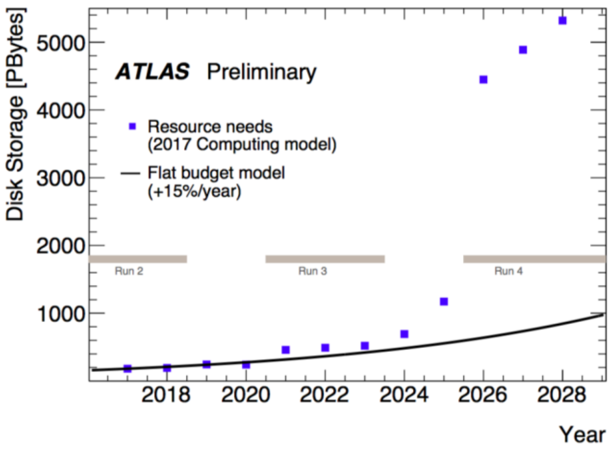
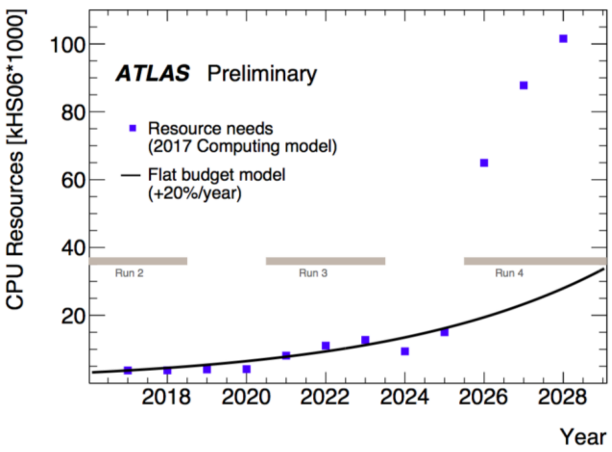
The WLCG data Lakes

Simone Campana (CERN)

Thanks for the invitation. I would love to be there in person, but unfortunately I could not make it this time. I'll try harder next time if I have the opportunity

WLCG needs manage and contain the cost of HL-LHC computing



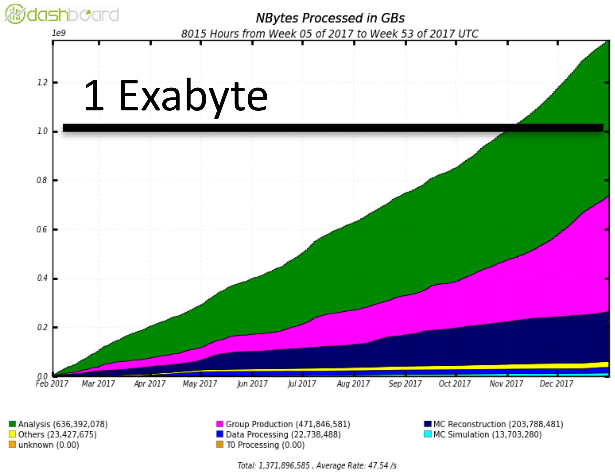


The cost comes both in terms of hardware (left) and operations

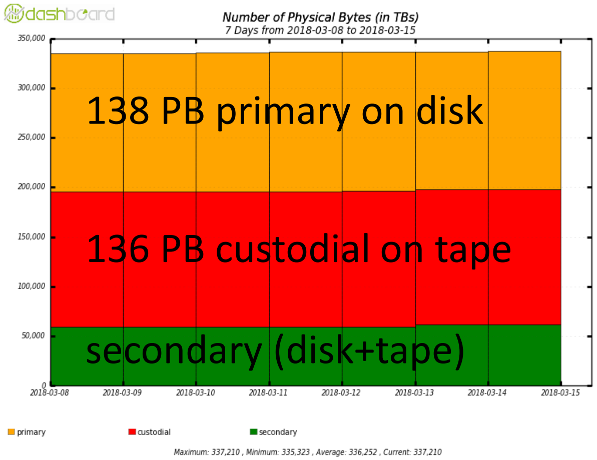
Facts:

- Storage today is the major hardware cost in most countries. Disk costs 4x more than tape per TB
- Storage is also the main operational cost at sites according to a recent (2015) survey

NB of bytes read (PanDA jobs) in 2017


1 Exabyte

NB of bytes currently stored


138 PB primary on disk

136 PB custodial on tape

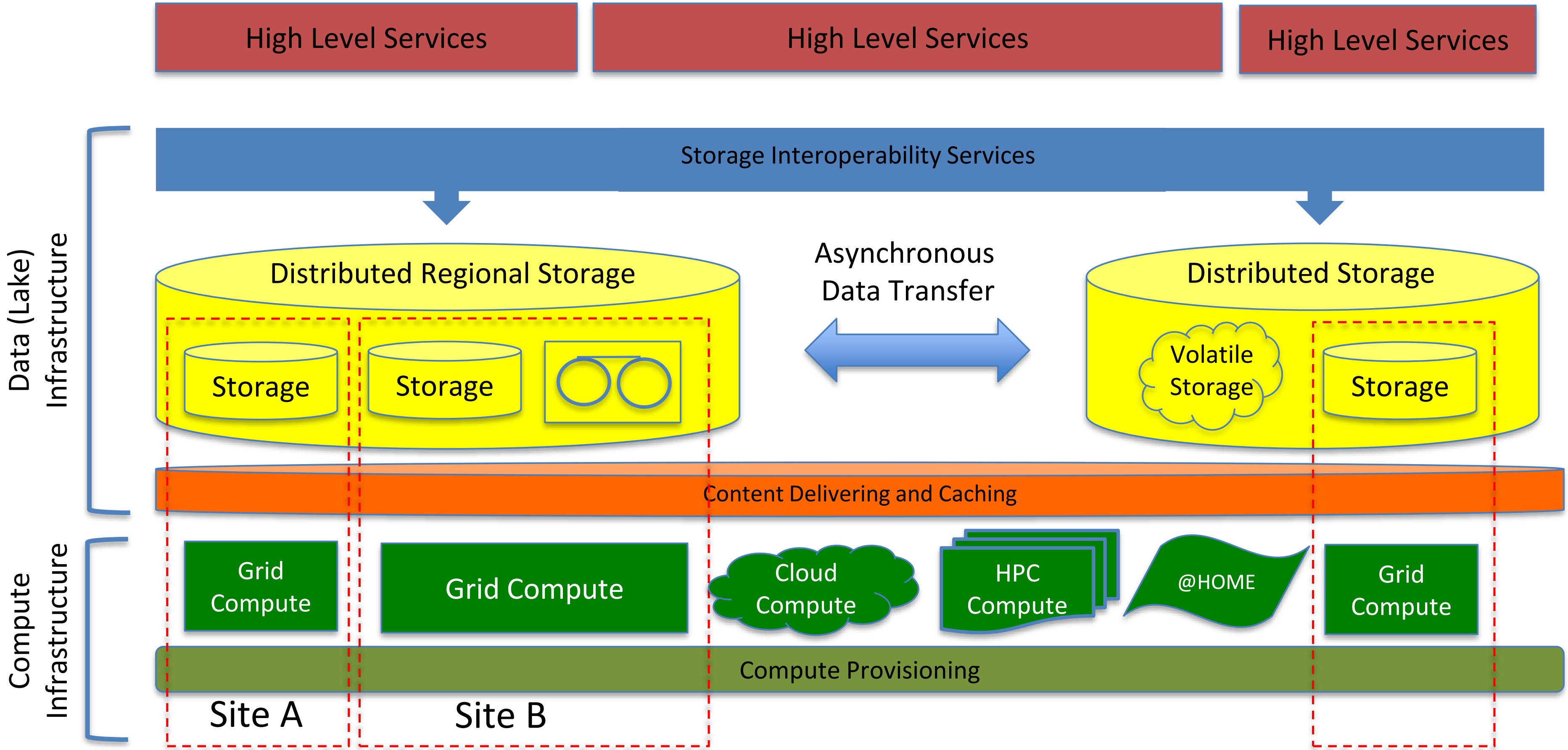
secondary (disk+tape)

Facts and Opinions:

- LHC data is rather “cold”. E.g. PanDA jobs access 1.4 EB/year of data (**). There are 165 PB of data on pledged disk (and 172PB on tape). Each file on disk is accessed O(10) times.
- Most of the data is accessed in a scheduled way (reconstruction/derivations). Data access patterns are extremely workflow dependent

(**) Actually, less than that, as not the full event information is accessed

Evolution of Data and Compute Infrastructures



Cost Model and Metrics

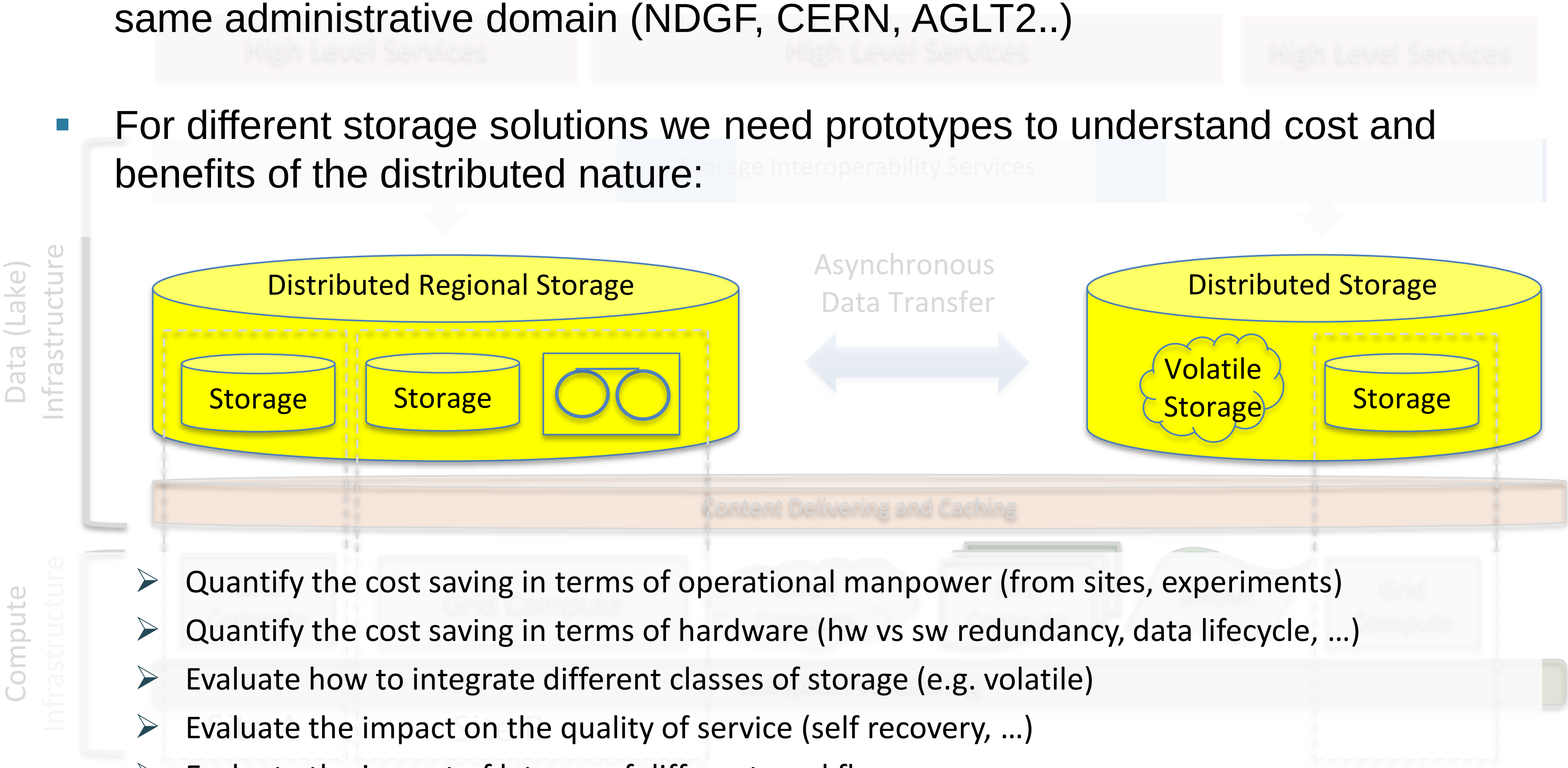
Before/while we start prototyping any change in the infrastructure, services, computing models, we need the following:

- A cost model, telling us if what we are doing is really going to reduce cost.
  - There is a WLCG working group on this
- An understanding of which workflows we should be looking at and which metrics characterize them
  - Regular meetings between ATLAS and IT-WLCG for this purpose
- A set of tools to measure those metrics
  - Tools such as Hammercloud Monitoring and Analytics do exist
  - We need to make sure they can do what we need
  - There is work ongoing on that as well

Storage Consolidation

- We have of course experience with distributed storage instances under the same administrative domain (NDGF, CERN, AGLT2..)

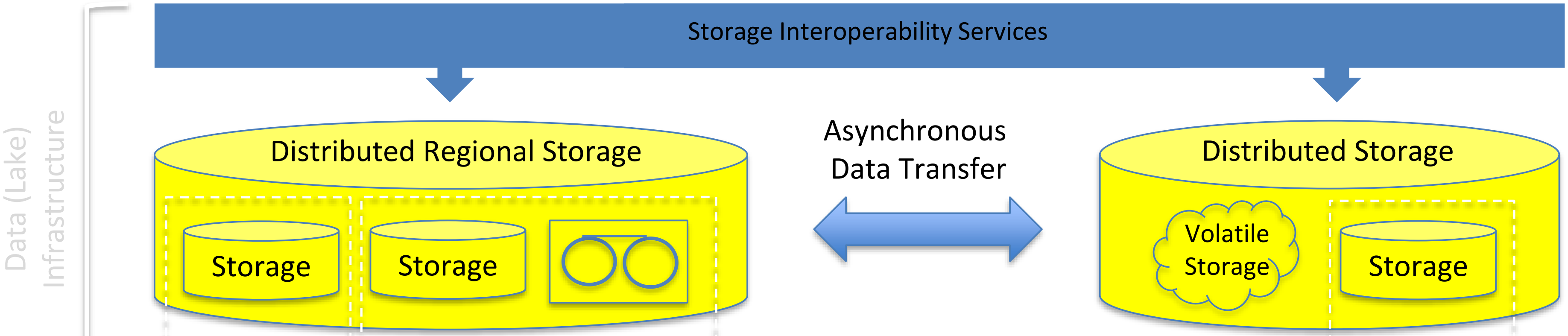
- For different storage solutions we need prototypes to understand cost and benefits of the distributed nature:


High Level Services
High Level Services
High Level Services
Storage Interoperability Services
Data (Lake) Infrastructure
Distributed Regional Storage
Storage
Storage
Asynchronous Data Transfer
Distributed Storage
Volatile Storage
Storage
Content Delivering and Caching
Compute Infrastructure

- ➢ Quantify the cost saving in terms of operational manpower (from sites, experiments)
- ➢ Quantify the cost saving in terms of hardware (hw vs sw redundancy, data lifecycle, …)
- ➢ Evaluate how to integrate different classes of storage (e.g. volatile)
- ➢ Evaluate the impact on the quality of service (self recovery, …)
- ➢ Evaluate the impact of latency of different workflows

Storage Interoperability

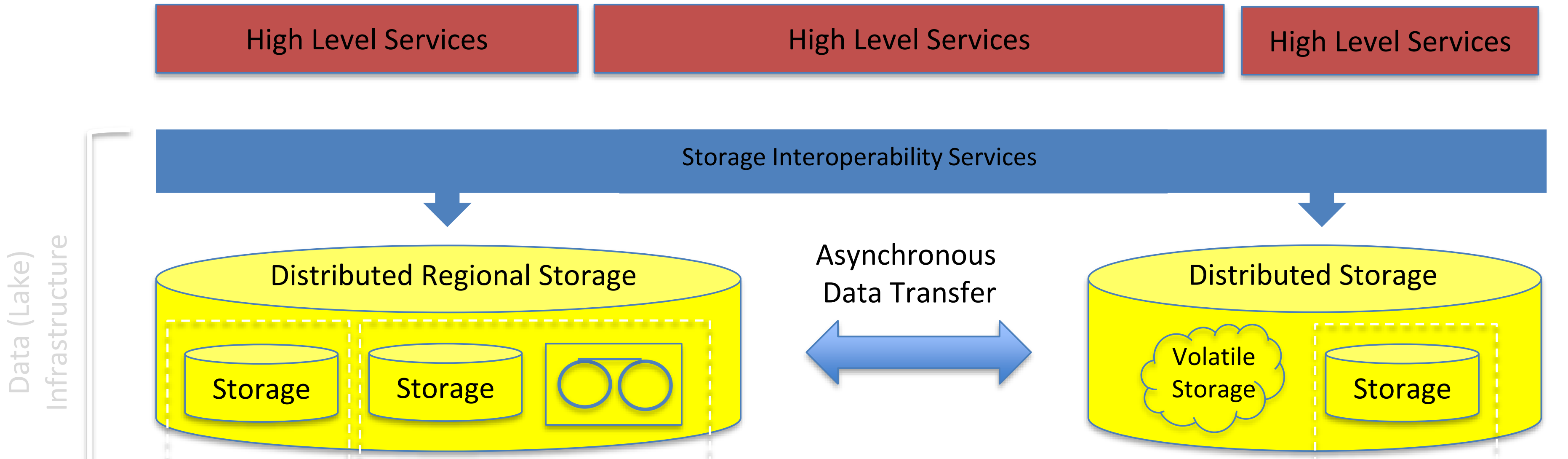
- Very unlikely we will end up with one distributed storage spanning all WLCG. Storages will need to interoperate. A similar model to Amazon’s regions.


Data (Lake) Infrastructure
Storage Interoperability Services
Distributed Regional Storage
Storage
Storage
Asynchronous Data Transfer
Distributed Storage
Volatile Storage
Storage

- We need to understand how to obtain the benefits in the previous slide, in this regional model. Develop the required services

- We will need to asynchronously exchange large volumes of data across regions.
  - Evolution of File Transfer Systems
  - 3rd party transfer protocols and a different Authentication/Authorization model
  - How to shape networks within the storage and between storages

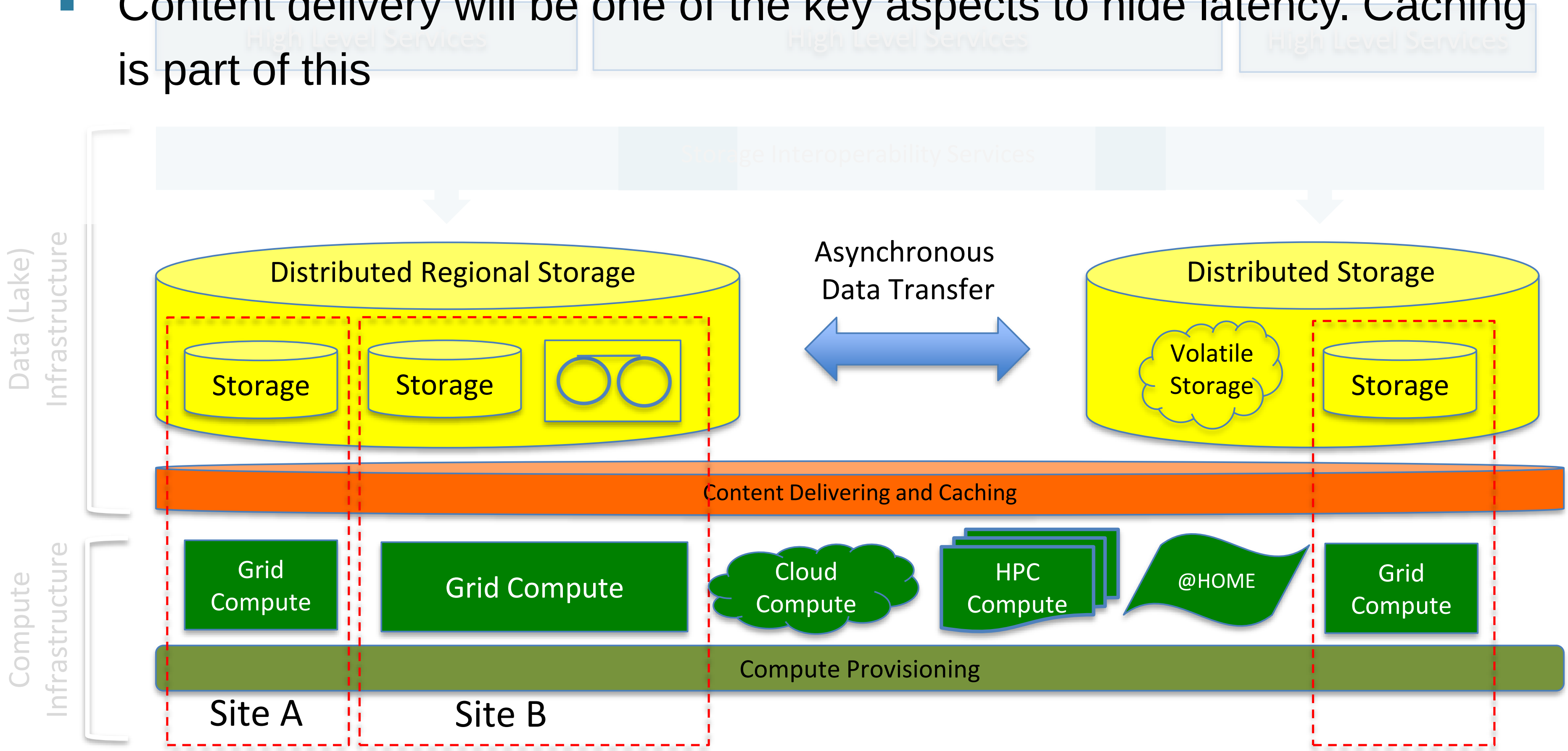
Interfaces


High Level Services
High Level Services
High Level Services
Storage Interoperability Services
Distributed Regional Storage
Storage
Storage
Asynchronous Data Transfer
Distributed Storage
Volatile Storage
Storage
Data (Lake) Infrastructure

- Commonalities at the level of the high level services will be the key of simplification
- Ideally we would have one Data Management and one Workflow Management system
- Realistically, we should try to implement the main functionalities in common services and rely on a thin specific layer on top of it

Content Delivery and Caching

- Content delivery will be one of the key aspects to hide latency. Caching is part of this



Content Delivery and Caching

- Evaluate different caching technologies and methodologies
  - Caching needs to be workflow aware and workflows need to be cache aware
- Work needs to be done in the area of organized data processing from archival media
  - How to efficiently store data to facilitate recall
  - How to schedule recalls based on the workflow and how to organize the workload based on the recalls
  - Understand how the archival storage technologies need to evolve and at which scale of resources
- Data organization (datasets), storage (files) and processing (sub-events) work at different granularities. Leverage that rather than suffer from that
  - Interesting prototype suggested by UChicago. Decouples storage and compute representation of data. Refer to Rob/Iljia

Conclusions

- The solution to the storage problem in HL-LHC is very simple:
  - Consolidate storage in few administrative domains and reduce operation/hardware cost
  - Store everything on archive media and recover the factor 4 in cost
  - Stage the data in an organized campaigns on little but very fast disk
  - Through a reliable content delivery system hide the impact of latency

- In fact, this requires a lot of R&D work in the next couple of years.
  - Some areas are more R ("understand") and some areas are more D ("prototype")

- We plan to organize a WLCG project covering all aspects above, to organize the discussion and measure the progress. TBD in Napoli next week.