Facing the computing challenge of HL-LHC

José M. Hernández
CIEMAT, Madrid

3rd RED LHC Workshop

UAM, Madrid
6-8 May 2019

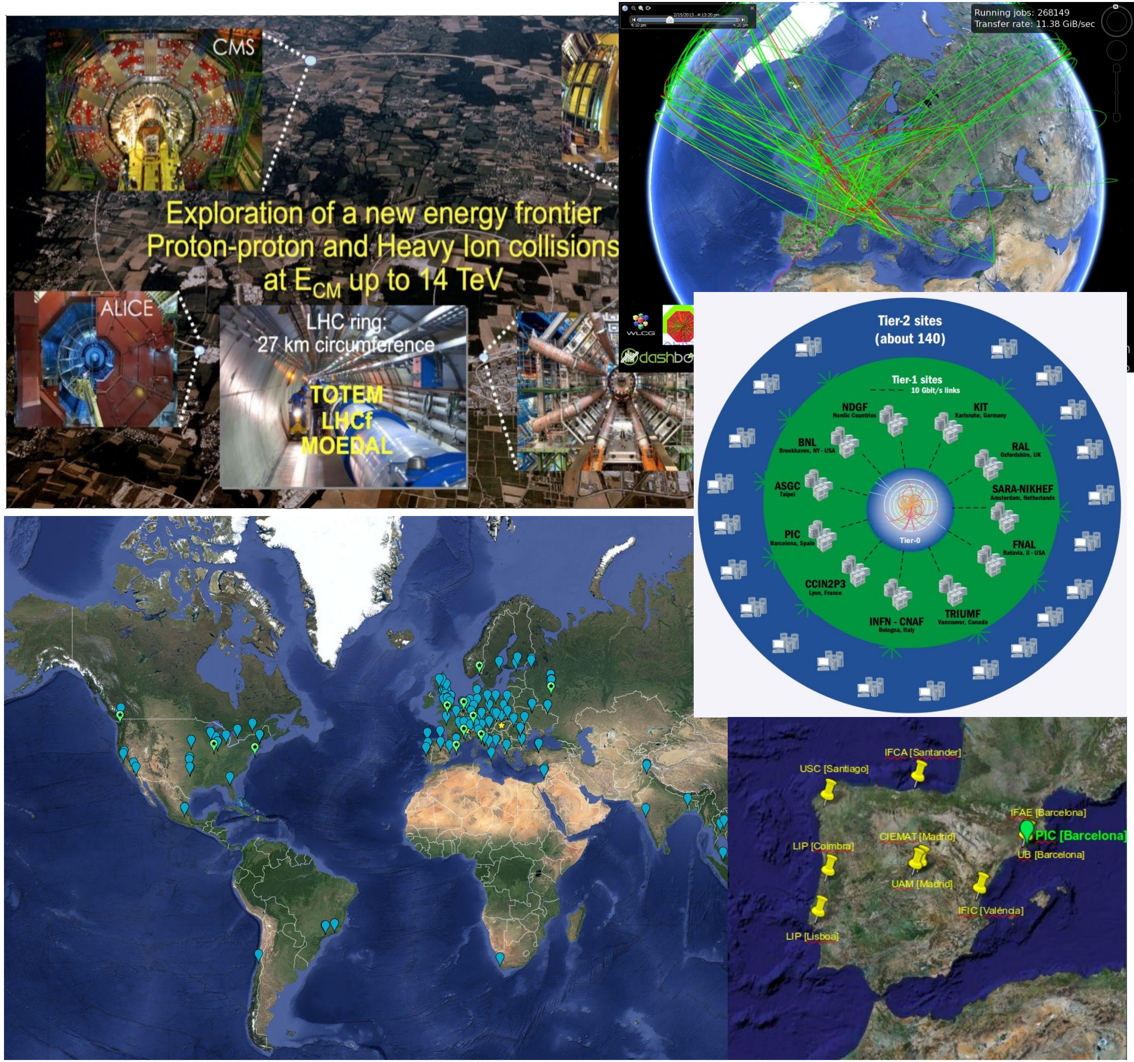
Worldwide LHC Computing Grid

Distributed high-throughput computing infrastructure to store, process & analyze data produced by LHC experiments

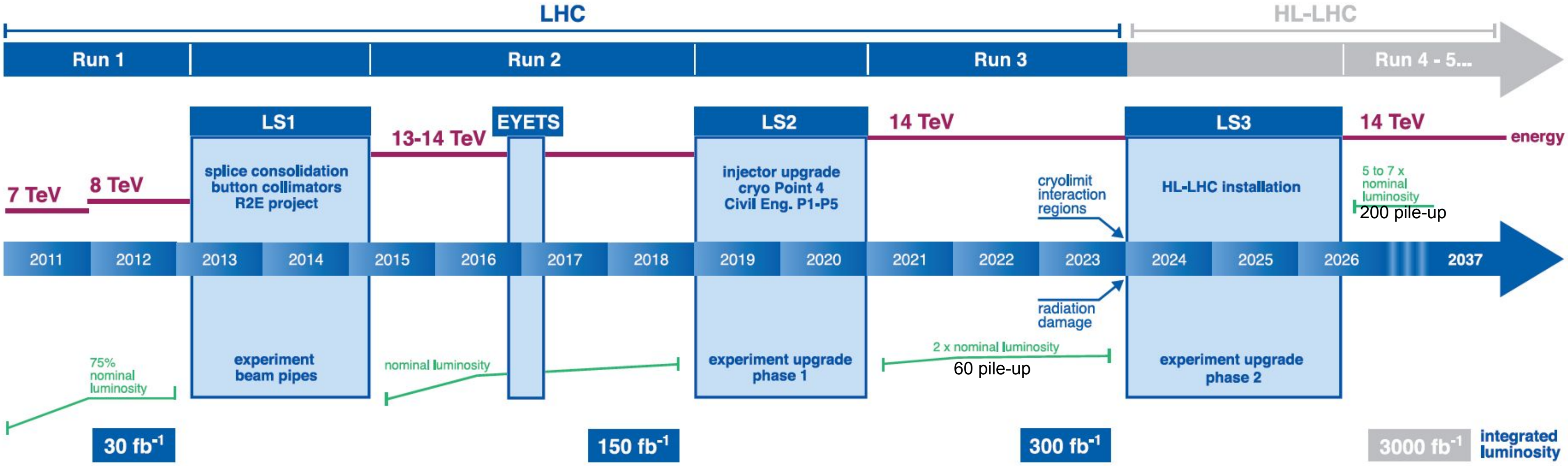
- 167 sites, 42 countries, 63 MoU's
- 800k cores
- ~500 PB disk storage
- ~750 PB tape storage
- Optical private network (LHCOPN) and overlay over NRENs (LHCONE) with 10/100 Gbps links

Spanish contribution:

- ~5% resources (MoU)
- 1 Tier-1 center (PIC CIEMAT/IFAE)
- 6 Tier-2 centers (1 en CIEMAT)

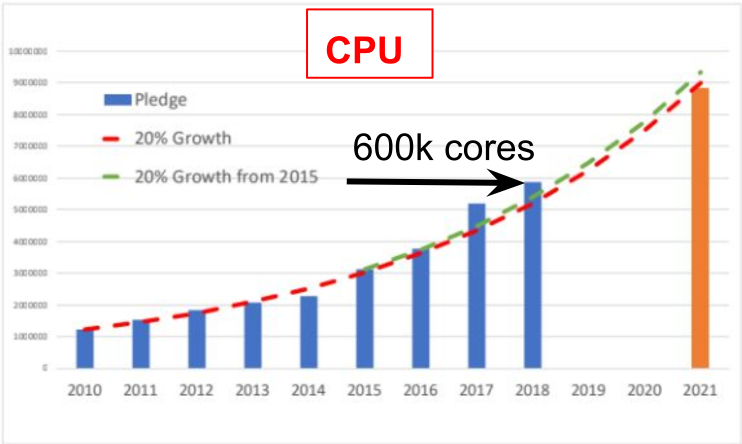


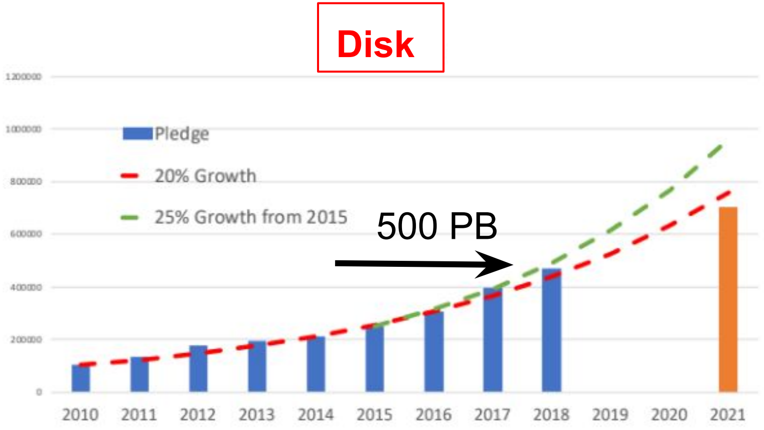
LHC / HL-LHC Plan

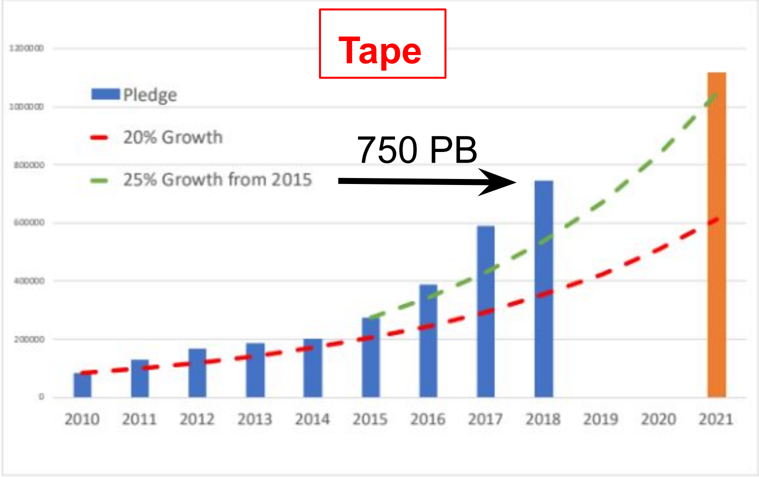


- Run 3 (2021-2023): **~2x more data**. **Evolutionary** changes in computing models
- Run 4 (HL/LHC, 2026+): **~20-30x more data**. **Revolutionary** changes required

Run 3 resource needs evolution



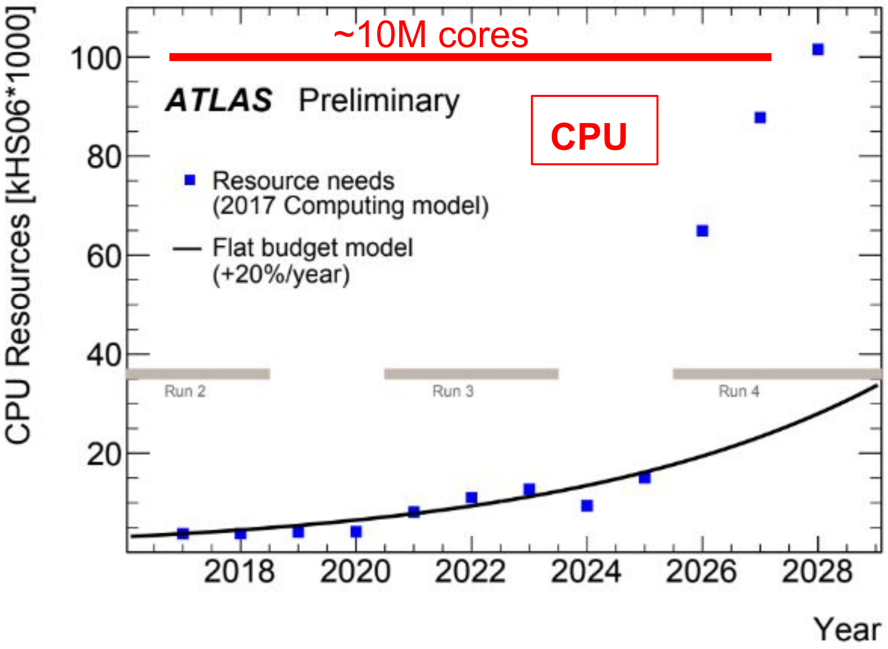


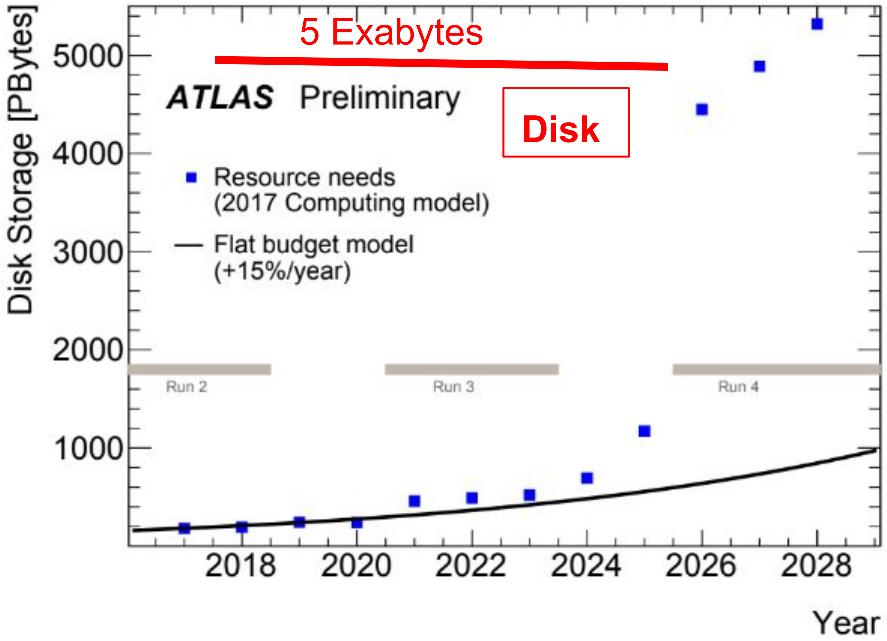


- 2010-2018 – pledges
- 2021 assume 1.5 x 2018

Overall, Run-3 resource needs look compatible with flat spending in the next years

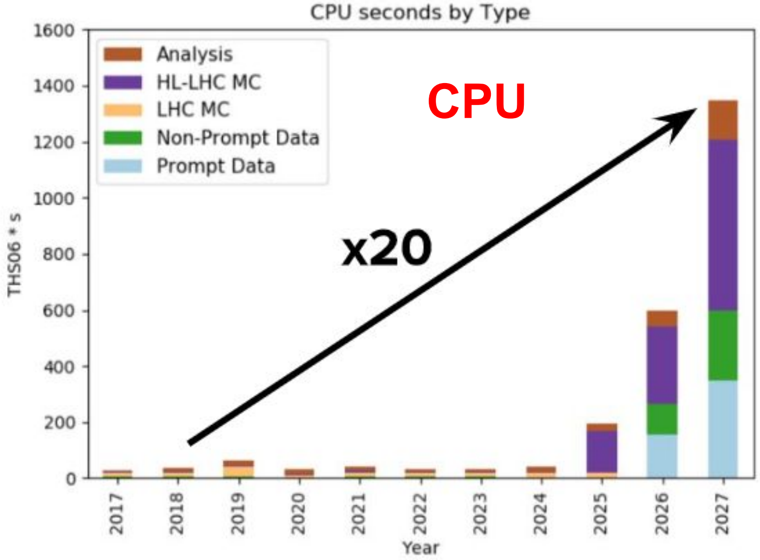
The HL-LHC computing challenge: ATLAS

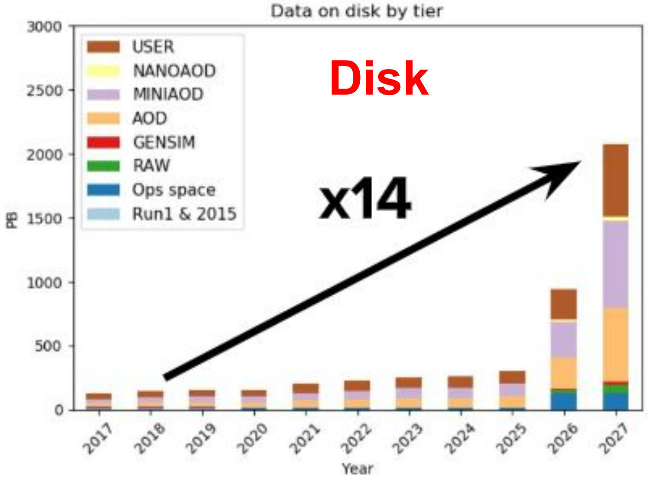


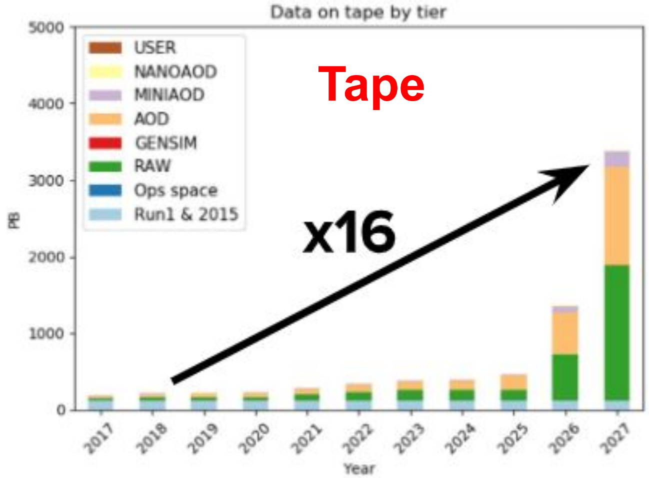


- **~4-5x gap** between "flat budget - 20% annual increase" and resource requirements for HL-LHC
- **Intense R&D** to reduce data and compute resource requirements

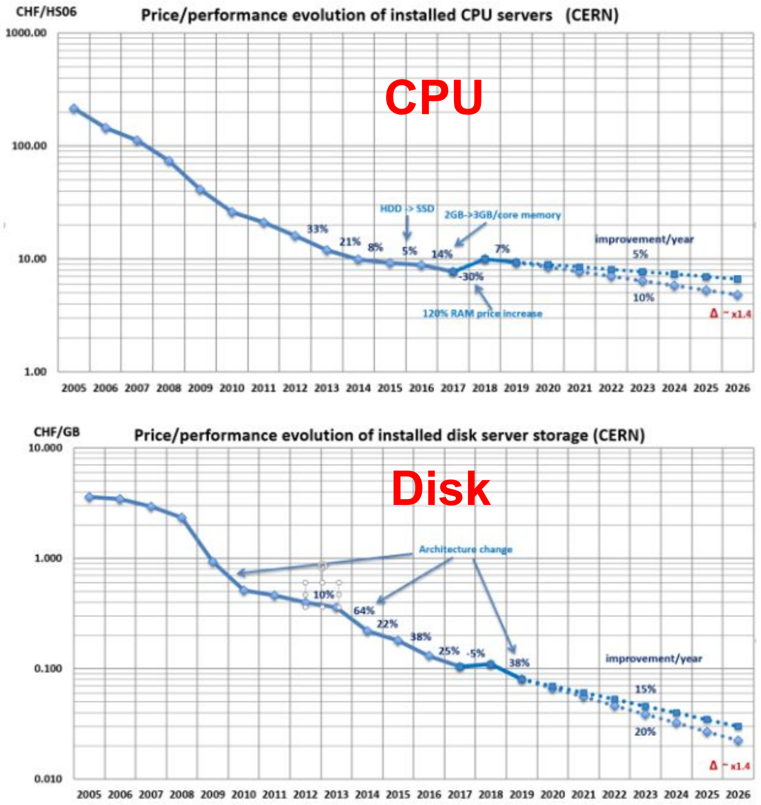
The HL-LHC computing challenge: CMS



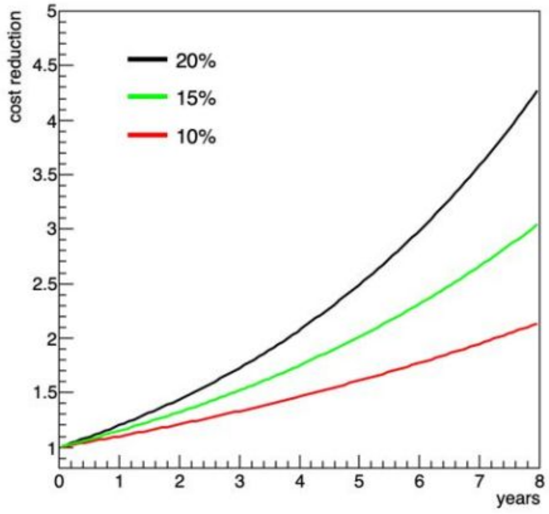




Cost evolution



- Unclear hardware cost evolution
    - Significant impact
- Current price reduction assumption:
    - 10% CPU, 15% disk, 20 tape



R&D for HL-LHC computing

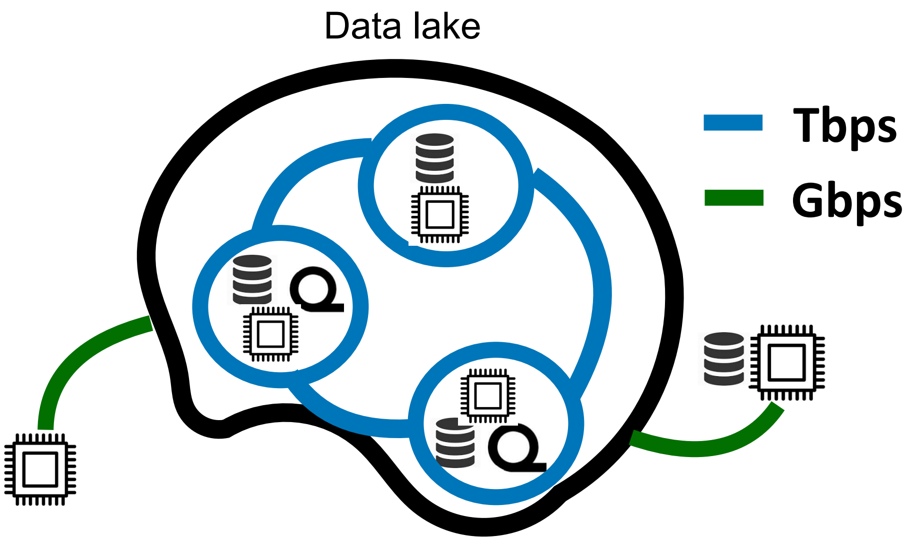
Towards a more efficient computing infrastructure

The data lake model

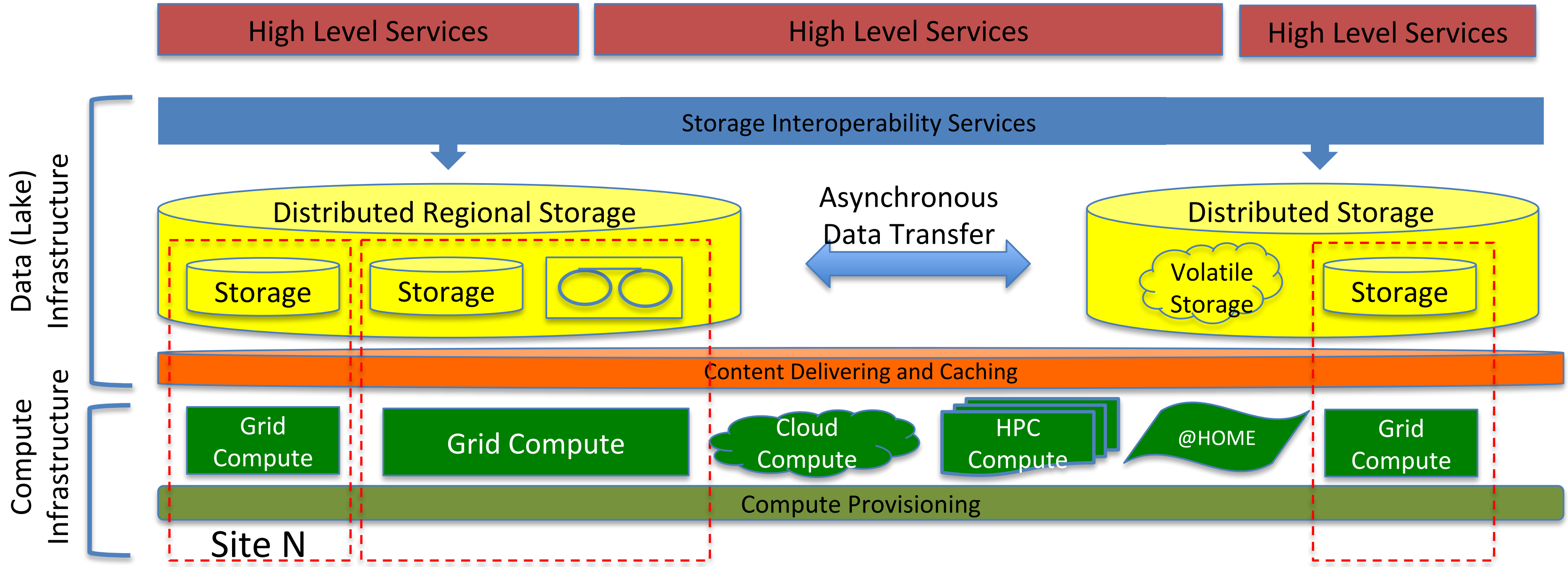
Current storage model

- Lots of sites (150+) with managed storage
- Mostly local data access
- High level of data replication

- Reduce operational cost: deploy fewer (larger & **federated**) storage services
    - Global redundancy, economy of scale
- Introduce **caching** layer to hide latency of **remote data streaming**
    - High bandwidth content delivery network
- Reduce hardware cost: introduce the concept of **QoS** (Quality of Service)
    - Data tiering to optimize access



Data and Compute Infrastructures



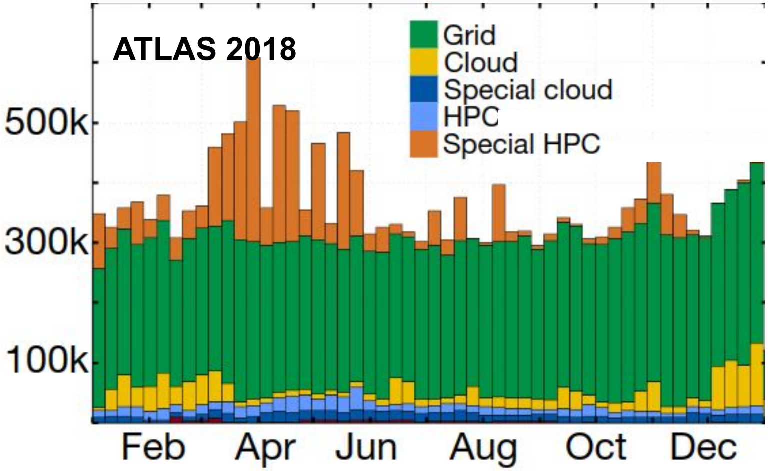
Use additional compute resources

Exploiting supercomputers for LHC

- Lot of **funding** worldwide in supercomputer (HPC) facilities
  - Defined roadmap towards ExaFlop machines
    - e.g. EuroHPC B€ funding: 2 ~200 PFlop machines by 2021, 2 exaFlop by 2024
  - Funding agencies pushing us to use those resources
- Data intensive computing with HPC facilities is a **challenge**
  - Limited/no network connectivity in compute nodes
  - Limited storage for caching input/output event data files
- Our applications are not really suited for HPC
  - No large parallelization (no use of fast node interconnects)
  - No substantial use of accelerators (GPU, FPGA)
- Substantial **integration** work to make HPC work for HTC
  - No one-fit-all solution: each facility is different
  - Little effort available in the LHC experiments
- Not suitable resource **allocation** model
  - We would need a guaranteed share of resources rather than apply for allocations

HPC usage in LHC

- ATLAS and CMS are using HPC centers in the US and Europe
    - NERSC (US), CINECA (IT), BSC (ES), Piz Daint (CH)
- Mostly for event generation and (geant4) simulation (**CPU-bound**)
    - ~20% of the ATLAS simulation, ~1% CMS simulation
- The prominence of GPUs is increasing in future HPC machines
    - Need to adapt workflows to these highly parallel architectures
- Important to influence the architecture of future HPC machines
    - Support for high throughput computing

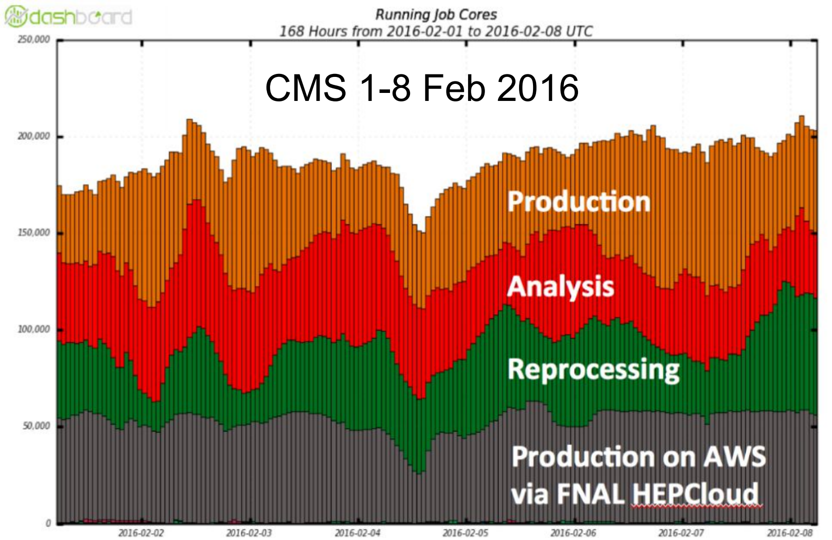


Barcelona supercomputer center (BSC)

- #25 in Top500
  - MareNostrum 4, 153k cores, 10 PFlops
- Bidding for EuroHPC pre-exascale machine
  - ~200 PFlops, 250 M€, 10 MW power
- ATLAS, through allocations granted to IFAE and IFIC has successfully used BSC
  - CMS, through a project led by CIEMAT, is adapting the workload management system
- **Agreement** being worked out with BSC to use resources for LHC simulation at large scale
  - Technical and policy questions under discussion
    - Accessibility, edge services, allocations

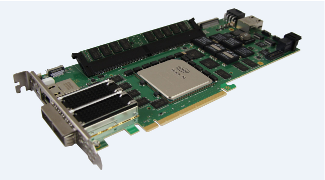
Use of commercial Cloud resources

- CMS and ATLAS have run large scale tests using Cloud compute nodes
  - Amazon AWS, Google Cloud, Microsoft Azure
  - ~50k cores running concurrently for few days
- Cost **not yet competitive**
  - Need to use spot market instances, much cheaper than on-demand resources
  - High storage and networking costs
- Currently essentially no commercial cloud use for LHC computing
- Potential future **opportunities**
  - E.g. the European Open Science Cloud (EOSC)
    - A EU model for use of cloud computing in the private and public sector



Use of compute accelerator cards

- Dramatic development of massively parallel architectures
  - Graphics Processing Units (GPU)
  - Field Programmable Gate Arrays (FPGA)
- Potential large speed improvement from hardware accelerated coprocessors
  - Large performance/€ and smaller electric consumption/performance
- **Difficult to use**
  - Need to re-engineer our codes to a massively parallel environment
  - Data ingestion can be a limiting factor
- Very suitable for certain applications
  - E.g., excel at training deep neural networks
- New HPC machines will bring a lot of these cards



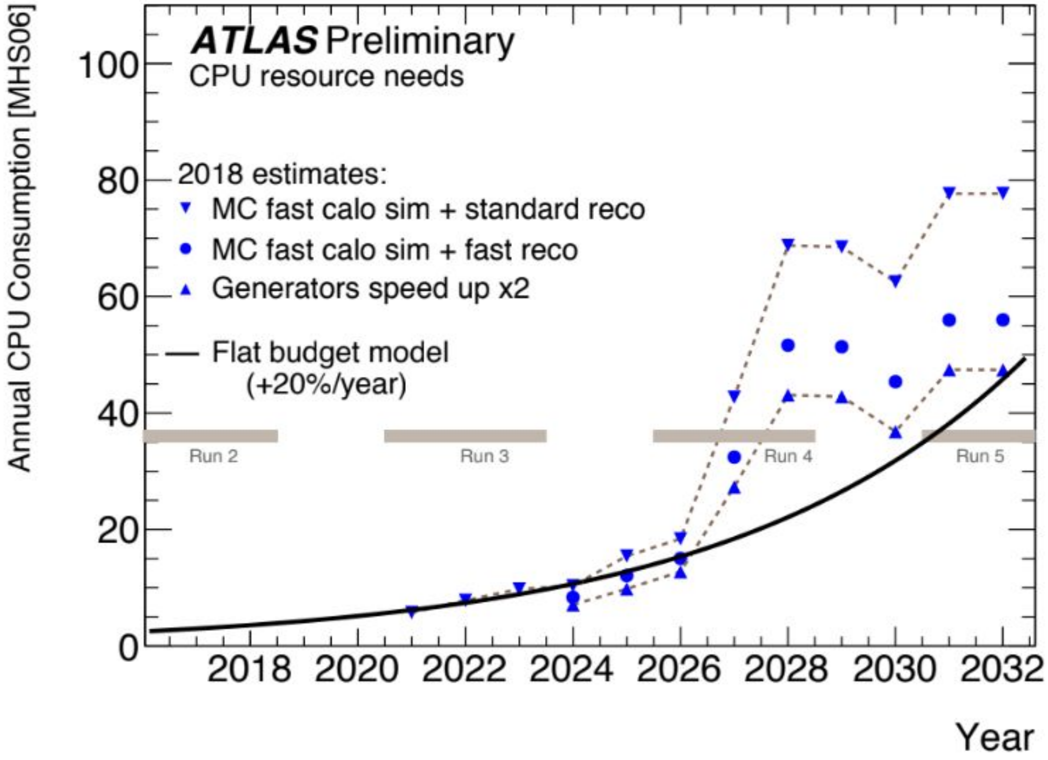
Software optimization

The solution could come from the software

- Recent initiatives
  - HEP Software Foundation (coordinate software R&D for LHC)
  - Institute for Research & Innovation in Software for HEP (IRIS-HEP); 25M$, 5 years
  - Proposal a EU scientific software institute
  - COMCHA forum in Spain
- Exploit **new hardware architectures**
  - High level parallelism, new instruction sets, non x86 processors
  - Support in software frameworks for **heterogeneous** hardware
    - Support for multi-threading, vectorisation, CPU/GPU orchestration
- **Innovative algorithms**
  - Machine/deep learning
  - Recast physics problem as machine learning problem vs re-rewrite physics algorithms for new hardware

ATLAS CPU needs reduction by using fastsim/fastreco

Faster physics algorithms: exploit more broadly fast simulation & reconstruction

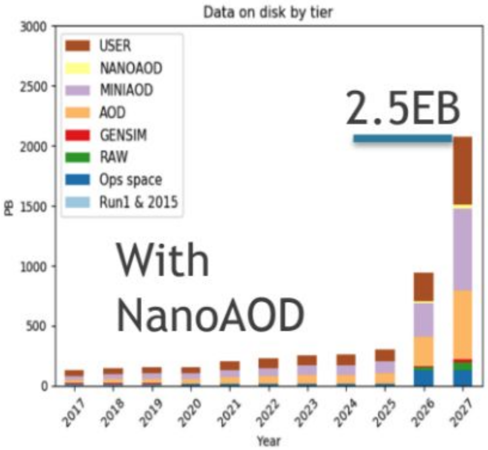


Less data

Reduce amount of data

- Less data ➦ less storage, less processing and analysis compute needs
  - Reduce **trigger** output rate (HL-LHC planned 7.5 kHz ➛ ?)
  - Reduce data **formats**
- **Impact** of physics?

- NanoAOD format in CMS
  - ~1 kB/event
  - Goal: to be used by 50% of physics analyses
  - ~Halves CMS storage needs for HL-LHC



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)
NANOAOD(SIM)	1 (50x reduction)

Analysis data formats

Summary and outlook

- HL-LHC poses a big computing challenge
  - Resources unaffordable with current computing models and flat funding
- Problem not solved yet but well underway
- The solution will most probably be a combination of new software and hardware technologies
  - Machine learning, accelerator cards, supercomputers, ...
- Intense ongoing R&D program
  - WLCG TDR by 2022
- Still 7 years to go. A lot in terms of technology evolution