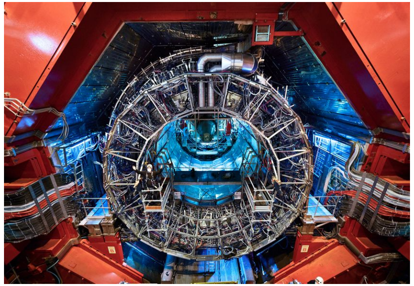
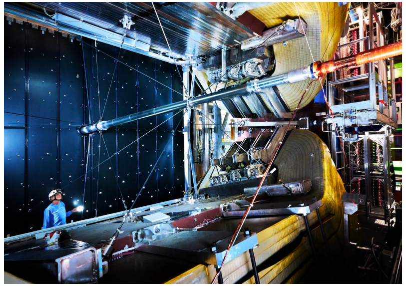
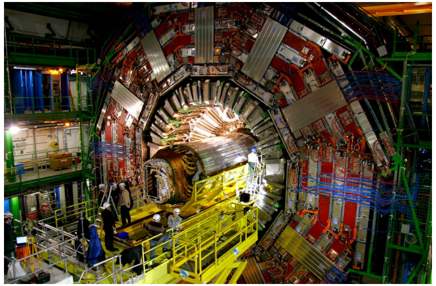





Computing For LHC Experiments

...with a special focus on ATLAS...

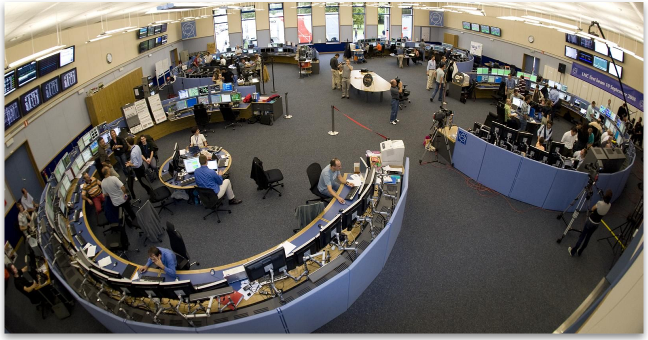
Dr. Mario Lassnig
CERN
Experimental Physics Dept.

CS3 2024 - Cloud Storage Synchronisation and Sharing
2024-03-12
https://indico.cern.ch/event/1332413

ATLAS EXPERIMENT

CERN

The Large Hadron Collider (LHC)



Exploration of the energy frontier in proton and ion collisions

- 27 km circumference, 50-175 metres below the surface
- More than 10'000 superconducting magnets, cooled down close to absolute zero (1.9K)

Also represents a new frontier in physics data volume

- Between them, the LHC experiments generate ~150 PB of collision data/year

The ATLAS Collaboration

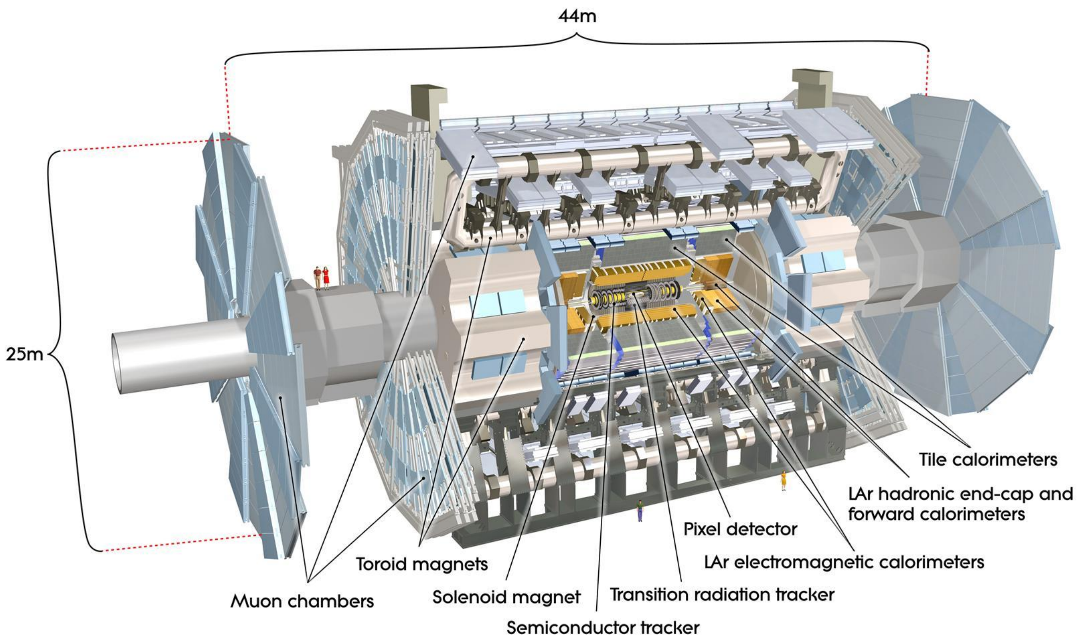
3000
Scientific authors

182
Institutions

42
Countries

1200
Doctoral students

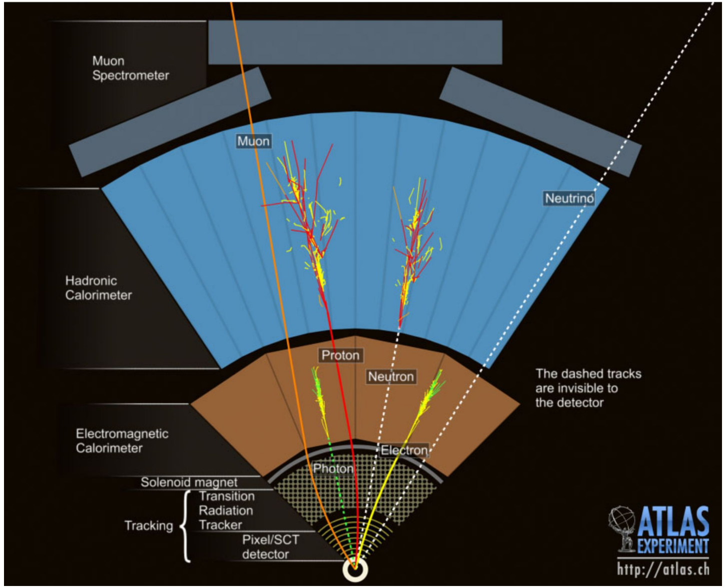
The ATLAS Detector



25 m diameter **44 m** length **7000 tons**

150 million readout channels

3 kHz event rate after filtering



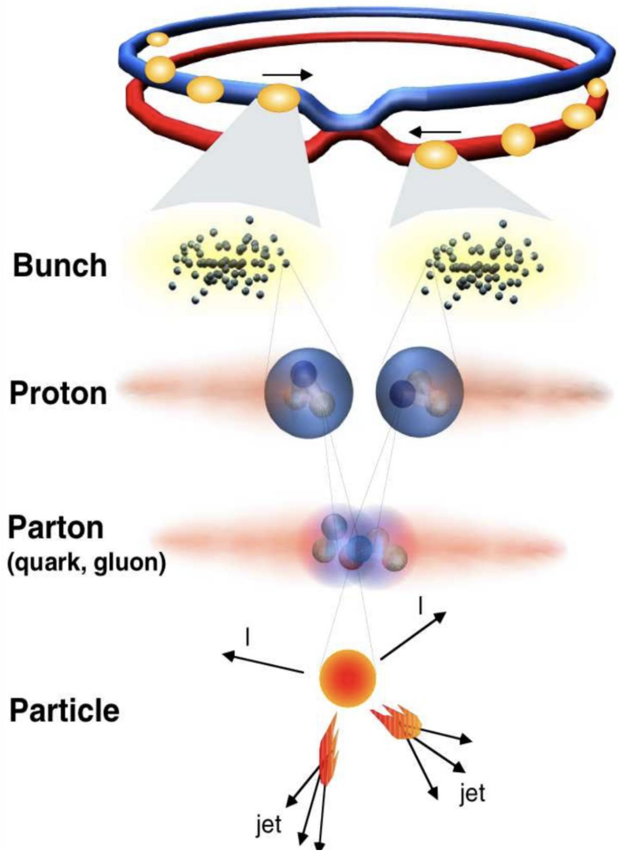
Sophisticated magnet system to constrain and bend the particle tracks

- Central Solenoid Magnet, Barrel Toroid, and End-cap Toroids

Specialised sub detectors arranged in layers

- Particle tracking (Pixel detector, silicon strip tracker, transition radiation tracker)
- Energy/momentum measurements (Liquid argon calorimeter, tile calorimeter, muon spectrometer)

LHC Collisions at ATLAS



Run 1 data (2011-2013)

Centre of mass energy **7-8 TeV**

Run 2 data (2015-2018)

Centre of mass energy **13 TeV**

Run 3 data (since 2022)

Centre of mass energy **13.6 TeV**

Resolution of 25ns

Nominal 10^{11} protons per bunch

Bunches cross at 40 MHz

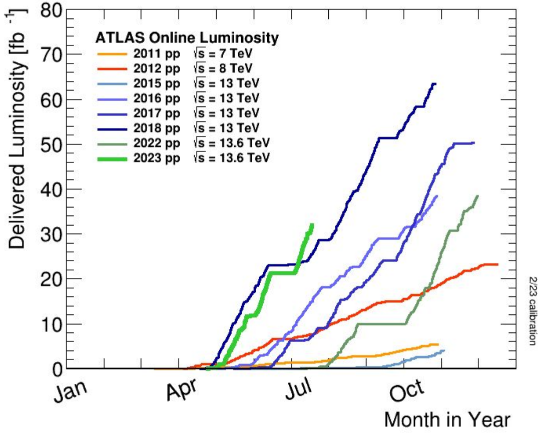
1.5B collisions / second

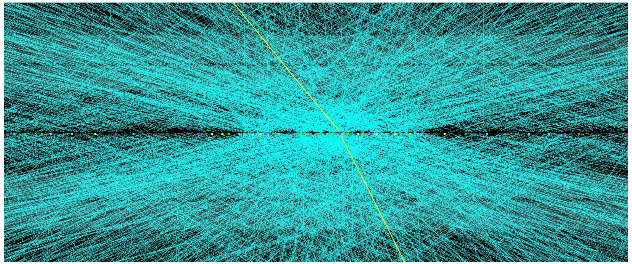
Pile-up

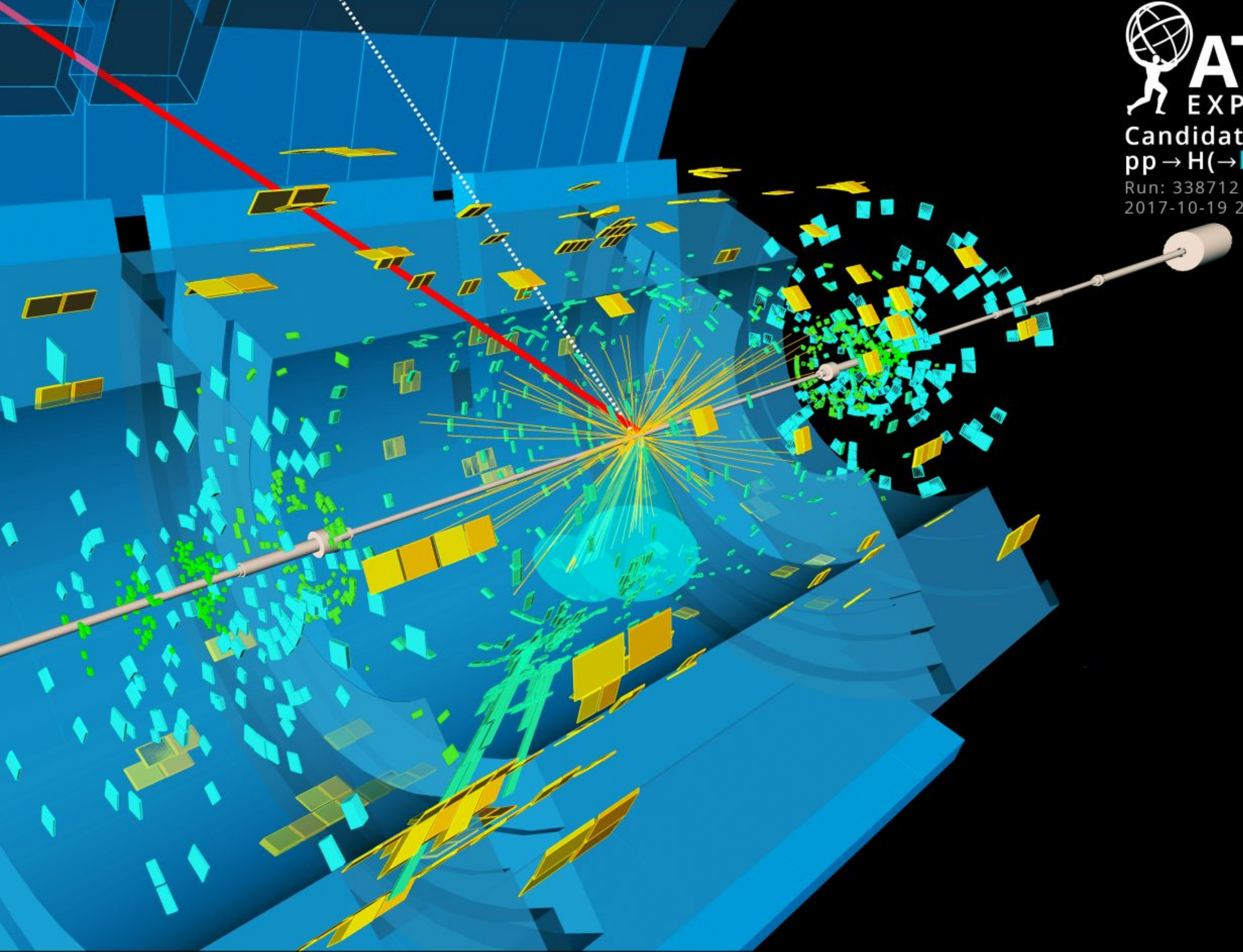
Nr collisions per bunch-crossing

Trigger and event selection

Suppression factor of up to 10^{-10}





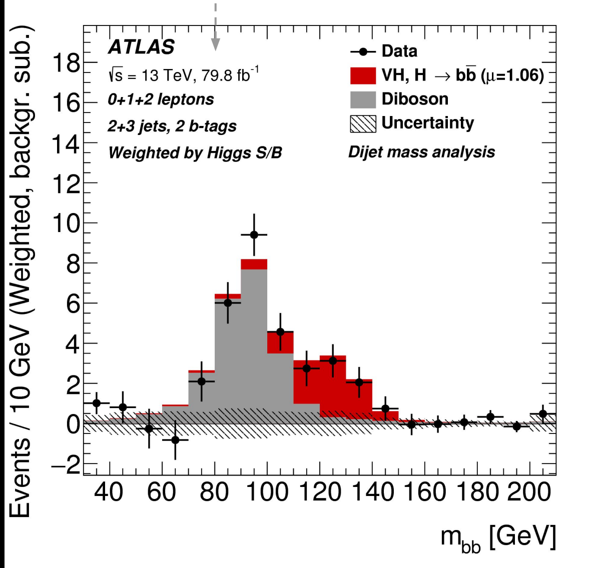


Candidate Event:
pp → H(→bb) + W(→μν)
Run: 338712 Event: 335908183
2017-10-19 23:31:18 CEST

1 Analysis

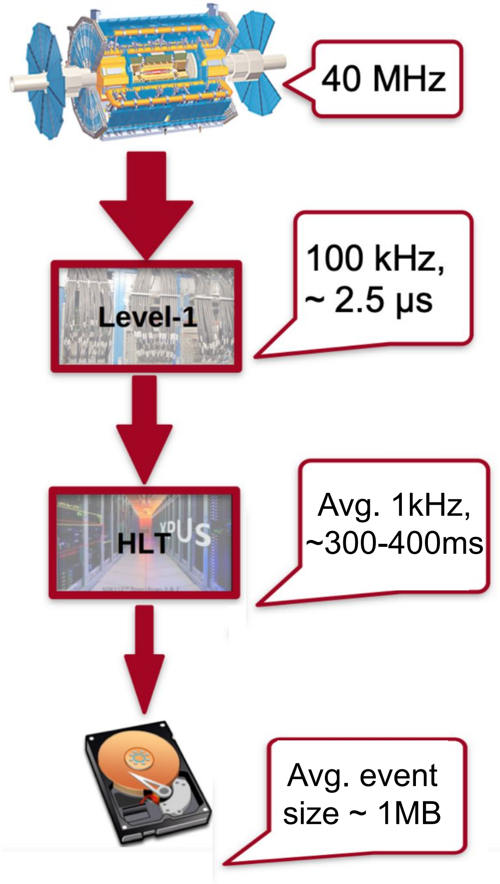
13 TeV detector data
- 8 quadrillion collision candidates
- 92 petabytes
- 130 million files

13 TeV simulation data
- 166 petabytes
- 544 million files



A candidate event display for the production of a Higgs boson decaying to two b-quarks (blue cones), in association with a W boson decaying to a muon (red) and a neutrino. The neutrino leaves the detector unseen, and is reconstructed through the missing transverse energy (dashed line).

Trigger and data acquisition in LHC Run-2



Level 1 Hardware Trigger 100 kHz

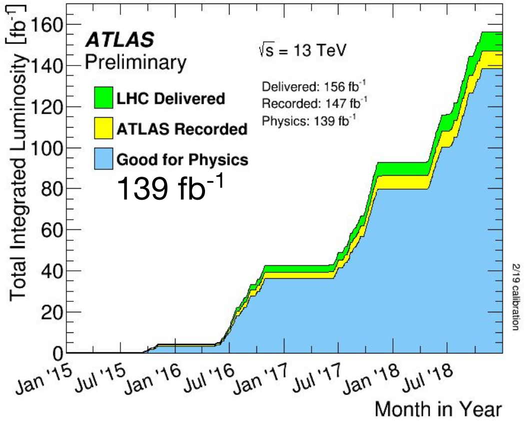
- First selection based on calorimeter and muon systems
- Rate / Latency limit from detector and trigger hardware

High Level Software Trigger 1 kHz

- Processing time of 300-400ms
- Size of HLT farm comprising ~100k cores
- Final output rate ~ 1kHz

In Run-3 this increased substantially

- Acceptance rate at 3 kHz
- Event size increased to 3MB



Main physics stream

Year	Raw events	SFO total volume	SFO event volume
2015	1,694,555,330	1.4 PB	828.2 KB
2016	5,387,420,813	4.9 PB	1004.8 KB
2017	5,649,311,254	5.5 PB	1 MB
2018	6,400,342,575	6.2 PB	1 MB

19 billion events collected by ATLAS

18 PB of raw data

The raw data is only the start

Raw instrument data

Sensor hits, energy deposits, timing information

Analysis Object Data (AOD)

4-vector momentum of tracks

Energy in jet clusters

Particle identification

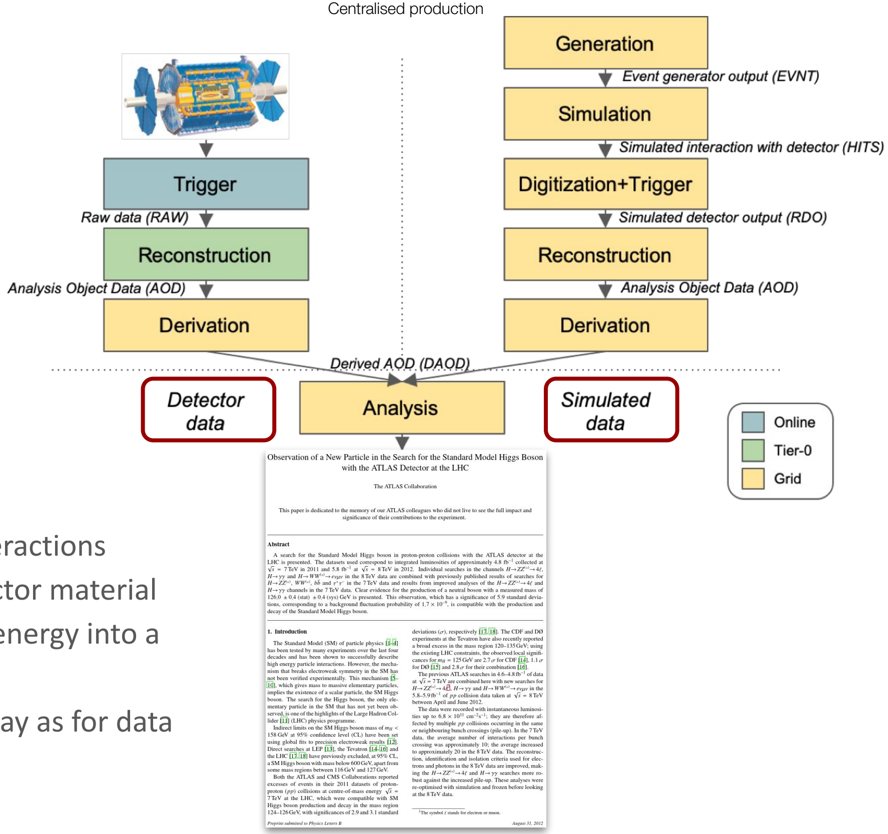
First calibrations

Derived AODs

Selected analysis level information with full calibration

Starting point for analysis

The Data Processing Chain



Monte Carlo Simulation

Event generation	EVNT	Calculated particle interactions
Simulation	HITS	Interactions with detector material
Digitization	RDO	Transforms simulated energy into a detector response
Reconstruction	AOD	Performed the same way as for data

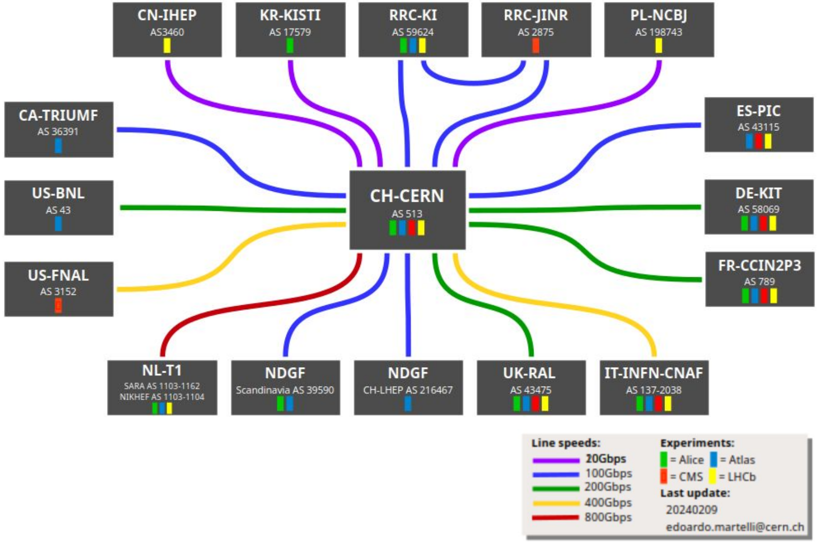
A global shared infrastructure

Worldwide LHC Computing Grid (WLCG)

- **Global collaboration** of 170 institutes & laboratories
- **Shared** across the experiments
- **Provides resources** to store and analyse all experiment data
- **Heterogeneous** installations in different administrative domains
- **Over 1 Million cores** of computational resources
- **2 Exabyte of data** stored across all experiments
- Long-term and forward looking **sustainable technology R&D programmes**

Terabit scale global network connectivity

- Supports the experiments **data needs**
- **Archival, transport, and processing**
- Dedicated **optical private networks**
- **Peered** with NRENs and commercial clouds
- **Overlays** available for all resources



Tuned to support complex data flows

- Long flows shipping multiple **Petabytes per day**
- **Latency-aware** remote interactive analysis

Zooming into ATLAS again

ATLAS Distributed Computing (ADC) comprises the hardware, software, and operations to

- Support **distributed computing activities** of the experiments
- Support the **evolving needs** of the experiment

Running 24 / 7 / 365

- **Computing never stops**
- 80+ people contributing centrally
- 50+ people across the WLCG

Four major areas

- Physics activities requiring computing
- Infrastructure & operations
- Data management
- Workload & workflow management

PHYSICS	FABRICS	DATA MANAGEMENT	WORKFLOW MANAGEMENT
Production Coordination	***Coordination***	***Coordination***	***Coordination***
M. Borodin	*V. Garonne*	*S. McKee, P. Vokac*	*R. Walker, F. Barreiro Megino*
Analysis Coordination			
A. Forti			
Centralised Production	**Infrastructure**	**System**	**System**
Monte Carlo Production	Tier-0	Rucio	Workflow Definition
Group Production	Grid		Workload Management
Data Reprocessing	HPC	**Operations**	Workload Execution
Physics Validation	Cloud	System Deployment	
HLT Reprocessing	BOINC	DDM Central Operations	**Operations**
	Analysis Facilities	Monitoring	System Deployment
Physics Analysis			Monitoring
User Analysis Tools	**Operations**	**Research**	
Analysis Model Group	Computing Run Coordination	Networks	**Research**
DAST	DA Operations	Caches	Data Analytics
	DPA Operations	Storage	Analysis Facilities
	Central Services	Cloud	Cloud
	CRIC		HPC
	HammerCloud		
	Monitoring		
	ADCoS		

Plus many task forces and working groups, e.g., HPC or monitoring

Computing and processing

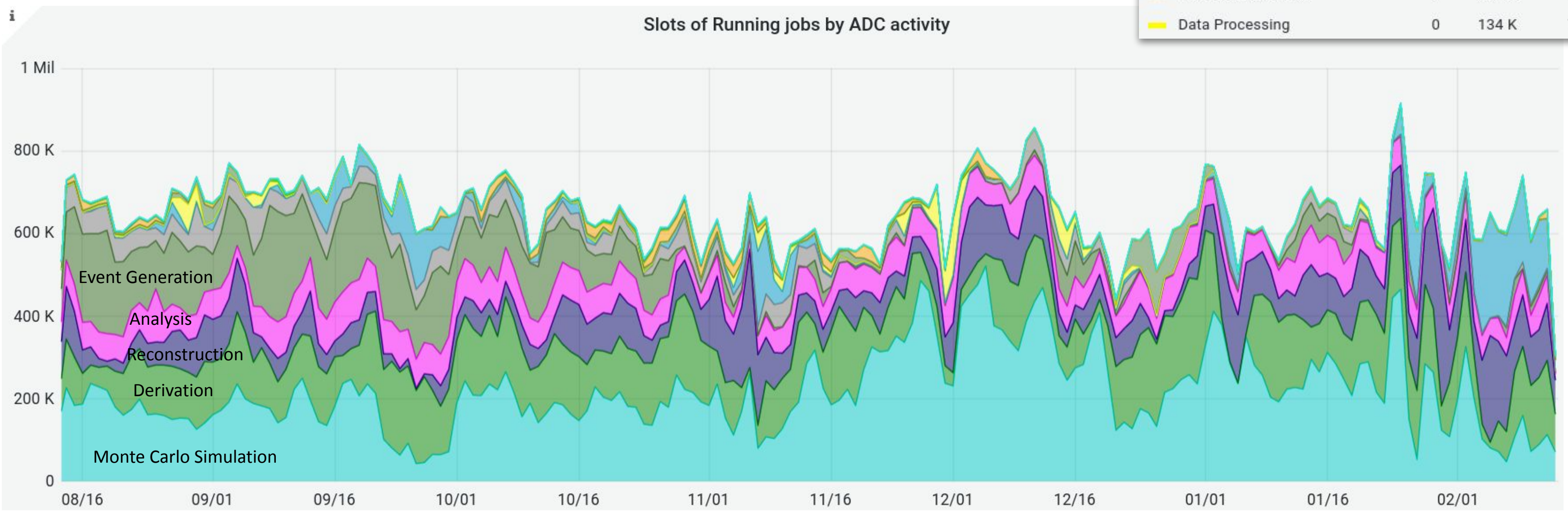
Global high-throughput computing system

- **Steady** 600k to 800k running jobs, with **full spread** of **experiment activities**
- Spread across ~250 clusters **worldwide**

Sophisticated scheduling system

- **Physics campaigns** spread across **processing tasks**
- Tasks are **split into jobs** based on available computational resources

	min	max	avg
MC Simulation Full	8.34 K	500 K	172 K
Group Production	1.40 K	279 K	129 K
MC Reconstruction	5.59 K	338 K	85.0 K
User Analysis	13.4 K	129 K	68.5 K
MC Event Generation	77.8	270 K	68.0 K
Group Analysis	277	116 K	28.7 K
MC Simulation Fast	0	251 K	28.2 K
Data Processing	0	134 K	7.32 K



Resource usage

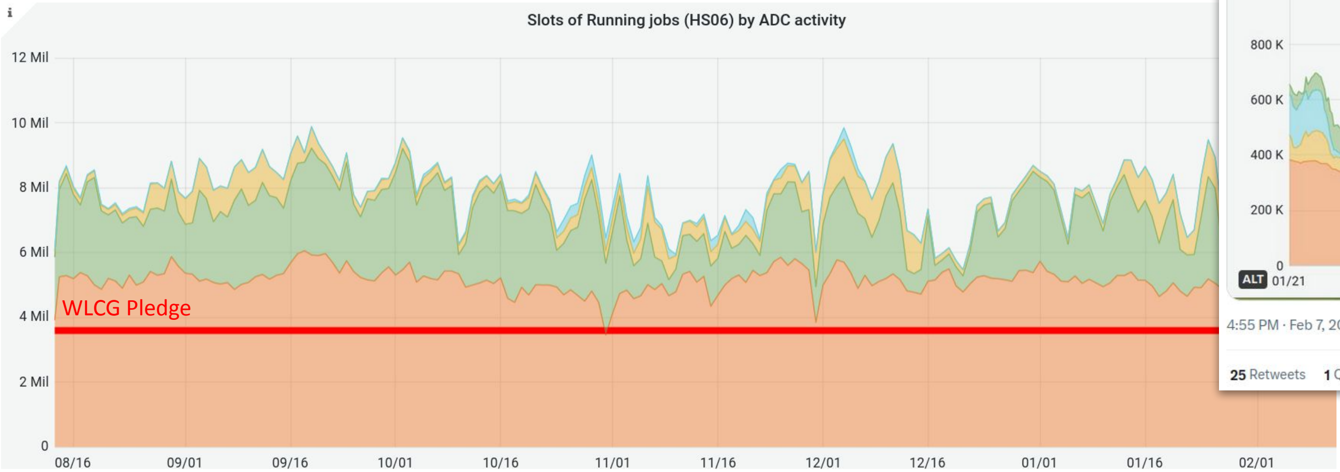
Computing power expressed in terms of HEPSPEC benchmark

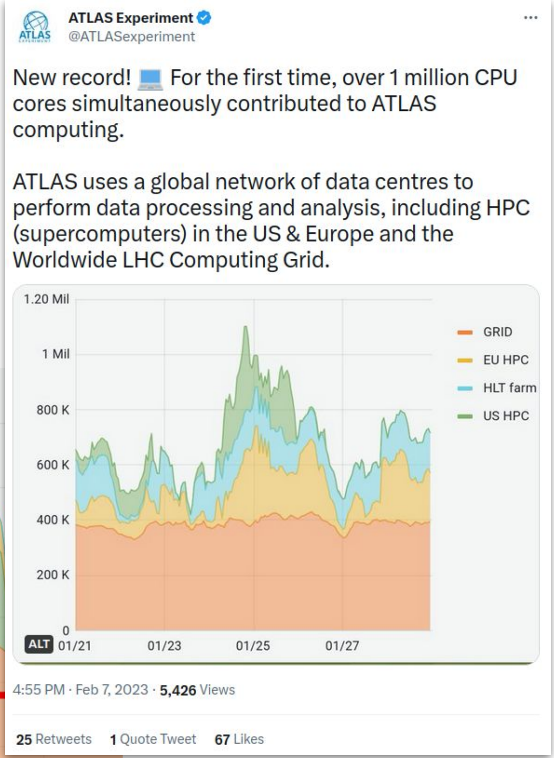
- 1 modern x86_64 core ≈ 10 HEPSPEC
- ATLAS-available infrastructure is **consistently over WLCG pledge**

Integration of new and/or opportunistic resources

- Integrating **special resources** offered to us, e.g. ARM cores or GPUs
- This brings interesting **challenges in resource accounting and scheduling**
- **Dynamic repurposing** of the online hardware during LHC downtimes

Significant contributions from **EuroHPC** and **US HPCs**





Experiment job mix

Globally configured shares are employed to allocate the available resources among the activities

- Done by **agreement** between the various physics groups
- **Hierarchical** implementation of the configuration parameters
- Related activities have the opportunity to **inherit idle resources**

Essentially two major categories of jobs

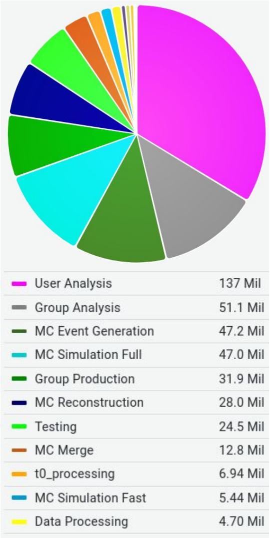
Production Data processing and reprocessing
Event Generation / Simulation / Reconstruction
Derivation

Analysis User Analysis
Group Analysis

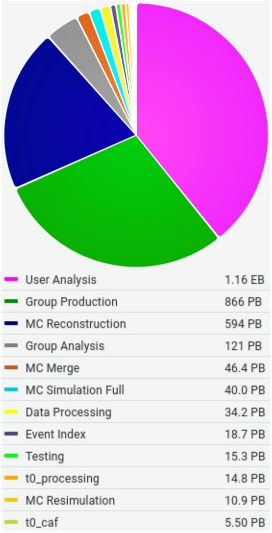
The main activity at a given time can depend on many things

- Data **reprocessing** or Monte Carlo **production** campaigns
- **Conference** deadlines, need for an increase for user analysis

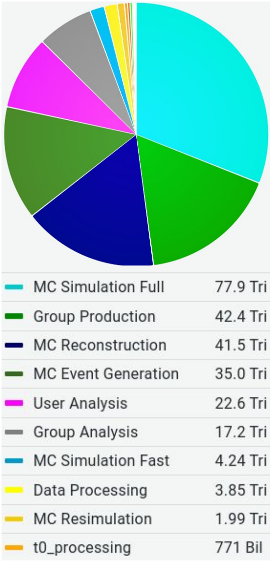
and … global **pandemics …**



Job Slots 2022



Processed Data 2022



Wall Time 2022

Helping with COVID research

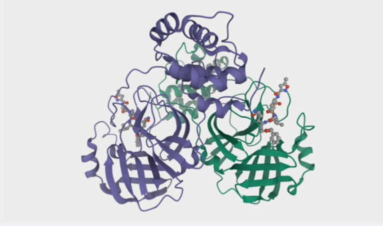
ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

BACK IN THE FOLD —

The coronavirus pandemic turned Folding@Home into an exaFLOP supercomputer

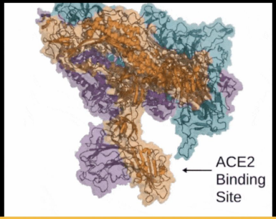
Folding@Home had settled into a low-profile niche. Then came COVID-19.

ANDY PATRIZIO - 4/14/2020, 5:15 PM



Enlarge

sciencealert Trending



SARS-CoV-2 protein spike in action. (Folding @ home/YouTube)

TECH

People Running Folding@Home Accidentally Created The World's Biggest Supercomputer

DAVID NIELD 17 APRIL 2020

You may have heard of Folding@home, the number-crunching app you can run on your computer to help researchers tackle certain medical problems, including the new coronavirus. In the past month, the network of volunteers who've installed it has become so vast, the platform is outperforming the most powerful supercomputers in the world.

topics

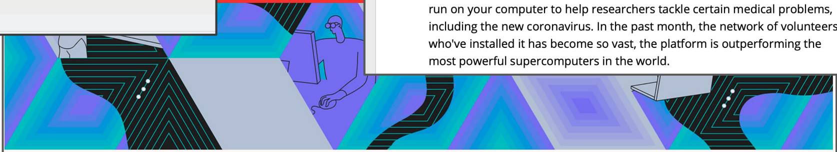


Illustration by Sandbox Studio, Chicago with Ana Kova

Physicists crowdsource pandemic problem-solving

05/04/20 | By Diana Kwon

The group Science Responds harnesses physicists' expertise in fields like data science, statistics and software development to support efforts to respond to COVID-19.

Team: CERN & LHC Computing

Date of last work unit	2020-05-26 07:16:26
Active CPUs within 50 days	1,228,373
Team Id	38188
Grand Score	25,931,972,247
Work Unit Count	7,067,253
Team Ranking	25 of 253595
Homepage	http://public.web.cern.ch/public/
Fast Teampage URL	https://apps.foldingathome.org/teamstats/team38188.html

Team members

Rank	Name	Credit	WUs
38	CMS-Experiment	10,290,021,099	2,059,008
56	ATLAS_CPU	8,347,461,690	2,028,906
366	LHCbHLT	1,825,988,340	287,261
397	ALICE-FLP	1,695,094,633	149,989
463	CERN_Cloud	1,495,243,514	675,048
1,093	DESY-ZN_GPU	699,405,834	5,197
3,119	UC_ATLAS-ML	229,128,604	115,034
3,889	CMSDCS	178,594,476	19,905
4,540	BNL_HPC_CPU	149,043,910	8,998
5,878	ALICE-CS	110,367,581	19,339
6,915	ANALY_MANC_GPU	92,839,127	4,360
9,999	Cloverfield	62,682,810	524
16,767	Pic	36,401,454	10,031
19,147	ALICE-CERN	32,702,236	53,682
21,035	ANALY_MWT2_GPU	29,835,413	1,346
21,243	TheLaboratoire	29,493,588	479
22,499	CERN_openlab	27,816,689	26,637
12,133	Alpinwolf	26,465,743	408
23,717	ANALY_LRZ_GPU	26,285,509	1,950
24,352	ryukisai	25,219,040	157

Integrated CERN and WLCG resources with Folding@Home project

- First with our **Tier-0** resources, then including the **trigger farm**
- Then included **CPU and GPU** resources from the WLCG
- **Analysis share backfill** pushed us beyond 60k concurrent jobs

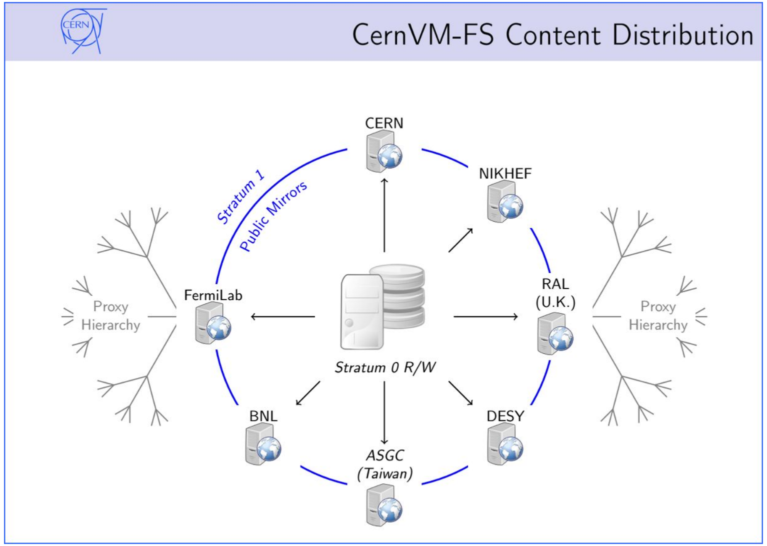
ATLAS Software distribution: CVMFS

ATLAS relies on CVMFS (CERN VM FileSystem)

- Network file system based on HTTP
- Optimized to deliver experiment software
- New SW pushed into the system at CERN Stratum-0
- Replicated to the Stratum-1 public mirrors
- Massive replication through set of Squids hosted at the sites
  - WN at sites accessing the site's squids
  - Resilient in case of Squid failures, retries going one level up

All standard ATLAS sites use CVMFS

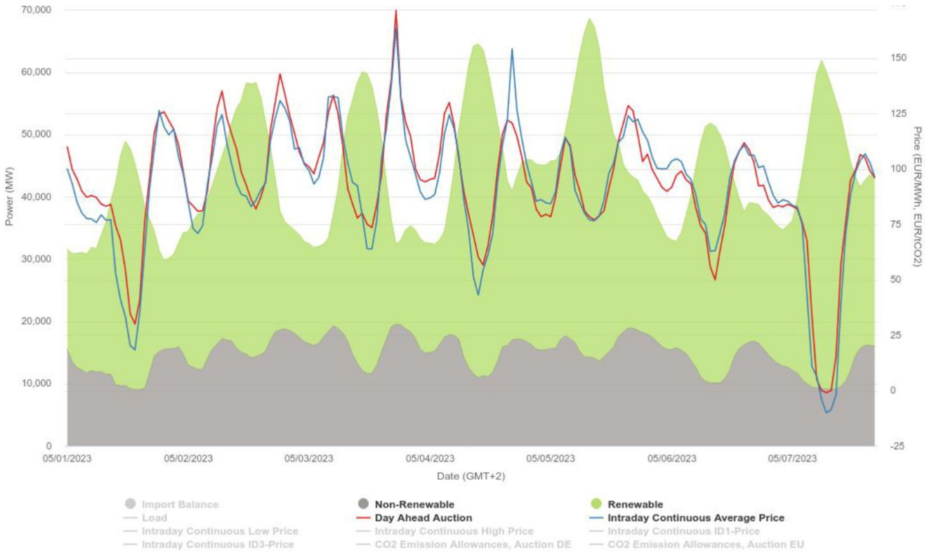
- Requires connection to the outside world
- Not suitable for most HPC due to connectivity



Carbon efficient computing

Significant enthusiasm in the community for addressing sustainability

- **Efficiency** and reliability of software and data centres
  - Bugs and failures correspond directly to **wasted CO2**
  - **Dedicated R&D** on improving site failures and user failures, retrial strategies, etc …
- **Electricity mix** and flexible demand
  - e.g., ARM using 40% less power per HEPSPEC overall than Intel
- **Flexible computing** demand
  - Price and gC02/kWh vary
  - Cheaper with renewables available
- **It matters WHEN the electricity is consumed!**



Data centre modulate power consumption?

- **Freeze** processes to let CPU sleep
- **Reduce** CPU frequency to minimum
- **Switch** to battery if available

Lots of R&D ongoing now!

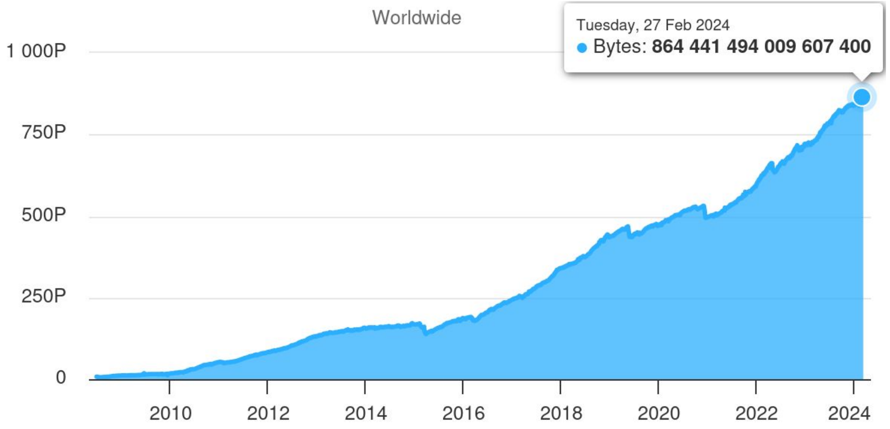
From computing to data

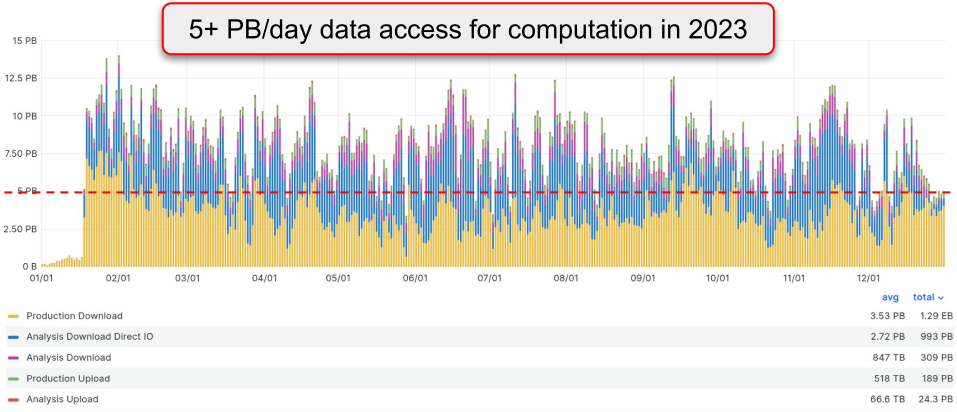
A few numbers about the ATLAS scale

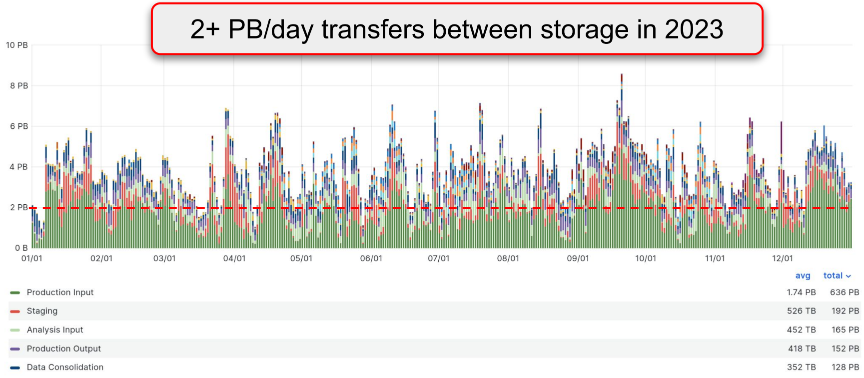
1B+ files, 850+ PB of data, 400+ Hz interaction rate
120 data centres, 5 HPCs, 3 clouds, 1000+ users
1.5 Exabytes/year transferred
3 Exabytes/year uploaded & downloaded

Efficient data management is the key

We have developed a system to do that, called **Rucio**







Rucio in a nutshell

Rucio provides a mature and modular scientific **data management federation**

- **Seamless integration** of **scientific and commercial** storage and their network systems
- Data is stored in a **global unified namespace** and can contain **any potential payload**
- Facilities can be **distributed at geographically independent locations** belonging to **different administrative domains**
- Designed with **more than a decade of operational experience** in very large-scale data management

Rucio is location-aware and manages data in a heterogeneous distributed environment

- Creation, location, transfer, deletion, annotation, and access
- **Orchestration of dataflows** with both low-level and high-level policies

Principally developed by and for the ATLAS Experiment, now with many more communities

Rucio is **free and open-source software** licenced under *Apache v2.0*

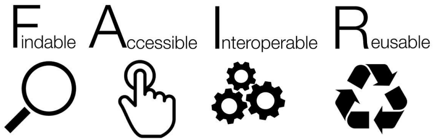
Open **community-driven** development process

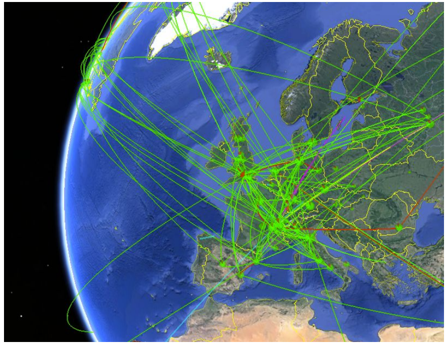
Rucio main functionalities

Provides many features that can be enabled selectively

More advanced features ↓

- **Horizontally scalable catalog** for files, collections, and metadata
- Transfers between facilities including **disk, tapes, clouds, HPCs**
- **Authentication and authorisation** for users and groups
- **Many interfaces** available, including CLI, web, FUSE, and REST API
- **Extensive monitoring** for all dataflows
- Expressive **policy engine** with rules, subscriptions, and quotas
- Automated **corruption identification and recovery**
- Transparent support for **multihop, caches, and CDN dataflows**
- **Data-analytics based flow control**





Rucio is not a distributed file system, it **connects existing storage infrastructure** over the network

- No Rucio software needs to run at the data centres **(!)**
- Data centres are free to choose which storage system suits them best - **No Vendor Lock-In (!)**

Declarative data management

Objective is to minimise human interaction as much as possible

Express what you want, not how you want it

e.g., *"Three copies of this dataset, distributed across MULTIPLE CONTINENTS, with at least one copy on TAPE"*
e.g., *"One copy of this file ANYWHERE, as long as it is a very fast DISK"*

Replication **rules**

Rules can be **dynamically added and removed** by all users, some pending **authorisation**
Evaluation **engine resolves all rules** and tries to satisfy them by requesting transfers and deletions
Lock data against deletion in particular places for a given lifetime
Cached replicas are **dynamically created replicas** based on traced usage over time
Workflow system can drive rules automatically, e.g., **job to data flows** or vice-versa

Subscriptions

Automatically generate rules for newly registered data matching a **set of filters or metadata**
e.g., *"All derived products from this physics channel must have a copy on TAPE"*

Full and generic **metadata support**

Allow Rucio to be connected to **different metadata backends** (JSON columns, MongoDB, external systems, …)

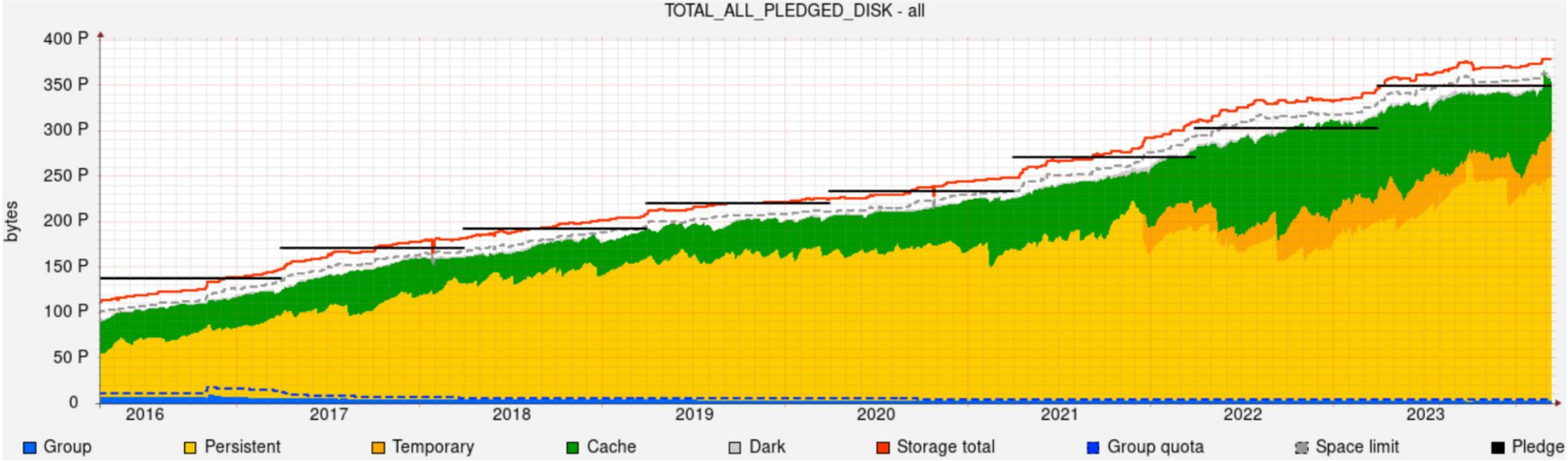
Keeping our storage under control

Our disks are constantly full, and that is good thing!

- Strive for a healthy **cached-to-persistent** ratio
- AOD and HITS volume is stable, DAOD grows from constant production and new physics requests

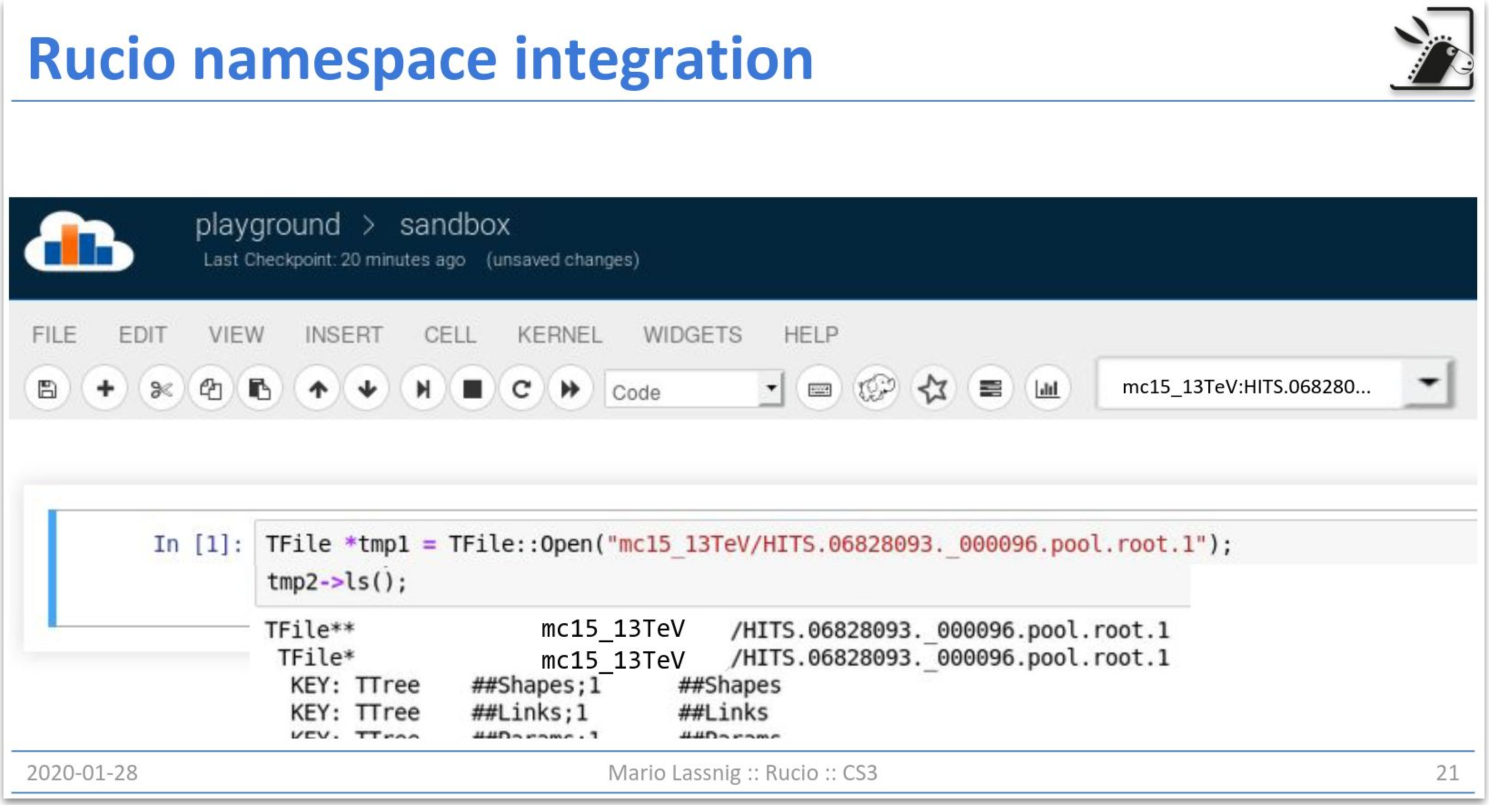
We are automating our data lifecycle as much as possible

- Migrate data to tape, recycle disk resident copies as cache for faster processing, …
- Lifetime exceptions from physics groups for special analyses, …



Speaking of analyses… CS3 & Rucio?

A great idea (or so I thought), and a bad photoshop



CS3 & Rucio!

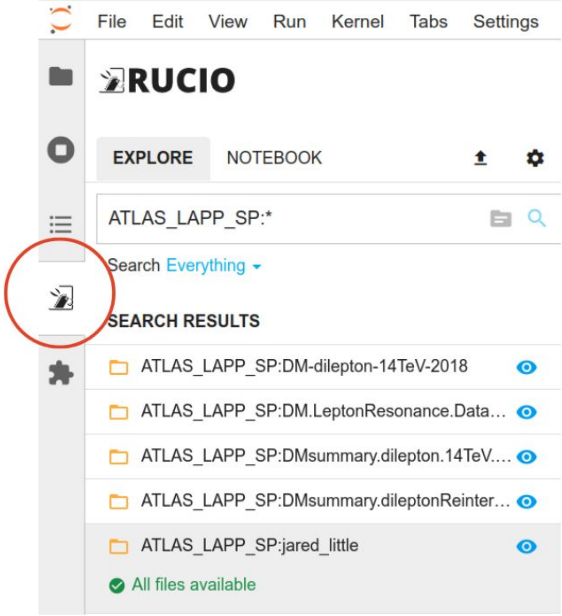
Led to an incredible development with the **Virtual Research Environment**

Data into the notebook

The **Jupyterhub Rucio extension** hides the complexity of the Data Lake and allows users to

- browse experiments' data catalogue
- authenticate with OIDC tokens to the Rucio infrastructure
- replicate data into the notebook
- import the data into the notebook by assigning a parameter to it
- run preliminary analysis to prototype code

RUCIO → jupyter



The Virtual Research Environment, E. Gazzarrini, CHEP 2023 18

A success story

SCIENTIFIC DATA MANAGEMENT
RUCIO

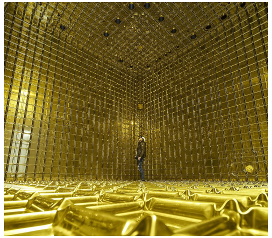
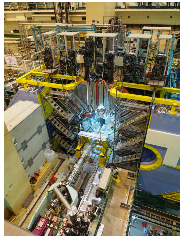
Rucio has become the **de-facto standard** for **open scientific data management**

Used by CERN-based experiments	ATLAS, CMS, AMS
And non-CERN experiments	XENON, Belle II, LBNF/DUNE, SBN/ICARUS, KIS Solar, LIGO/VIRGO/KAGRA, CTAO, Vera Rubin Observatory, …
Under evaluation by many others	Copernicus, SKA, EIC/ePIC, …

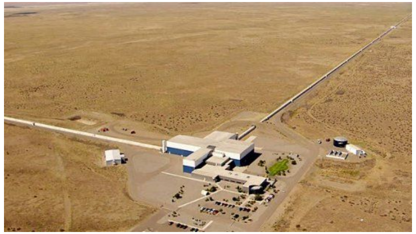
Free and **open-source software** with an **open community-driven** development process

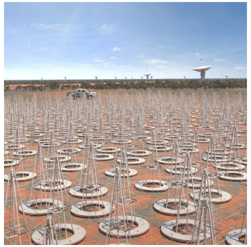
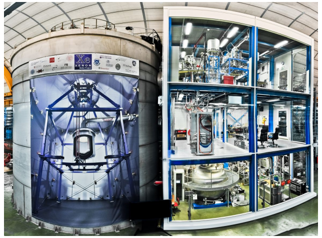
Find it here	https://rucio.cern.ch/
Read about it here	https://link.springer.com/article/10.1007/s41781-019-0026-3

open source initiative
Approved License





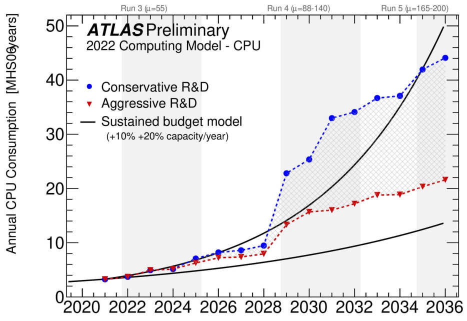



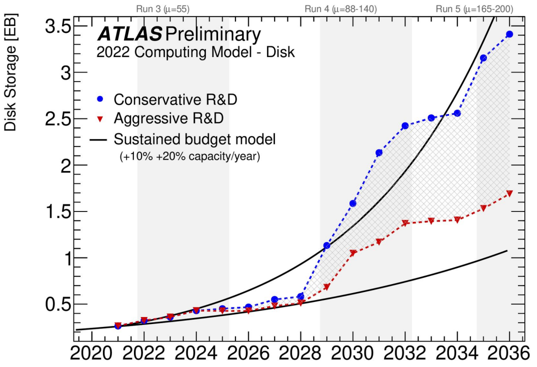

Our data challenges and opportunities

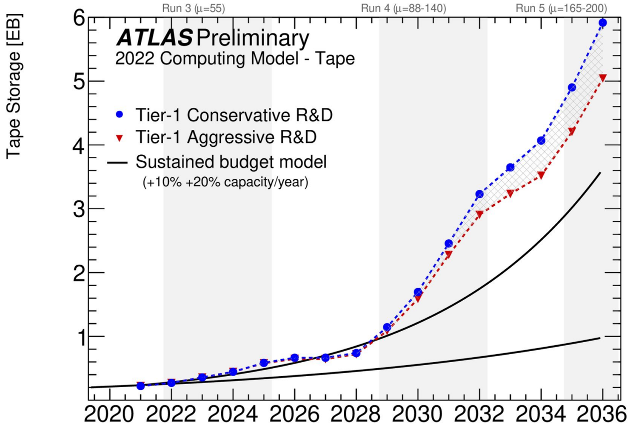
The High Luminosity upgrade to the LHC
- 10 times increase in **accelerator performance**
- Leads to more and bigger, complex events
- 10 times increase in **data volume/usage**
- In a very tight computing capacity envelope

We cannot compromise physics performance

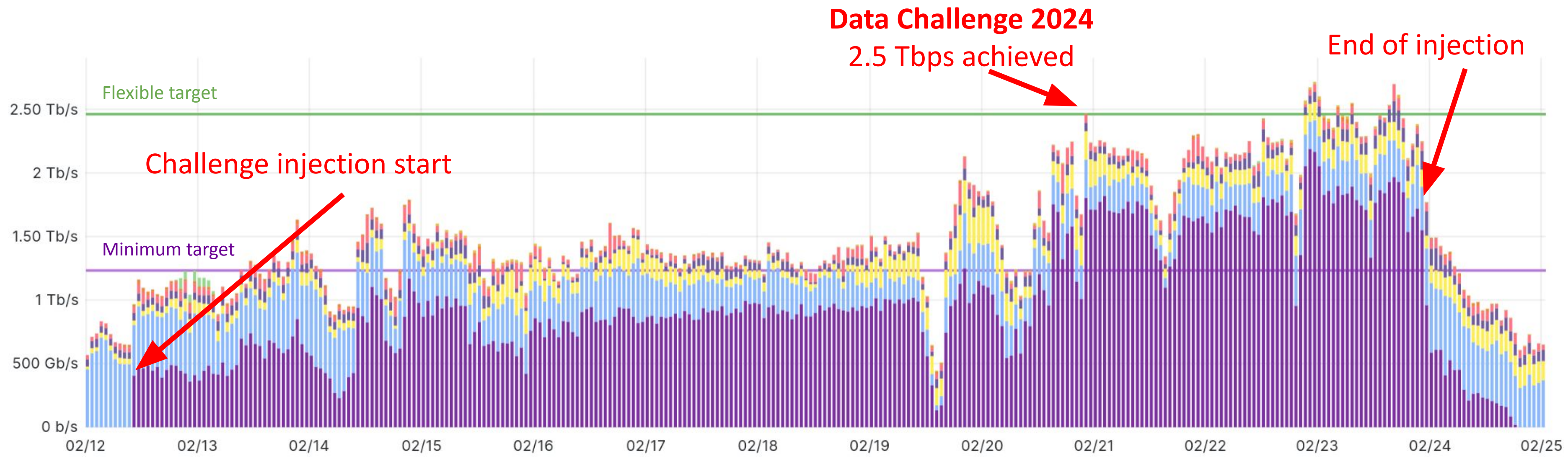
Long-term R&D programme to address the gap
- To support the **European Strategy for Particle Physics**
- **Community-driven** computing R&Ds
  - Advanced **software-defined networks** (SDNs)
  - **Smart content delivery** and caches
  - New analysis **data formats** and models
  - Integration of new **external developments**
  - **Industry collaborations** for new technologies
- **Collaborations** with other sciences
  - Shared infrastructure with other big communities
  - Prototype of a common European Data Infrastructure
- **MSc and PhD studies** to train our future computing engineers







HL-LHC Data Challenges



	max	avg
Data Challenge	2.19 Tb/s	987 Gb/s
atlas	706 Gb/s	316 Gb/s
alice xrootd	349 Gb/s	114 Gb/s
cms	271 Gb/s	56.8 Gb/s
cms xrootd	191 Gb/s	67.7 Gb/s
lhcb	83.1 Gb/s	2.35 Gb/s
belle	38.9 Gb/s	9.33 Gb/s
dune	28.6 Gb/s	5.47 Gb/s

2020 estimation of HL-LHC needs

- **4.8 Tbps** of total network capacity for the Run-2 computing model
- **9.6 Tbps** for the Run-3 (and beyond) computing model

Data Challenges until HL-LHC startup

- **Bi-annual steps** of 25% expected capacity
- With an **accompanying R&D** programme for software and hardware

ATLAS Google Project

Long-standing R&D cooperation with Google, in multiple phases

Phase 0 Demonstrate integration with ADC systems (2019-2022)

Phase 1 Investigate Google Cloud Platform as an ATLAS analysis facility (2021-2022)

Phase 2 Full integration and production usage (2022-2023)

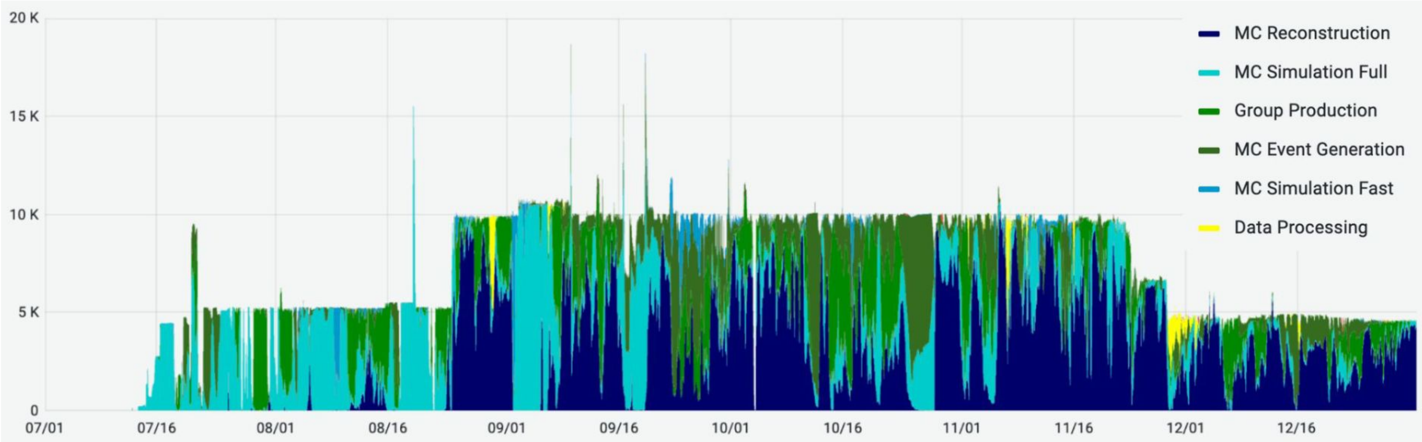
Evaluation **all ATLAS workflows**, including data reprocessing, and user analysis

Demonstrate rapid and **efficient bursting** to additional, large scale resources

Validation of ATLAS **software on ARM** resources

Data analysis using **parallel workflows on GPUs**

Evaluate **Total Cost of Ownership** of employing a commercial cloud site at scale



Cloud amusements

Commercial clouds need "creative care"

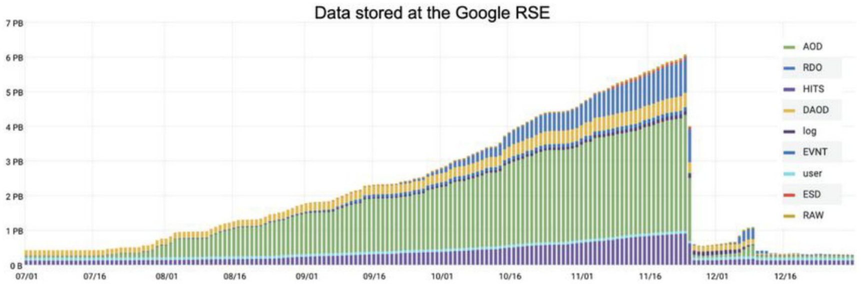
- On WLCG sites we leave data as cache
- But Google has infinite capacity?
- The more cache you have the more it's used…

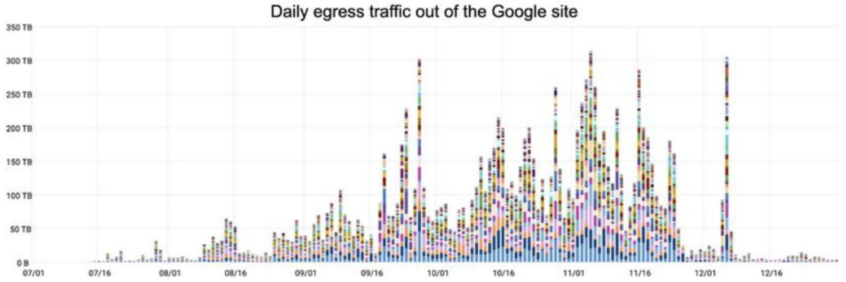
Peering with LHCOPN is crucial

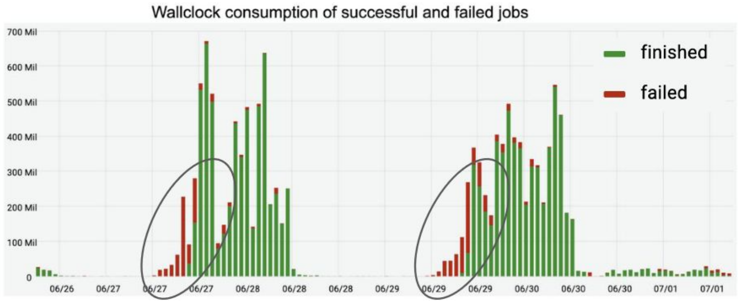
- Incurs extra charges for WLCG sites

Cloud bursting can be **fast**

- MC Full Simulation of 50M event sample in 24h
  - Control sample took 8 days
- Interesting walltime issues observed
  - Software deployment via CVMFS couldn't keep up
  - Node preemption in Google Belgium data centre





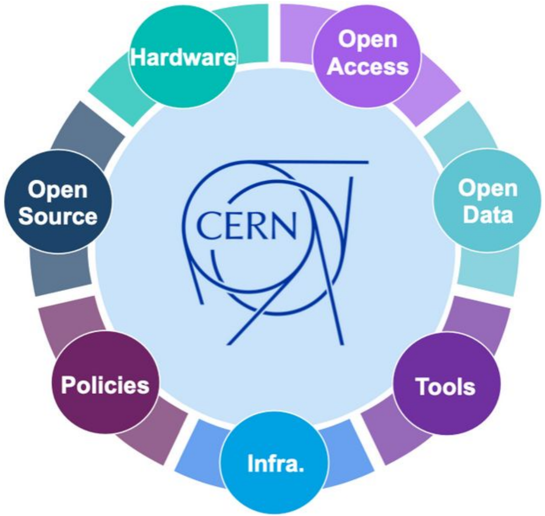


Open data for science

Open data at CERN is an organisational mission

Ecosystem of policies, initiatives, services, and technologies

Maximize the potential of **global impact** of CERN research



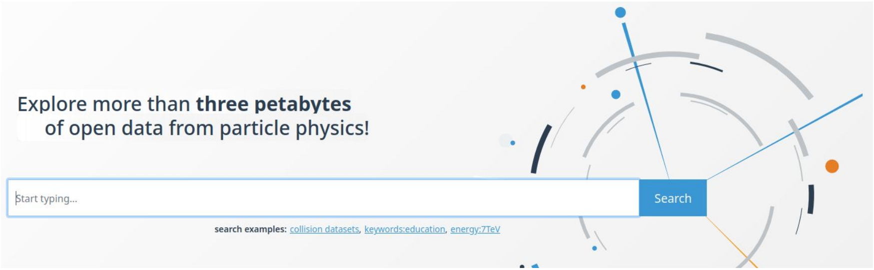
Our pillars of open data

Open access to publications and their data

Preservation through reusable and reproducible analyses

Open software and hardware

Training, outreach, and education



zenodo

Conclusions — From data to knowledge

LHC Experiments are working 24/7

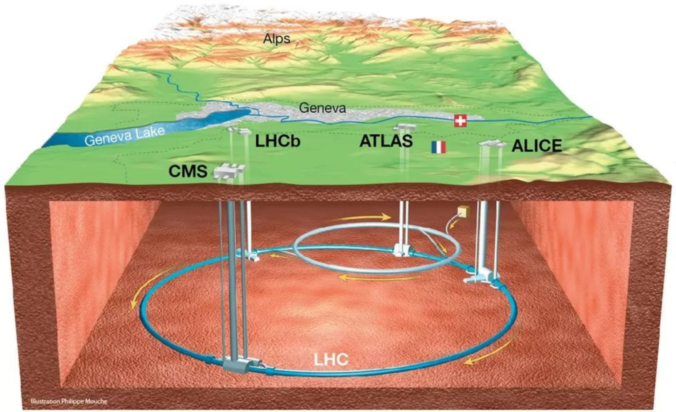
Petabytes per second of electronics readout

Triggers select the **interesting physics**

Physics data is written to the **CERN data centre**

And then **exported to our collaborating institutes**

Where it's **processed and analysed**



Data is our most precious resource

Large variety of analyses

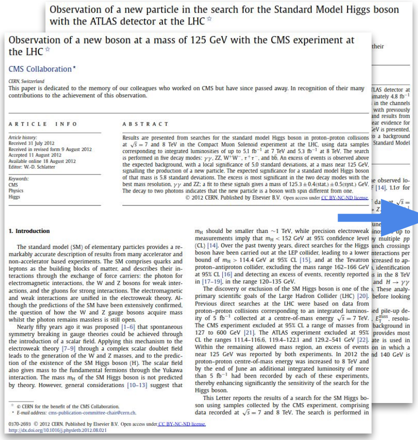
Focused searches

Precision measurements

Exotic particles

The **Unexpected**

Technology R&D


Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC

CMS Collaboration

A huge team effort!

CERN

ATLAS
EXPERIMENT