New challenges for distributed computing at the CMS experiment

Natascha Krammer
on behalf of the CMS Collaboration

Outline

- Present and future computing needs in
  - LHC Run-3 and HL-LHC

LHC RUN3

HiLumi HL-LHC PROJECT

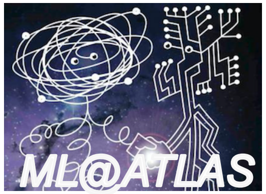
- How to full-fill the computing needs
  - Grid Comp., Cloud Comp., High-Performance Computing (HPC)

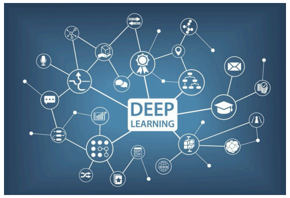
Google Cloud Platform

amazon web services™

HELIX NEBULA THESCIENCECLOUD

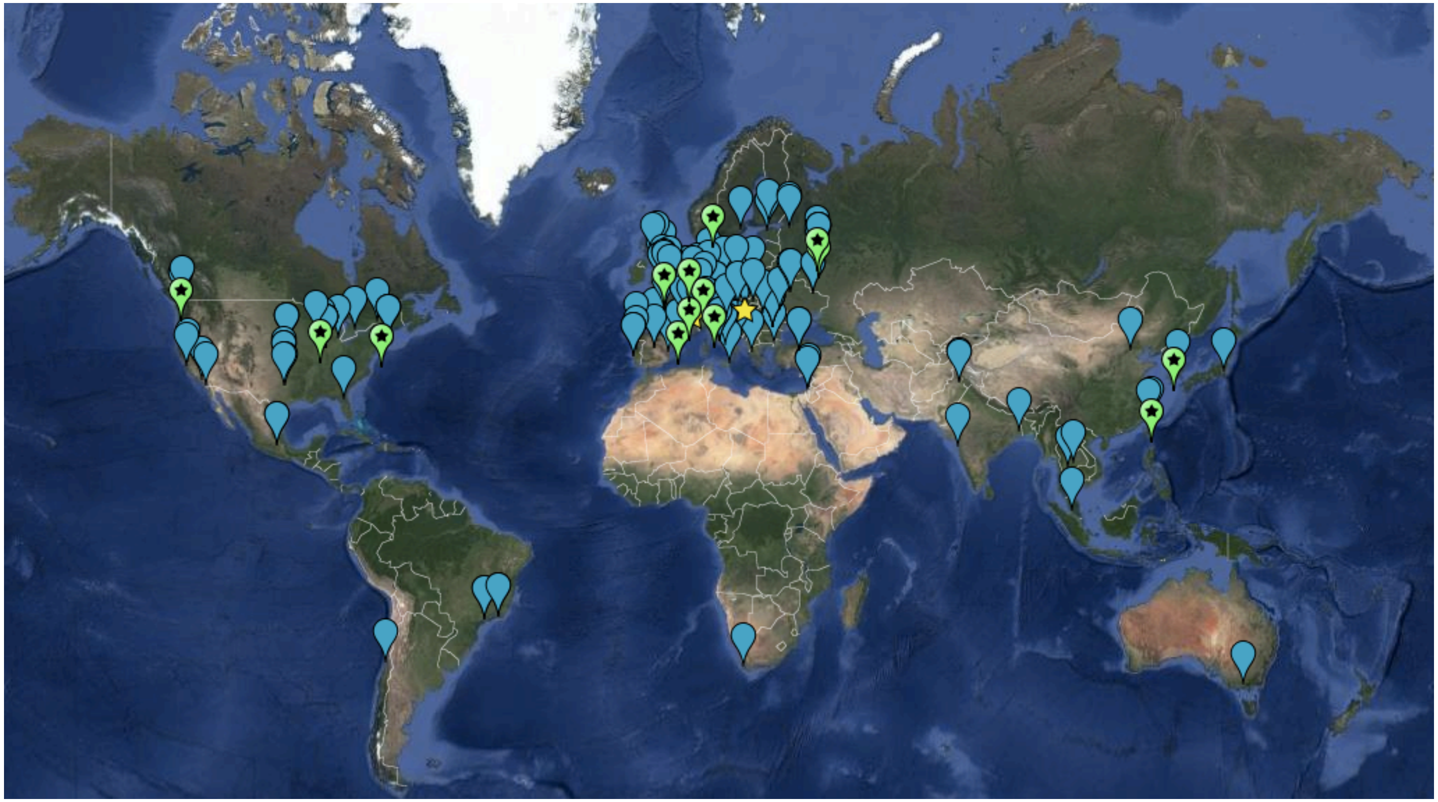
- Modern state-of-the-art methods in physics analysis and data science
  - Big-Data, Machine Learning - Deep Learning



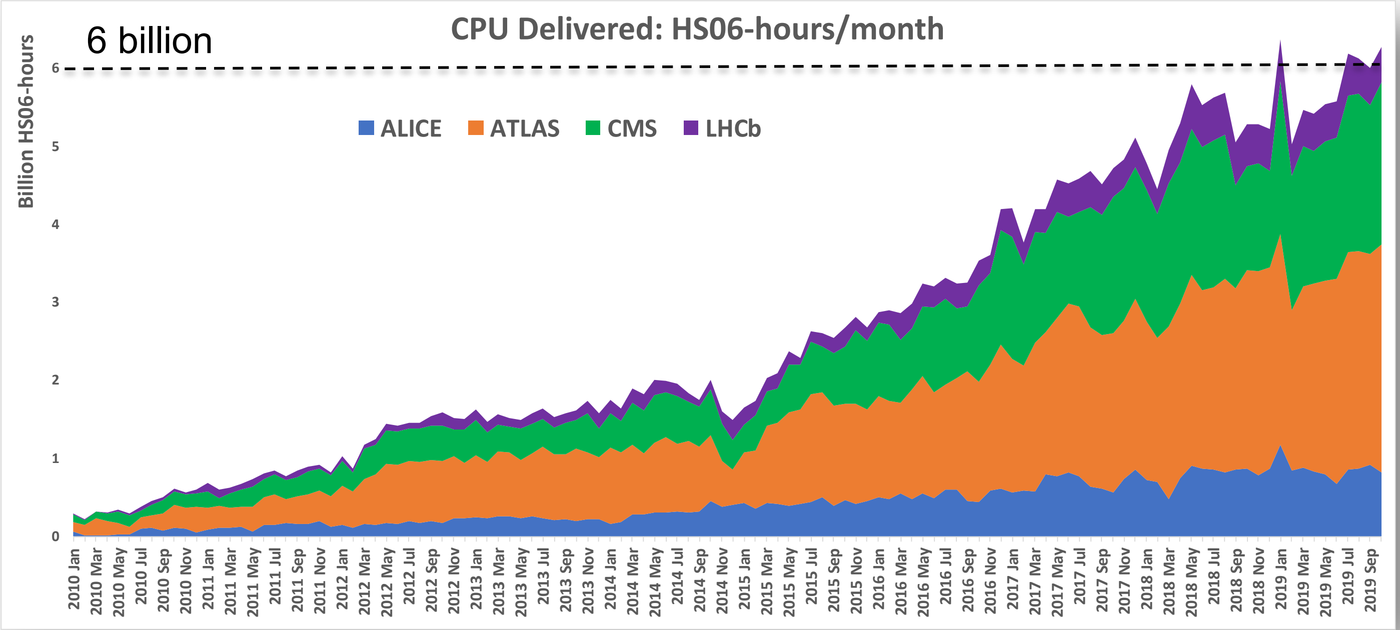


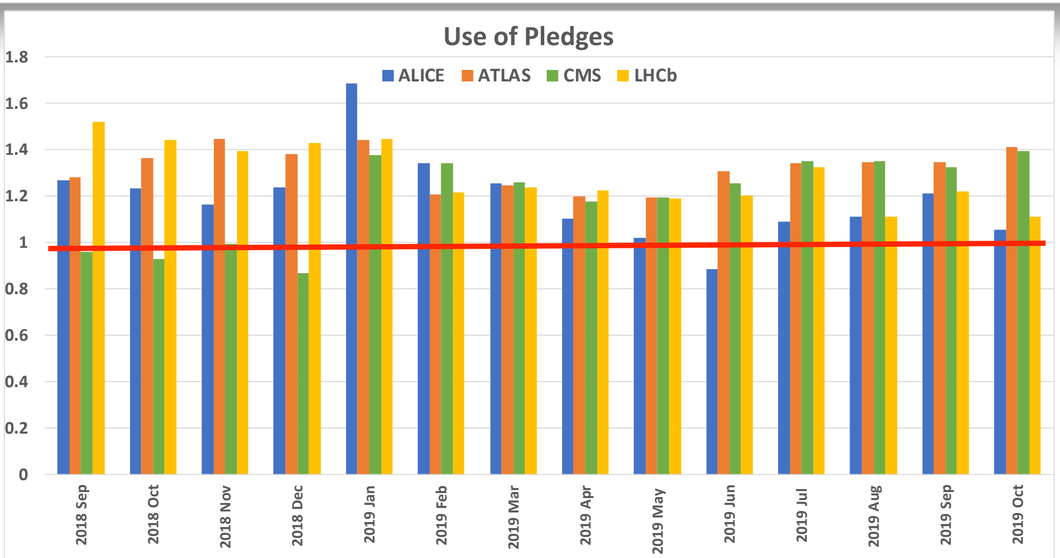
WLCG collaboration

- WLCG (Worldwide LHC Computing Grid) landscape
  - Over 170 computing centers in 42 countries
  - Computing power in 2020
    - CPUs: 6.500.000 of today's fastest cores (6.5 million)
    - Storage: Disk: 575 PB, Tape: 800 PB



LHC experiment resources



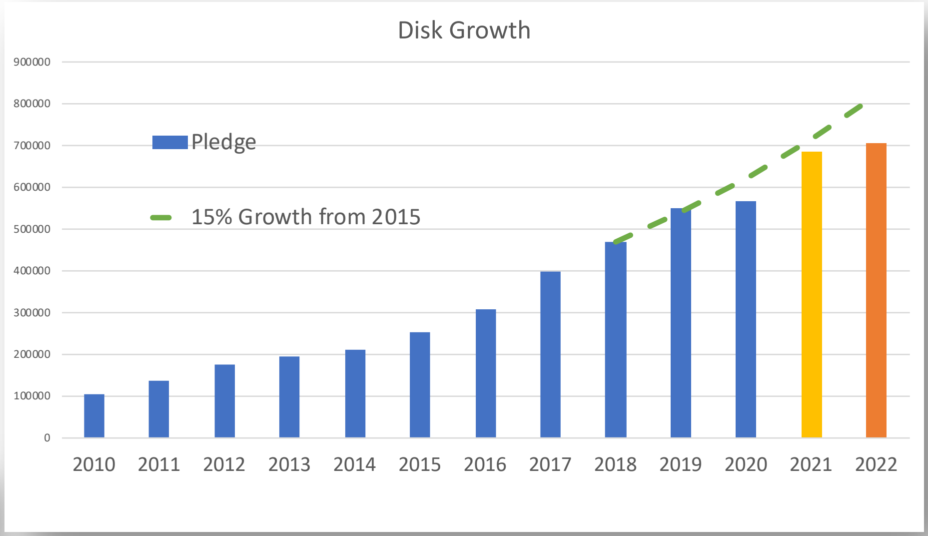


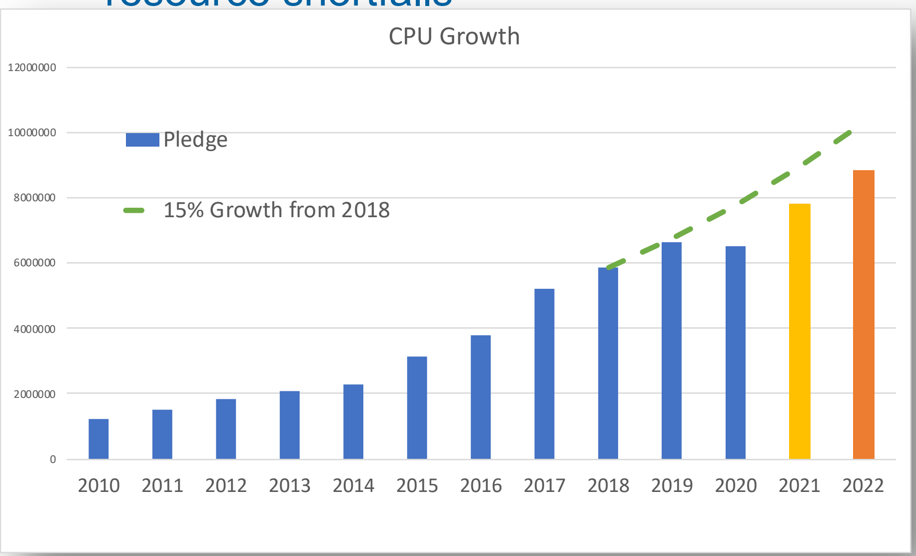
- Pledged (Grid) resources fully used by all experiments

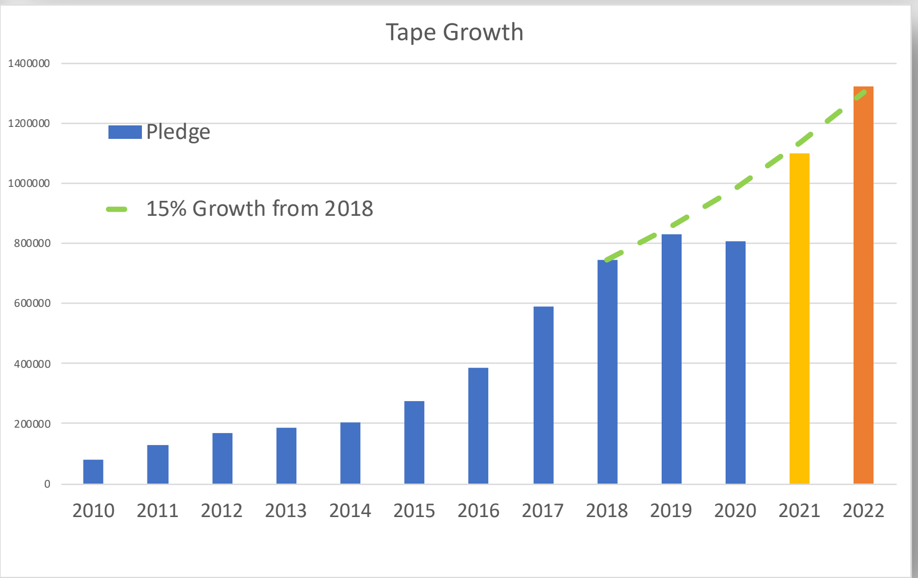
Volker Beckmann, WLCG Referees Report, LHCC, 20. Nov. 2019

LHC Run-3 resource evolution

- Run-3 probably manageable overall, but constant budget growth until Run-4 is essential for HL-LHC
- 15-20% increase in technology maybe an optimistic assumption – many indications that flat budget is much worse (~ 10%)
- Short extension of LS2 does mitigate resource shortfalls



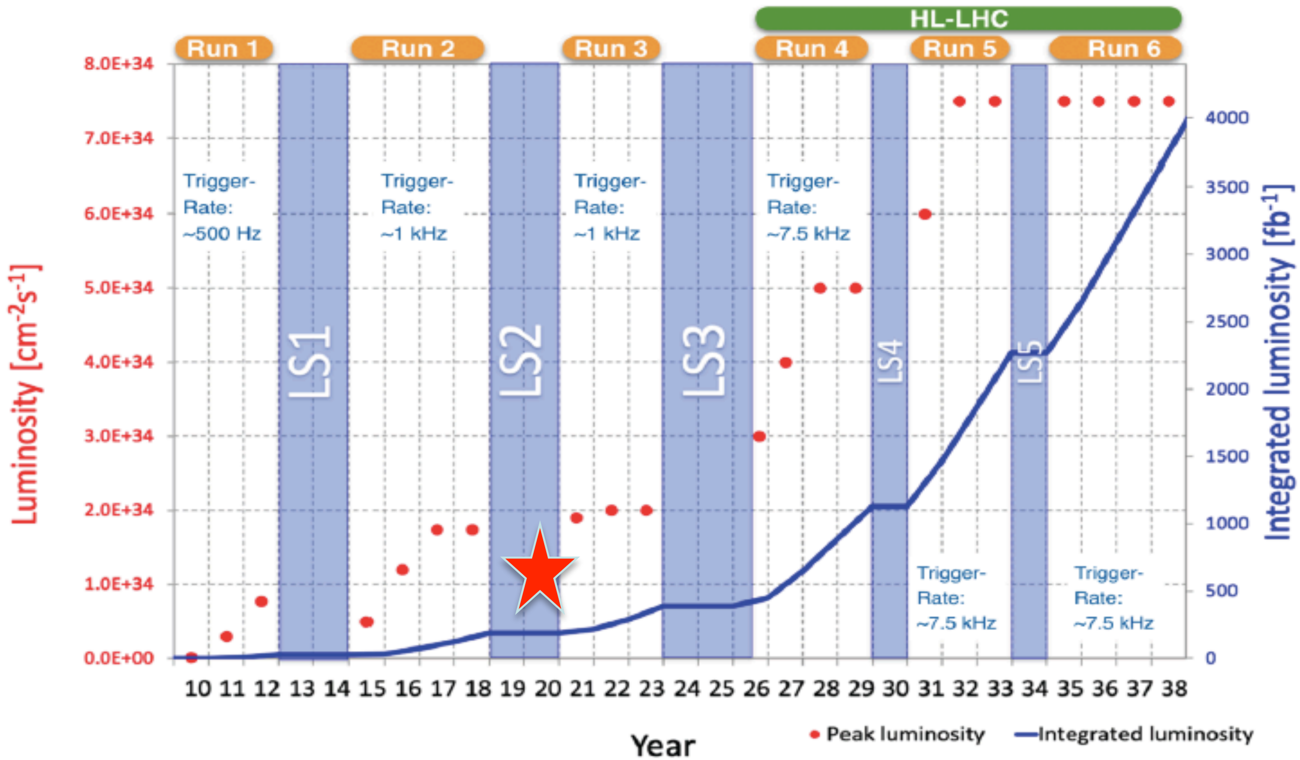




Actual installed
2021 request to C-RSG
2022 = 1.5*2018

Volker Beckmann, WLCG Referees Report, LHCC 20. Nov. 2019

HL-LHC operation schedule

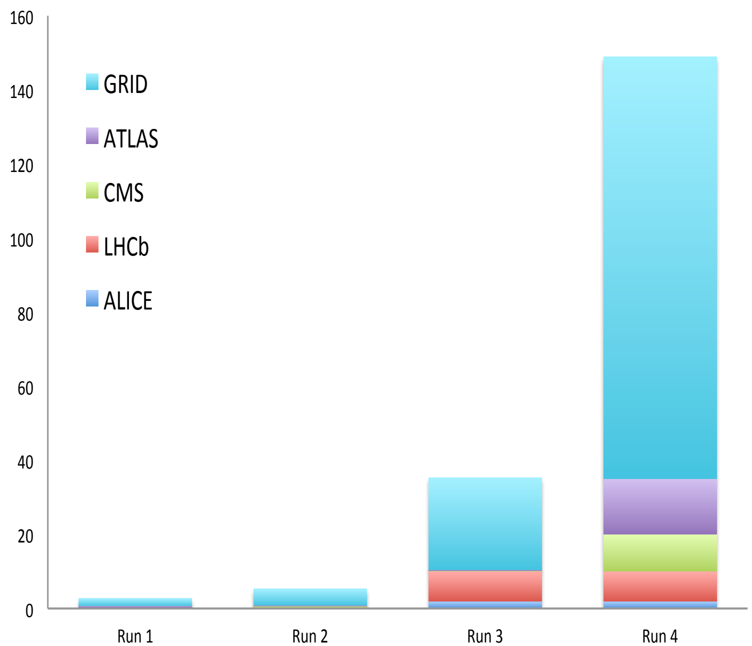


Modeling of the CMS HL-LHC computing system David Lange CHEP2019 04. Nov. 2019

- HL-LHC data analysis
  - To extract physics results requires to handle/analyze a lot more data!
  - Tests and the usage of new technologies are ongoing
    - Big Data technology (new toolkits and systems to support analysis of datasets)
    - Cloud Computing and High Performance Computing (HPC)
  - Educates our community to use industry-based technologies
  - Use tools developed in larger communities reaching outside of our field

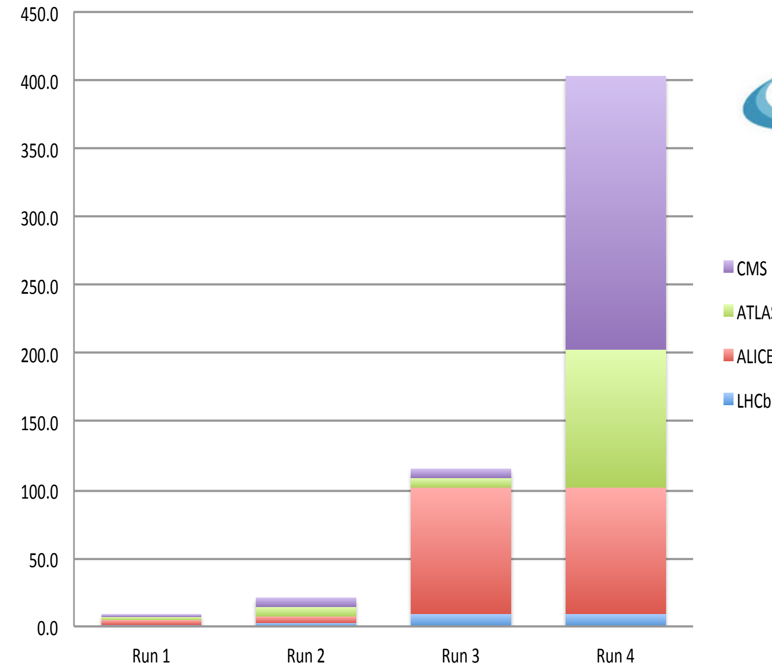
HL-LHC computing needs

Computing: Growth > x 60



2010 2015 2021 2025 (HL-LHC)

Data: ~50 PB/year → 600 PB/year, x 12

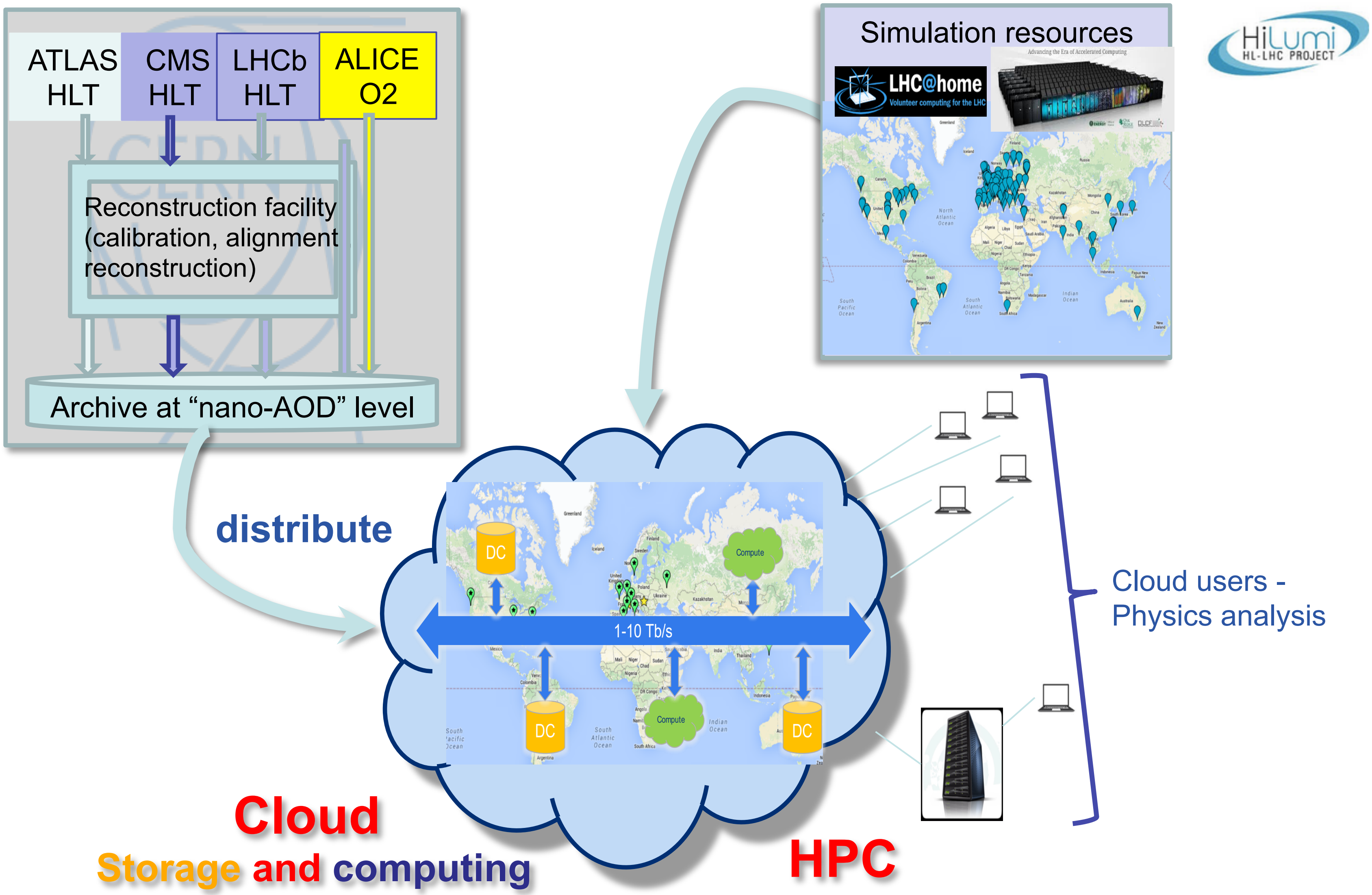


2010 2015 2021 2025 (HL-LHC)

Frédéric Hemmer ,CERN School of Computing, Oct. 2018

- Simple model based on today's computing models, but with operating parameters (pile-up, trigger rates, etc.)
- Technology at ~20%/year will bring x 6-10 in 10-11 years
- At least x 10 above is realistic to expect from technology with reasonable constant cost

HEP computing model for HL-LHC



Frédéric Hemmer, CERN School of Computing, Aug. 2016

➢ Classes of Resource Providers

GRID

Virtual Organizations (VOs)
➔ Pledges

Cloud

Community Clouds and Commercial Clouds
➔ Rented resources

HPC

Researcher granted access to High-Performance Computing installations
➔Exploit opportunistic resources

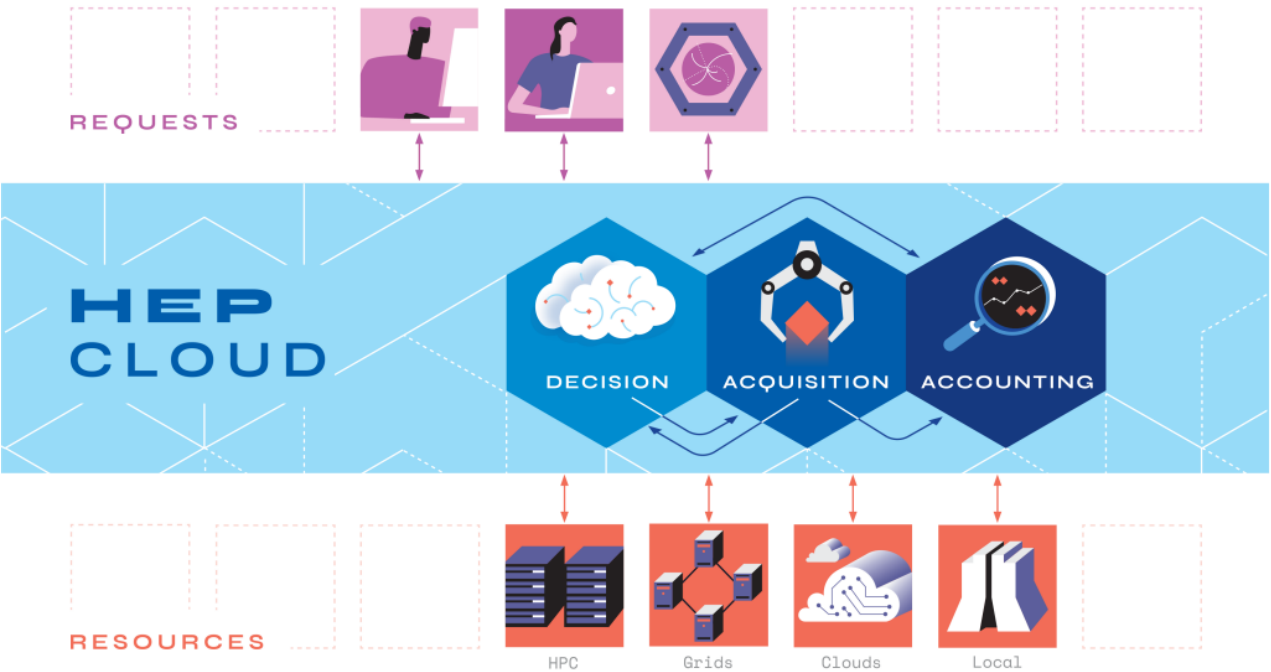
Helix Nebula Science Cloud

- European hybrid cloud platform that will support high-performance, data-intensive scientific computing
- sponsored by 10 of Europe's leading public research organizations and co-funded by the European Commission (H2020). Procurers: CERN, CNRS, DESY, EMBL-EBI, ESRF, IFAE, INFN, KIT, SURFSara, STFC
- for end-users from many research communities: High-energy physics, astronomy, life sciences,…
- Funds, manpower, use-cases with applications & data, in-house IT resources

Ian Bird, Helge Meinhard, CWP Workshop, Jan. 2017

HEPCloud @ Fermilab

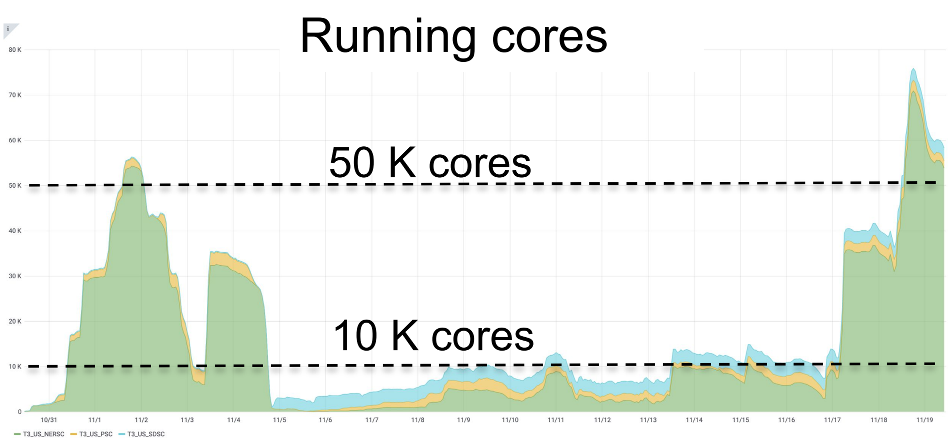
- routes jobs to local or remote computing resources accessing the various resources
- expands the resources available to include HPC centers and commercial cloud resources
- routes the jobs to the best resources available based on the requirements



https://hepcloud.fnal.gov/

Usage of opportunistic CPU resources @ CMS experiment

- Use them at the smallest cost possible for computing operations
- BEER (running jobs parasitically on resources for CERN IT services, EOS):
  - 1.7k additional cores in 2020
- HPCs: lots of effort for using such machines
  - CMS is doing its best to profit from HPCs, not possible to replace WLCG resources with HPCs



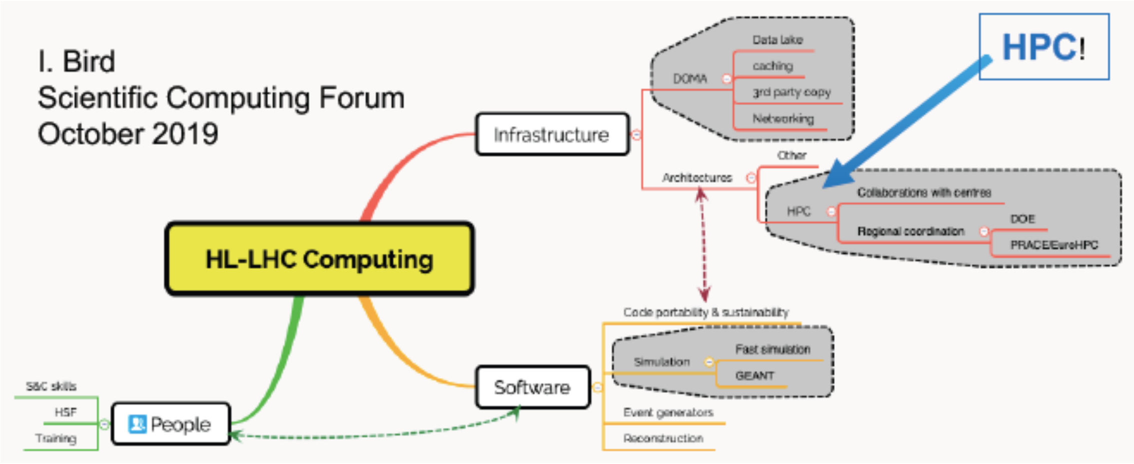
HPCs, CMS multithreaded jobs through HEPCloud

T3_US_NERSC T3_US_PSC T3_US_SDSC

Offline Software and Computing for Run3, M. Klute, D. Piparo, CMS Week Bangkok, 18. Dec. 2019

Motivation for use of HPC resources @ CMS experiment

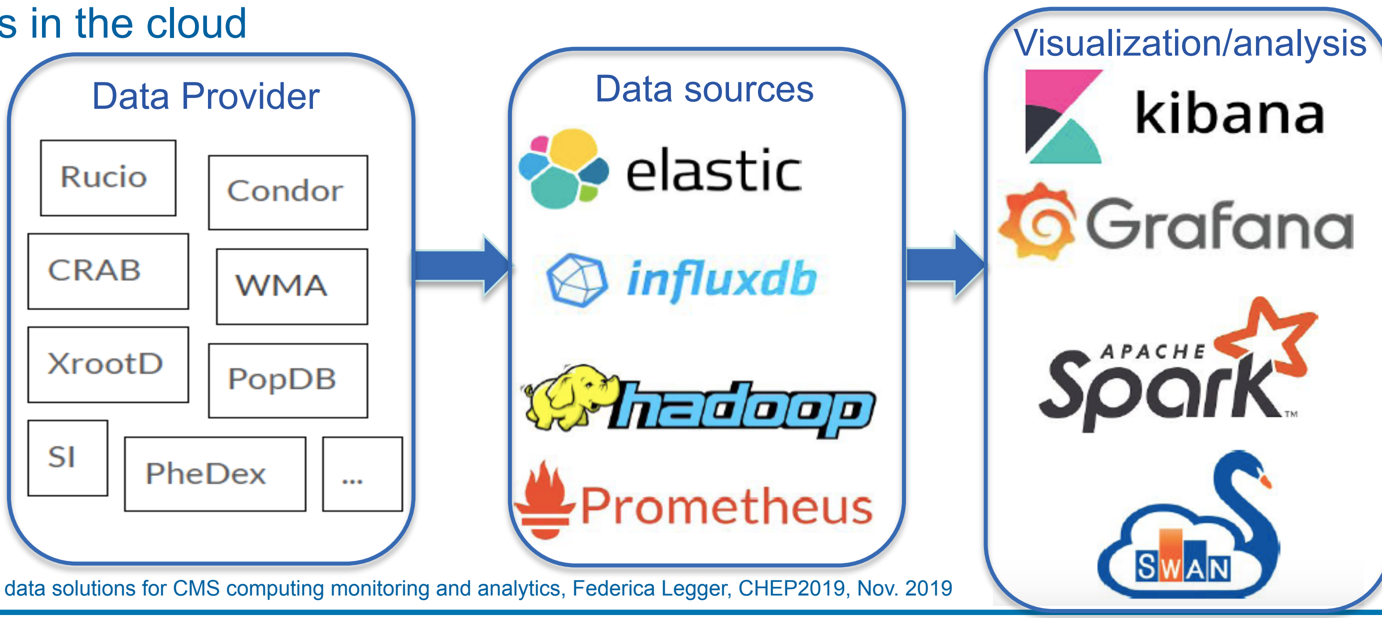
- CMS aims towards increasing the usage of HPC resources in the mid to long term future (Run-3 & HL-LHC)
- Growing funding in HPC infrastructures looking onwards to deploying Exascale machines (one exaFLOPS, a billion billion calculations per second)
- Countries/Funding agencies pushing HEP communities to make use of these resources
- Interest in HEP experiments to access best technologies available, usually employed at HPC sites
- HPC future contribution regarded as integral part of WLCG strategy towards HL-LHC



CMS Strategy for HPC resource exploitation, Antonio Perez-Calero Yzquierdo, CHEP2019, 04. Nov. 2019

Analytics and Monitoring Infrastructure @ CERN (Aka Monit)

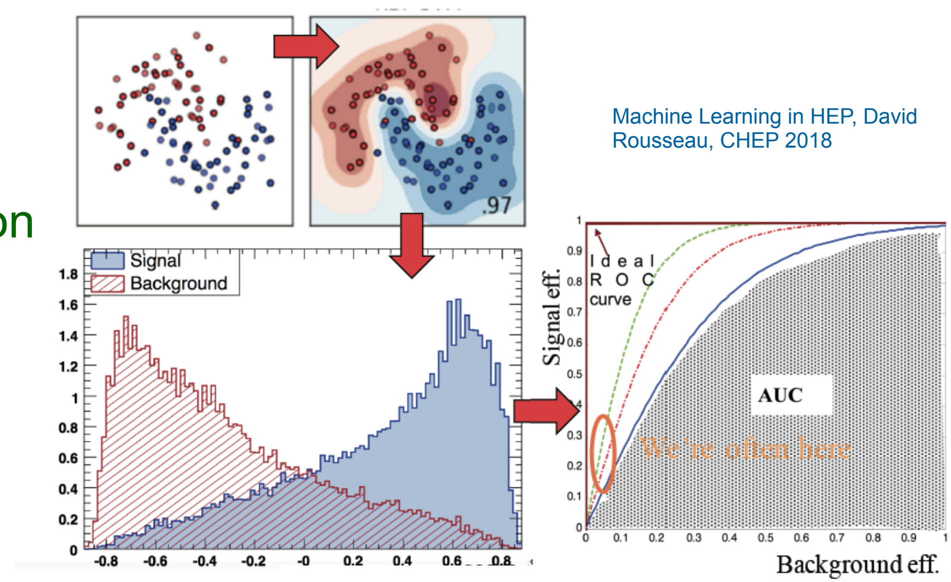
- common big data solutions based on open-source, scalable, and no-SQL tools, such as Hadoop, InfluxDB, and ElasticSearch, available through CERN IT infrastructure
- monitoring and accounting applications using visualisation tools - Kibana and Graphana
- alarms can be raised when anomalous conditions in the monitoring data are met, and the relevant teams are automatically notified
- data sources from different subsystems are used to build complex workflows and predictive analytics (data popularity, smart caching, transfer latency) and for performance studies
- exploiting scalable solutions based on Spark (unified analytics engine for large-scale data processing); SWAN (Service for Web based ANalysis) - platform to perform interactive data analysis in the cloud



Big data solutions for CMS computing monitoring and analytics, Federica Legger, CHEP2019, Nov. 2019

Machine Learning (ML) in HEP

- HEP (High Energy Physics) has a long history of using Machine Learning (since 2013)
- mostly used for signal vs. background classification
  - with Neural Networks and Boosted Decision Trees
  - train on signal and background Monte-Carlo
  - learn the separation between signal and background distribution
  - apply on test sample
  - apply on data
- used in many other fields
  - tools for object reconstruction
  - tagging (deep jet tagging)
  - fast simulation (calorimeter showers, TPC cluster)

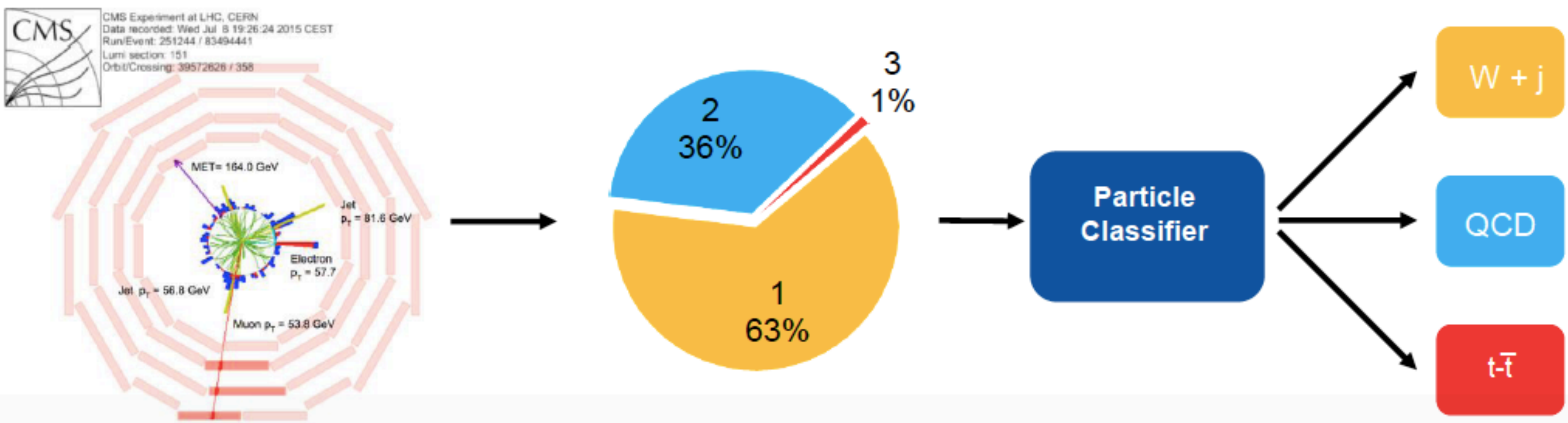


Machine Learning in HEP, David Rousseau, CHEP 2018

Particle Classifier Using Neural Networks

R&D to improve the quality of filtering system

- Develop a "Deep Learning classifier" to be used by the filtering system
- Goal: Identify events of interest for physics and reduce false positives
  - False positives have a cost - wasted storage, bandwidth and computing

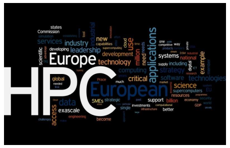


Big Data Tools and Pipelines for Machine Learning in HEP, Luca Canali, 04. Dec. 2019

Summary

- **Challenges on HL-LHC computing**
  - HEP computing much more capacity is needed
  - New computing models and more efficient software have to be developed

- **Additional resources are needed – Cloud Computing, High-Performance Computing**
  - Cloud resources are much more competitive in terms of cost than in the past



  - Increasing usage of HPC resources in the mid to long term future; usage of best technologies available
  - An important resource as supplement to the existing resources

- **Modern tools and methods are used - Big-Data, Machine Learning - Deep Learning**
  - Used in many ranges of application in HEP (monitoring, analysis optimization, particles classification)