Tackling Computing Challenges at CERN

Maria Girone
CERN openlab CTO

LHC

- 50-175m underground
- 27 km circumference tunnel
- Four giant experiments
- Particles travelling at 99.9999991% the speed of light
- 11245 turns every second

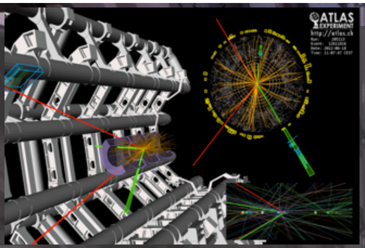
The most powerful discovery machine at CERN is the Large Hadron Collider

The LHC Data

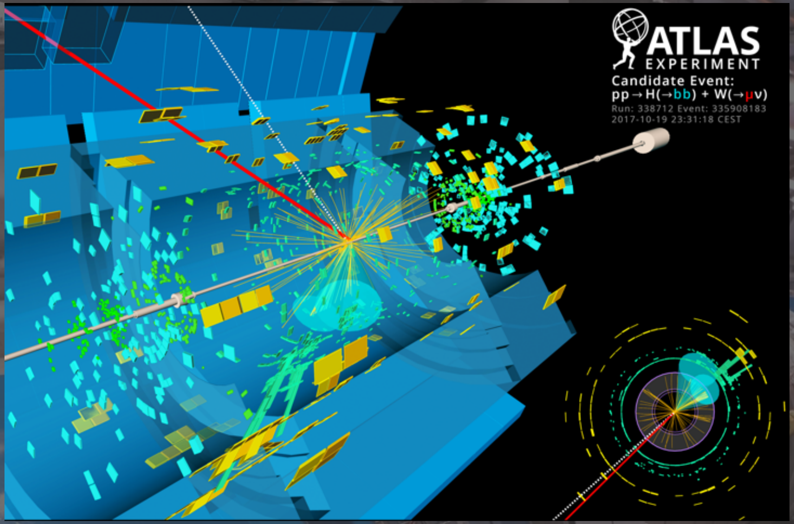
- Raw data:
  - Was a detector element hit?
  - How much energy?
  - What time?

- Reconstructed data:
  - Momentum of tracks (4-vectors)
  - Origin
  - Energy in clusters (jets)
  - Particle type
  - Calibration information
  - Analysis Objects
  - …

- 150 Million sensors deliver data …
  40 Million times per second



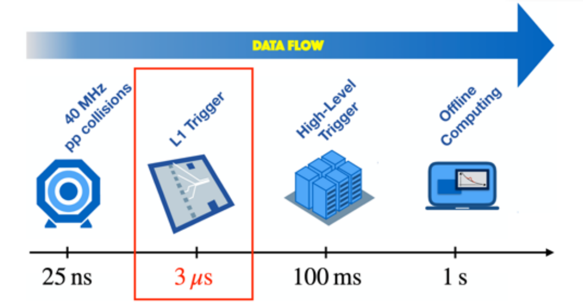
- Generates ~ 1 PB per second

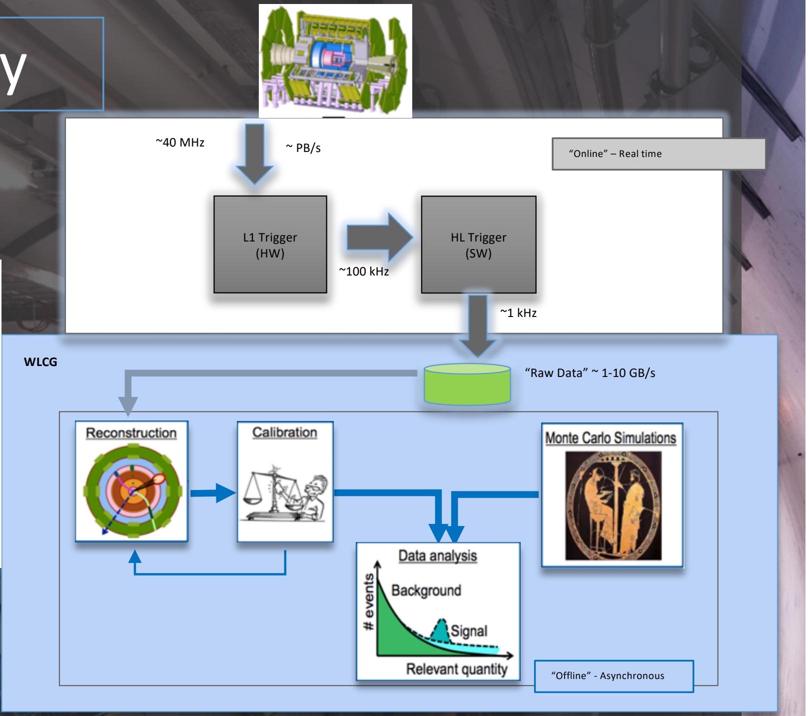


Data processing and analysis drive physics discovery

Software

- 50M lines dominated by C++ and Python
- Contributions by hundreds of scientists
- Much 20+ years old




~40 MHz

~ PB/s

"Online" – Real time

L1 Trigger (HW)

~100 kHz

HL Trigger (SW)

~1 kHz

WLCG

"Raw Data" ~ 1-10 GB/s

Reconstruction

Calibration

Monte Carlo Simulations

Data analysis

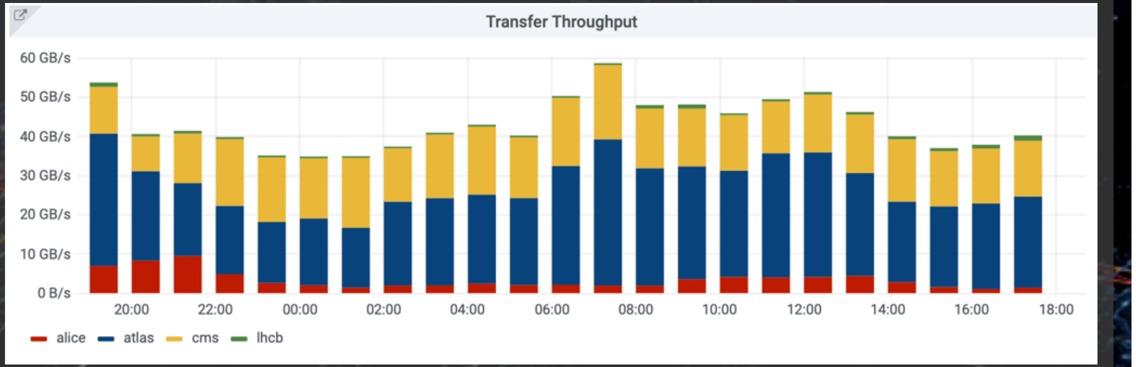
"Offline" - Asynchronous

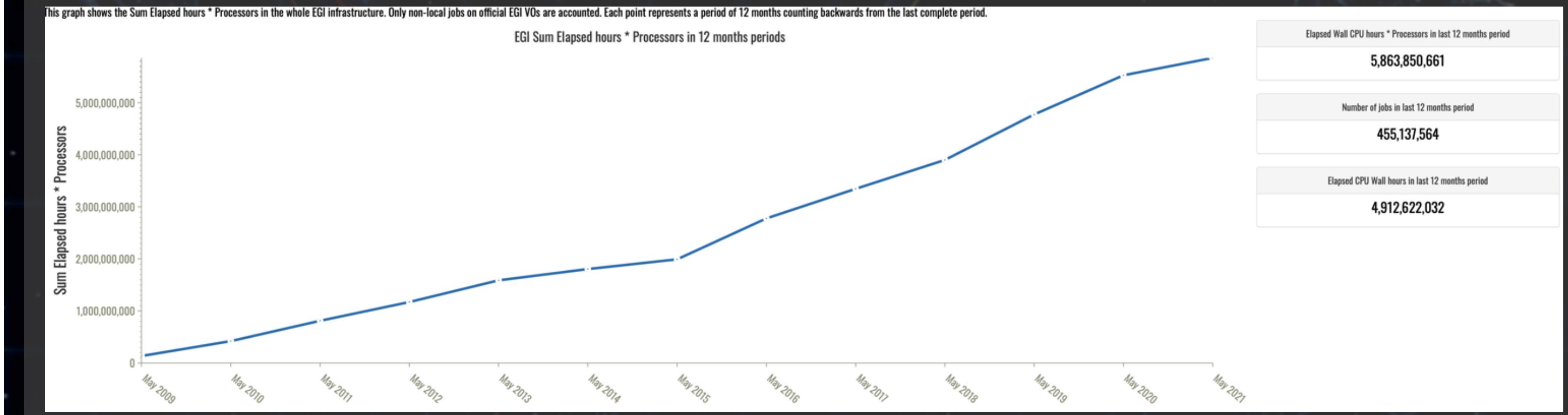
Dataflow at LHC

Maria Girone
CERN openlab CTO

Hardware

- Primary computing resources is the WLCG, a globally distributed storage and processing infrastructure
- 167 sites over 42 countries
- **~1M CPU cores** and **~1 exabyte** of storage (disk and tape)

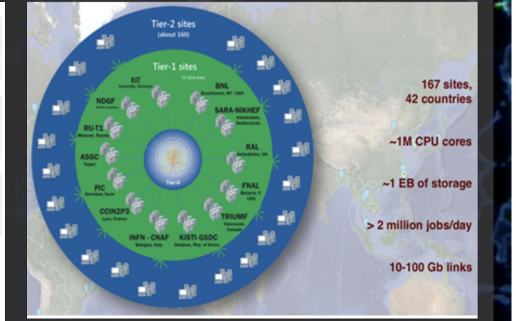



This graph shows the Sum Elapsed hours * Processors in the whole EGI infrastructure. Only non-local jobs on official EGI VOs are accounted. Each point represents a period of 12 months counting backwards from the last complete period.

Elapsed Wall CPU hours * Processors in last 12 months period
5,863,850,661

Number of jobs in last 12 months period
455,137,564

Elapsed CPU Wall hours in last 12 months period
4,912,622,032

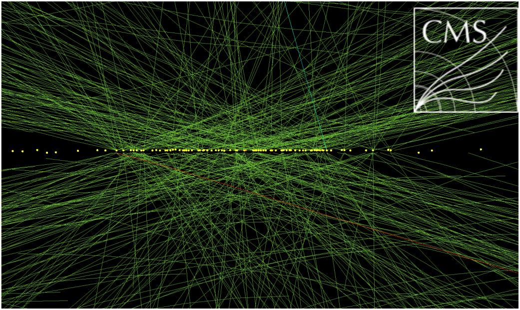


The Worldwide LHC Computing Grid

Maria Girone
CERN openlab CTO

The Challenges of HL-LHC

Scheduled Upgrades of LHC Program

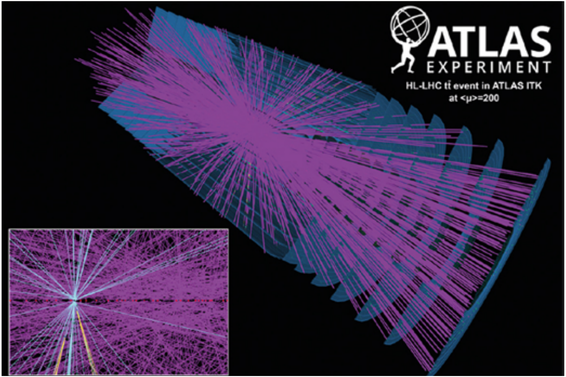


Run2 – Average 40 collisions per crossing

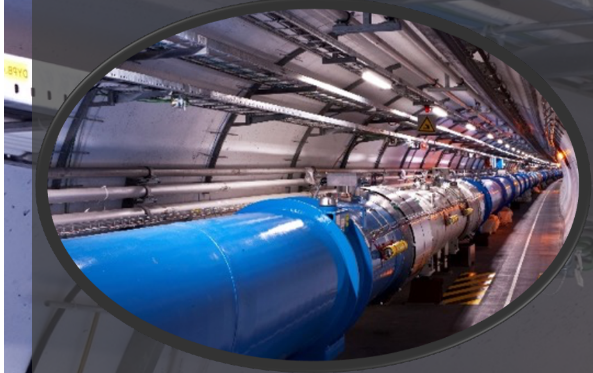
LHCb and ALICE will be upgraded for Run3 and will collect much more data

The LHC will be upgraded for Run4 to the High Luminosity LHC (HL-LHC). This will deliver:

- x10 increase in luminosity over LHC design
- great increase in event complexity
- more collisions and more complex data will result in a compute challenge at the Exascale level



Run4 – Average 200 collisions per crossing



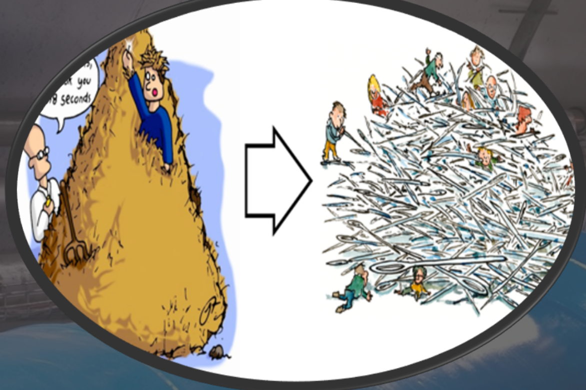
Upgraded Accelerator

- Higher Luminosity

Upgraded Detectors

- Higher Granularity
- Higher Occupancy

- Changing Filtering Paradigms
  - Higher Sensitivity
  - Higher Data Rates



New Computing Challenges

Upgraded Program = New Challenges

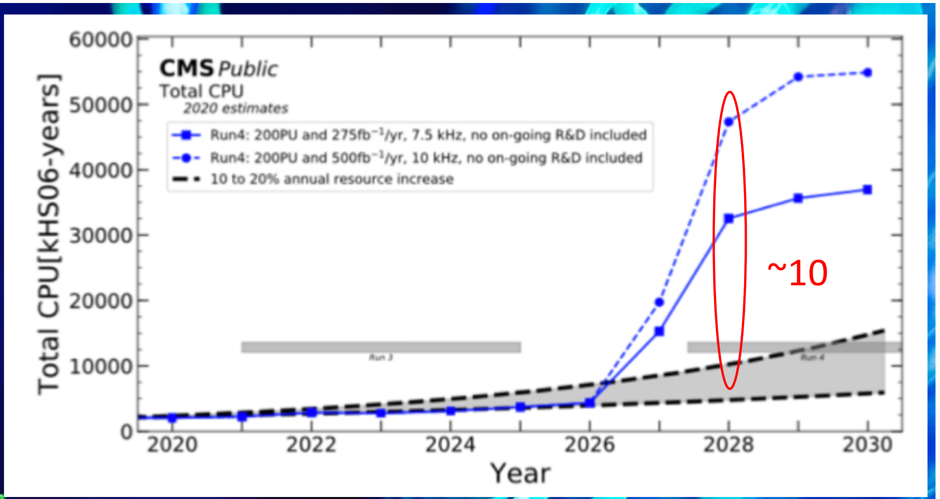
Computing Challenges

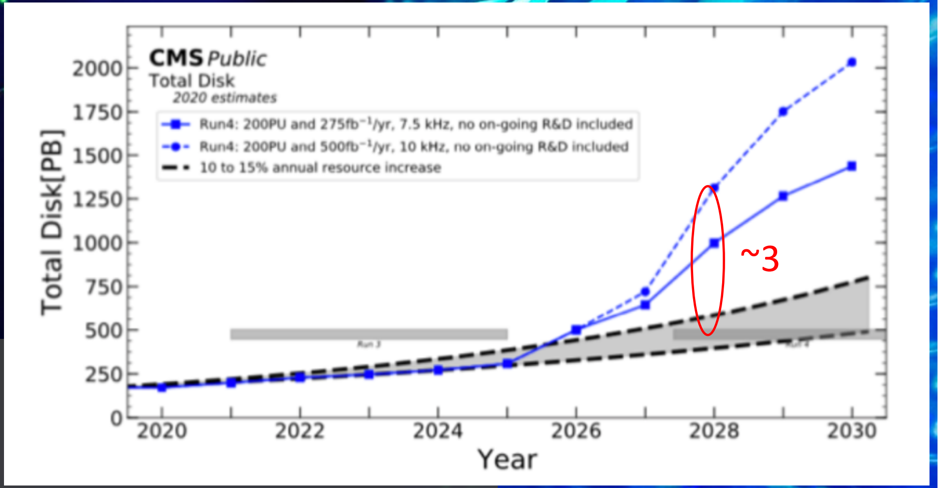
Gap between needed/available computing resources

- CMS estimate computing power needs to be ~6-10x higher
- CMS estimate disk needs to be ~3-5x times larger
- ATLAS estimates are similar

Investments for R&D in

- Code modernization and optimization
- Adapting code to hardware accelerators and HPC
- Reducing storage needs
- New techniques like AI and ML





General purpose CPU performance increases have slowed

Optimized heterogenous architectures have evolved faster, **HEP Is investing heavily in development to use new hardware resources**

- **GPUs** are the most common
- **FPGAs** currently used mostly in low latency applications
- **TPUs** and specialized ASICs are available

Accelerated Reconstruction And AI/ML

Low latency Online applications

		Accelerator				
		Intel	NVidia	AMD	FPGA	Other
CPU	Intel	Aurora	Cori Piz Daint Tsukuba Mare Nostrum		Tsukuba	
	AMD		Perlmutter JUWELS Booster	Frontier El Capitan LUMI		
	IBM		Summit Sierra Mare Nostrum			
	ARM		Wombat			Astra Fugaku

General Purpose X86 processing resources

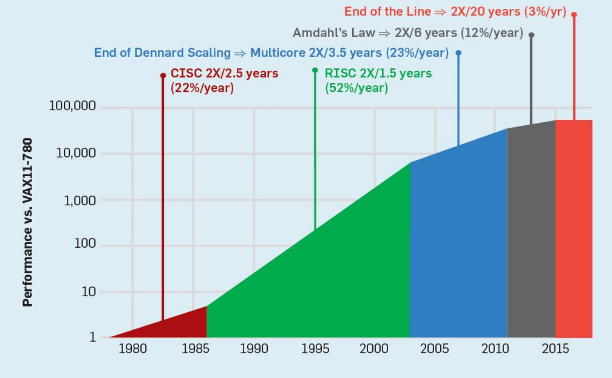
Code ported to Power

Low power highly parallelized

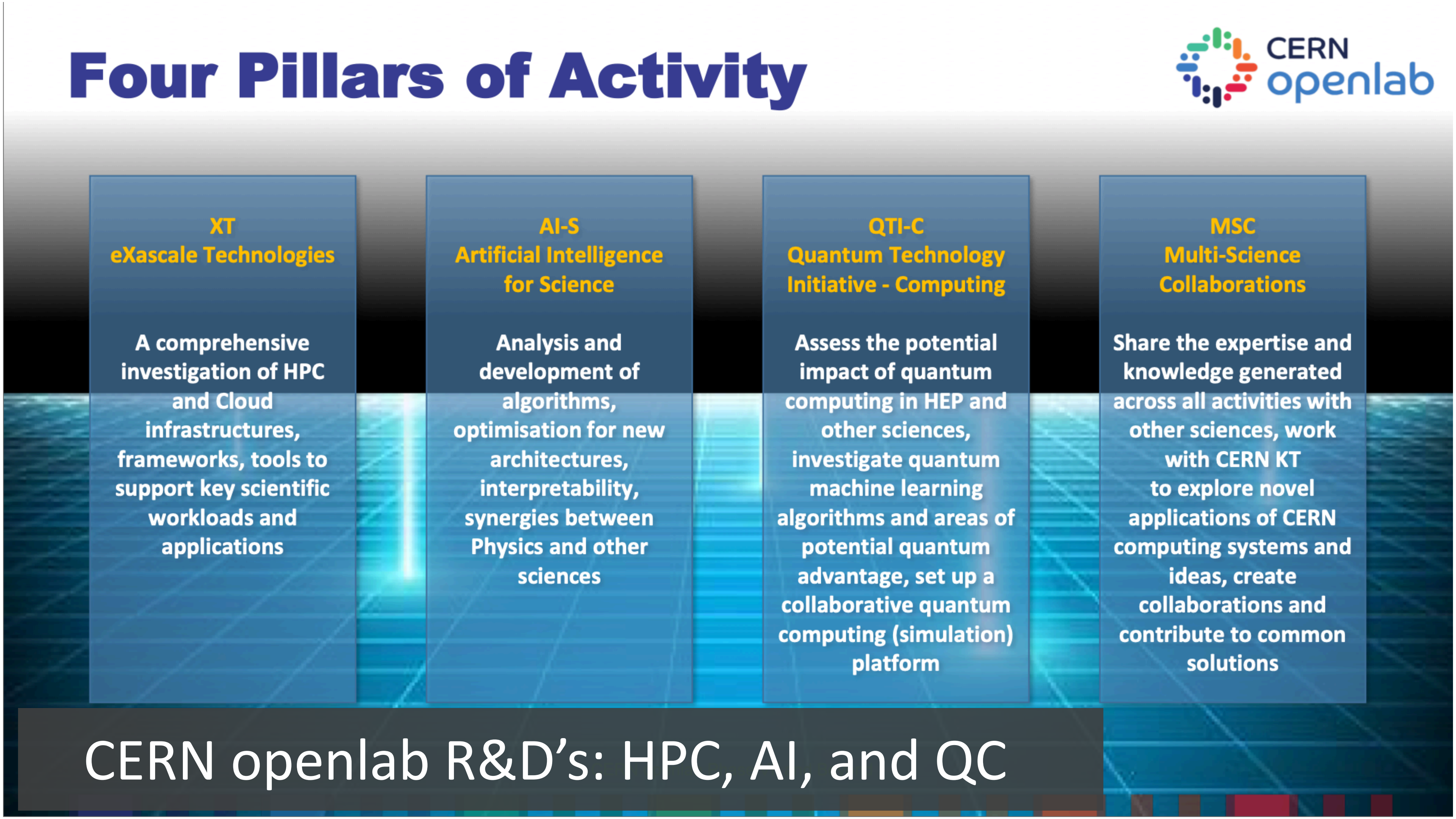
Amazon
Graviton2
Google Cloud TPU
Microsoft Azure
Intel DevCloud

HEP and the Computing Landscape

https://cacm.acm.org/magazines/2019/2/234352-a-new-golden-age-for-computer-architecture/fulltext



CERN openlab R&D’s: HPC, AI, and QC



Four Pillars of Activity

XT
eXascale Technologies

A comprehensive investigation of HPC and Cloud infrastructures, frameworks, tools to support key scientific workloads and applications

AI-S
Artificial Intelligence for Science

Analysis and development of algorithms, optimisation for new architectures, interpretability, synergies between Physics and other sciences

QTI-C
Quantum Technology Initiative - Computing

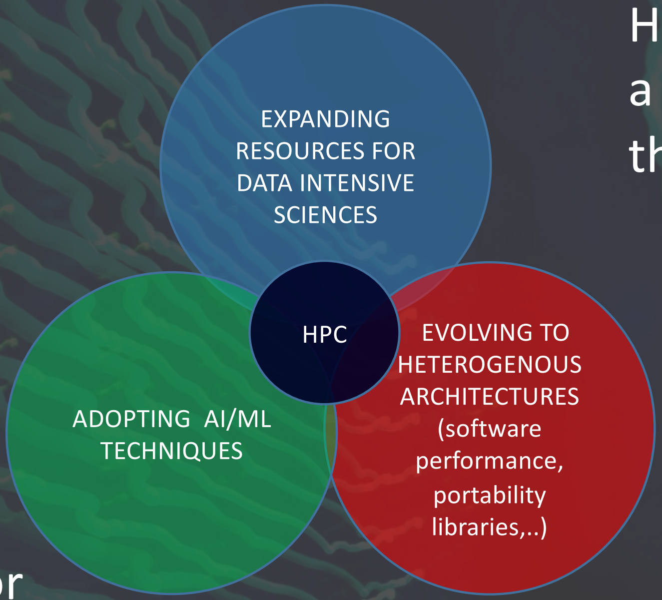
Assess the potential impact of quantum computing in HEP and other sciences, investigate quantum machine learning algorithms and areas of potential quantum advantage, set up a collaborative quantum computing (simulation) platform

MSC
Multi-Science Collaborations

Share the expertise and knowledge generated across all activities with other sciences, work with CERN KT to explore novel applications of CERN computing systems and ideas, create collaborations and contribute to common solutions

High Performance Computing

HPC falls at the intersection of several important R&D areas



EXPANDING RESOURCES FOR DATA INTENSIVE SCIENCES

HPC

ADOPTING AI/ML TECHNIQUES

EVOLVING TO HETEROGENOUS ARCHITECTURES (software performance, portability libraries,..)

Engagement with the HPC Community can be a catalyst for progress

HPC Supercomputers will grow by a factor of 10 on the time scale of the HL-LHC

A thorough R&D program has been established

Unified programming models facilitate HPC adoption

Working Together: The HPC Collaboration

- **An HPC Collaboration agreement was signed by CERN, SKAO, GÉANT and PRACE, CERN and SKA on 22.07.2020**
  - Engages at the community level
    - Bringing together data intensive sciences, high-performance computing infrastructures and networking
- Collaboration built around 4 pillars
    - Building a common centre with expertise to support heterogenous hardware
    - Benchmarking Demonstrator
    - Data Access Demonstrator
    - Authentication and Authorization Demonstrator

An Exascale project for an Exascale problem

High Performance Computing (HPC) Collaboration Agreement (CA)

This CA is made on 22 July 2020

between

PRACE, GÉANT, CERN and SKA

Each a "Party" and together the "Parties"

1. DEFINITIONS

Background IPR	means any IPR controlled or owned by any Party prior to the entry into force of this CA or IPR generated by any of the Parties independently of this CA and controlled or owned by that Party;
Confidential Information	has the meaning given to it in Section 7 of this CA;
Creating Party	has the meaning given to it in Section 8 of this CA;
Foreground IPR	means any IPR arising from tasks as described in the Work Plan and carried out in the course of this CA by any of the Parties;
IPR	means any and all intellectual property rights anywhere in the world whether registered, are registerable or otherwise, including but not limited to patents, trademarks, registered designs, domain names, applications for any of the foregoing, trade or business names, copyright and rights in the nature of copyright, design rights, rights in databases, moral rights and know-how;
Party in Default	has the meaning given to it in Section 14 of this CA;
Results	means all information, know-how, results, inventions, software and IPR identified or first reduced to practice or writing in the course of the tasks as described in the Work Plan;
Work Plan	means the tasks to be performed during the course of this CA, as annexed thereto.

1

Eckhard Elsen (top left), Director for Research and Computing at CERN; Philip Diamond (top right), SKA Director-General; Erik Huizer (bottom left), Chief Executive Officer of GÉANT; and Philippe Lavocat (bottom right), PRACE Council Vice-Chair, signed the agreement for the new collaboration.

https://home.cern/news/news/computing/cern-skao-geant-and-prace-collaborate-high-performance-computing

CERN, SKAO, GÉANT and PRACE to collaborate on high-performance computing

The next generation of high-performance computers holds significant promise for both particle physics and astronomy but key challenges remain to be addressed

22 JULY, 2020 | By Andrew Purcell

Collaborations

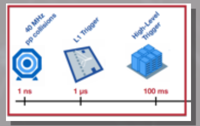
Maria Girone, CERN openlab Technical Workshop 2021

Participation on the path to exascale

Proven CERN capability

Use case specific

Fast ML

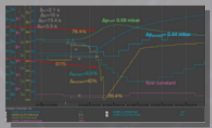


Ultra-fast on-edge inference under strict latency constraints

Anomaly detection

Object identification, classification, anomaly detection in big and noisy data sets

Industrial controls

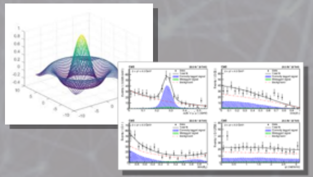


Machine efficiency and predictive maintenance with industrial control systems

Distributed computing

Optimization of distributed computing, storage, and networks; fast I/O for large files

Large scale, science grade data analytics and visualization

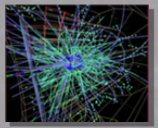


Cross use case

- Optimization and evaluation for science-grade precision of large data sets using advanced data analytics
- Data visualization, interactive plotting (e.g., statistical visualizations, uncertainties, distributions), model visualization
- Large-scale, quality-controlled CERN data as testbed/ benchmark (e.g., single data set with 100m examples, >1TB)

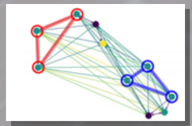
In development, opportunity for joint R&D

Simulation

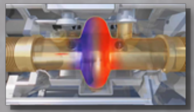


Simulation and reconstruction with generative DL for efficient computation

Graphs



Exploring Graph NNs for high-multiplicity problems with non-linear distances

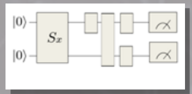


Determining optimal machine design and component configuration

ML in Robotics

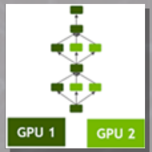
Remote maintenance and safety with autonomous robots and computer vision

Quantum ML



Research quantum algorithms to solve pattern recognition, classification and generation problems

Computing parallelization

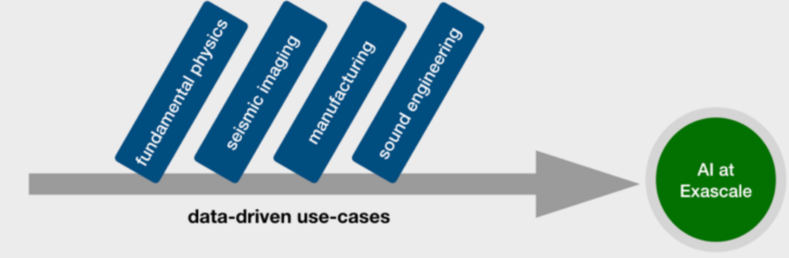


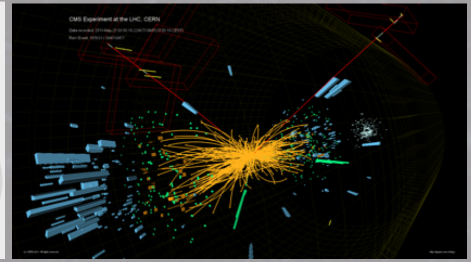
Training and optimization of complex NNs on parallelized GPU infrastructure

Progress on AI/ML Capabilities

Launched in January the RAISE Center of Excellence enabled researchers from science and industry to develop novel, scalable Artificial Intelligence technologies towards Exascale along representative use-cases from Engineering and Natural Sciences

CERN is leading the leading the data driven use-cases



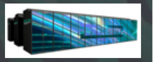


AI/ML Projects

RAISE Center of Excellence

HEP, HPC, and Commercial Clouds

Challenges

Software and Architectures		Supercomputers are early adopters of heterogenous architectures
Benchmarking and Accounting		Performance on diverse architectures needs to be understood
Data Processing and Access		Enormous data volumes to stage, process, and export
Authorization and Authentication		Strict cyber security
Runtime Environments and Containers		Resources are shared, environment needs to be brought with the workload
Provisioning		Resources allocated for periods of time through allocations
Wide and Local Area Networking		Processing and storage resources are separate

Challenges in HPC Integration

The common challenges for HPC integration into LHC Computer were described in an engagement document

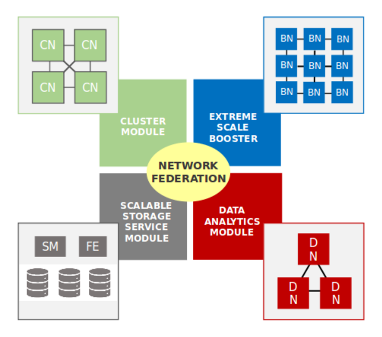
https://zenodo.org/record/3647548#.YBnA1y2cbVs

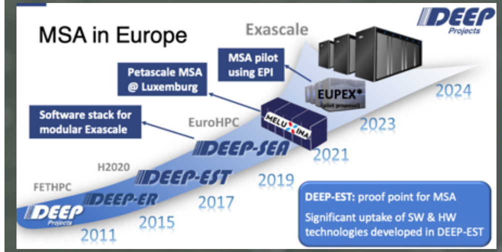
Optimising HEP applications towards Exascale

Develop an energy-efficient system architecture that fits HPC and HPDA workloads

Build a fully working Modular Supercomputer Architecture prototype to Exascale

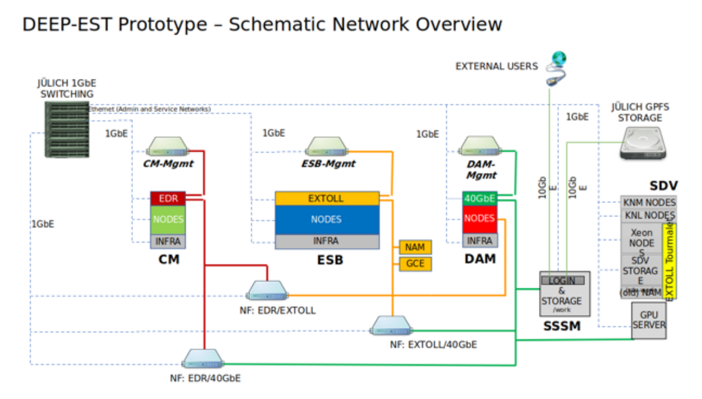
Large variety of hardware available supporting the different requirements of HPC, Big Data Analytics and Machine Learning with highest efficiency and scalability





DEEP-EST: proof point for MSA

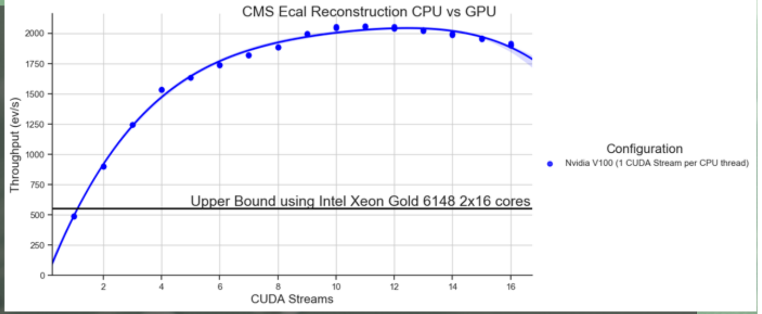
Significant uptake of SW & HW technologies developed in DEEP-EST

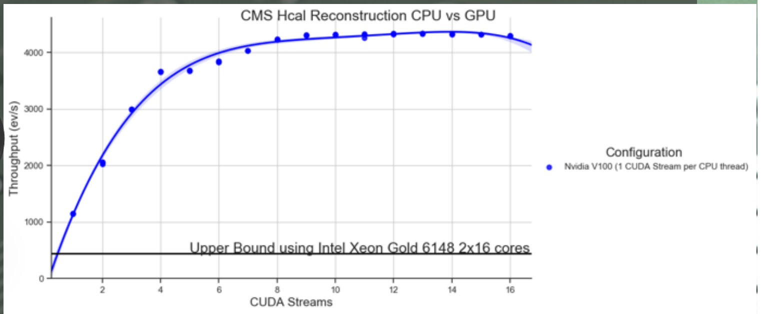


Progress Using Heterogenous Architectures

Within the DEEP-EST project, re-engineered CMS ECAL and HCAL local reconstruction workloads to use GPUs using CUDA

- Achieving between 3x(ECAL) and 8x(HCAL) using Nvidia V100 vs filling in 2-socket Intel Xeon Gold 6148
- Now integrated with CMS framework and will be used in the CMS HLT reconstruction for Run3





Results on Open Data:
http://opendata.cern.ch/record/12303

Performance studies on the HEP MC generator code on GPUs using CUDA

- The idea is to enable utilization of heterogenous architectures for MC generation as well

NTNU
HPC-Lab

We are investigating **unified programming models** to create sustainable code that can be supported on multiple architectures

Maria Girone
CERN openlab CTO

How do we bring large datasets to supercomputers

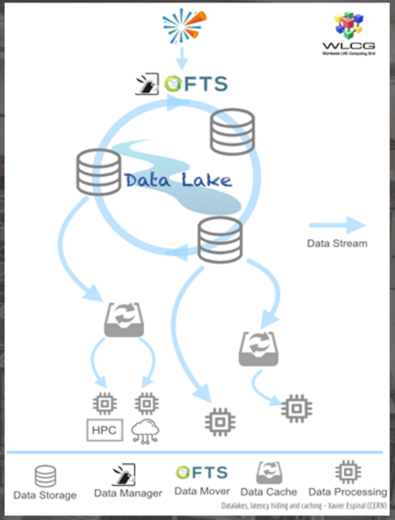
WLCG *Data Lake* model separates storage and processing functionality. HPC will be a part of the *Data Lake* model

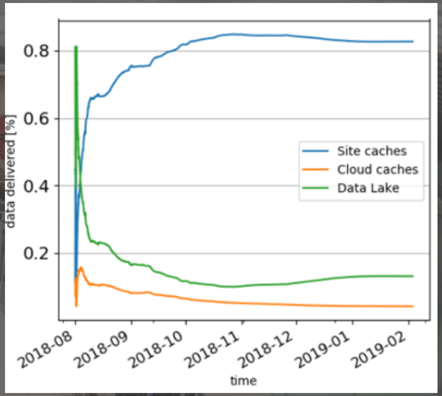
- Relies on caching and networking
- EuroHPC will have significant WAN connectivity and disk space

Technical Activities

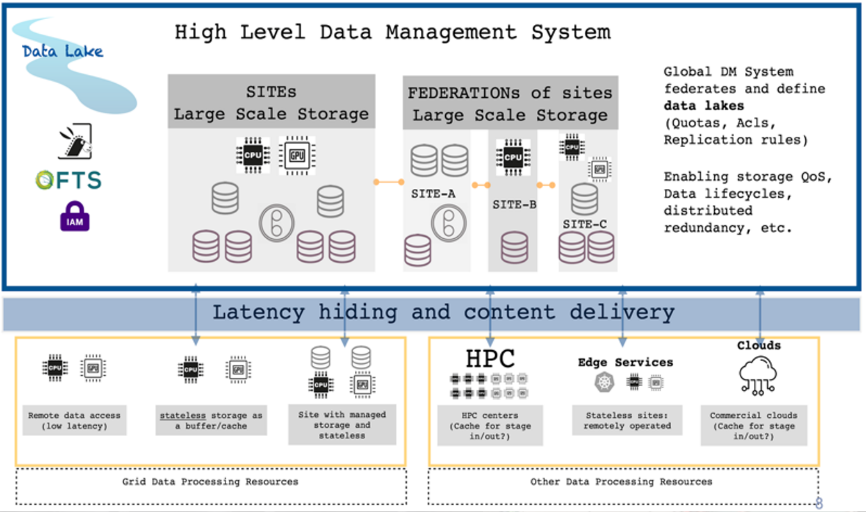
Execute a series of data challenges to demonstrate the feasibility of the *Data Lake* model on a path to Exascale

Data Access – *Data Lakes*





Emulation of cache delivery vs. time including regional caches



HEP Data in HPC

HEP data is primarily stored as files, optimized for highly parallel HTC

- **ROOT** is the HEP analysis framework
- **ROOT** defines columnar data layout tailored for HEP: extreme throughput compared to alternatives
- https://root.cern

ROOT Challenges

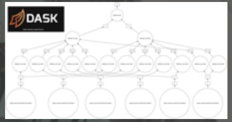
- Maximize throughput I/O and optimize for HPC
- Optimize **persistent** data layout to facilitate conversion for CPU, GPU, SIMD (LLAMA), read patterns, and storage backend

Ongoing R&D, bringing >4GB/s from off-the-shelf desktop to HPC

ROOT team bringing heterogenous computing and environments to physicists

- **Declarative programming:** physicists define data + analysis flow; "kernel graph" built behind the scene with runtime-detected input data types
- **Transparent acceleration:** algorithms (modelling / minimization) in multi-arch libraries selected at runtime, covering architectures' SIMD to GPU
  - Optimal abstraction? Autovec + CUDA / std::simd / alpaka SYCL
- **Enabling feature:** C++ just-in-time compilation (cling); also supports runtime-CUDA. Use of C++ automatic differentiation (clad) for minimization
- **Scaling:** multi-threaded (>200 cores), distributed backends (dask / spark /...)

APACHE Spark



Courtesy of A. Naumann

ROOT
Data Analysis Framework

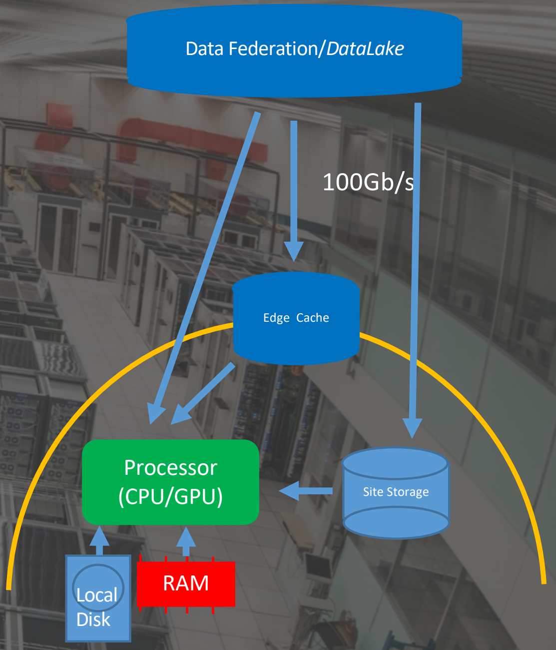
Studying Data Access

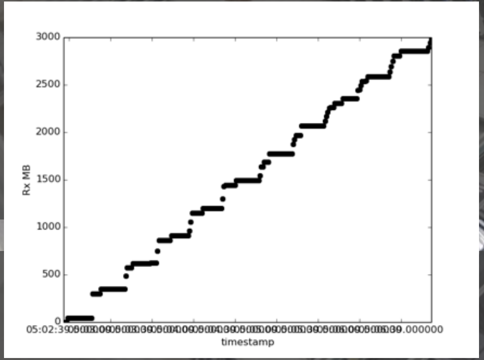
Establishing a data access framework

- Multiple containerized applications with different IO profiles
- An aggregator to analyse, study, and optimize facility and workflow deployments

Studying workflow performance progressively farther from the processing resources

- Goal is a series of data processing challenges that will eventually reach 10PB a day





For Exascale HPC O(1M) cores
-> O(10K) nodes -> O(150) GB/s

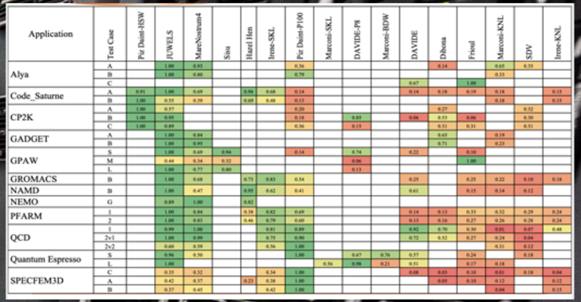
Courtesy of V. Khristenko

Benchmarking Demonstrator

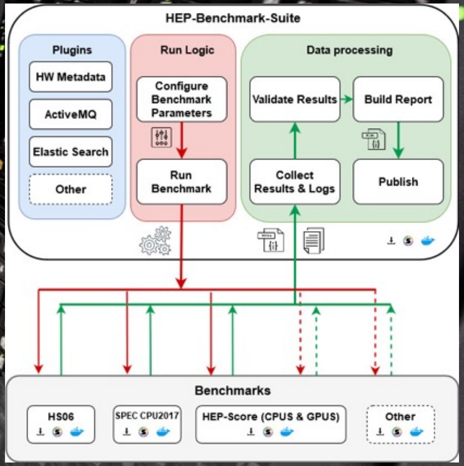
Benchmarking Activities

- PRACE-CERN-GÉANT-SKAO collaboration brings opportunity to expand capabilities using tools already developed for HPC sites by each community:
  - Unified European Applications Benchmark Suite (UEABS)- 13 workloads for HPC
- CERN is evolving the approach to benchmarking in HEP to embrace HPC:
  - Builds on experience from WLCG computing environment tools
  - Developed with secure, self-contained workload images (Singularity)
  - Assumes no privileges, no docker, limited/restricted node connectivity

PRACE
Unified European Applications Benchmarking Suite



Application	Test Case	Piz Daint-HSW	JUWELS	MareNostrum4	Sisu	Hazel Hen	Irene-SKL	Piz Daint-P100	Marconi-SKL	DAVIDE-P8	Marconi-BDW	DAVIDE	Dibona	Frioul	Marconi-KNL	SDV	Irene-KNL
Alya	A		1.00	0.93				0.36					0.14		0.65	0.35	
	B		1.00	0.80				0.79							0.33		
	C											0.67		1.00			
Code_Saturne	A	0.91	1.00	0.69		0.96	0.68	0.14				0.14	0.18	0.19	0.18		0.15
	B	1.00	0.55	0.39		0.69	0.48	0.13							0.18		0.15
CP2K	A	1.00	0.97					0.20					0.27			0.32	
	B	1.00	0.95					0.18		0.85		0.06	0.53	0.06		0.30	
	C	1.00	0.89					0.36		0.15			0.31	0.31		0.31	
GADGET	A		1.00	0.84									0.65		0.19		
	B		1.00	0.95									0.71		0.23		
GPAW	S		1.00	0.69	0.94			0.14		0.74		0.22		0.10			
	M		0.44	0.34	0.32					0.06				1.00			
	L		1.00	0.77	0.80					0.13							
GROMACS	B		1.00	0.68		0.73	0.83	0.54				0.25		0.25	0.22	0.10	0.18
NAMD	B		1.00	0.47		0.95	0.62	0.41				0.61		0.15	0.14	0.12	
NEMO	G		0.89	1.00		0.82											
PFARM	1		1.00	0.84		0.38	0.82	0.69				0.14	0.13	0.33	0.32	0.29	0.24
	2		1.00	0.83		0.46	0.79	0.60				0.13	0.16	0.27	0.26	0.28	0.24
QCD	1		0.99	1.00			0.81	0.89				0.92	0.70	0.30	0.01	0.07	0.48
	2x1		1.00	0.99			0.75	0.90				0.72	0.52	0.27	0.24	0.04	
	2x2		0.60	0.59			0.56	1.00							0.31	0.12	
Quantum Espresso	S		0.96	0.50				1.00		0.67	0.76	0.57		0.24		0.18	
	L		1.00						0.56	0.98	0.21	0.51		0.17	0.18		
SPECFEM3D	C		0.35	0.32			0.34	1.00				0.08	0.03	0.10	0.01	0.18	0.04
	A		0.42	0.37		0.23	0.38	1.00					0.05	0.10	0.12		0.12
	B		0.37	0.45			0.42	1.00							0.04		0.15

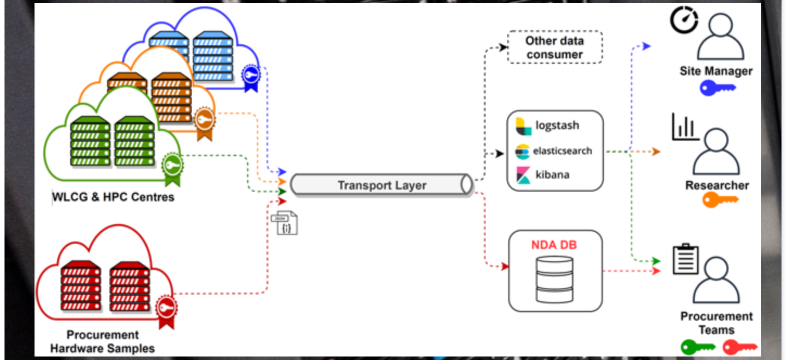


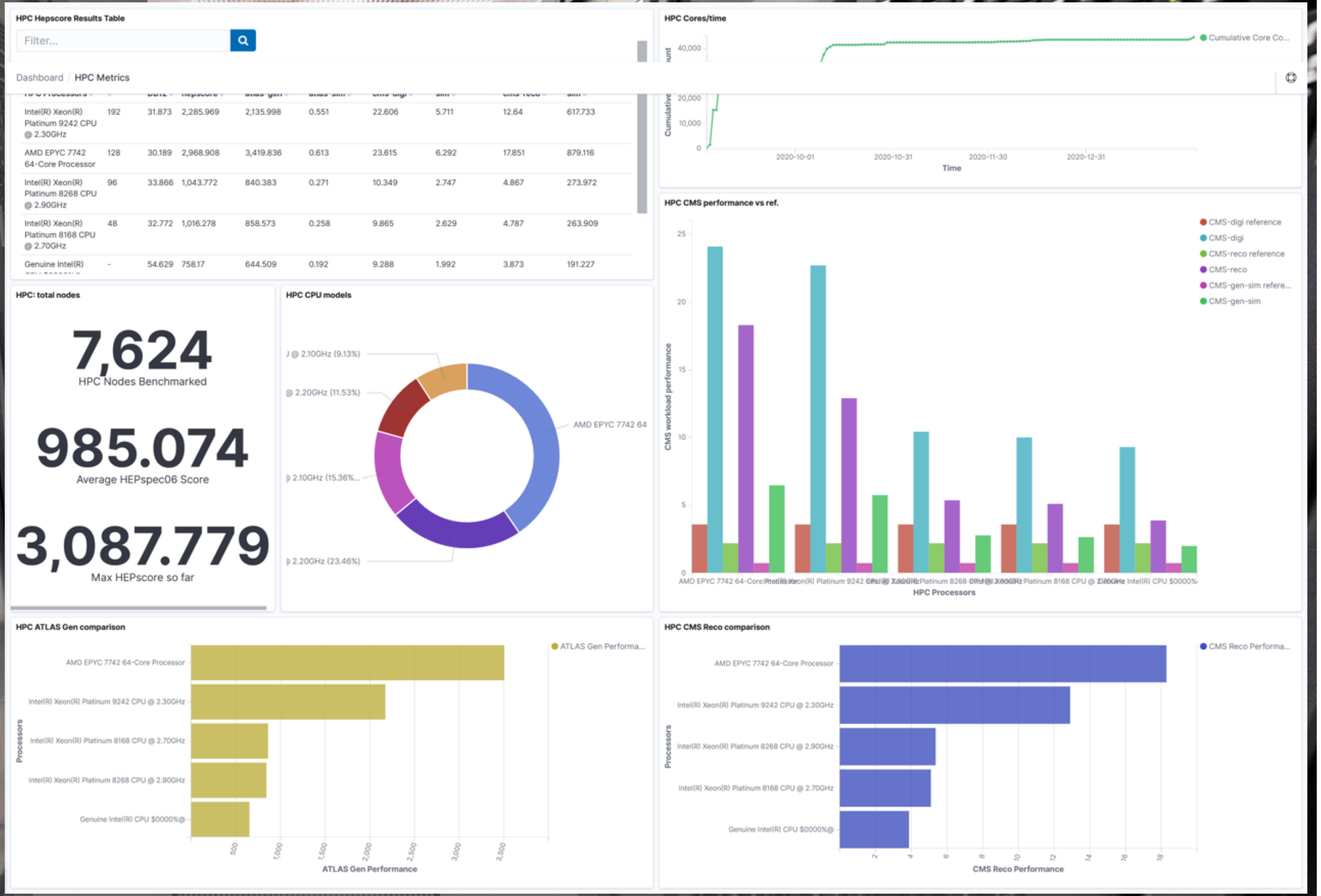
HEP Workflows on HPC

Benchmarking Heterogenous architectures

- Multi-architecture as workflows become available (ARM, IBM Power)
- GPU accelerators (NVIDIA, AMD)

Automated collection and aggregation





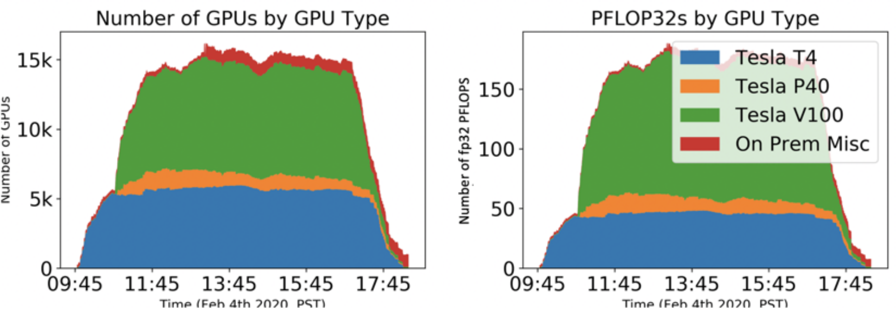
Intel(R) Xeon(R) Platinum 9242 CPU @ 2.30GHz	192	31.873	2,285.969	2,135.998	0.551	22.606	5.711	12.64	617.733
AMD EPYC 7742 64-Core Processor	128	30.189	2,968.908	3,419.836	0.613	23.615	6.292	17.851	879.116
Intel(R) Xeon(R) Platinum 8268 CPU @ 2.90GHz	96	33.866	1,043.772	840.383	0.271	10.349	2.747	4.867	273.972
Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz	48	32.772	1,016.278	858.573	0.258	9.865	2.629	4.787	263.909
Genuine Intel(R)	-	54.629	758.17	644.509	0.192	9.288	1.992	3.873	191.227

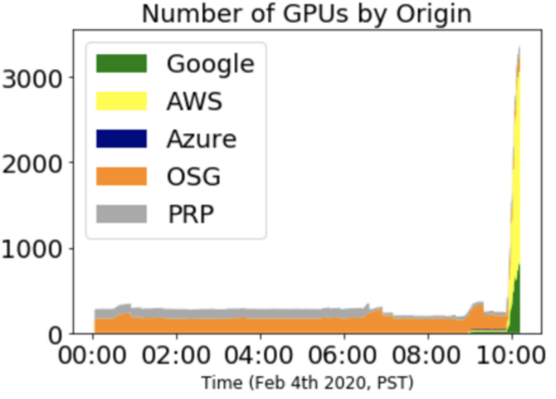
https://gitlab.cern.ch/hep-benchmarks

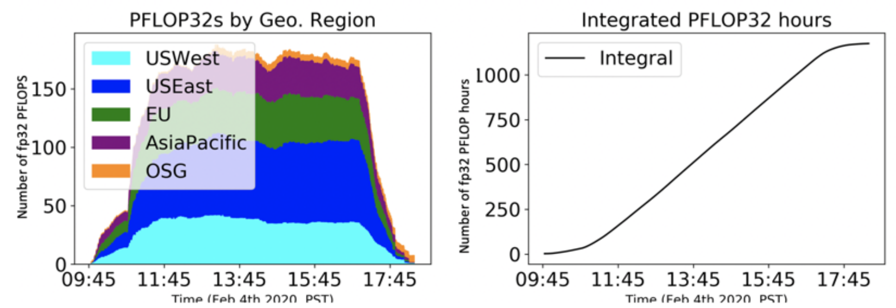
Courtesy of D. Southwick

IceCube with OSG and UCSD have been producing simulation with GPUs using the major commercial cloud providers

- Over the 9h test reached 1 exaflop/hour delivered
  - 150PF/h





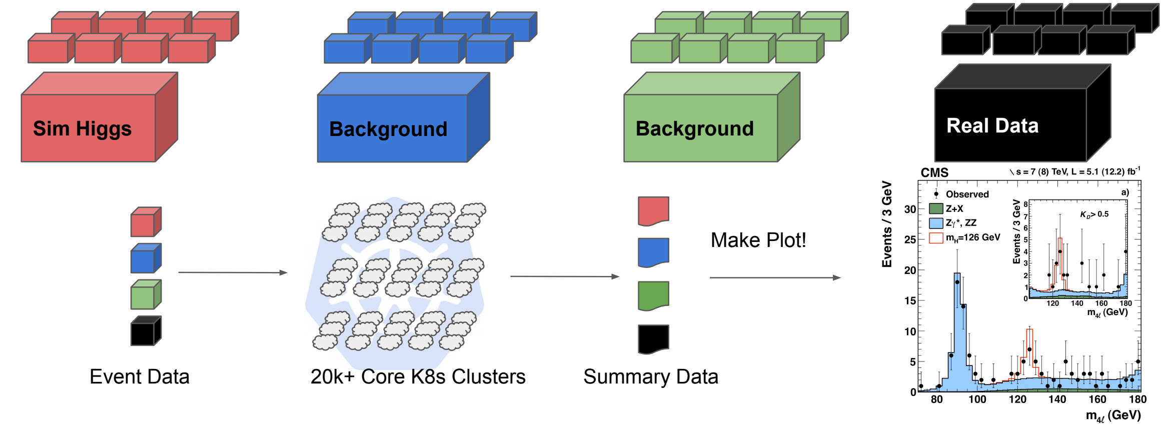


Simulation on Commercial Clouds

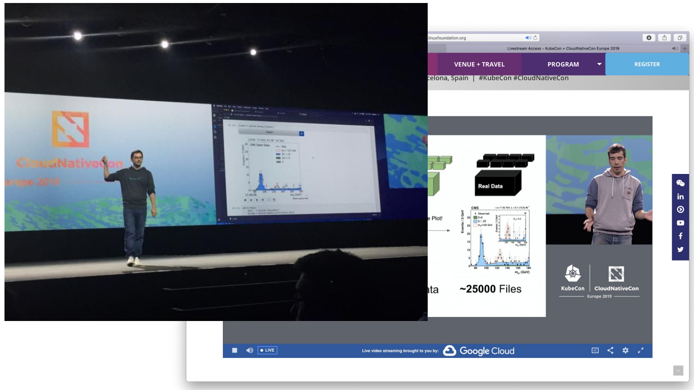
https://www.linkedin.com/pulse/cost-effective-exaflop-hour-clouds-icecube-igor-sfiligoi

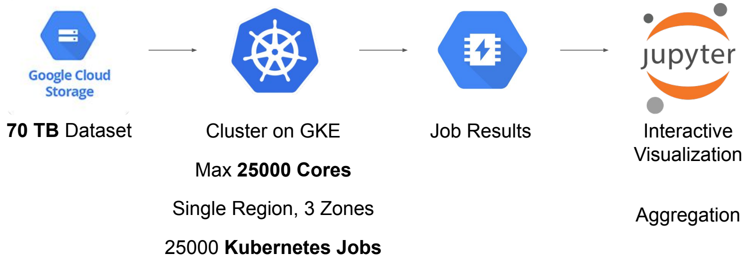
Demonstration of cloud analysis access was shown on CMS open data during CHEP, Adelaide 2019

- Analyzing 70TB of data and generates the Higgs discovery plot in about 20 minutes



70 TB of Physics Data **~25000** Files





70 TB Dataset

Cluster on GKE

Max **25000 Cores**

Single Region, 3 Zones

25000 **Kubernetes Jobs**

Job Results

Interactive Visualization

Aggregation

Cloud Use for Analysis

https://www.youtube.com/watch?v=CTfp2woVEkA

KIT Site extension

1k

ForHLR2

T3

Experiment Activities in HPC and Clouds

Experiment Use of HPC

ATLAS Computing Model (CM) designed to use distributed computing centers. CM based on three main pillars : Data Management (Rucio), Workload Management (PanDA) and monitoring

- ATLAS decided to add HPCs and to integrate High Throughput Computing (Grid) with HPC. HPCs integrated into the production, analysis and data management systems (also to monitoring and accounting) in 2016

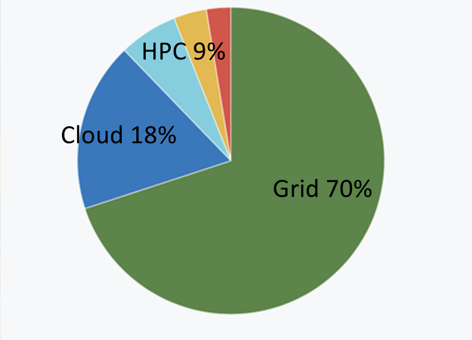
- Over the past 7 years, the ATLAS experiment collaborated with many large HPC sites for full integration into ATLAS distributed computing

CMS continuously invests effort to build up expertise on HPC resources integration

- Given the unicity of the HPC Machines multiple approaches have been successfully commissioned:
  - HEPCloud: US-CMS gateway to provide access for CMS to US HPC
  - Site extension: mechanism exploited at CINECA Marconi A2 and ForHLR2(KIT) and CLAIX(RWTH Aachen)
- Working on the exploitation of CPU resources at HPC centres where compute nodes do not have external network connectivity.

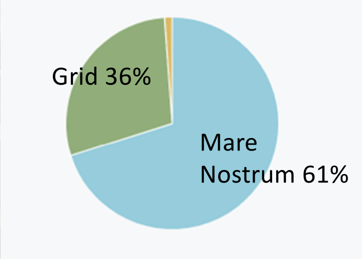
ATLAS CPU Resource Mix

- HPCs delivered 9% of ATLAS normalized wallclock usage in the past year
- A large number of HPCs contributed worldwide. Clear demonstration that we can integrate diverse mix of HPC systems to enable LHC physics
  - *Incomplete list of integrated centers : CSCS, MareNostrum, OLCF, ALCF, NERSC, TACC, LRZ, Nordic HPCs, RU NRC KI, iT4I, ….*
- ATLAS focused on enabling all HPC centers available to ATLAS into the distributed computing system
- New opportunities with EuroHPC project :
  - New ways how physics analysis and data processing will be done



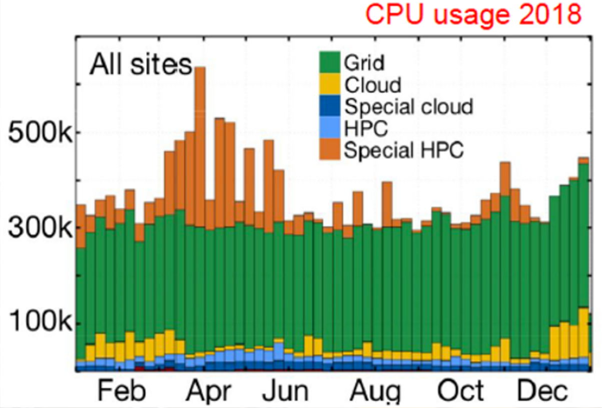
ATLAS jobs normalized wallclock consumption. All resources
Jan 2020 - Jan 2021

ATLAS jobs normalized wallclock consumption at Tier-1 in Barcelona
Jan 2020 - Jan 2021



ATLAS weekly CPU consumption in 2018 (LHC Run2)

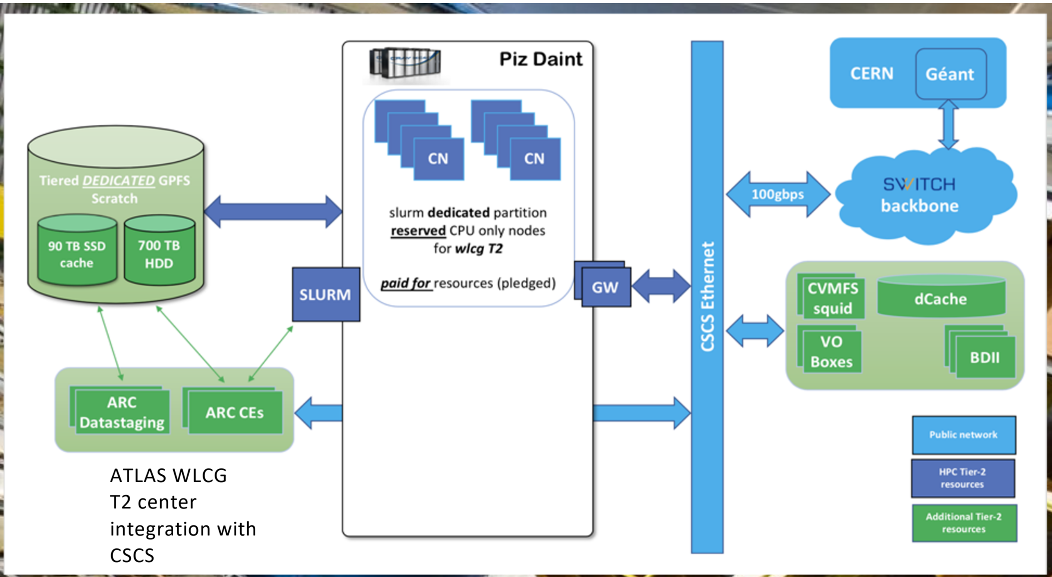
- *HPC delivered 700 million CPU wallclock hours for Monte-Carlo simulations*



Courtesy of D. Benjamin A. Filipic, A.Klimentov

EU-HPC Integration with ATLAS Computing

- ATLAS CH Tier 2 is integrated with Piz Daint
  - First use of HPC as a WLCG Tier 2 Center
  - Using ARC-CE and ARC cache
- MareNostrum
  - Served by ARC-CEs located in Spanish Tier 1/2 WLCG centers)
  - Using singularity container (no internet from WN) for ATLAS SW and databases (O(100GB))



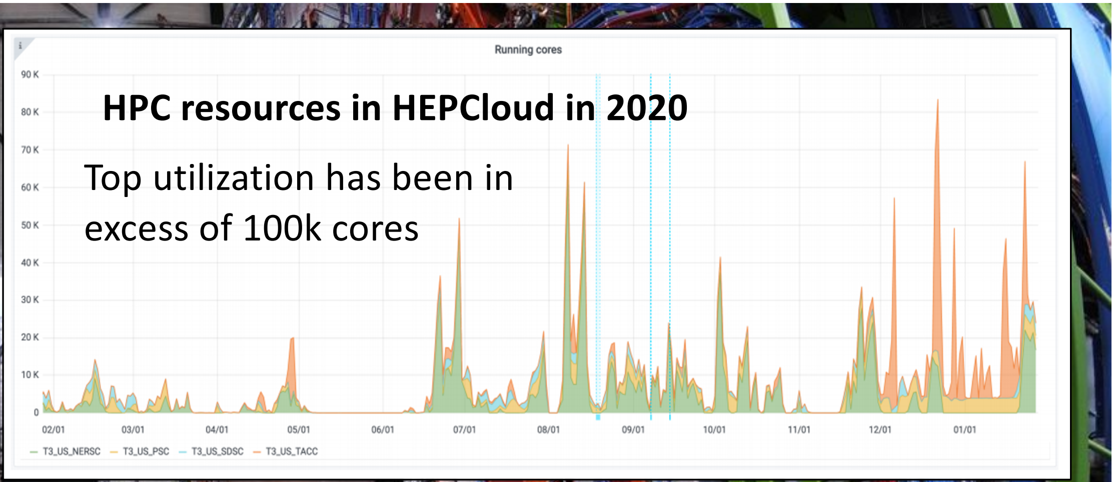
ATLAS WLCG T2 center integration with CSCS

Software developed to address HPC challenges :

ATLAS software in containers, granular data processing, seamless integration with grid, preemption, backfill mode, ...

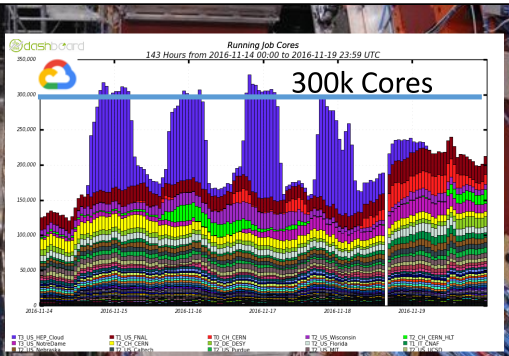
Courtesy of D. Benjamin A. Filipic, A.Klimentov

CMS has demonstrated co-located HPC access as well as the use of HEPCloud for accessing HPC and Commercial Clouds

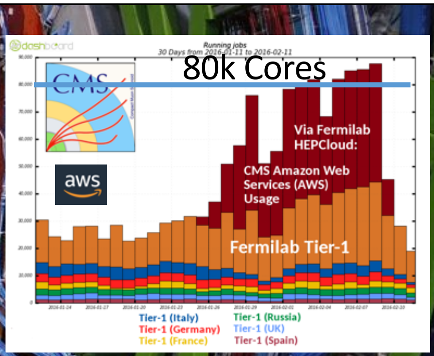


HPC resources in HEPCloud in 2020

Top utilization has been in excess of 100k cores



300k Cores



80k Cores

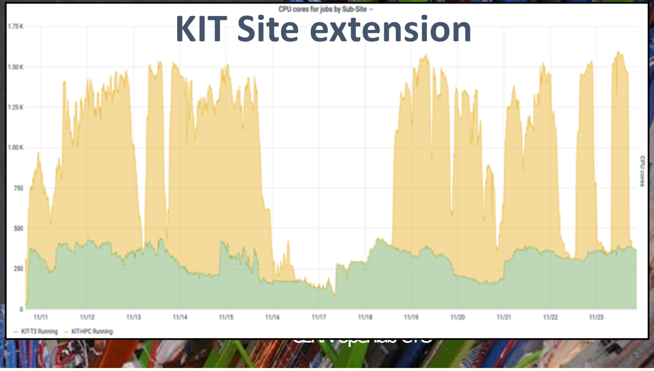
Via Fermilab HEPCloud:

CMS Amazon Web Services (AWS) Usage

Fermilab Tier-1

Courtesy of D. Spiga, C. Wissing

CMS Resource Usage



KIT Site extension

Outlook

HEP is facing an unprecedented computing challenge from the Exabytes of data expected from the HL-LHC

- We have successfully operated our distributed computing environment, the WLCG, for more than a decade and exploited globally distributed computing resources to realize the scientific potential of our data

Looking forward, we will need more resources and opportunities with HPC/commercial clouds may play an important role

- We are involved in projects to exploit exascale capabilities for data-intensive science
  - Continuing explorations of heterogenous architectures
  - Building the expertise in AI/ML
  - Increasing scale of processing and data access solutions
  - Liaising to technology providers through CERN openlab

HPC Collaboration

We are working with to establish enablers for data intensive science using HPC and commercial cloud resources