

L'evoluzione dei Computing Model negli esperimenti a LHC

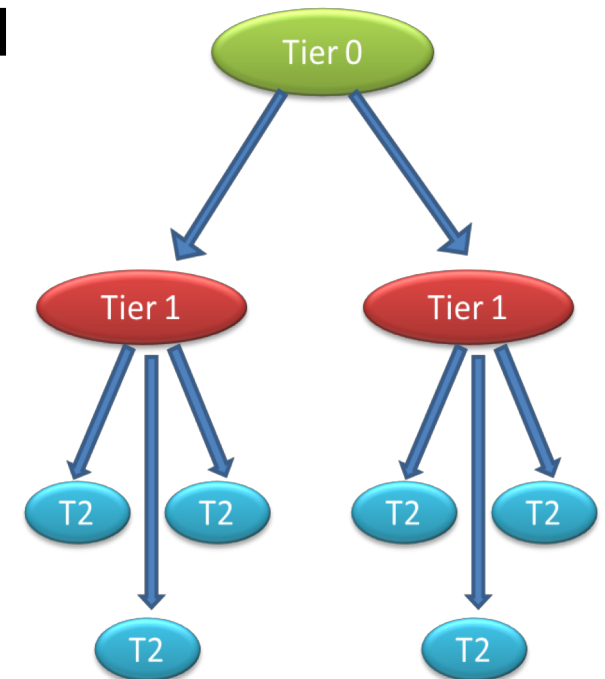
G. Carlino – INFN Napoli

12 Luglio 2019

Recas - Bari

- In 1998 MONARC project defined tiered architecture deployed later as LHC Computing Grid

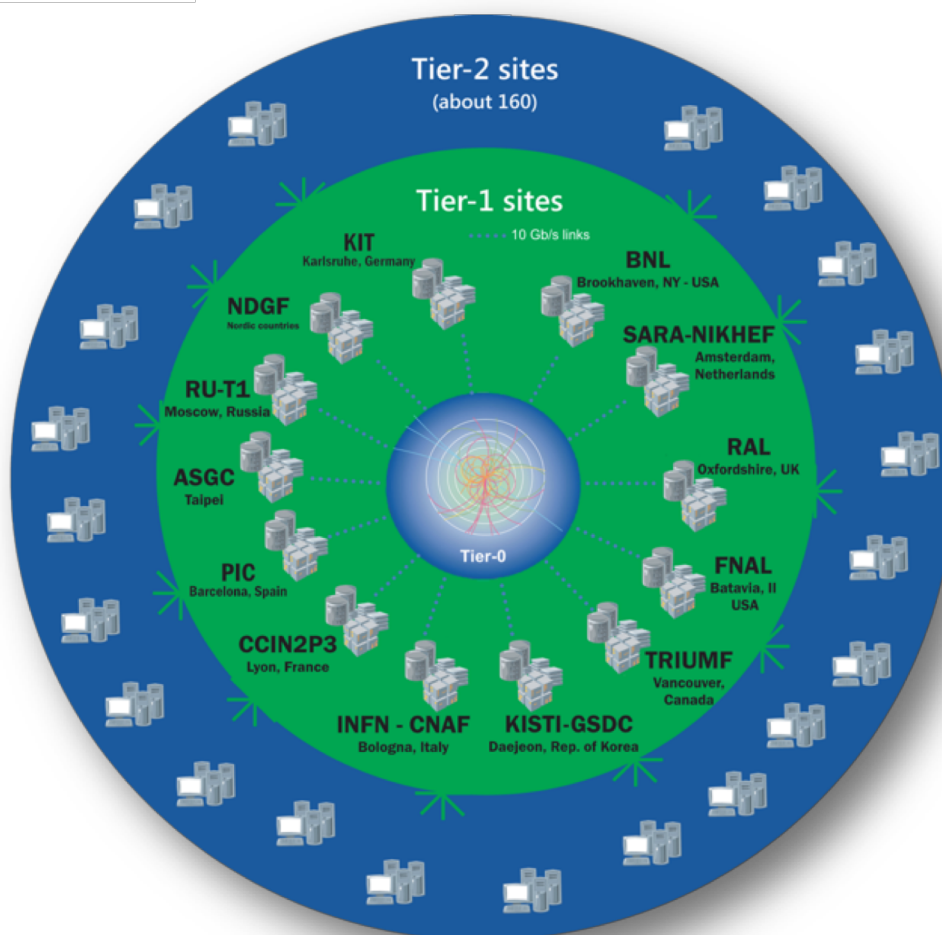
- a distributed model
 - Federate national and international grid initiatives
 - Integrate existing centres, department clusters, recognising that funding is easier if the equipment is installed at home
 - local physics groups have more influence over how local resources are used, how the service evolves
- a multi-Tier model
 - Static strict hierarchy. Multi-hop data flows
 - Network costs favour regional data access. Lesser demands on Tier2 networking
 - Static data pre-placement



Hierarchy in data placement.
Data flow via the hierarchy

The LHC WLCG Tier Model

WLCG is a distributed computing infrastructure to provide since early 2000 computing and storage for the LHC experiments



167 sites in 44 countries

Tier-0 (CERN):

- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):

- Permanent storage
- Re-processing
- Analysis

Tier-2 (~160 centres):

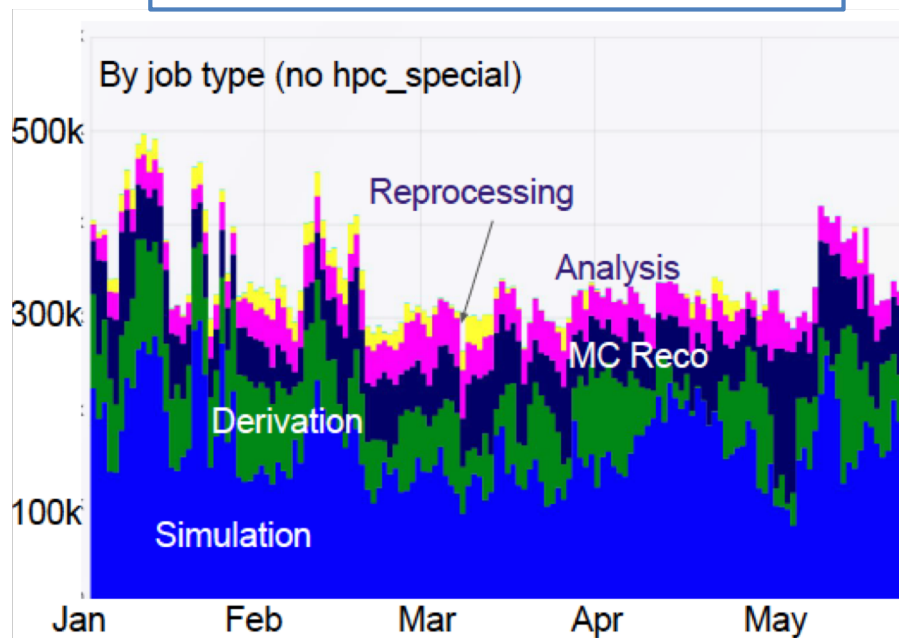
- Simulation
- End-user analysis
- Permanent and secondary storage

~ 1 M CPU cores - ~ 1 EB storage – 10/100 Gbps links - > 2 M jobs/day

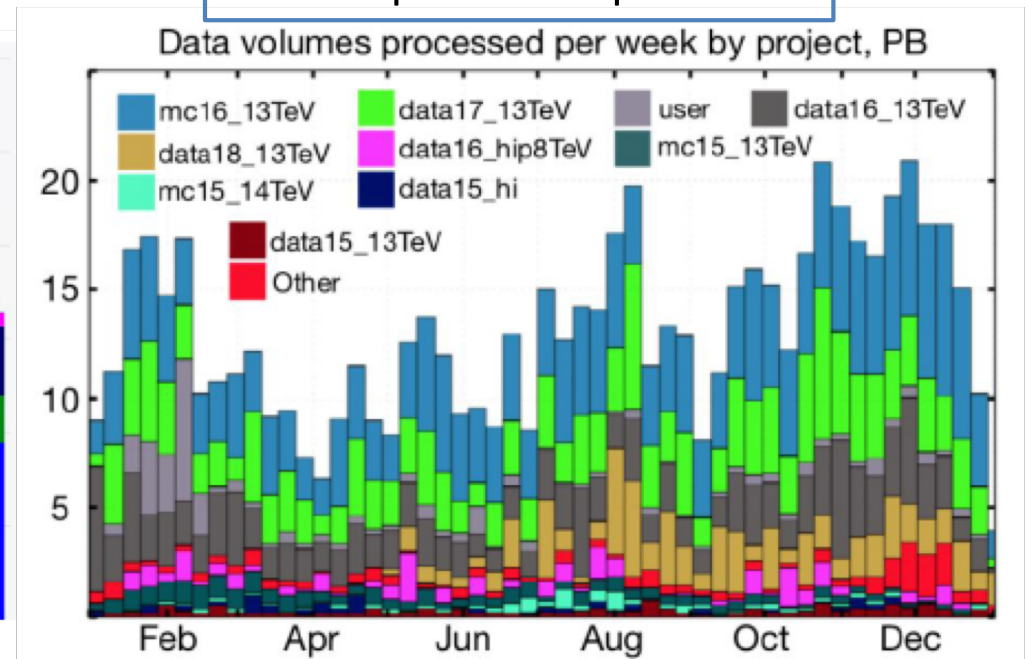
The Evolution of the CMs

- CM are not static. Continuous evolution
 - since the beginning of the data taking, the “ideal” CMs have been replaced by realistic ones exploiting the technology and infrastructure improvements
- In Run-1 and Run-2 the LHC experiments have been able to cope with an unforeseen amount of data transferred and analysed

> 300k concurrent jobs per day



> 10 PB processed per week



Evolution in Networking

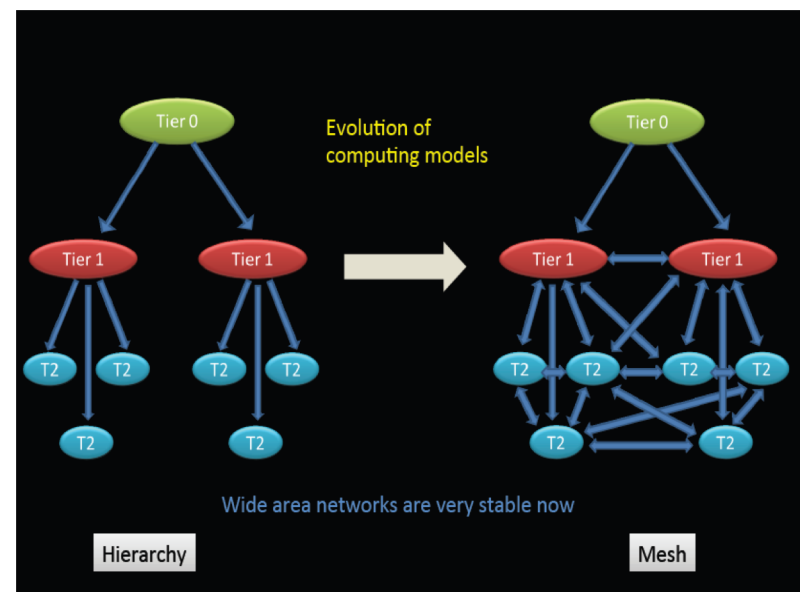
Network is as important as site infrastructure:

Key point to optimize storage usage and jobs brokering to sites

- At the beginning network was the bottleneck. The hierarchical model was based on the assumption of a rather limited connectivity between computing centres. Only links between well connected sites (Tier0 and Tier1s) were dedicated to cover fundamental roles.

Network capacity improved very fast

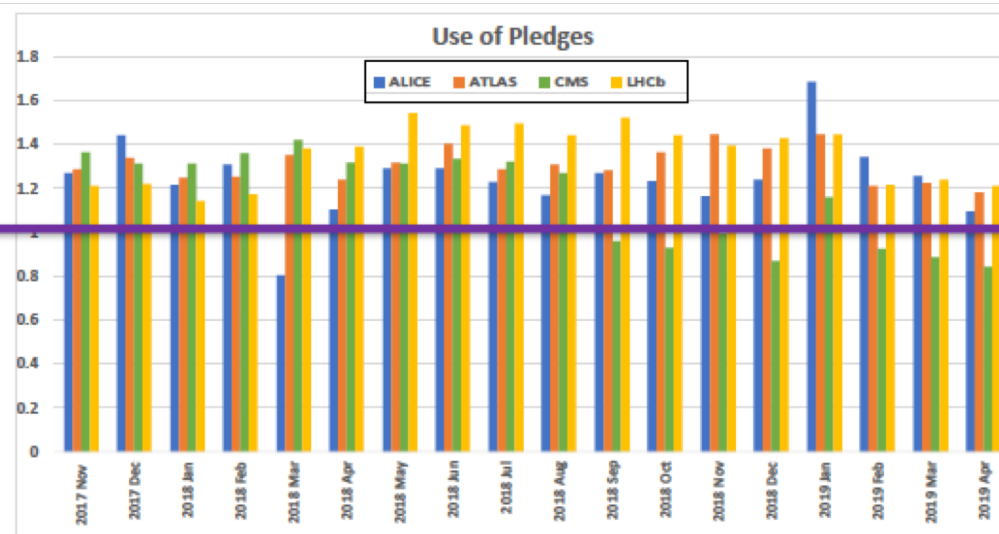
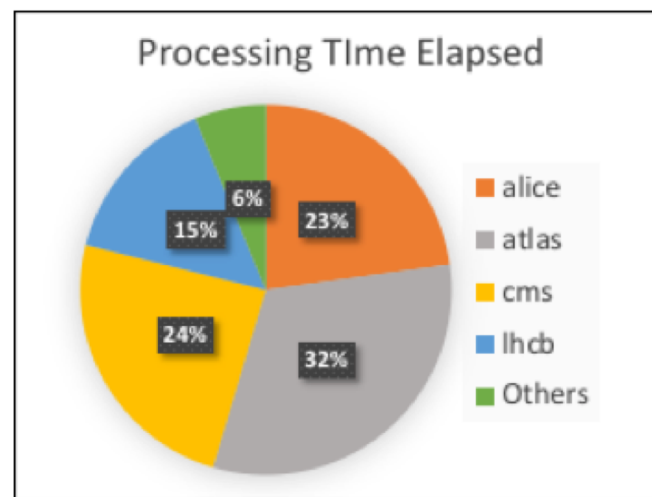
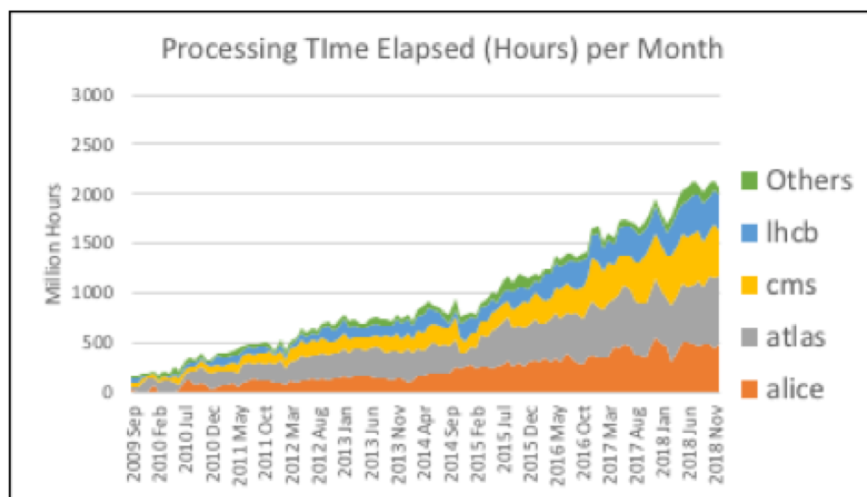
- WAN is very stable and performance is very good
 - It allows to relax MONARC model: migration from hierarchy to full mesh model: sites are all **directly interconnected** and **independent** of the Tier1s
- Data management based on popularity concept
 - Dynamic storage usage
 - Reduction of data replicas. Only data really needed is sent (and cached)
- Network awareness
 - Workload management systems and data transfers will use networking status/performance metrics to send jobs/data to sites



**This is what happened
in Run-2 (2016-2018)**

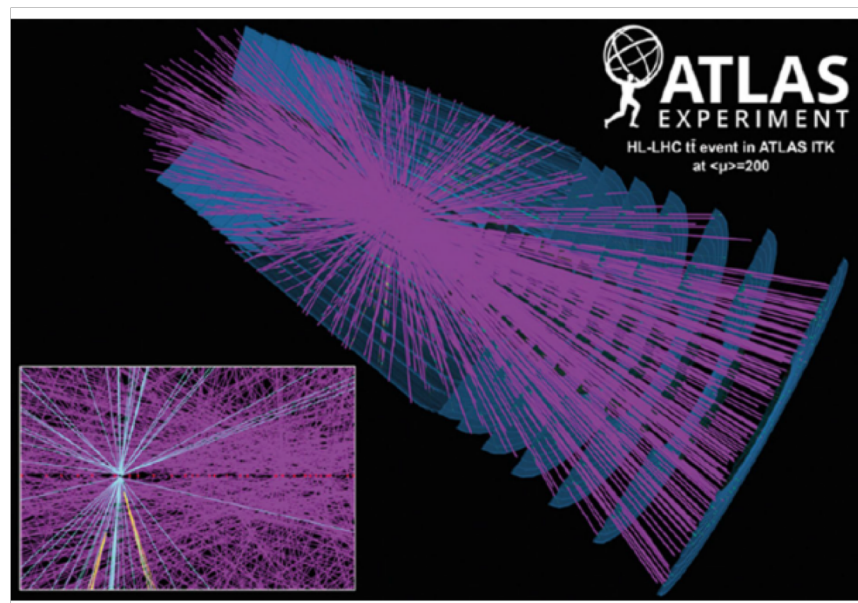
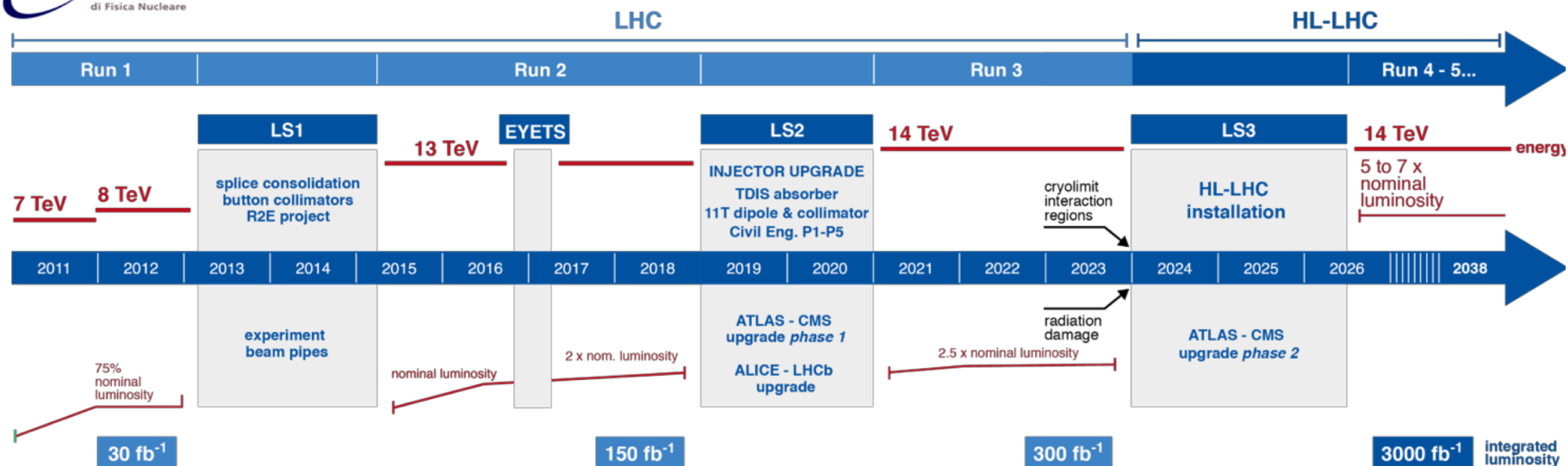
HEP Computing today

- LHC is the main consumer of the grid resources shared among many experiment (> 90% of the accounted computing capacity)

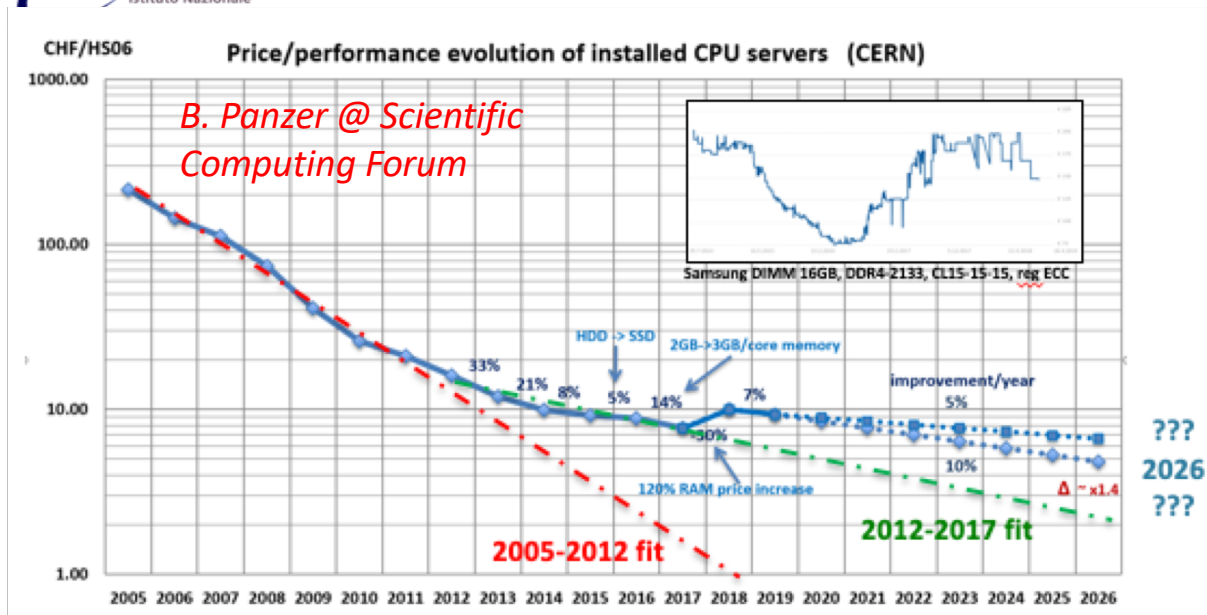


Pledged resources fully used.
Opportunistic ones available:
extra grid resources +
HPCs/Clouds/HLTs

LHC / HL-LHC Plan

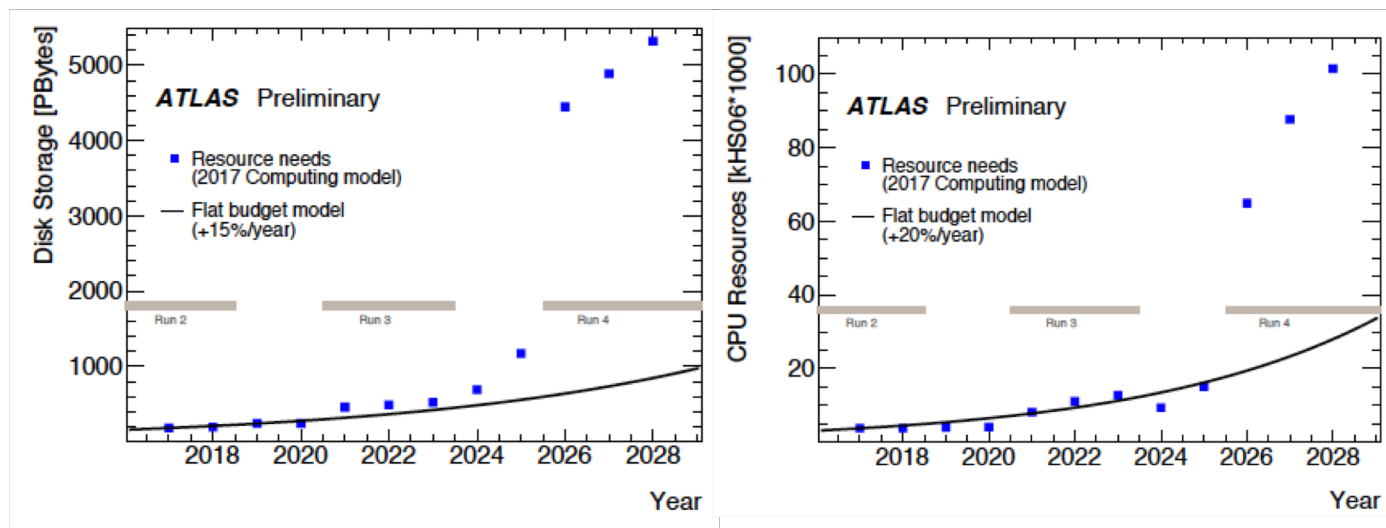


Challenges



- Cost of hardware decreasing vs time but much less than in the past
- Trends driven by market rather than technology

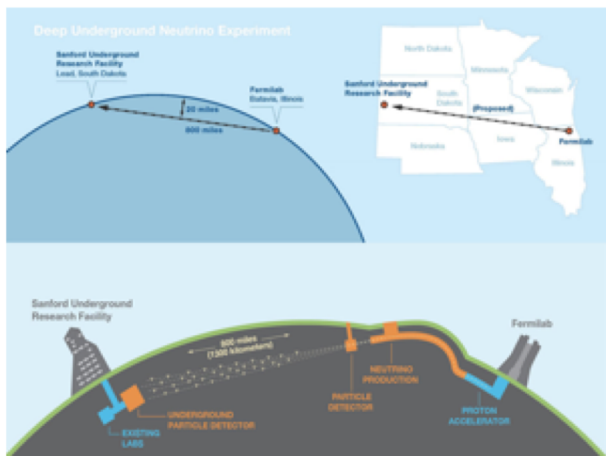
Long term predictability not reliable



Resource needed for Run-4 by ATLAS and CMS well beyond flat funding scenario

Future is not only LHC

S. Campana – ESPP 2019



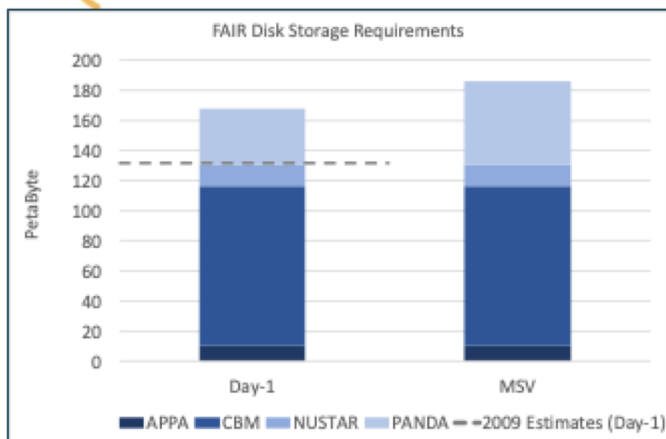
DUNE DEEP UNDERGROUND
NEUTRINO EXPERIMENT

DUNE foresees to produce
~70PB/year in the mid 2020s

Several experiments will require relatively large
amount of compute and storage resources.
Several factors less than HL-LHC

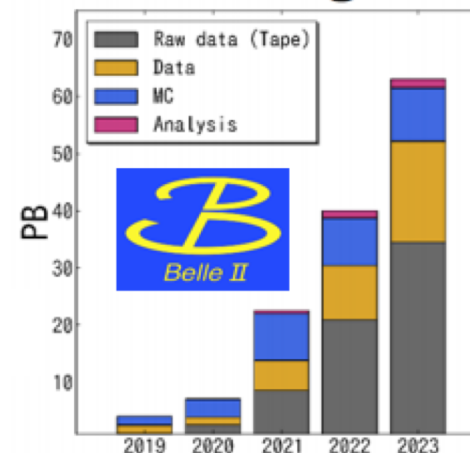


Comparable data
volume to LHC



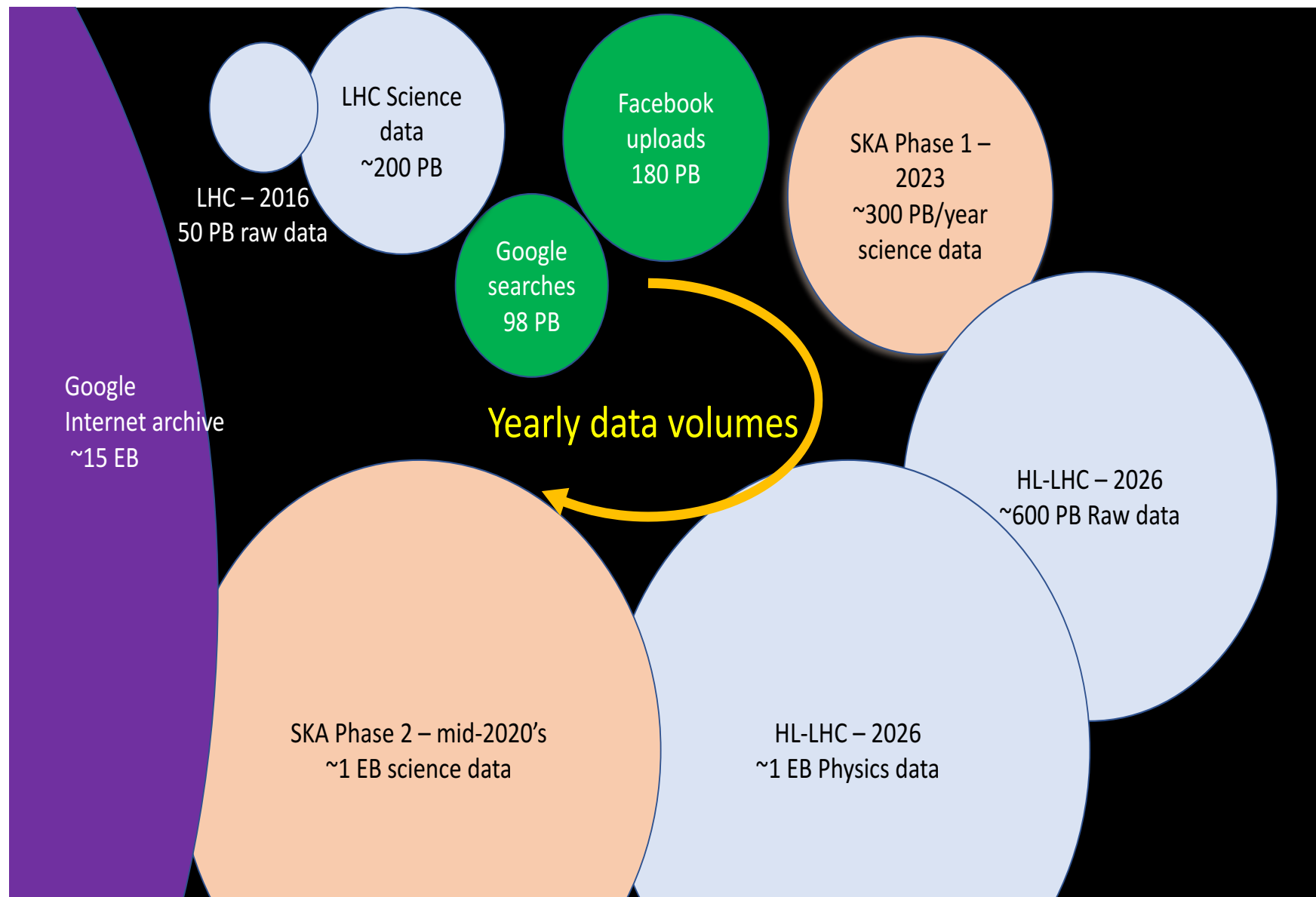
J. Eschke @ ESCAPE kick-off

Storage



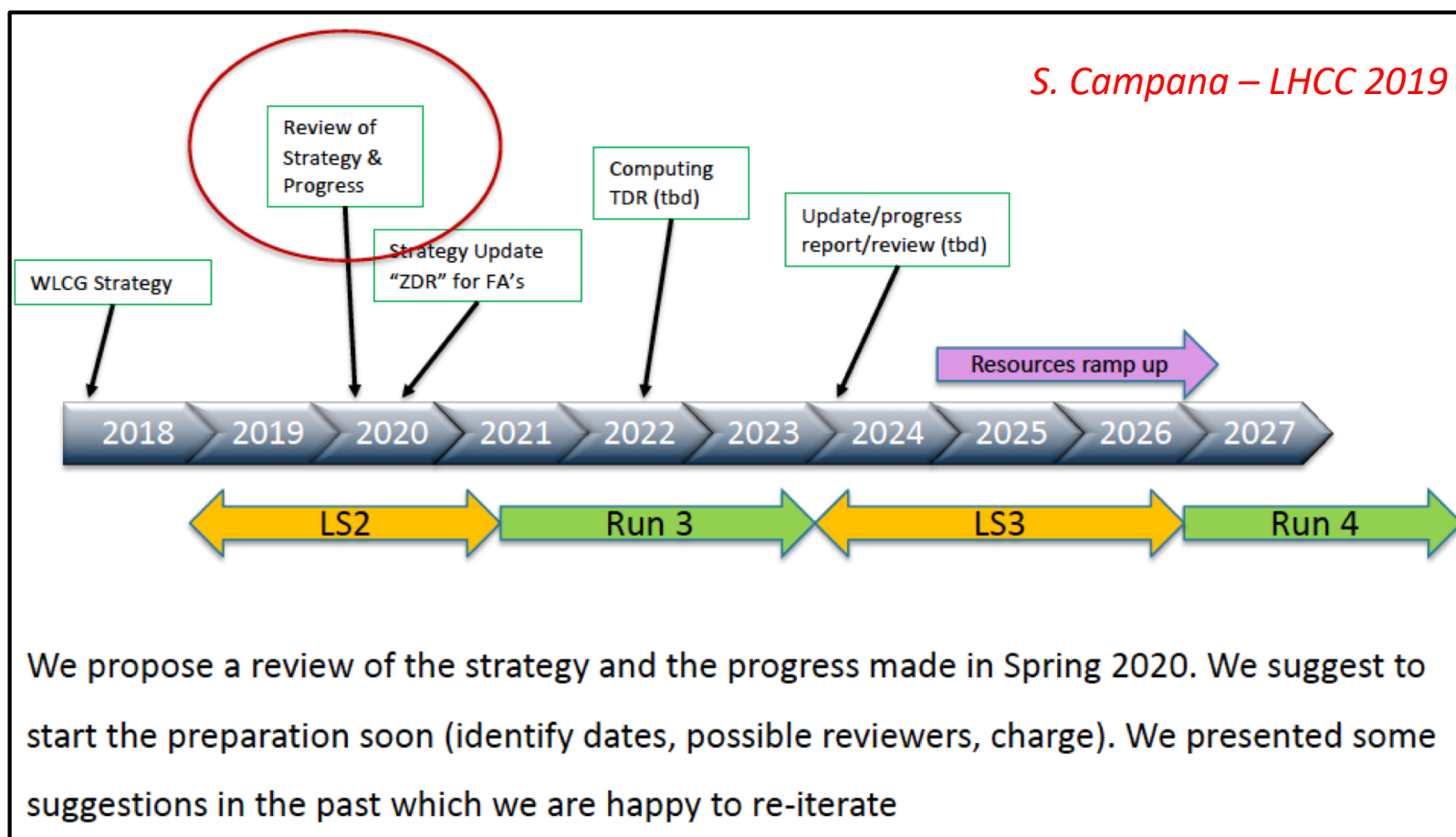
Y. Kato @ HOW 2019

Future is not only LHC



Computing Road toward HL-LHC

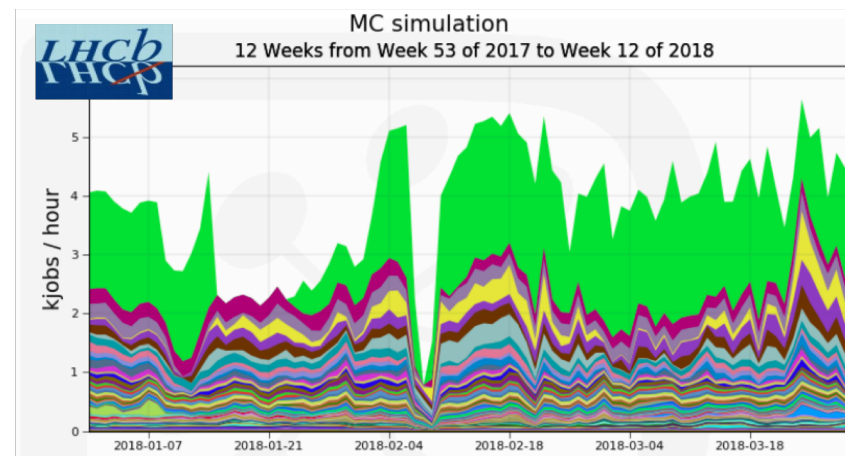
- HEP Software Foundation delivered in 2018 a Community White Paper identifying the main challenges in HEP computing and defining the roadmap for Software and Computing evolution
- WLCG Strategy Document for HL-LHC proposed by WLCG in 2018



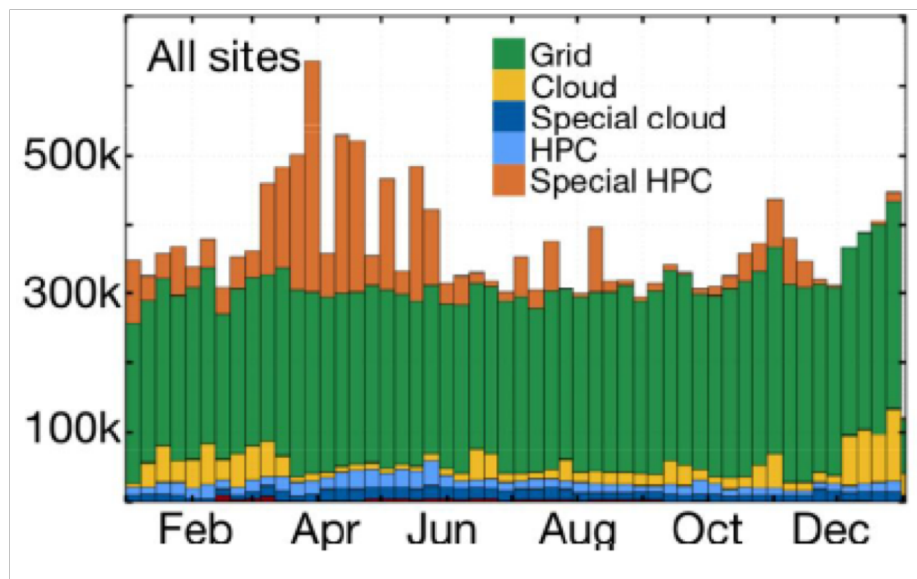
Heterogeneous Resources

- Modern hardware landscape is heterogeneous. Efficient usage will be mandatory in the future
- So far, up to 25% of resources used by LHC experiments are not GRID:
 - Cloud Computing (commercial)
 - High Performance Computing
 - HLT farms
- Challenges: heterogeneous interfaces, network, operational effort, data management

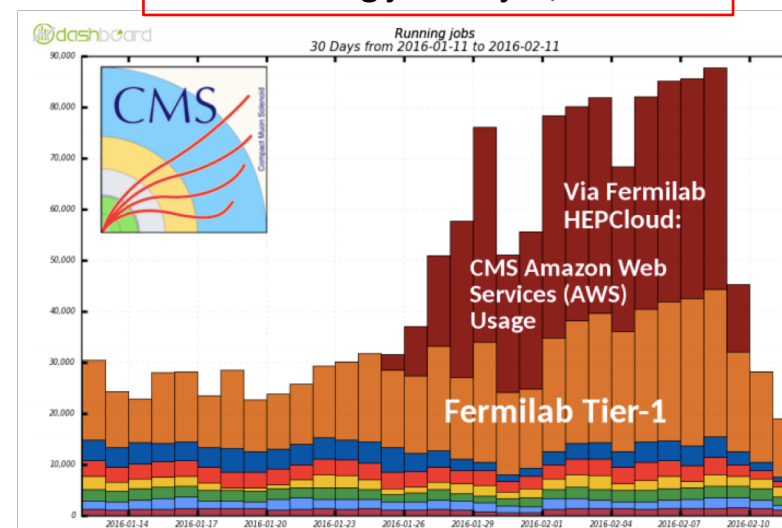
LHCb HLT farm in 2018 shutdown



ATLAS CPU usage in 2018



CMS running jobs in jan/feb 2016



Google, Amazon via HepCloud

Azure via DoDas

High Performance Computers

HPCs are already in HEP computing and the usage will grow because of major Funding Agencies requirements (HPC as pledge capacity) and EU projects funding (EuroHPC)

This has a cost:

- HPC systems heavily rely on the accelerator power. No use of accelerators, no running on exascale machines
- HPC systems heterogeneous themselves
- sw issues: portability to non X86, usage of accelerators
- Experimental framework evolution to offload code on CPUs and accelerations transparently
- Dedicated investment of effort on edge services and tools (sw distribution, workload and data management)

Cost effective if we are able to have stable allocations not just backfill



But ..

From the discussion summary:

S. Campana – LHCC 2019

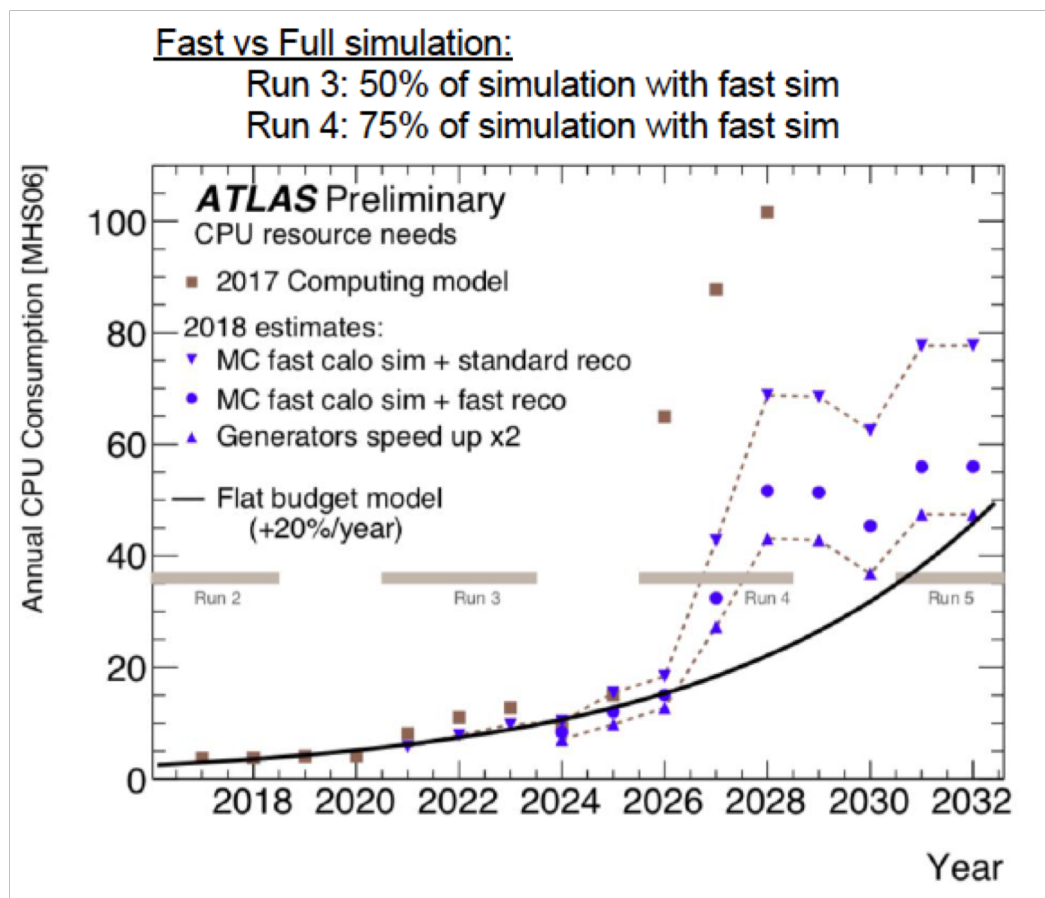
We experiments think that, while on our side we are committed to do our best, we should be facilitated by the FAs in these aspects. The best approach would be to have HEP experts to be part of the definition process of both architectures and policies, having hence the HEP use case as a first-class citizen for HPCs. Sadly, this is not happening today, with HEP entering the game only when HPC machines have already been built and put into production.

sw and model evolutions

Usage of new heterogeneous resources is not enough

Evolution of experiment software and models

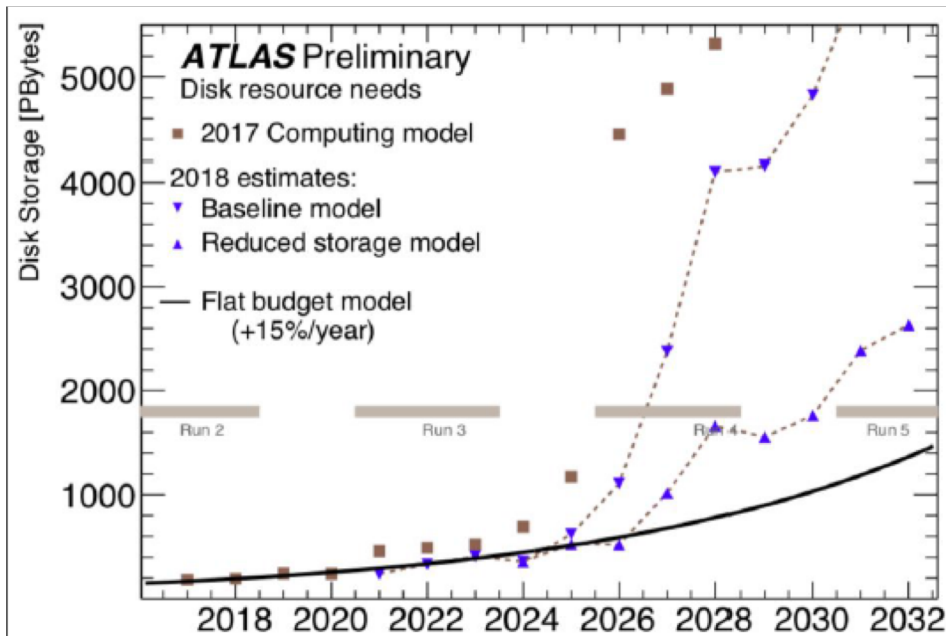
Simulation and Reconstruction evolution:
fast vs full



Experiment sw and model evolutions

Opportunistic disk storage not available


New Storage Model



Solution from physics

- Analysis models
 - Most disk space used by analysis formats
- Data Model

T. Boccali, M Klute



Data Tier	Size (kB)
RAW	1000
GEN	< 50
SIM	1000
DIGI	3000
RECO(SIM)	3000
AOD(SIM)	400 (8x reduction)
MINIAOD(SIM)	50 (8x reduction)
NANOAO(SIM)	1 (50x reduction)

Analysis data formats

- Mini and Nano AODs the only format on disk with multiple copies
- CMS using them since Run2

50 PB of MINI and few PB of NANO per year

ATLAS model for Run3/4:

- Instead of a plethora of analysis format, only DAOD_PHYS (miniAOD): 50 kB/event and DAOD_PHYSLITE (nanoAOD): 10 kB/event
- Only 1 DAOD for MC and Data (only 20 for CP and sys studies)
- AOD on tape (data carousel)

Storage consolidation

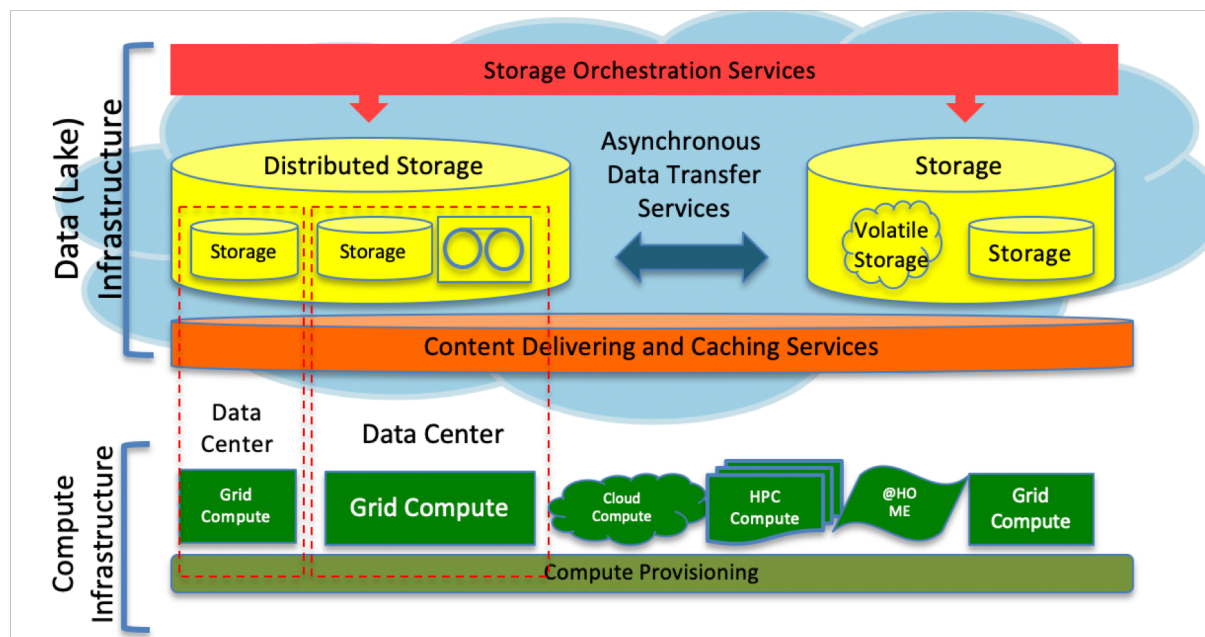
Opportunistic disk storage not available - Data is the main asset of HEP !

Computing Model evolution in order to introduce changes in the way we use and manage storage.

- Resource optimization to improve performance and efficiency and simplify operations

Storage consolidation based on a Data Lake Model

- Separation of compute and storage services. Higher specialization and improvement in reliability and operations
- Storage costs reduction: global redundancy vs local redundancy, QoS, economy of scale



DOMA projects – Data organization, Management, Access

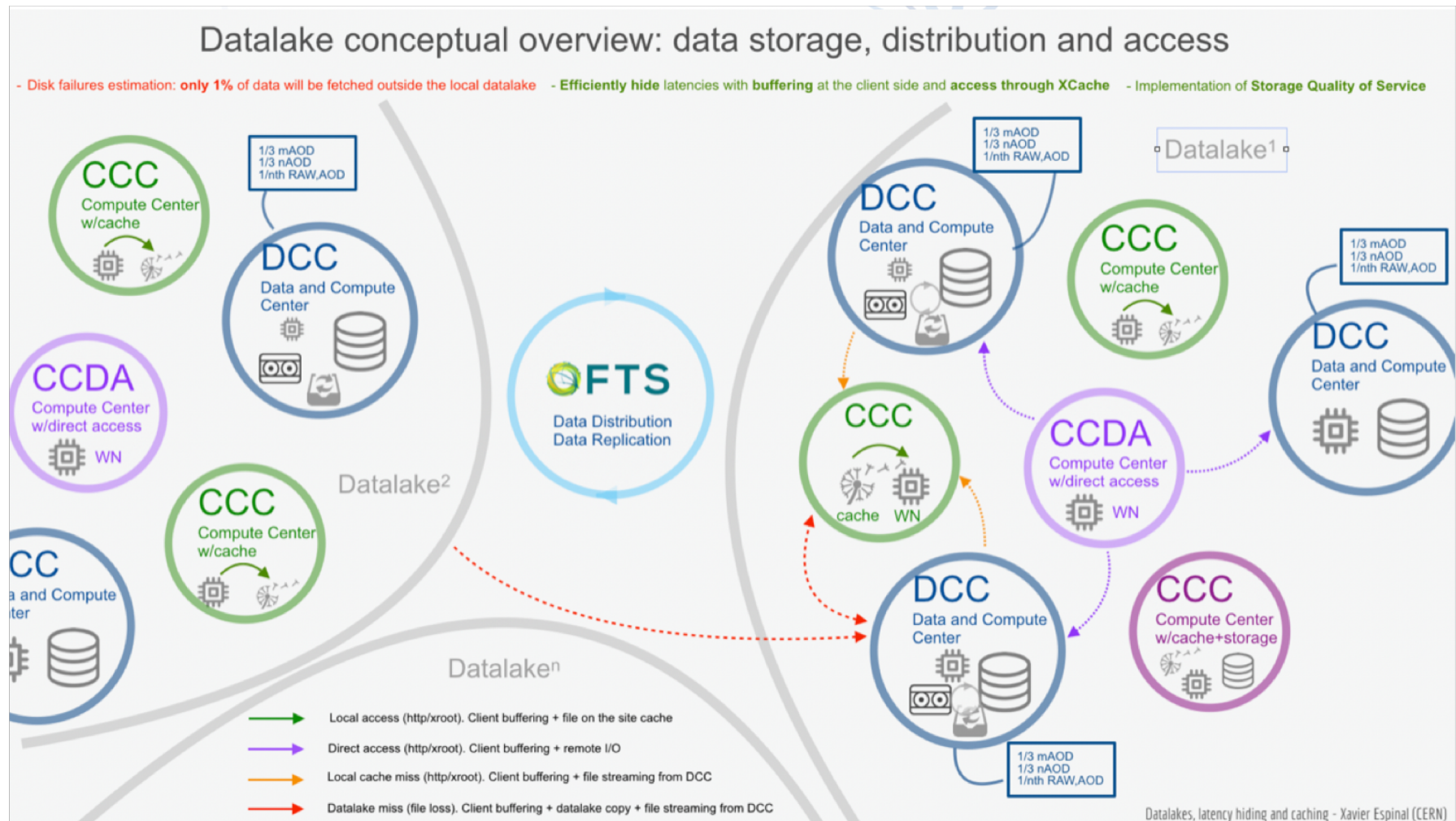
- A set of R&D activities evaluating components and techniques to build a common HEP data cloud

The different Data Lakes worldwide forms an universal storage infrastructure

- A Data Lake is composed of **Compute Centers**, **Data** and **Compute Centers** and at least one **Archive Center**.
 - 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.

X. Espinal – HSF & WLCG Workshop 2019

Storage Consolidation



Some Data Lakes are foreseen (>1 – USA, EU, Asia Pacific at least)

Computing and Storage needs of LHC, HEP and non HEP experiments in the next decade will be one order of magnitude higher of what current computing models and technology evolution can guarantee

- The WLCG and HSF communities are working hard to:
 - Evolve compute models
 - Optimize experiment software and middleware
 - Use efficiently heterogeneous resources in transparent ways
- Role of Italy
 - Active participation in international activities (experiments, EOSC, Escape,)
 - Italian projects (IDDLS, PON IBISCO,) to evolve the Italian scientific computing infrastructure towards the HL-LHC models