
40% of accesses and 55% of traffic could have been served from 50TB cache.

Would probably be best served by two nodes (40Gbps NIC, with 5TB of SSD and 16 HDD each).

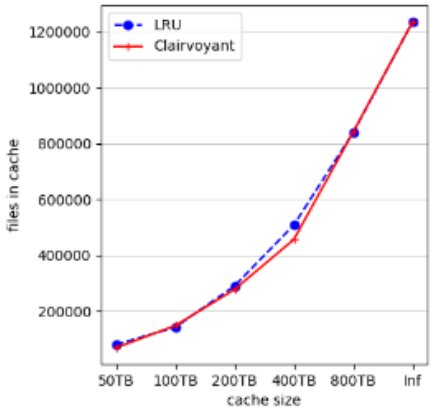


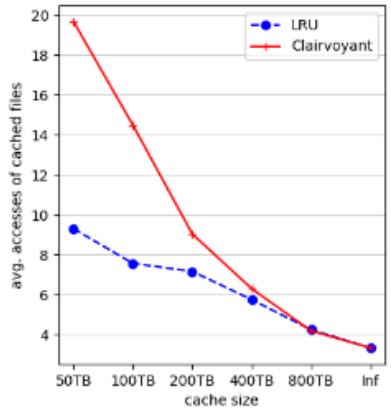




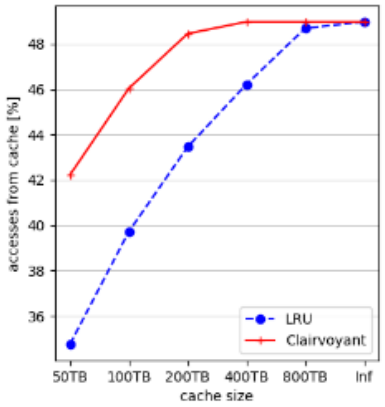


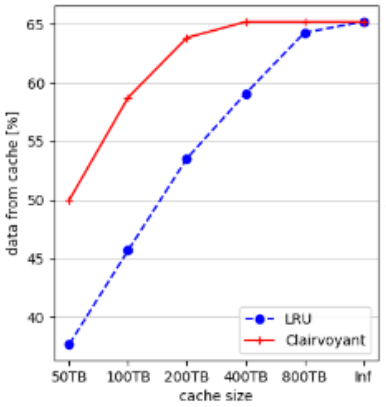
Production

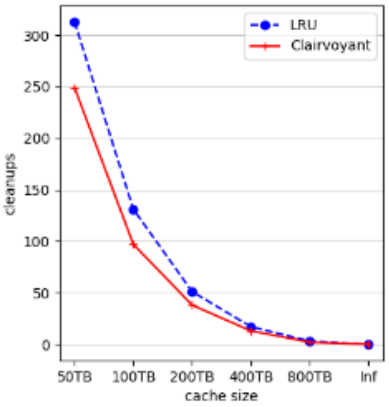


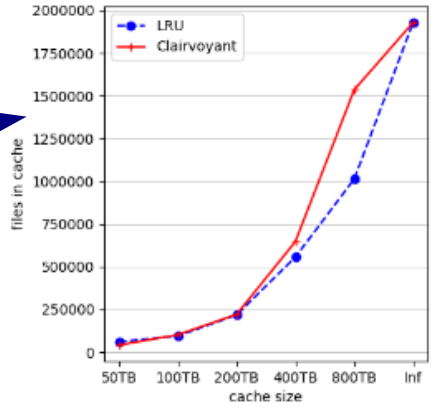


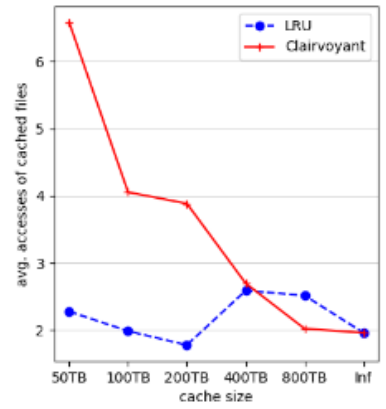
Analyse











Caching specific files?

Ideal files to cache: frequently used and re-used, small

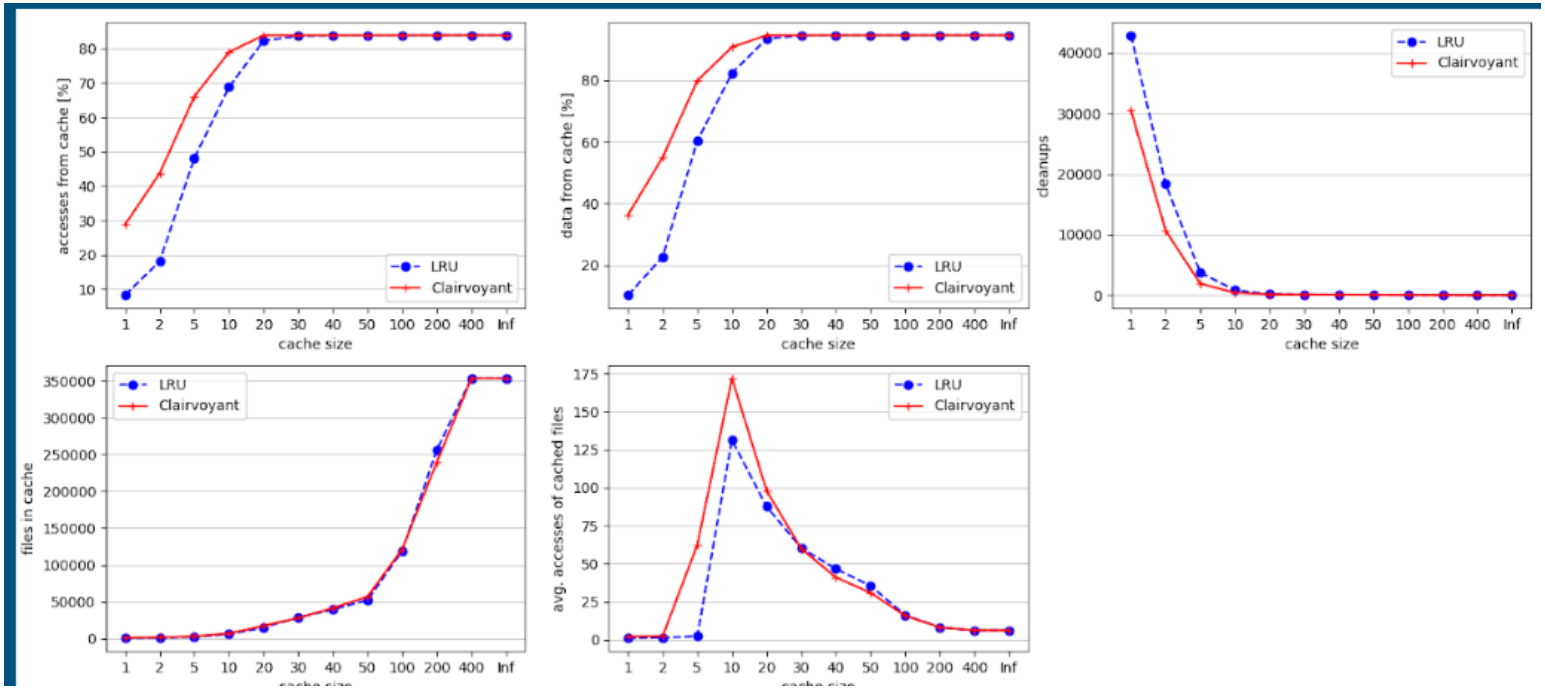
Significant reuse but large size.

Small but no reuse at all.

Small not much reuse but reuse is maybe concentrated?.

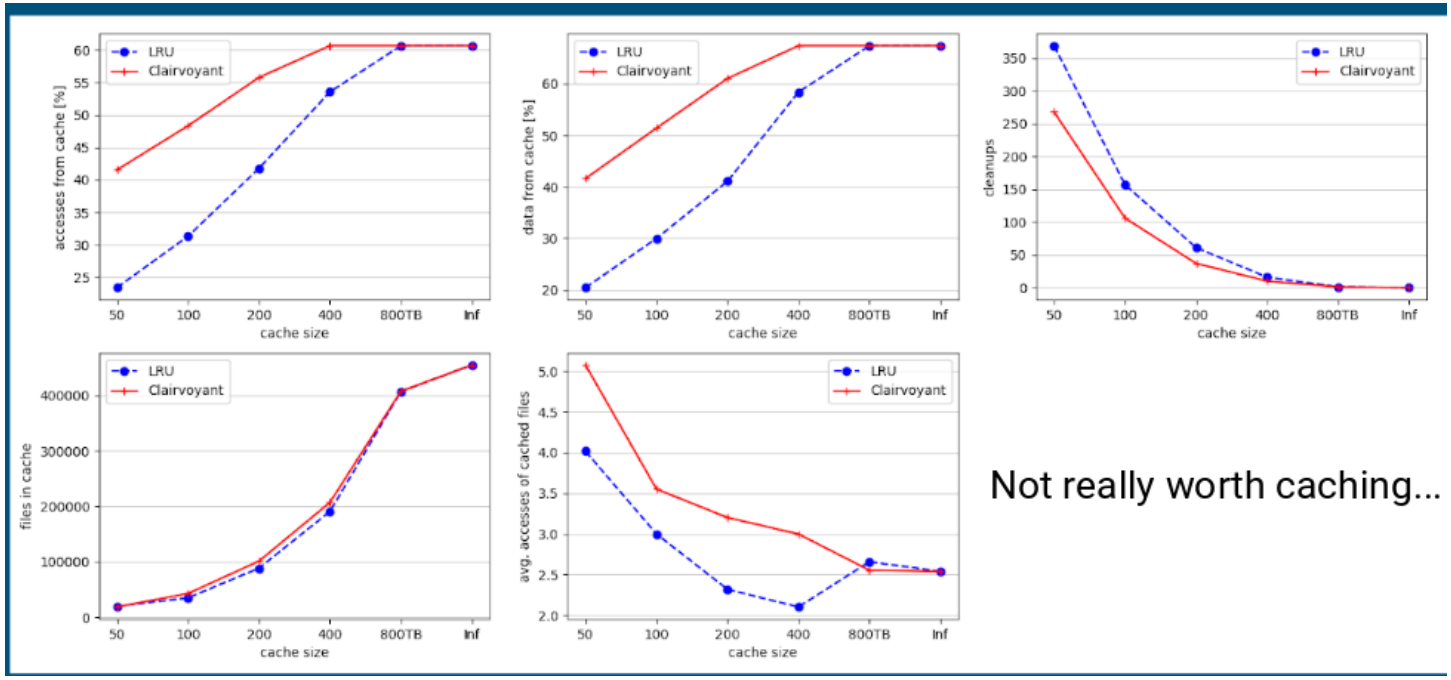
Small files, large reuse.

filters	Count	Average filesize	unique files	Sum of filesize
HITS*	2,188,827	2.2GB	351,380	4.6PB
AOD*	1,156,100	2.1GB	451,091	2.32PB
DAOD*	985,921	1.31GB	400,588	1.23PB
panda.um.*	906,694	8.55MB	900,922	7.4TB
TXT*	726,415	31.78MB	251,770	22.02TB
EVNT*	633,834	30.38MB	558,428	18.36TB
*.lib.tgz	544,510	53.52MB	13,186	27.79TB



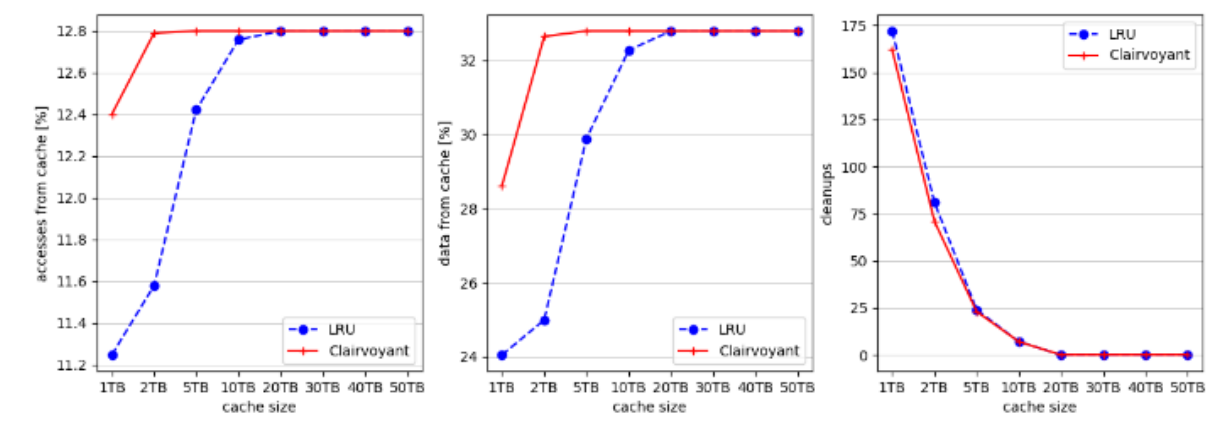
HITS

AOD



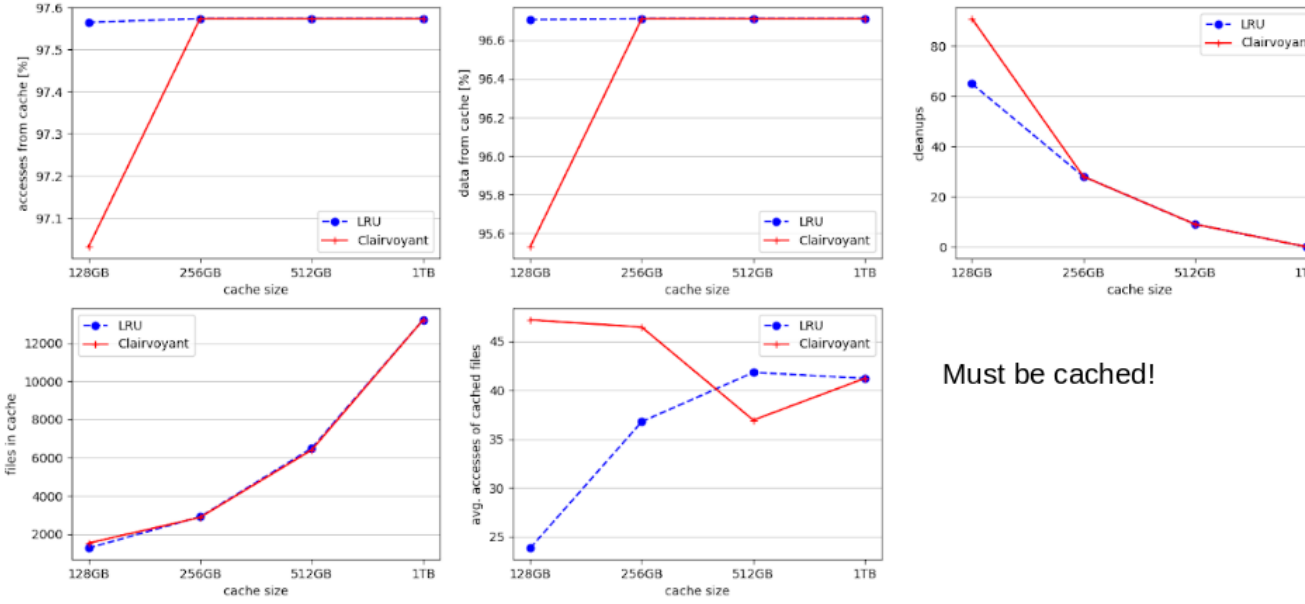
Not really worth caching...

EVNT



filesize	Count
0B	585,942
64MB	540
128MB	1,369
192MB	12,194
256MB	144
320MB	3,665
384MB	28,003
448MB	1,976
576MB	1

lib*.tgz



Fort taux de réutilisation mais ces fichiers sont créés sur SE local, jamais copiés d'ailleurs

Must be cached!

Mes remarques sur les simulations

- Les conclusions peuvent varier un peu selon le site : plus ou moins de production, type de production, rôle spécial d'un site etc.
- Les simulations sont basées sur le système actuel où la production pré-positionne des fichiers qui seront naturellement plus souvent utilisés. Plus généralement, les jobs sont envoyés vers les fichiers, ce ne sera pas forcément le cas pour des fichiers en cache (sauf si l'information est remontée centralement).
- Dans le workflow d'ATLAS, les fichiers HITS sont pré-positionnés sur les sites, on voit qu'ils feraient de bons candidats pour être en cache, du moins ceux qui sont réutilisés couramment (pileup).

Quel genre de cache?

Options

How much we gain from cache will depend how we use it. Different use cases will need different developments and integrations in existing systems. Some developments can be used in multiple options.

- Node local cache
- Local storage speedup
- Site with no managed storage capacity
- Full stack caching (FB model)

Goal here is to investigate these options and consider operational impacts in production environments.

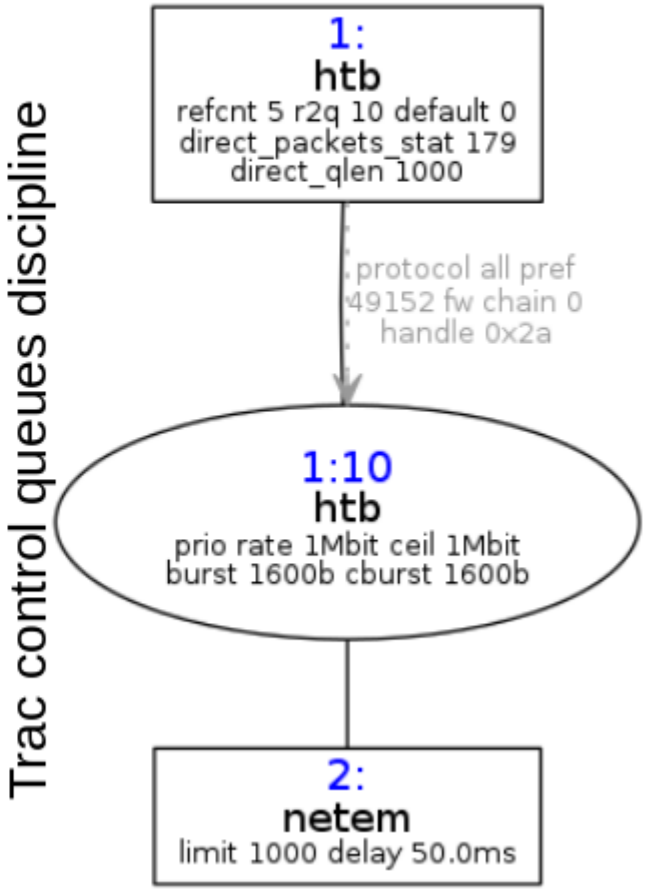
2

Avec des outils pour simuler latence ou limitations réseau

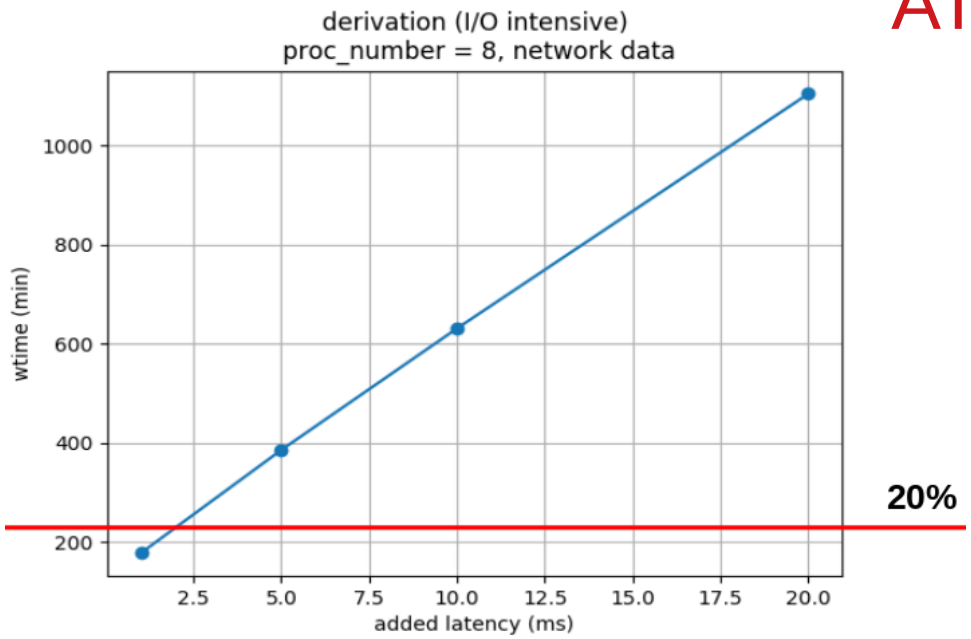
@CERN avec étudiant d'été Présentation

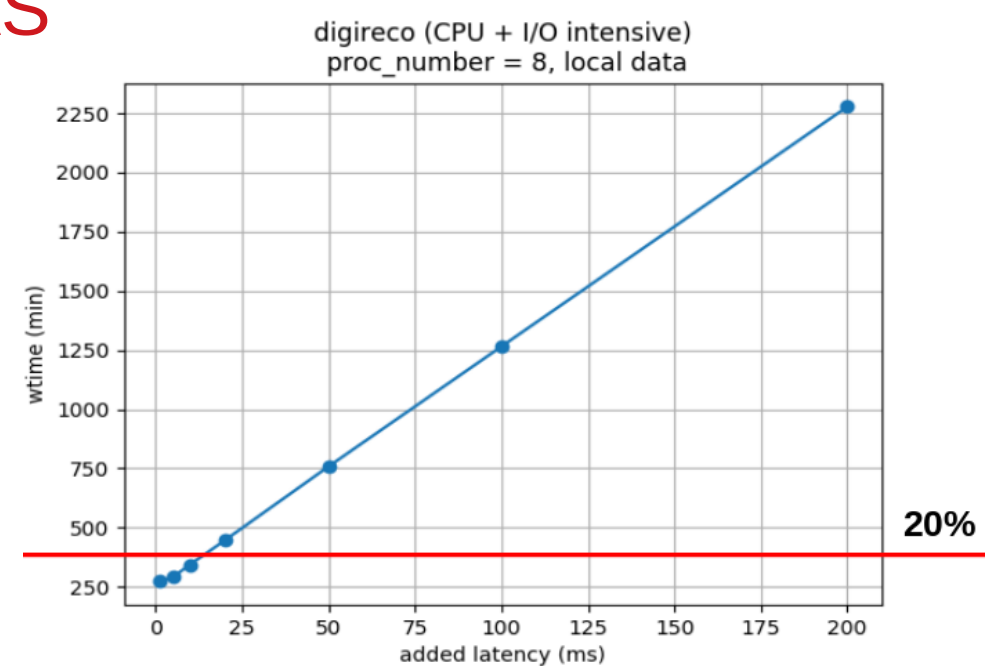
Adding Latency/Bandwidth Limitations

- Cgroup: mark egress and ingress packets from a group of processes
- Egress trac : Trac Control to add rate limit and/or latency
- Ingress trac: iptables module HASHLIMIT to drop packets above a given rate
- Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz 32 cores

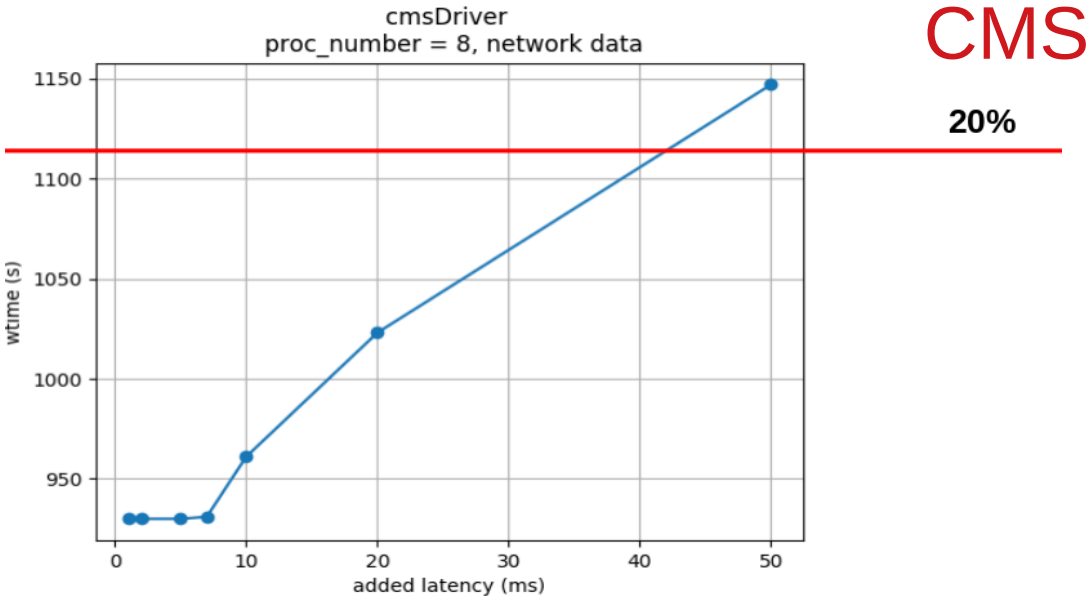


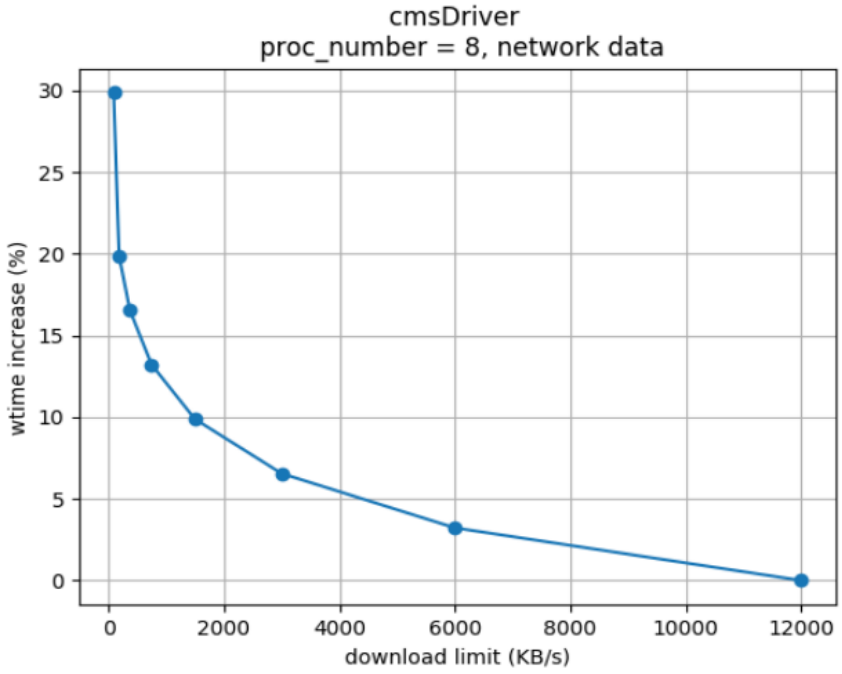
ATLAS


derivation (I/O intensive)
proc_number = 8, network data
wtime (min)
added latency (ms)
1000
800
600
400
200
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
20%


digireco (CPU + I/O intensive)
proc_number = 8, local data
wtime (min)
added latency (ms)
2250
2000
1750
1500
1250
1000
750
500
250
0
25
50
75
100
125
150
175
200
20%

CMS


cmsDriver
proc_number = 8, network data
wtime (s)
added latency (ms)
1150
1100
1050
1000
950
0
10
20
30
40
50
20%


cmsDriver
proc_number = 8, network data
wtime increase (%)
download limit (KB/s)
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000
12000

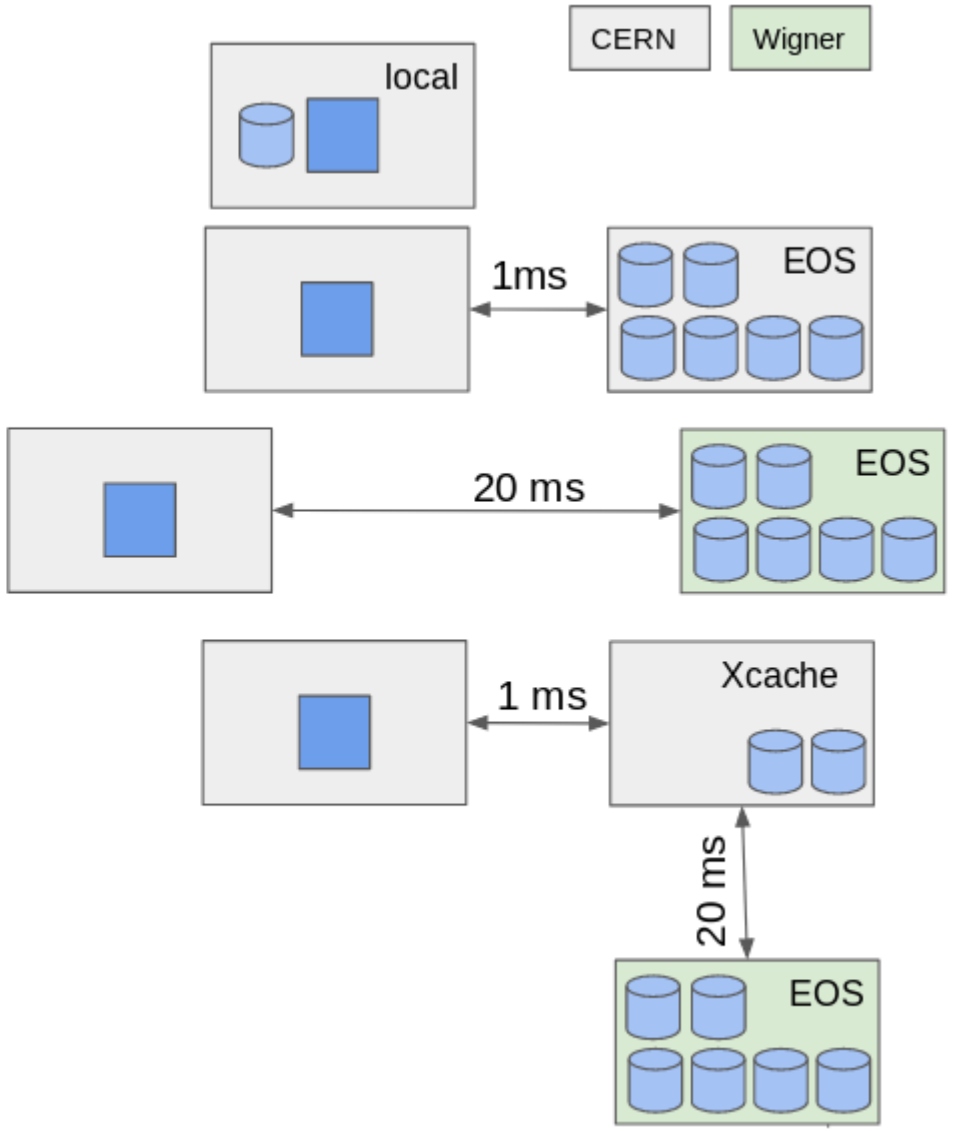
Xcache et latence

@CERN avec étudiant d'été

Présentation

Data Caching: Xcache

- Jobs:
  - ATLAS digi-reco 28GB input
  - ATLAS derivation Job 45GB input
- Setup
  - Data on WN (local)
  - Data on the same site (remote close)
  - Data at Wigner (remote far)
  - Xcache server at the same site (<1ms)
- Goal
  - Measure the impact on throughput



Xcache et latence

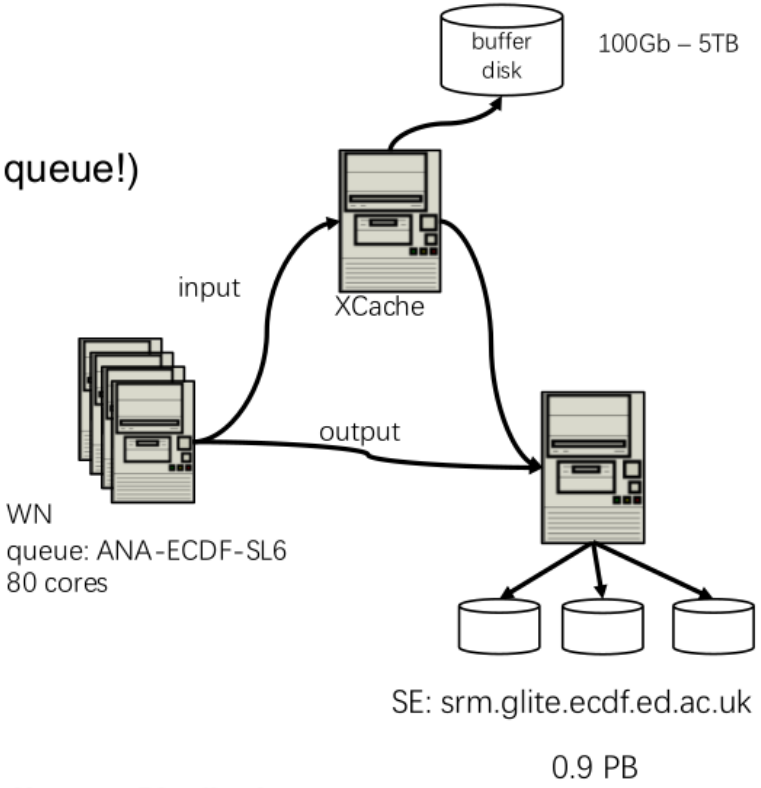
Measurements (Preliminary!!!)

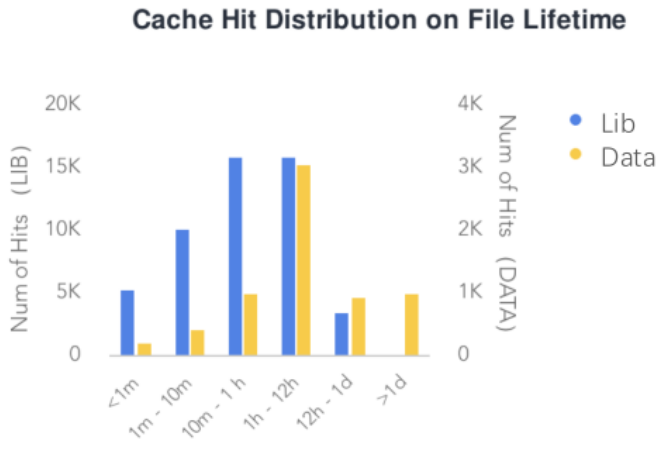
Job type	Run's conditions	Run time	Relative Run time
Atlas-mcdigi-reco	Local Data	240m19s	1.00
	Remote Far	480m28s	1.9
	Empty Cache	261m39s	1.08
	Populated Cache	249m43s	1.04
Atlas-derivation	Local Data	147m19s	1.00
	Remote Far	1217m14s	8.26
	Remote Close	151m24s	1.02
	Empty Cache	155m17s	1.05
	Populated Cache	152m44s	1.03

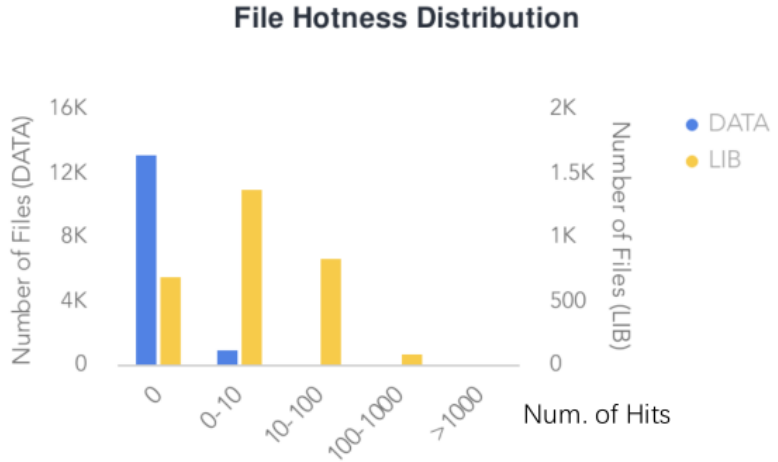
Lesson learned: xcache can, for the tested workloads, hide latency efficiently

Xcache testing at ECDF

- 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





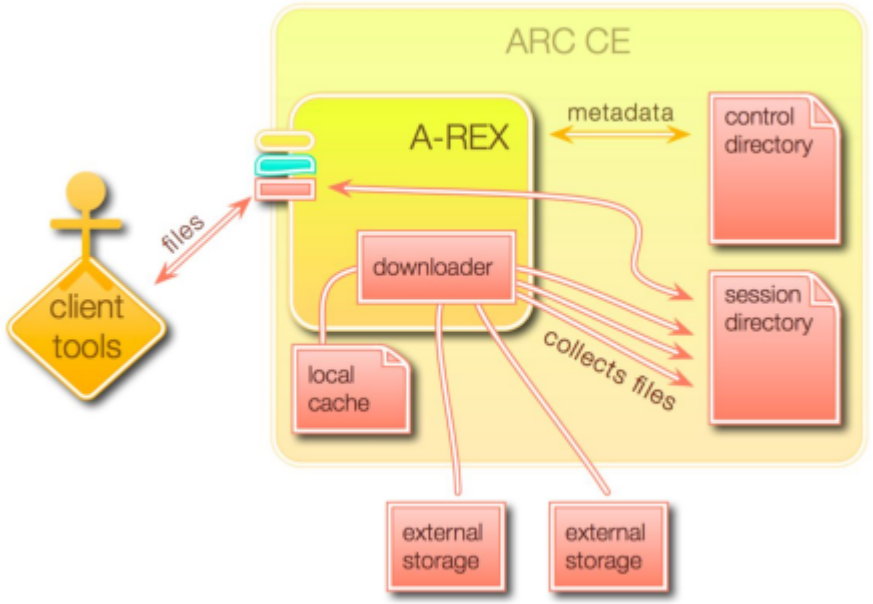


Remarques :

- ECDF commence a utiliser la simulation pour comparer à ses mesures, tout n'est pas parfaitement compris. Simulation également de la queue de production.
  - Quels types de fichier garder en cache ?
- Dans une perspective où le site deviendrait cache-only, le caching des lib*.tgz devient pertinent
- Même si les cas étudiés ne sont pas les mêmes, il est surprenant que l'optimum de taille de cache soit aussi différents des optimum dans le cadre des simulation (GB vs TB!).
- Xcache reste une technologie à valider pour une utilisation à grande échelle.

ARC-CE cache

- Le CE pré-remplit le cache avec les fichiers d'input avant de soumettre le job.
- Utile pour exploiter des sites qui n'ont pas de SE : un filesystem distribué sur les WN (e.g. CEPH, GPFS, …) suffit.
  - Clusters non grille
  - HPC



Californifederation

- Utilisation des options de configuration de xcache pour partager un lots de datasets utiles pour l'analyse locale entre UCSD et CalTech
  - Règles sur les URL à « cacher »
  - Tests de performance, ça marche bien
  - Possibilités d'extension à d'autres sites
- Utilisation de cache pour gestion « automatisé » d'un espace T3

Mes conclusions

- Des outils de simulation sont mis en place
  - Basés sur l'historique d'utilisation dans les sites
- Intérêt démontré d'un cache pour accès distant :
  - Par ex. cas des fichiers HITS dans la simulation ATLAS
  - Xcache semble utile sur les use case ATLAS testés même quand le fichier n'est pas dans le cache (masque la latence réseau tant que le taux d'I/O est plus petit que la bande passante)
  - Cela pourrait être utile dès aujourd'hui !
  - Faut-il que le contenu du cache soit publié centralement ?
  - Tout ne vaut pas le coup d'être mis en cache !
- Intérêt de caching de block de fichier ?
- Utilisation de cache « managé » pour des sites non grille → ARC-CE
- Sites pur cache → fédération californienne
  - Solution pour remplacer un SE dans un petit site ?

Backup

Mandat du groupe

Mandate

- Study Data Access by applications including content delivery services and networks, smart clients, caching solutions for LHC experiments and sciences with similar challenges.
- Provide a forum to share experience gained on improving remote and local data access by the Experiments' current and future workloads, taking into account developments such as machine learning and compact analysis data formats.
- Gather measurements and metrics on the different data access studies and compile quantitative information with the primary goal to be used by the WLCG DOMA activities.
- Based on the shared experience and the needs of WLCG-DOMA this WG will identify, by open discourse, areas where further R&D is required and prioritise these topics
- This process is intended to stimulate collaboration between different parties and foster increased commonalities between experiments, storage solutions and site infrastructures
- This WG will track and regularly report on the progress of the identified topics
- This WG will maintain close links to relevant the DOMA project WGs and activities: TPC/Protocols, Networking, QoS, Authorisation and Authentication, etc.

Liste des sujets abordés

- Xcache
  - Utilisation dans une « fédération » de sites californien
  - Utilisation au T2 d'Edinbourg (EDCF)
  - Latence, bande passante et workflow
- Simulations :
  - Simulation de cache basé sur les log du T2 de Prague
  - Simulation de (x)cache basé sur les log ATLAS
- ARC cache
- Rucio et les caches (dont volatile RSE)
- Cas d'utilisations de caches dans ATLAS
- OSG data federation (~ distribution des données via CVMFS)