Data Access in DOMA

Frank, Markus, Ilija, Stéphane and Xavier

Motivation (1/4)

- HL-LHC will be a multi-Exabyte challenge where the envisaged Storage and Compute needs will be a factor 10 above what can be achieved by the evolution of current technology within a (not anymore) flat budget.
- The WLCG community needs to evolve current computing models in order to introduce changes in the way we use and manage storage. Focus on resource optimization to improve performance and efficiency and simplify operations.
- We think storage consolidation based on a Data Lake model is a potential good candidate for addressing HL-LHC data access challenges.

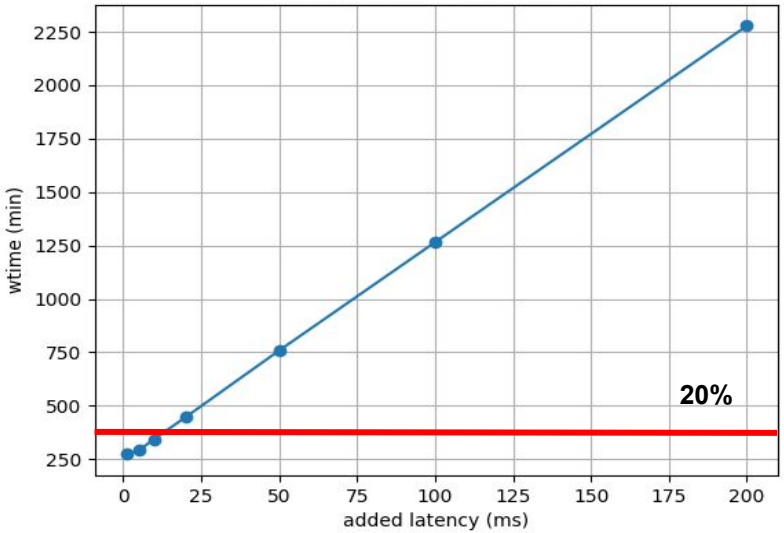
Motivation (2/4)

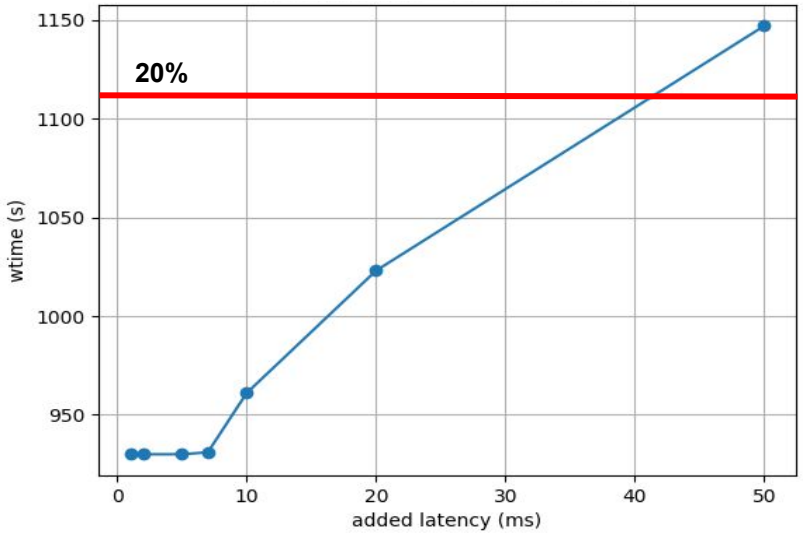
Some points in favor of the Data Lake model as an evolution of the current infrastructure:

- Storage costs reductions: global redundancy instead of local redundancy, dynamic adaptation of QoS, economy of scale.
- Conceptual separation of compute and storage services allows higher specialisation, leading to improvements in reliability and operations.
- Compute sites can provide increased resources with less effort
  - The effort for storage operations will be drastically reduced and sites can re-orient manpower and budget to provide *deeper* services to their users: support for enduser analysis, training in advanced analysis techniques, support for machine learning infrastructure, etc.

Motivation (3/4): latency hiding

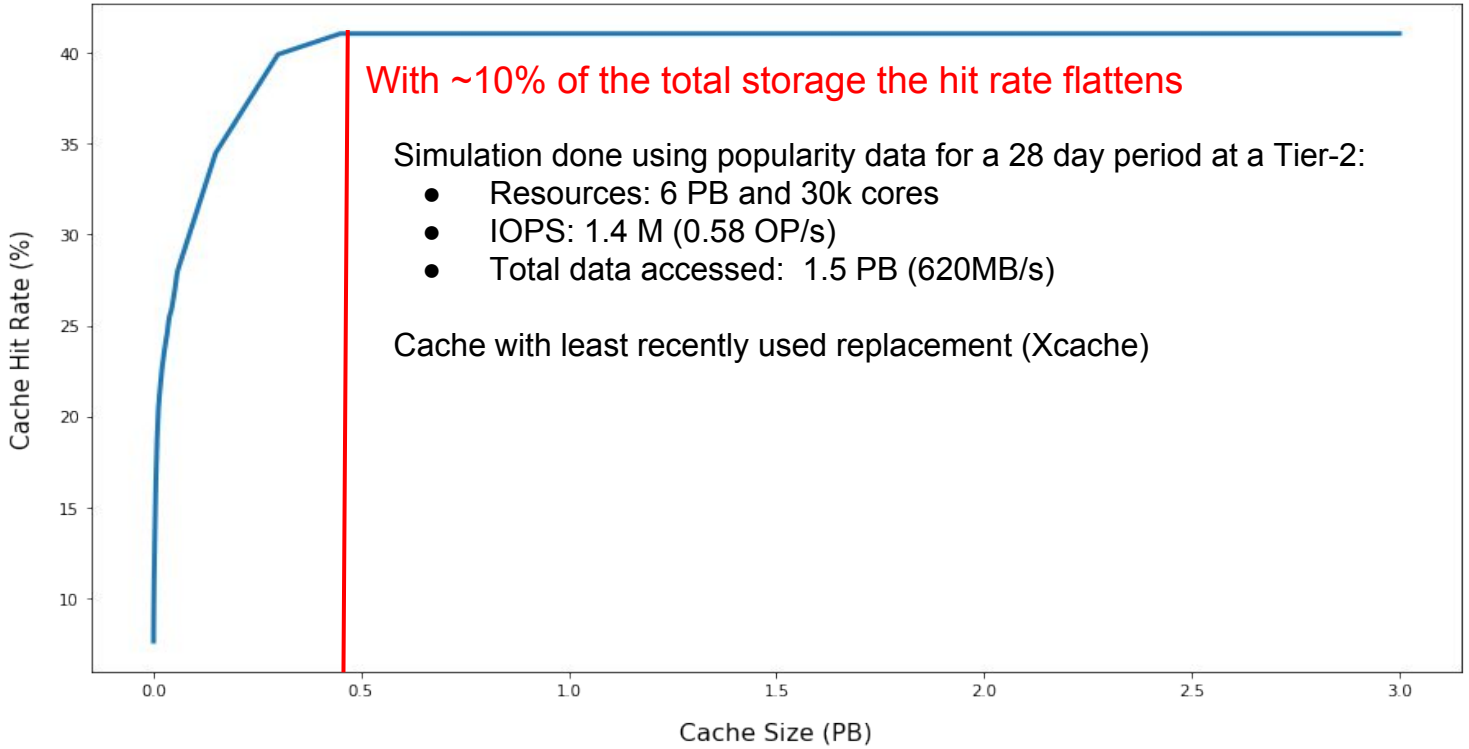
Latency hiding effects on moderate I/O workloads





Motivation (4/4): storage consolidation

Cache simulation using ATLAS popularity data for a reference Tier-2



DOMA ACCESS Working Group mandate (1/2)

- To address and evaluate future data access needs for HL-LHC and sciences with related challenges.
- To provide a forum to share experience on improving remote and local data access by the Experiments.
  - Through caching solutions, smart clients/endpoints, content delivery services and networks.
- To provide a forum where the current and future workload models are discussed taking into account developments such as the compact analysis data formats.
- To collect and compile quantitative information with the primary goal to be used by the WLCG DOMA project.

DOMA ACCESS Working Group mandate (2/2)

- To 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
- To track and report on the progress of the identified topics.
- To maintain close links to relevant WGs and activities: protocols, networking, authorisation and authentication.

Topics addressed so far (from Sep-18 to Mar-19)

- Status of caching activities in ATLAS and CMS: https://indico.cern.ch/e/757175/
- Content delivery and caching studies at CERN: https://indico.cern.ch/e/760850/
- Xcache (status, simulations, performance and federations):
  - https://indico.cern.ch/e/763847/
  - https://indico.cern.ch/e/764785/
  - https://indico.cern.ch/e/769500/
  - https://indico.cern.ch/e/769502/
- XrootD proxy and ARC cache: https://indico.cern.ch/e/764782/
- Ideas on data models and formats in ATLAS and CMS for HL-LHC:
  - https://indico.cern.ch/event/764786/
  - https://indico.cern.ch/e/769501/
- DPM volatile pools: https://indico.cern.ch/e/764785/
- Virtual placement and scheduling: https://indico.cern.ch/e/769502/
- Smart caching, dCache and Data Lakes: https://indico.cern.ch/e/769502/
- Data Access on a Data Lake Straw Model: https://indico.cern.ch/e/767211/

Presentations dominated by ATLAS/CMS but original contributions from Belle-II and beyond HEP (Ligo and Biology)

Data Format Strategy (1/2)

- To minimize disk costs and fit within flat budget, ATLAS and CMS strategies are similar : implement reduced data format.
  - CMS already implemented for Run2 analysis:
    - A single MINI format for the entire collaboration at ~40 kB/event.
      - 90-95% of all analysis activity uses this format in Run2.
    - A single NANO format that has been used for the first few public results at ~ 1-2 kB/event
  - ATLAS : First step foreseen for Run3
    - Few DAOD_PHYS format (similar to CMS MINI)
    - Convince physicists to this model under way

Data Format Strategy (2/2)

- ATLAS and CMS strategy towards HL-LHC rely on these kind of data formats
  - ~50PB of MINI per year and ~few PB of NANO per year
  - As compared to an exabyte of RAW and AOD combined

- Challenges to this vision:
  - Experiments need larger formats to develop object definitions for smaller formats
  - Need to define small subsets of the total trigger for this development, or some other way of providing support for detector commissioning and object development & evolution
    - E.g. ATLAS data carousel

Data Access on a Data Lake: strawman model* (1/4)

- We are preparing a strawman model for a Data Lake based infrastructure to explore how data usage and access can be optimized for analysis.

> A straw-man proposal is a brainstormed simple draft proposal intended to generate discussion of its disadvantages and to provoke the generation of new and better proposals. The term is considered American business jargon, but it is also encountered in engineering office culture. Wikipedia

- Aimed to address the **analysis data** use case while the **production data** workflows are to be managed by higher level infrastructure: Data Lake.

* This model is work in progress and yet being defined and modeled hence not final nor endorsed by the working group (please see, read and contribute here: document)

Data Access on a Data Lake: strawman model (2/4)

The different Data Lakes worldwide forms an universal storage infrastructure

- A Data Lake is composed of **C**ompute **C**enters, **D**ata and **C**ompute **C**enters and at least one **A**rchive **C**enter.
  - DCC provide large disk storage without the need for local redundancy. Implement QoS endpoints.
  - CC provide computing resources and access data from the Data Lake through:
    - Cache: data is accessed through a latency hiding cache, all data flow through this cache (ie. proxy behavior)
    - Direct Access: data is accessed directly relying on latency hiding capabilities at the client-side (ie. read-ahead)
  - AC provide tape or tape-equivalent-QoS able to provide long term data archive and a proportional Staging Area.

Data Access on a Data Lake: strawman model (3/4)

Analysis data in the Data Lake:

- A Data Lake hosts a distributed working set of analysis data.
- This comprises an experiment's full set of mini/nano-AODs (or equivalents) of a given campaign/period.
  - Based on CMS' forecasts O(50) and O(1) PB are required respectively for mini/nano AODs.
  - If not affordable, the analysis working set is distributed across more than one lake.
- The experiment's workload management system will allocate jobs within the matching Data Lake.
  - Popular datasets may be hosted in more than one Data Lake.
  - To allow for non-local redundancy at least 2 copies have to be available globally.
- Caches are used to reduce the impact of latency and reduce network load.

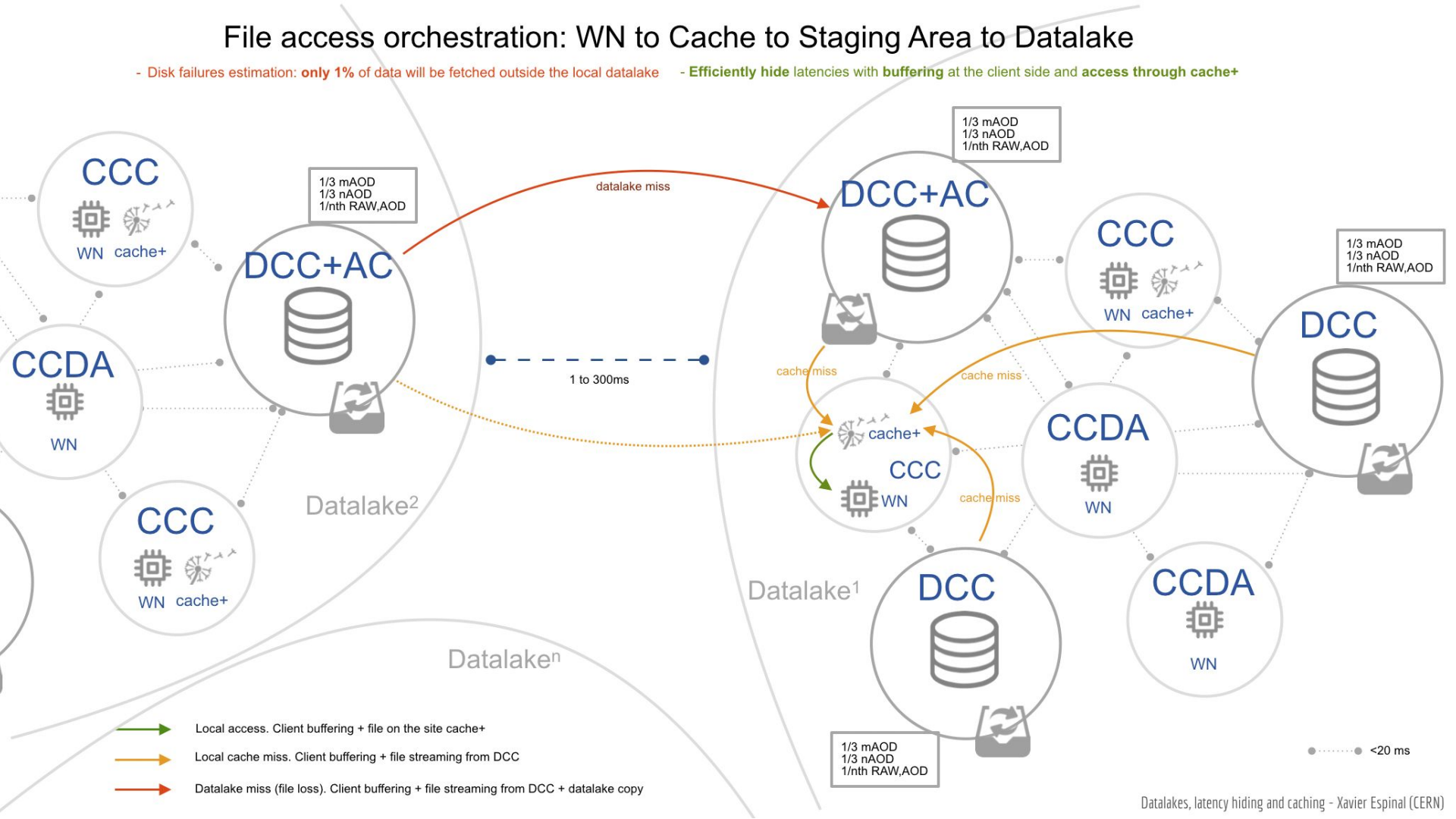
Data Access on a Data Lake: strawman model (4/4)

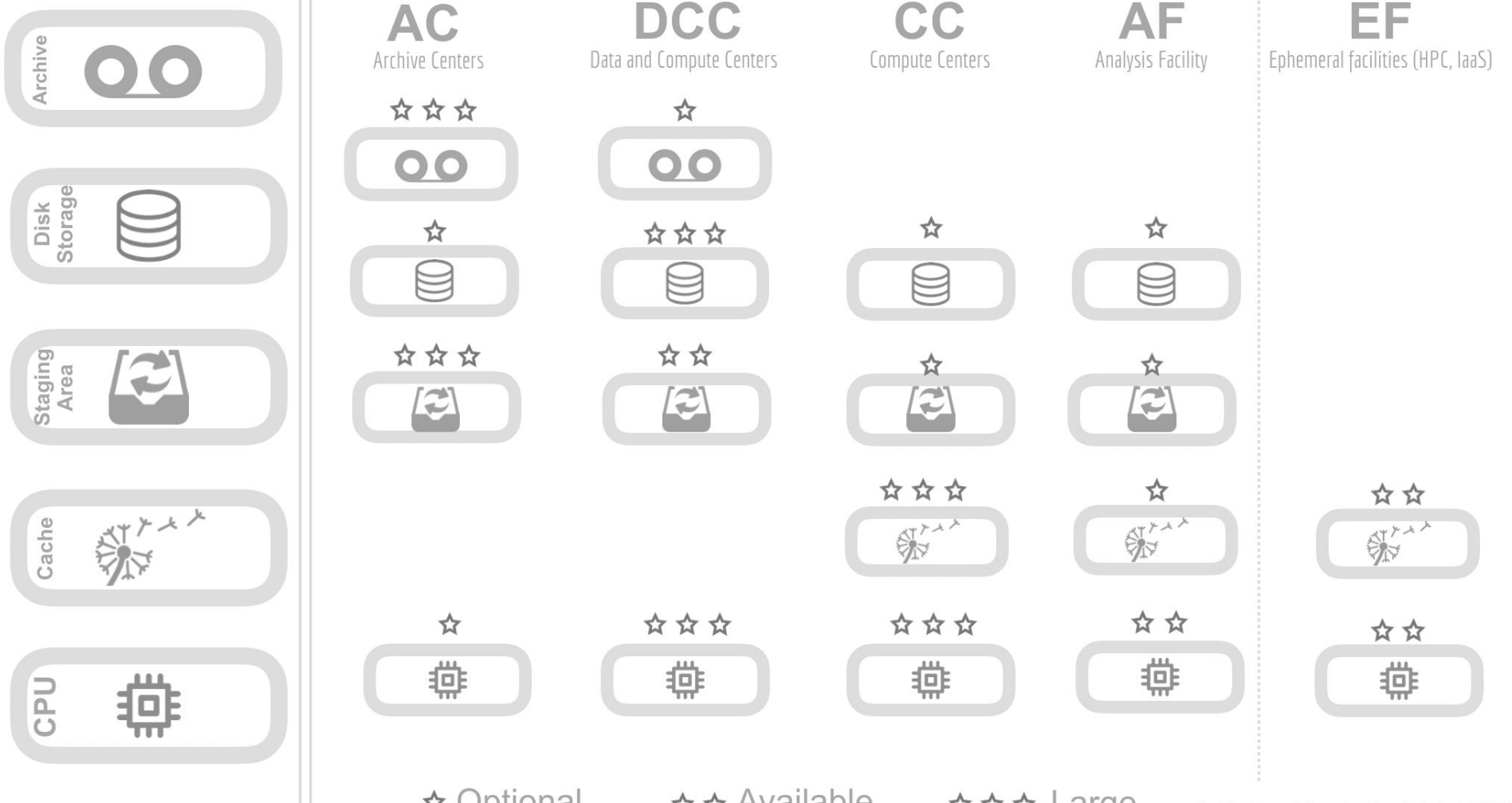
The cache at the Computing Centers:

- The cache is a self-managed stateless storage with **streaming** ability and providing **read-ahead** functionality: we named it **cache+**
  - The content of the **cache+** is only known by the **cache+** while for users and data management services the cache is fully transparent (discussions ongoing).
- The role of **cache+** is to:
  - **Reduce wide area network bandwidth** by holding frequently used files in the cache
  - Ability to **read ahead to reduce the impact of latency** and peak bandwidth requirements for the first reading of the file
- **cache+** can be located close to the WN (need for dedicated nodes) or can be distributed across the worker nodes of the Compute Center (no need for dedicated nodes)

File access orchestration: WN to Cache to Staging Area to Datalake

- Disk failures estimation: **only 1%** of data will be fetched outside the local datalake
- **Efficiently hide** latencies with **buffering** at the client side and **access through cache+**




Archive
Disk Storage
Staging Area
Cache
CPU
AC
Archive Centers
DCC
Data and Compute Centers
CC
Compute Centers
AF
Analysis Facility
EF
Ephemeral facilities (HPC, IaaS)
☆ Optional
☆☆ Available
☆☆☆ Large

Results from XCache deployments and studies performed

ATLAS all-parallel approach:

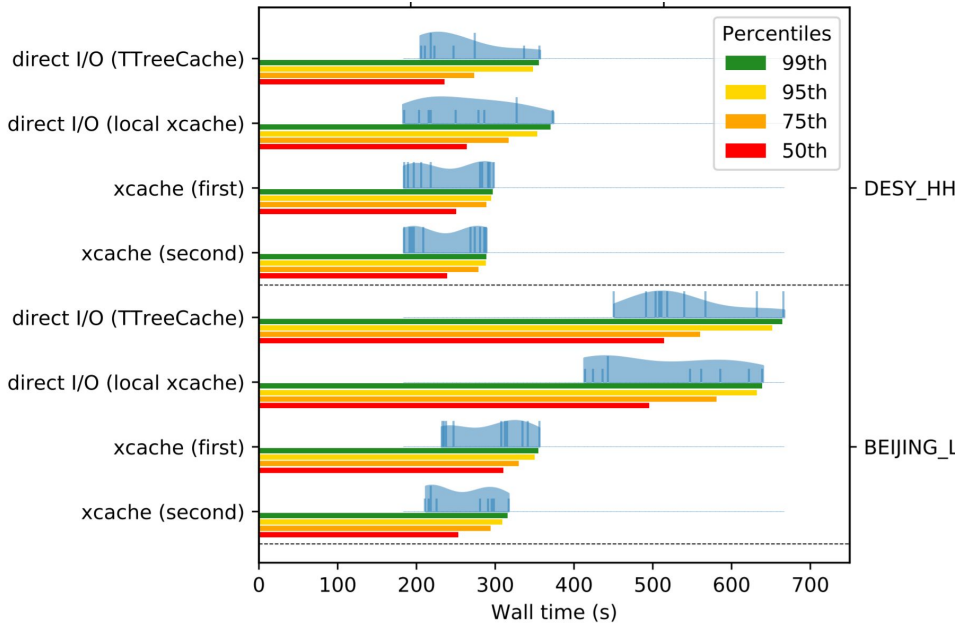
- Evaluate stability and performance of XCache server implementation(s).
- Develop and test centrally managed network of independent caching servers (using SLATE platform)
- Emulate and evolve different models of data distribution and job scheduling. Find ones that optimally use both CPU and cache while keeping network requirements at a reasonable level. Testing it at small scale still some time in future.

XCache deployment model

- Learned a lot from a federated cluster approach (FAX).
- Utilize SLATE platform to install/update/monitor all XCaches. Done by a single service administrator. Completely transparent for sites.
- Deployments are multiple instances of independent servers.
- All access paths algorithmically created by RUCIO, thus removing all the searches.

XCache stability/performance

- Production deployments at MWT2, AGLT2, test deployments at LRZ-LMU, BNL
- Corner case stability issues are being addressed one-by-one.
- Still searching for the best performance configurations (both hardware & software), preliminary base configs ready.
- Realistic analysis job performance evaluated at LRZ-LMU. Results show direct access through XCache hides latency as well as ROOT TTC with AsynchronousPrefetch.



Conclusions

- DOMA ACCESS working group is addressing future data access needs for HL-LHC and sciences with related challenges by:
  - Optimizing data storage and data access for analysis workflows.
  - Optimizing storage resources focusing on performance, efficiency and operations simplification.
  - Proposing a starting point for an evolution of the current site topology towards a global Data Lake infrastructure.
- A straw man model for data access has been presented (*document*). This enables to start discussions, to fit recommendations and to set priorities.
- Studies performed so far show encouraging results towards a Cache based model for analysis within a Data Lake storage infrastructure.
- Collaboration, understanding and discussion with experiments, storage and software providers and sites is fundamental in the next months. **Get involved.**

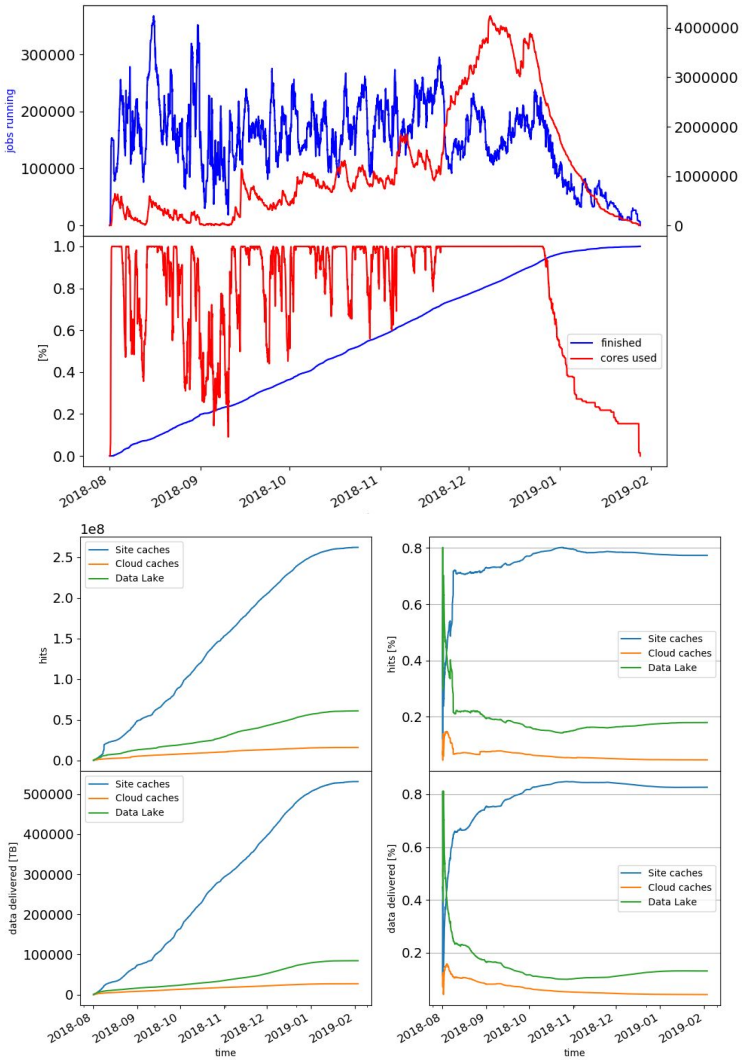
WLCG DOMA ACCESS working group

- Fortnightly meetings (Tue 17:30 GVA time): https://indico.cern.ch/category/10828/
- Working group Twiki: https://twiki.cern.ch/twiki/bin/view/LCG/ContentDeliveryCaching
- *Data Access on a Data Lake straw model* (document wip)
- *Data Access on a Data Lake straw model* (DOMA presentation)

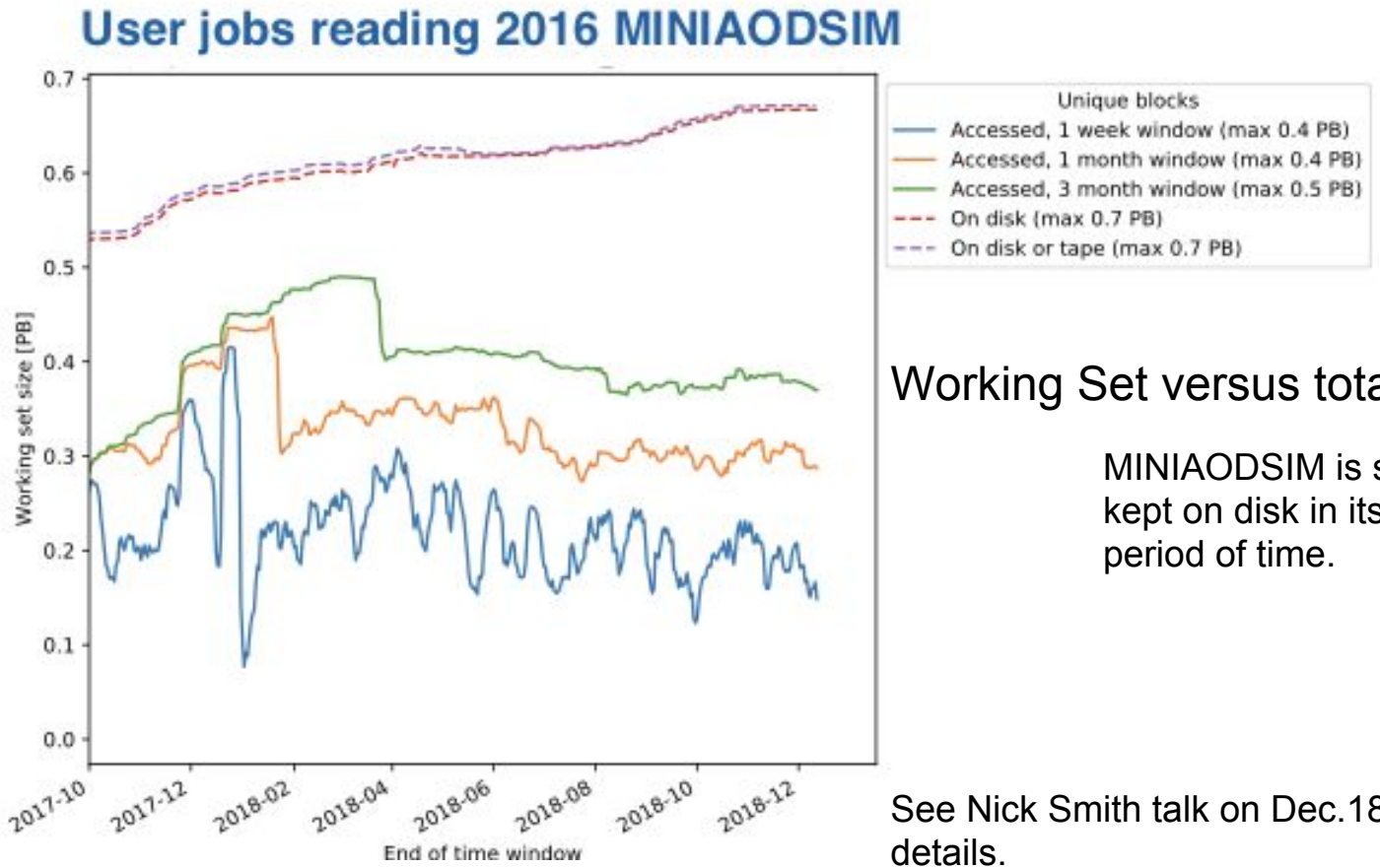
Backup slides

Scheduling jobs in cache aware way

- With all the details on ATLAS tasks easily available we can quickly simulate
  - Different scheduling models and policies
  - Number, size and placement of DataLakes
  - Number and size of caches, hierarchy of caches, caching strategies.
  - Failure modes (sites/storage/network)
- Multivariate optimization
  - Max CPU usage
  - Min Time-To-Complete
  - Min data movement
- Currently testing Virtual Placement Model



More on storage consolidation



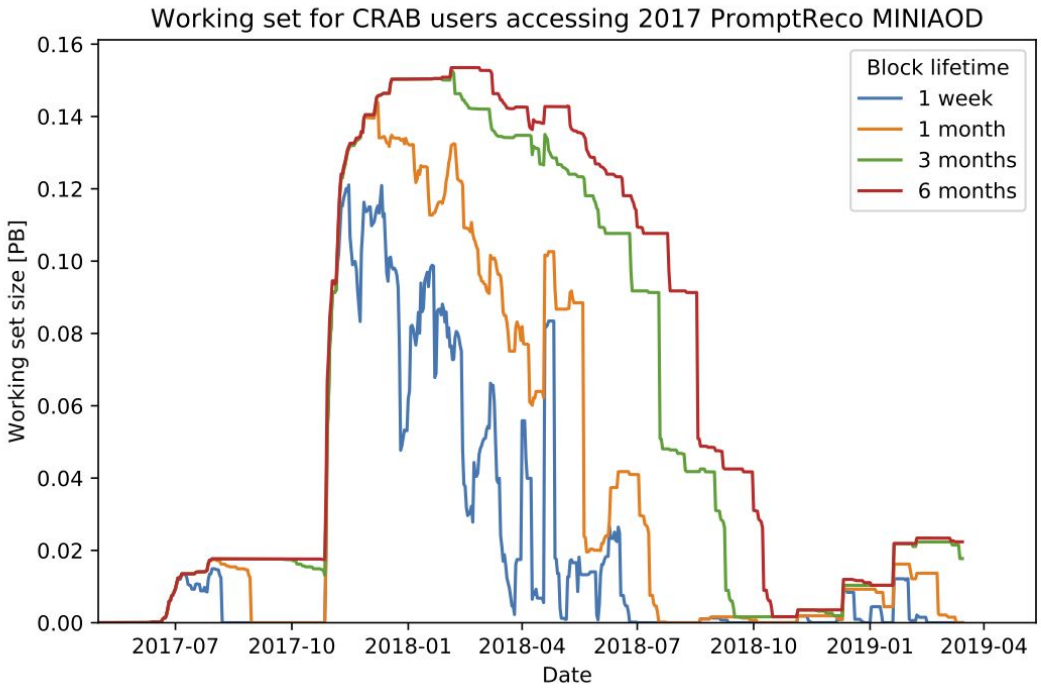
Working Set versus total disk and tape CMS

MINIAODSIM is small enough that it is kept on disk in its entirety for extended period of time.

See Nick Smith talk on Dec.18th 2018 for details.

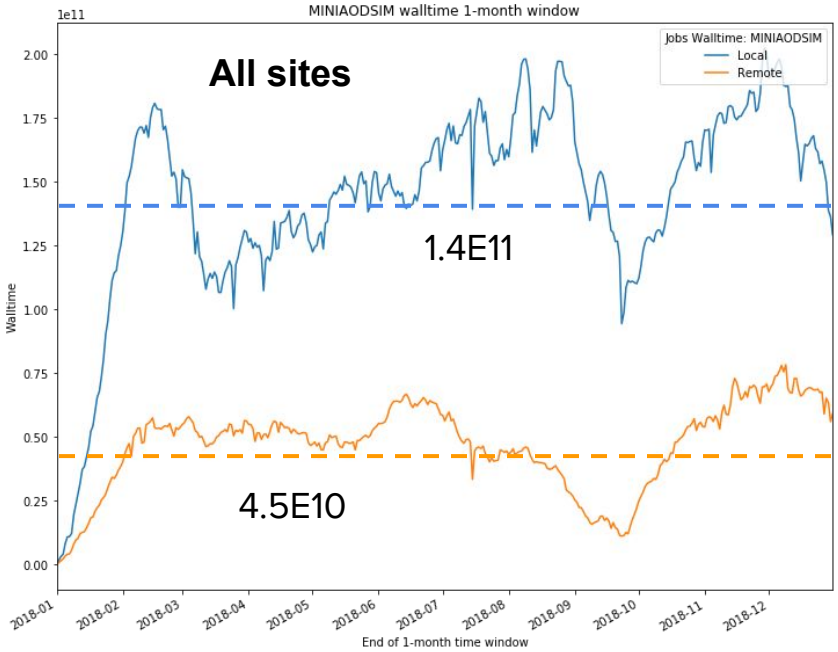
More on storage consolidation

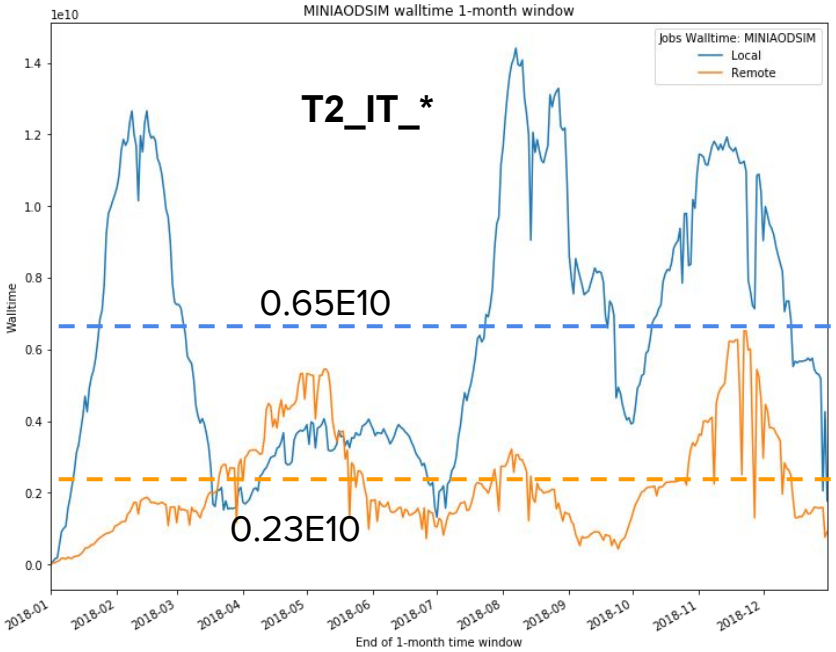
Lifetime for MINIAOD Working Set for CMS



- The figure shows the finite lifetime of MINIAOD for prompt reconstructed data..
- CMS reprocesses data annually after the end of the run.
- The reprocessed data then replaces the prompt reco as the primary version of the data that is accessed.

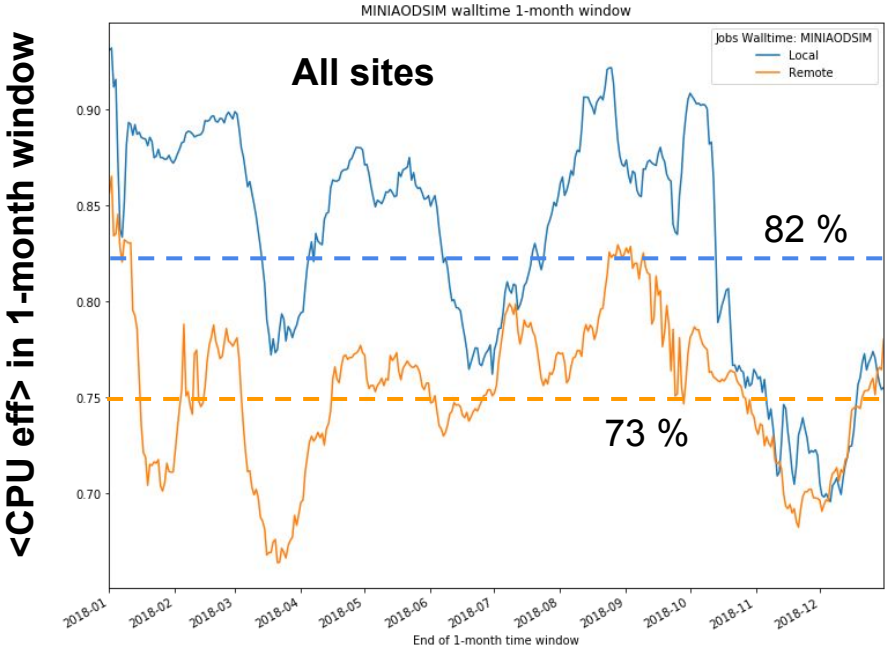
Time spent on remote read vs local one

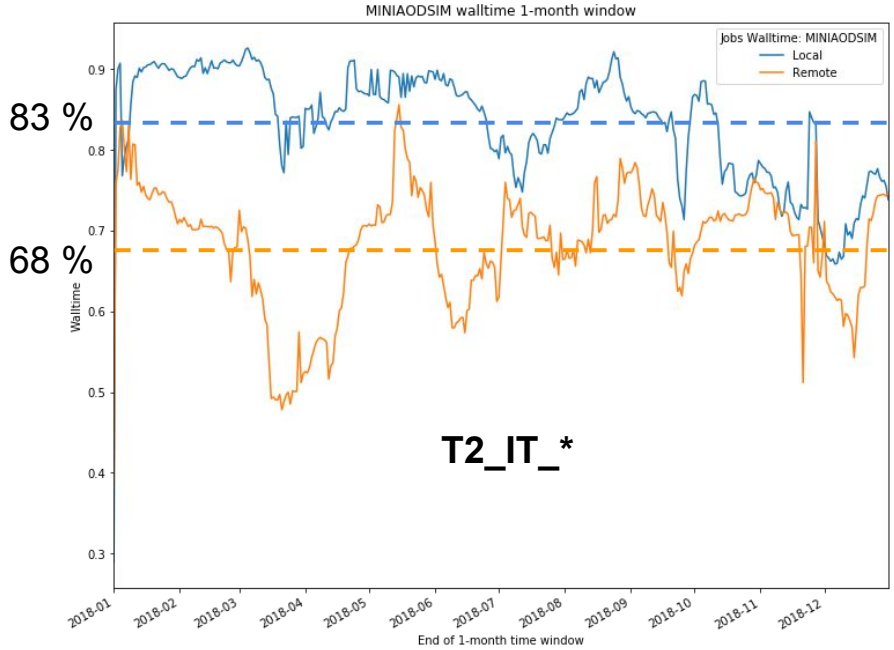




Remote/Local ~20% globally, from 10 to 120% at regional scale (34% on avg)

CPU loss when reading from remote





Remote read CPU efficiency from few percent to more than 11% loss.

19% on avg for T2_IT_*

Unique tasks hits over 1 month

T2_IT_*