Laboratoire d'Annecy de Physique des Particules

DOMA ContentDeliveryCaching : Cache

S. Jézéquel, I. Vukotic

ADC Jamboree

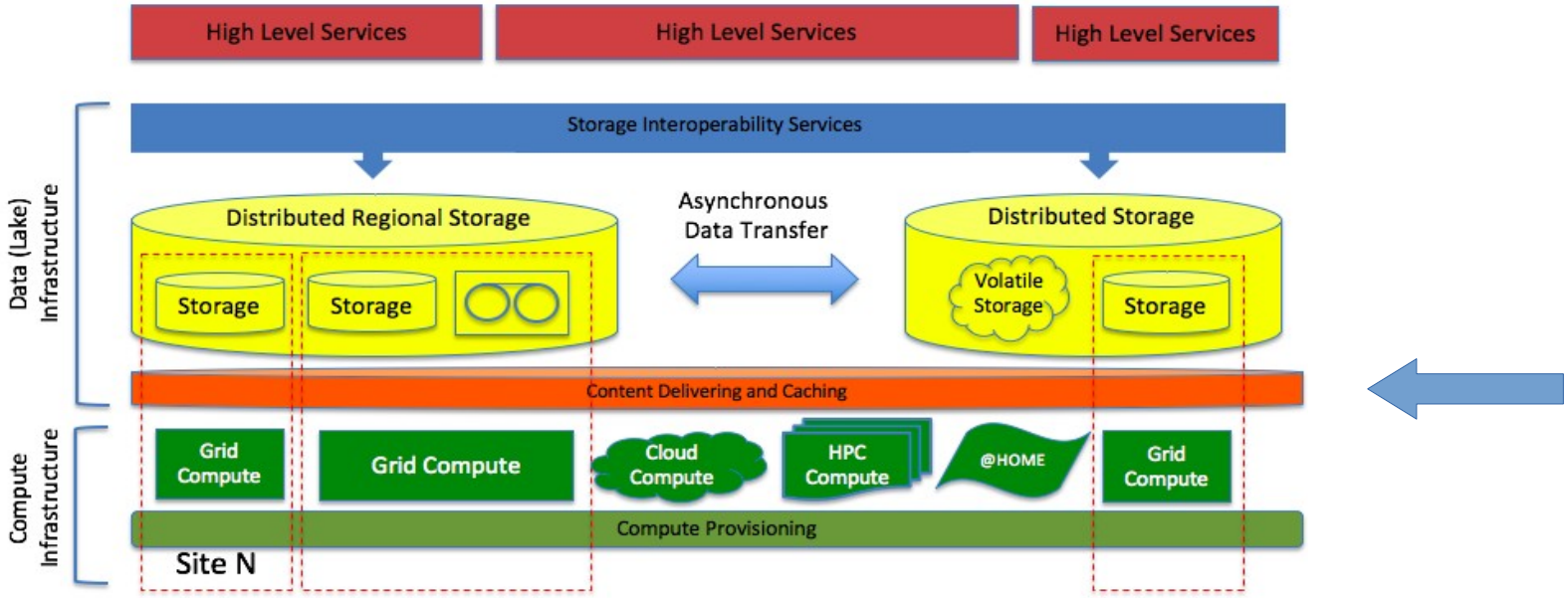
7 March 2019

cnrs

UNIVERSITÉ SAVOIE MONT BLANC

DOMA ContentDeliveryCaching

- Twiki : https://twiki.cern.ch/twiki/bin/view/LCG/ContentDeliveryCaching
- R&D for HL-LHC
  - With smooth transition over Run3 to validate and correct options
  - Computing TDR ~2022
  - General DOMA-Access meetings : each 2nd tuesday (agendas)
  - ATLAS-DOMA Access meetings on friday afternoon (agendas)



Interest of caching technology

- Simpler technology to copy/access files
  - Minimal maintenance by site admins (could be even delegated)
  - Deletion is handled locally
  - Adapt hardware to match speed access requests (SSD, HD)

- Keep popular files
  - Optimise data reuse by Grid or local batch without WAN trafic
  - Standalone process to decide files to delete and do it
  - Data potential reusage evaluated through simulation of data access
    - Accessible through past access pattern stored in Elastic Search (U. Chicago) (more in Analytics presentation)

- Speed up remote access
  - Use read ahead mechanism optimal for long distance
  - Partially hide possible network issues (keep fraction of file already transfered)

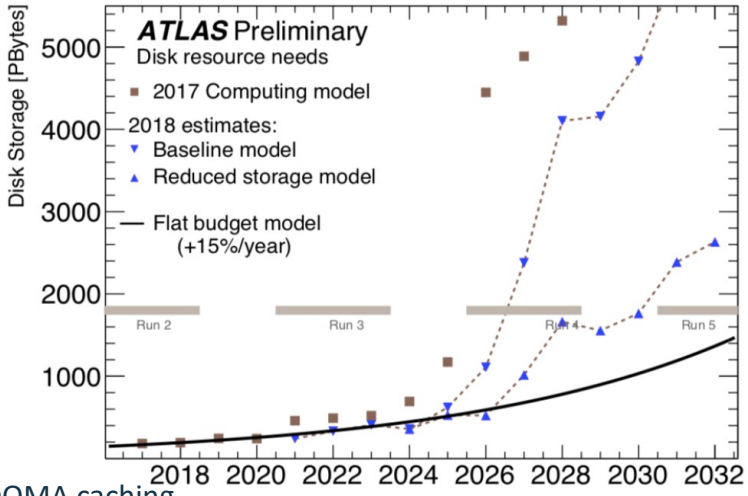
Not solving all problems

- Not used to write job output
  → Request ‘standard’ storage (Grid or other) to export data (trafic 5 times less than input transfer rate)

- Current evaluated model assumes no publication of content to Rucio
  - Job brokering should be smart to adapt to potential location (example : **Scheduling with Virtual Placement DDM’**)

Interest of caching technology : Dream

- Host permanent files on high latency (= cheap) storage (nowdays=TAPE)
  - Less pressure to permanently identify files to clean
- Host small and popular datasets for analysis on cache
  - Unused datasets are cleaned automatically replaced by recently accessed data
- Would avoid the complain that, in some sites, only small fraction of stored files on DISK are used over 6 months (**M. Schulz presentation**)
- Could fit with large amount of HL-LHC data with flat budget
- Compatible with acceptable data availability ? .



Data access popularity : Central production

- Permanent copy of input files only hosted in nucleus sites (except PU)
- Reusage on the Grid
  - Triggered by different campains using same datasets
    - EVNT for simulation : Many times per year (fullsim, fastsim for syst. studies)
    - HITS for digi+reco : Different PU conditions (Ex : mc16c and mc16d)
    - Derivation : Process AOD each month
  - Within same production campain
    - EVNT : 5-10 times (single job too short to process all events)
    - AOD → DAOD : 10-15 times
      - Aim to make fat train to reduce nb of accesses to 1-2

Data access popularity : Grid analysis

- Compiled library : Input panda*.lib.tgz
  - Specific to the site → always stored on local/associated Grid storage
  - Small size (O(MB))
  - Usage restricted to few days (created with lifetime)

- Input datasets (AOD or DAOD) local (mandatory in the past) or remote:
  - Popularity depends on the user

- Question : Optimal way to migrate datasets on high latency but cheap storage and automatically keep only usefull data on disk (à la cvmfs)
  - Could be done with Grid SE for sites keeping storage

Data access popularity : DAOD ?

* DAOD : Perfect candidate for heavy data usage BUT
  - Many formats in Run2 : ~100 formats

    → single format accessed by small community
    - Actions under way
      - Run 3 : Strong reduction of DAOD format **AMSG-R3 report**
      - HL-LHC : Reduction to DAOD_phys (50 kB/evt) and DAOD_Lite (10 kB/evt) currently favored **ATLAS-DOMA talk**
  - Users typically transform them into
    - NTUP (filtered events and variables) to data+MC on their local analysis farm
    - Another format to be processed by Machine Learning

Data access popularity : Local analysis

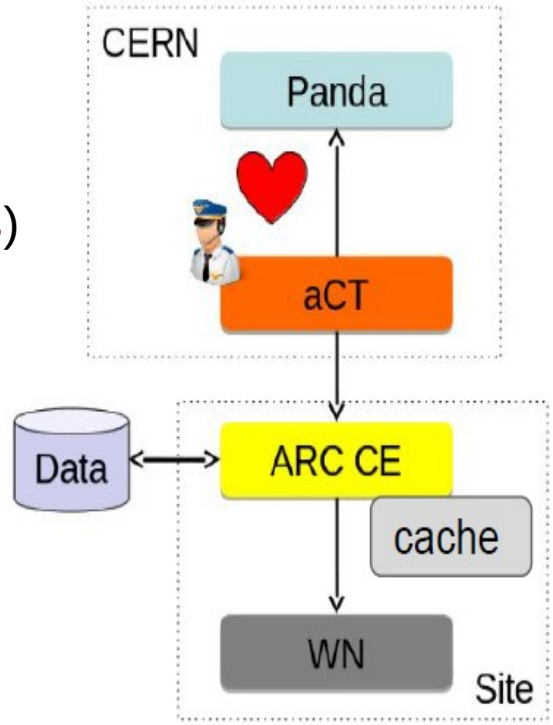
- Would avoid LOCALGROUPDISK to collect data from Grid storage
  - Could even complement long term archival to avoid to micro-manage quotas per user
  - But request to publish hosted data in Rucio
- Gives transparent access to remote data (scalability issue with network connection ?)

Caching technology

- Arc-cache :
  - Sites with Arc-CE and

    no Grid storage (**Presentation in DOMA Access**)
  - Running in production over many years

    in NDGF computing sites and many HPC in Europe
  - R&D :
    - Integration in Rucio
    - Arc-cache in a pilot model (without Arc-CE)



Caching technology (2)

* Xcache :
  - Many presentations from ATLAS and CMS in DOMA Access meetings
  - Stress tested in US, Germany, UK and evaluated in Italy
  - Tested in production environment in US and Germany
    - Reliability under load to be consolidated before being ready for production
  - Evaluation of cache on internet backbone
  - Possibility to install/manage remotely : Positive feedback from LRZ/ US sutes
  - Includes read-ahead for root format

→ Possibility to access fraction of files

→ Optimise read access for remote files (beyond Ttreecache optimisation)

  - Potential interface with Eos

Caching technology (3)

- Caching as extension of Grid storage (dcache, DPM,..) :
  - Potential interest for sites already hosting Grid storage
  - Ex : DPM presentation in DOMA Access : **Link**
    - Tested in Italy and similar test expected in France
    - Limitations :
      - File access possible only when full file is fully transfered
      - Current cleaning algo based replica creation date

Message to sites

- One of DOMA mandate is to evaluate caching : Computing model and technology
  - Bonus in case of data reusage which strongly depends on experiment workflow (production, analysis) and data format (single for all phys groups or not)
  - Potentially reduce network bandwidth and sensitivity to network unstability/congestion
  - Possibility to optimise remote access (read ahead)
  - Different caching technologies under evaluation (xcache most popular)

  - 2019 : Still R&D activity
    - Enough sites for the moment for a first evaluation
    - More volunteering/reactive sites could be called (especially to validate deployement model)
  - 2020 : If interest/reliabilty confirmed, general deployment could start

Backup

Current data pre-placement policy

- Production
  - Input/Output files transfered between source and local **Grid SE** through **Rucio+FTS**
  - Similar preplacement withing NorduGrid (ARC-CE cache)
  - Files kept 2 weeks
  - Over last 10 years
  - Few diskless sites which read/write to remote Grid SE
- Analysis
  - Historically : Brokering jobs close to data
  - Recently : Input files are transfered to Grid location with free CPU
  - Few diskless sites (Italy)

Caching vs FTS

- Caching :
  - Good : Caching mechanism ensures that file will be transfered even in chuncks → All transfers will go through
  - Optimal in a model with temporary copies
  - Requires (remote) Grid storage to write output
- FTS :
  - Transfer files between Ses (requires SE with some Grid components
  - Should have the global picture of all transfers to be done
  - Transfers always restared from scratch