Big Data al CERN:

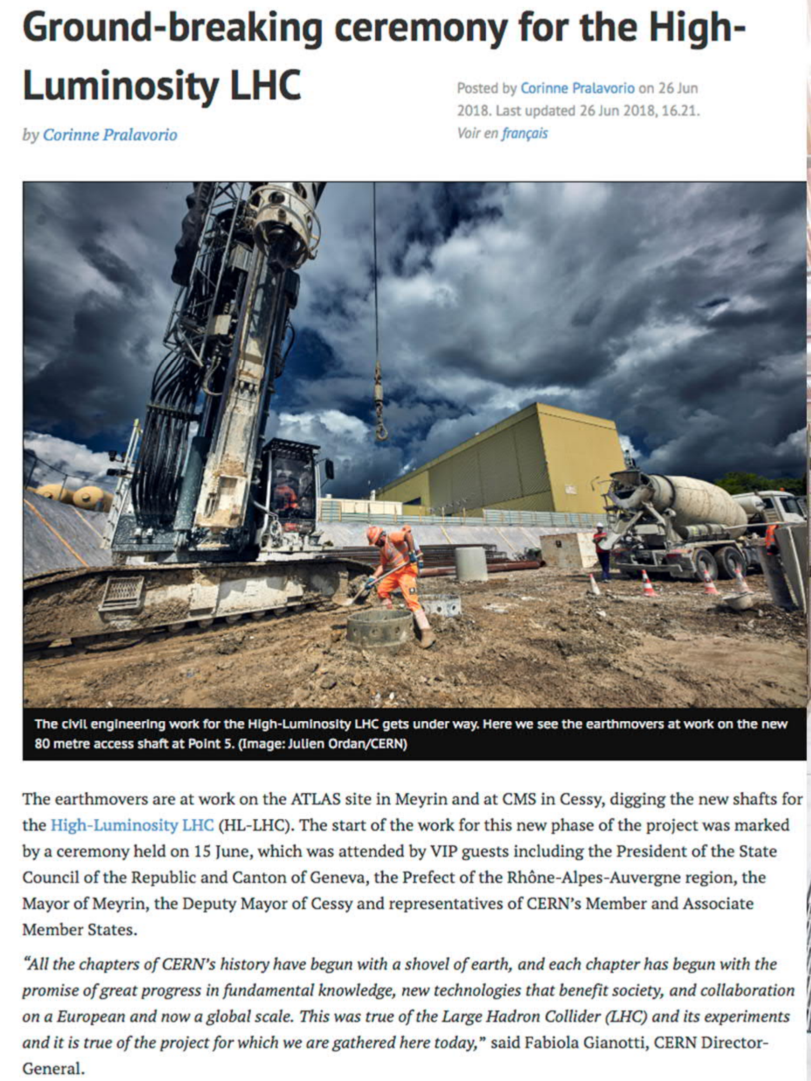
come si gestiscono i dati prodotti da grandi esperimenti di Fisica

Giuseppe Lo Presti
CERN IT Department

Seminario UniFe, 29/03/2019

Agenda

- The Big Picture
  - Computing and Data Management at CERN
- Big Data at CERN and outside
- Future Prospects
  - The “High-Luminosity” LHC and its challenges



Ground-breaking ceremony for the High-Luminosity LHC

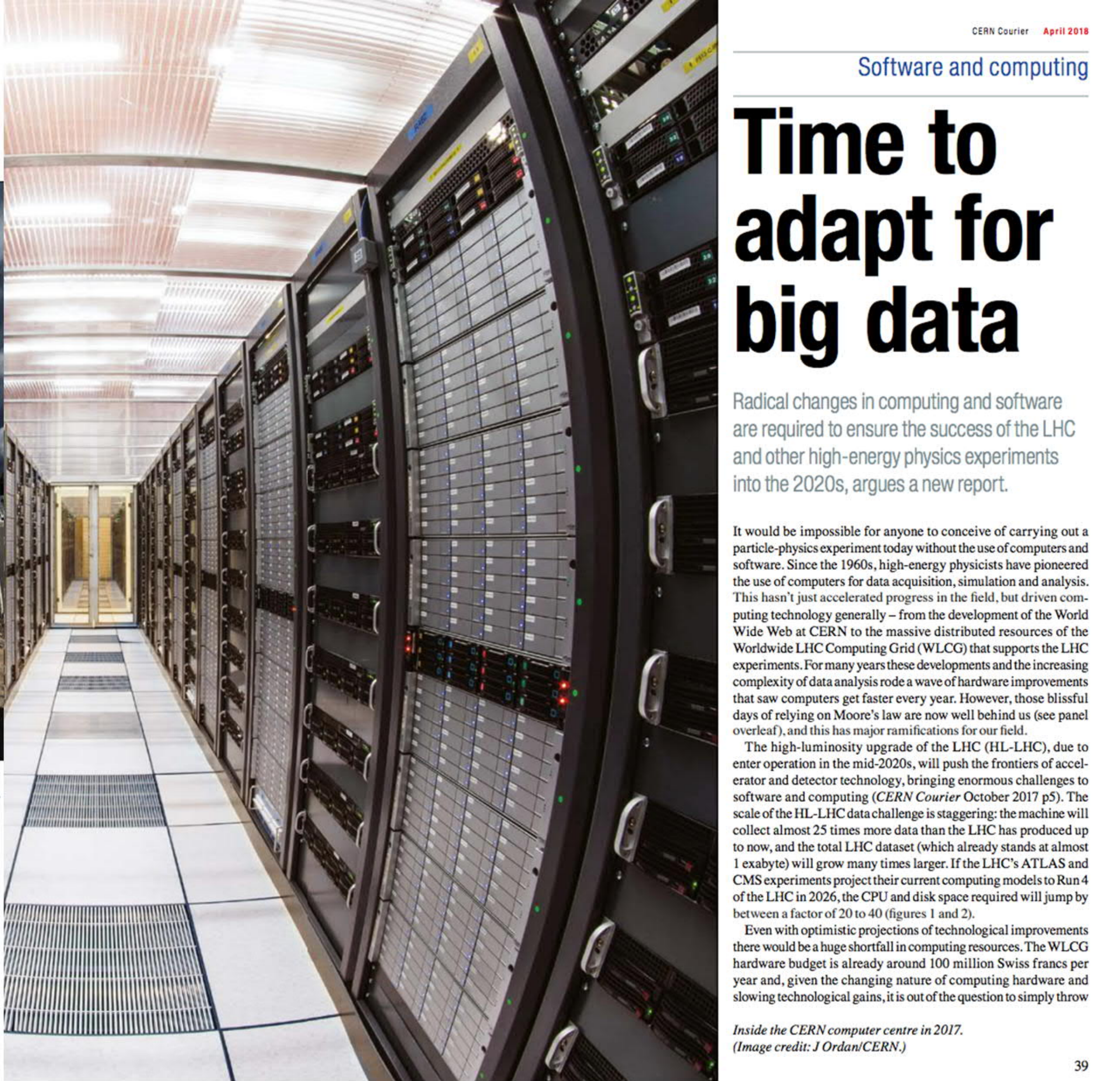
by Corinne Pralavorio

Posted by Corinne Pralavorio on 26 Jun 2018. Last updated 26 Jun 2018, 16.21.
Voir en *français*

The civil engineering work for the High-Luminosity LHC gets under way. Here we see the earthmovers at work on the new 80 metre access shaft at Point 5. (Image: Julien Ordan/CERN)

The earthmovers are at work on the ATLAS site in Meyrin and at CMS in Cessy, digging the new shafts for the High-Luminosity LHC (HL-LHC). The start of the work for this new phase of the project was marked by a ceremony held on 15 June, which was attended by VIP guests including the President of the State Council of the Republic and Canton of Geneva, the Prefect of the Rhône-Alpes-Auvergne region, the Mayor of Meyrin, the Deputy Mayor of Cessy and representatives of CERN's Member and Associate Member States.

"All the chapters of CERN's history have begun with a shovel of earth, and each chapter has begun with the promise of great progress in fundamental knowledge, new technologies that benefit society, and collaboration on a European and now a global scale. This was true of the Large Hadron Collider (LHC) and its experiments and it is true of the project for which we are gathered here today," said Fabiola Gianotti, CERN Director-General.



CERN Courier April 2018

Software and computing

Time to adapt for big data

Radical changes in computing and software are required to ensure the success of the LHC and other high-energy physics experiments into the 2020s, argues a new report.

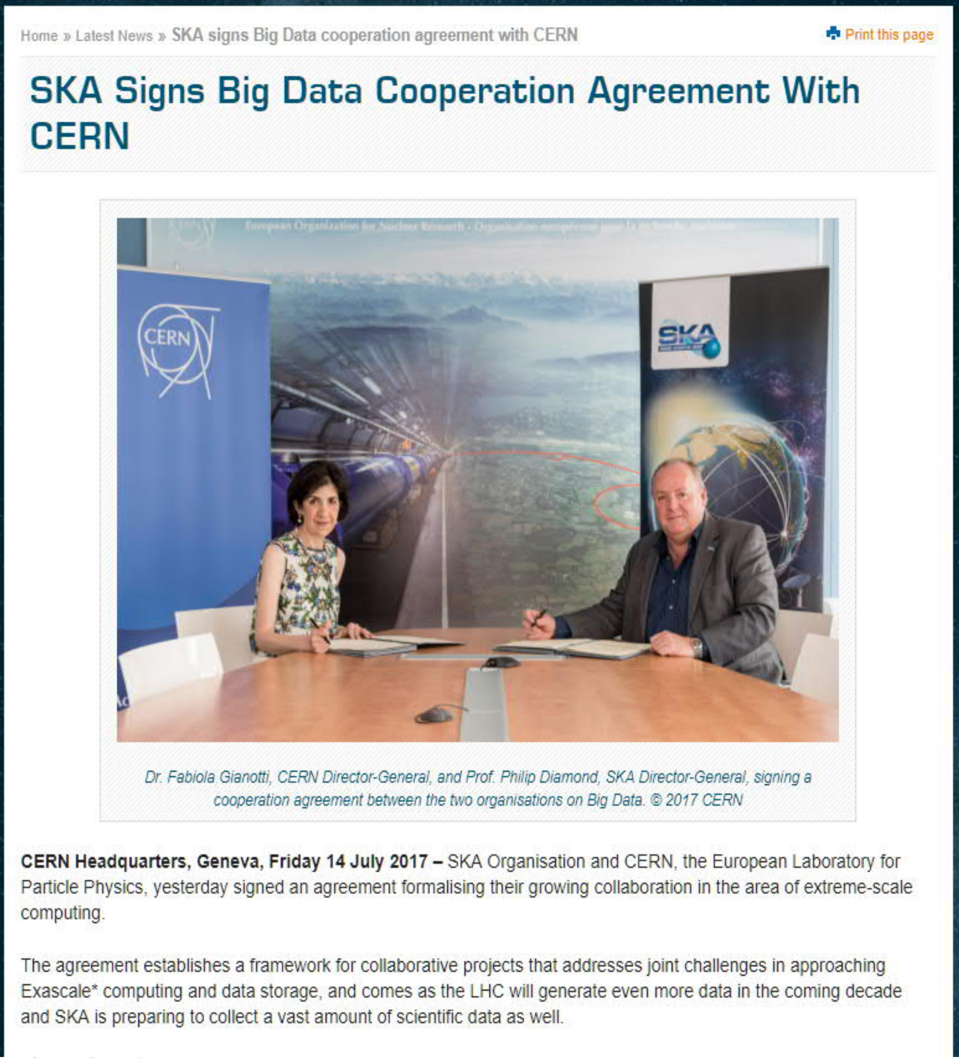
It would be impossible for anyone to conceive of carrying out a particle-physics experiment today without the use of computers and software. Since the 1960s, high-energy physicists have pioneered the use of computers for data acquisition, simulation and analysis. This hasn't just accelerated progress in the field, but driven computing technology generally – from the development of the World Wide Web at CERN to the massive distributed resources of the Worldwide LHC Computing Grid (WLCG) that supports the LHC experiments. For many years these developments and the increasing complexity of data analysis rode a wave of hardware improvements that saw computers get faster every year. However, those blissful days of relying on Moore's law are now well behind us (see panel overleaf), and this has major ramifications for our field.

The high-luminosity upgrade of the LHC (HL-LHC), due to enter operation in the mid-2020s, will push the frontiers of accelerator and detector technology, bringing enormous challenges to software and computing (*CERN Courier* October 2017 p5). The scale of the HL-LHC data challenge is staggering: the machine will collect almost 25 times more data than the LHC has produced up to now, and the total LHC dataset (which already stands at almost 1 exabyte) will grow many times larger. If the LHC's ATLAS and CMS experiments project their current computing models to Run 4 of the LHC in 2026, the CPU and disk space required will jump by between a factor of 20 to 40 (figures 1 and 2).

Even with optimistic projections of technological improvements there would be a huge shortfall in computing resources. The WLCG hardware budget is already around 100 million Swiss francs per year and, given the changing nature of computing hardware and slowing technological gains, it is out of the question to simply throw

Inside the CERN computer centre in 2017. (Image credit: J Ordan/CERN.)

39



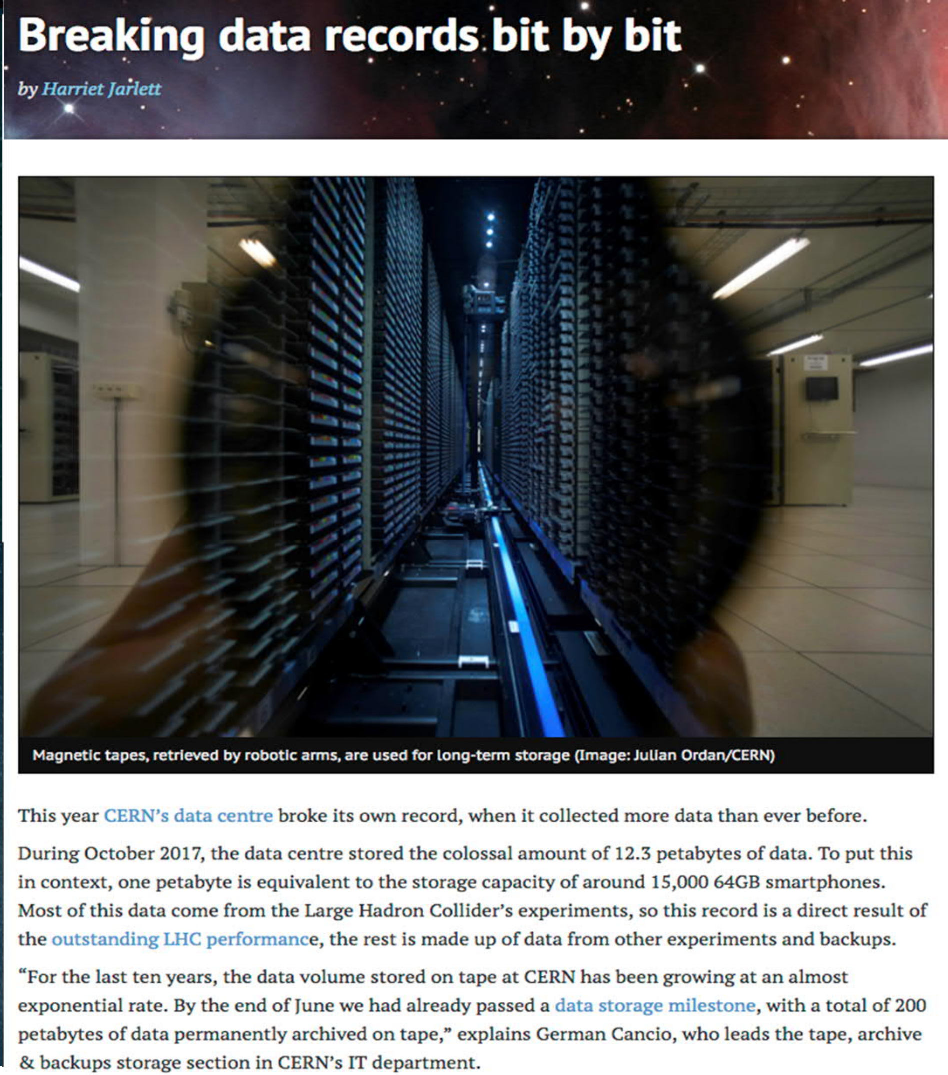
Home » Latest News » SKA signs Big Data cooperation agreement with CERN
Print this page

SKA Signs Big Data Cooperation Agreement With CERN

Dr. Fabiola Gianotti, CERN Director-General, and Prof. Philip Diamond, SKA Director-General, signing a cooperation agreement between the two organisations on Big Data. © 2017 CERN

CERN Headquarters, Geneva, Friday 14 July 2017 – SKA Organisation and CERN, the European Laboratory for Particle Physics, yesterday signed an agreement formalising their growing collaboration in the area of extreme-scale computing.

The agreement establishes a framework for collaborative projects that addresses joint challenges in approaching Exascale* computing and data storage, and comes as the LHC will generate even more data in the coming decade and SKA is preparing to collect a vast amount of scientific data as well.



Breaking data records bit by bit

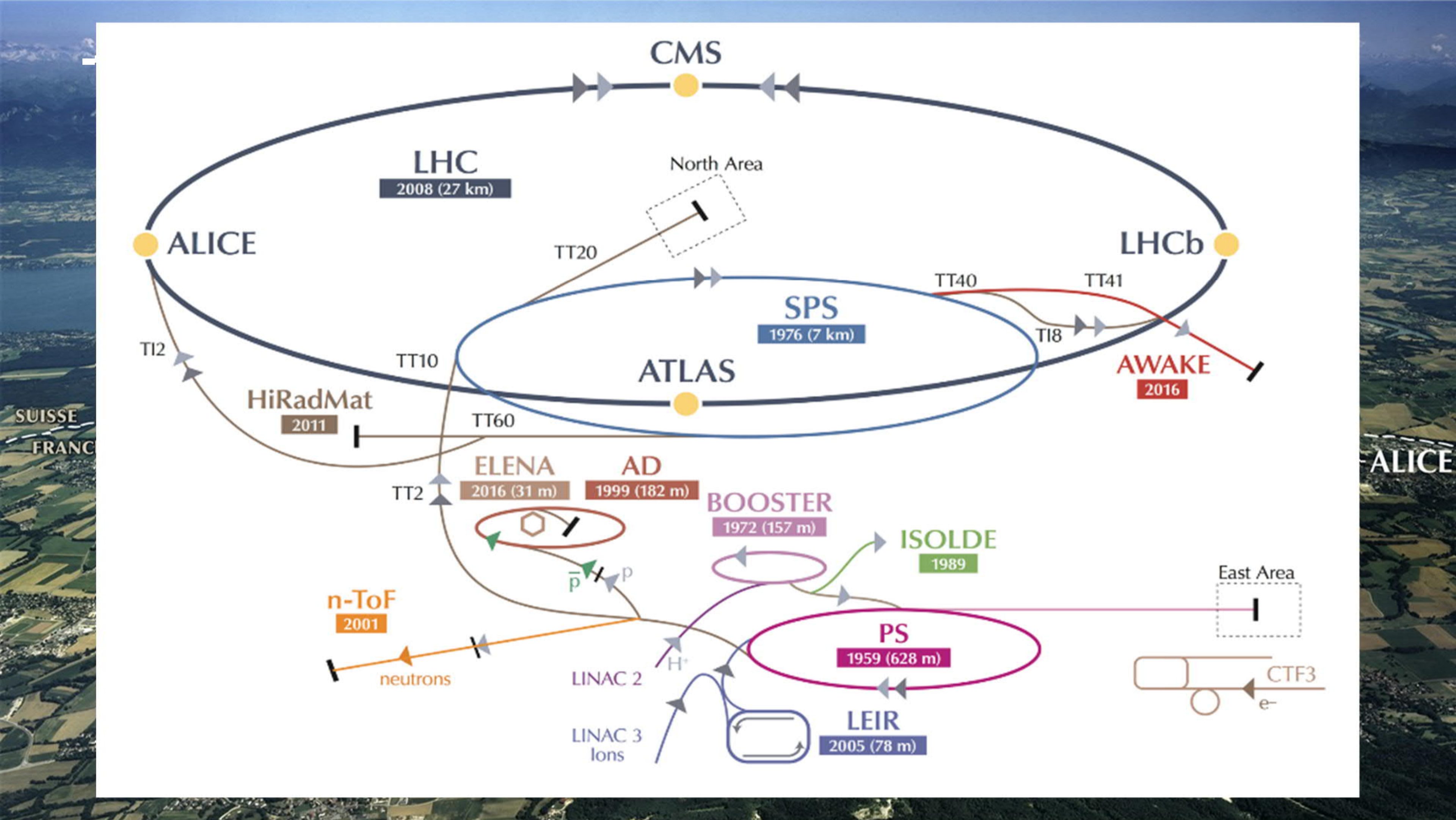
by Harriet Jarlett

Magnetic tapes, retrieved by robotic arms, are used for long-term storage (Image: Julian Ordan/CERN)

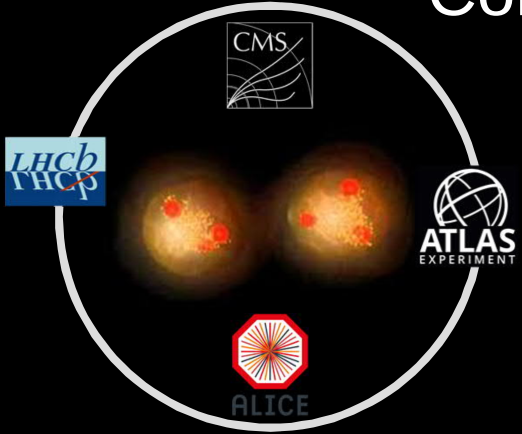
This year CERN's data centre broke its own record, when it collected more data than ever before.

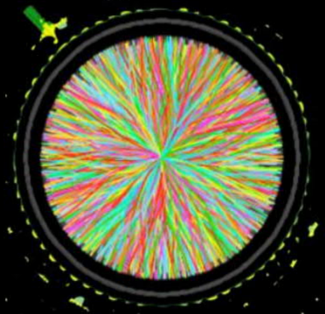
During October 2017, the data centre stored the colossal amount of 12.3 petabytes of data. To put this in context, one petabyte is equivalent to the storage capacity of around 15,000 64GB smartphones. Most of this data come from the Large Hadron Collider's experiments, so this record is a direct result of the outstanding LHC performance, the rest is made up of data from other experiments and backups.

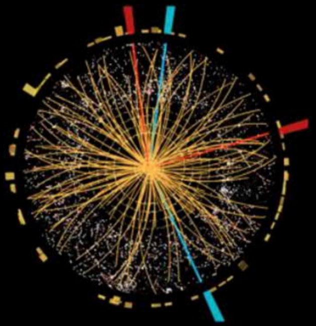
"For the last ten years, the data volume stored on tape at CERN has been growing at an almost exponential rate. By the end of June we had already passed a data storage milestone, with a total of 200 petabytes of data permanently archived on tape," explains German Cancio, who leads the tape, archive & backups storage section in CERN's IT department.

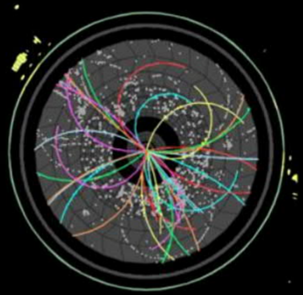

CMS
LHC
2008 (27 km)
North Area
ALICE
LHCb
TT20
TT40
TT41
SPS
1976 (7 km)
TI8
TI2
TT10
ATLAS
AWAKE
2016
HiRadMat
2011
TT60
SUISSE
FRANCE
ALICE
ELENA
2016 (31 m)
AD
1999 (182 m)
TT2
BOOSTER
1972 (157 m)
ISOLDE
1989
$\bar{p}$
p
East Area
n-ToF
2001
PS
1959 (628 m)
neutrons
H^+
LINAC 2
CTF3
e^-
LEIR
2005 (78 m)
LINAC 3
Ions

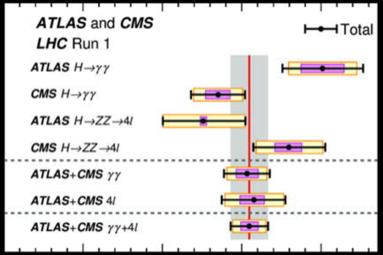
Computing at CERN: The Big Picture





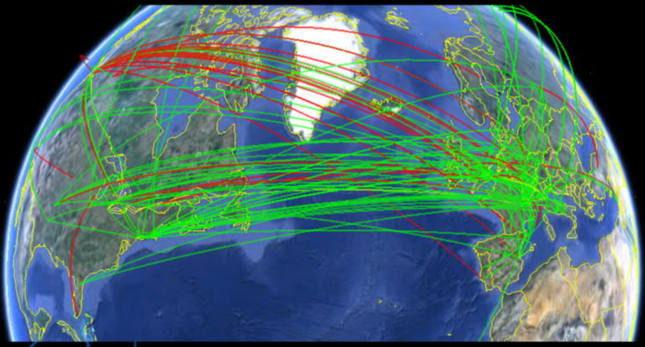






Data Storage - Data Processing - Event generation - Detector simulation - Event reconstruction - Resource accounting

Distributed computing - Middleware - Workload management - Data management - Monitoring



GAUDI-LHCb

ATHENA-ATLAS

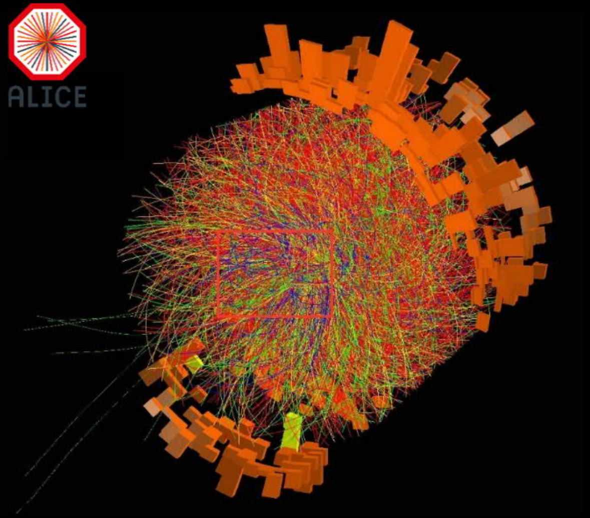
CMSSW-CMS

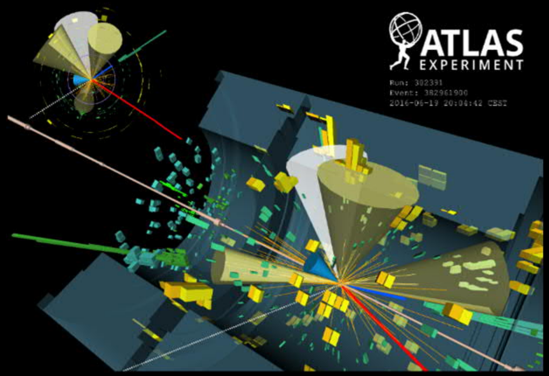
GEANT4
A SIMULATION TOOLKIT

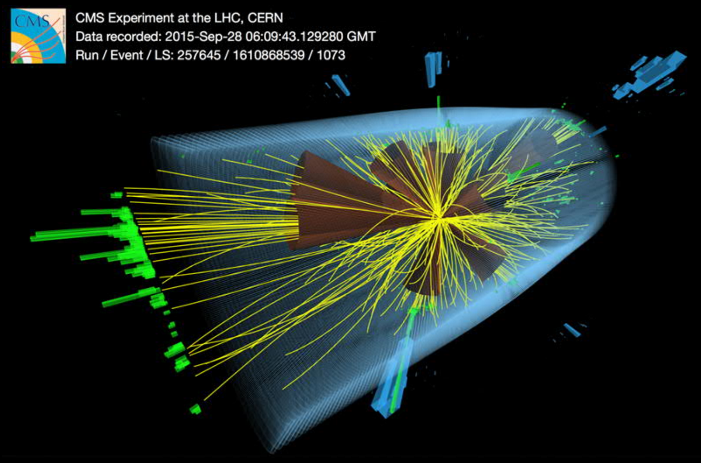
SHERPA

PYTHIA

From the Hit to the Bit: DAQ



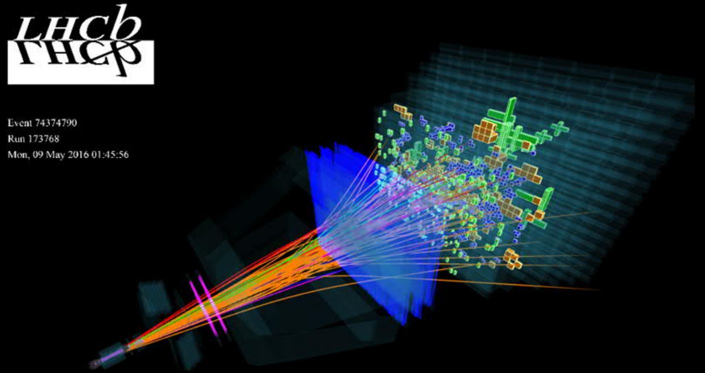



CMS
CMS Experiment at the LHC, CERN
Data recorded: 2015-Sep-28 06:09:43.129280 GMT
Run / Event / LS: 257645 / 1610868539 / 1073

100 million channels

40 million pictures a second

Synchronised signals from all detector parts



From the Hit to the Bit: event filtering

L1: 40 million events per second

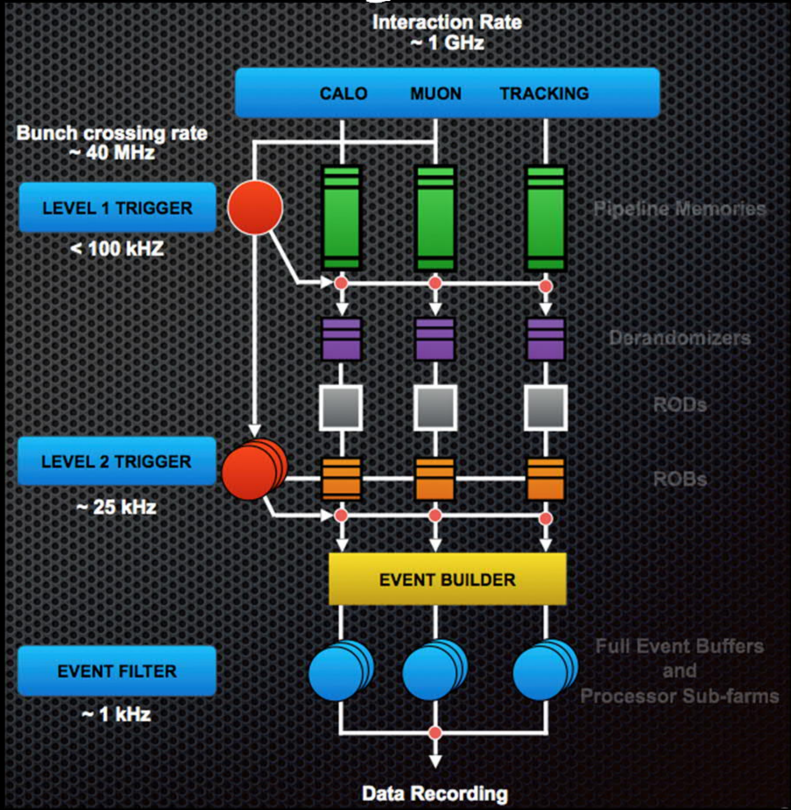
- Fast, simple information
- **Hardware** trigger in a few micro seconds

L2: 100,000 events per second

- Fast algorithms in local computer farm
- **Software** trigger in <1 second

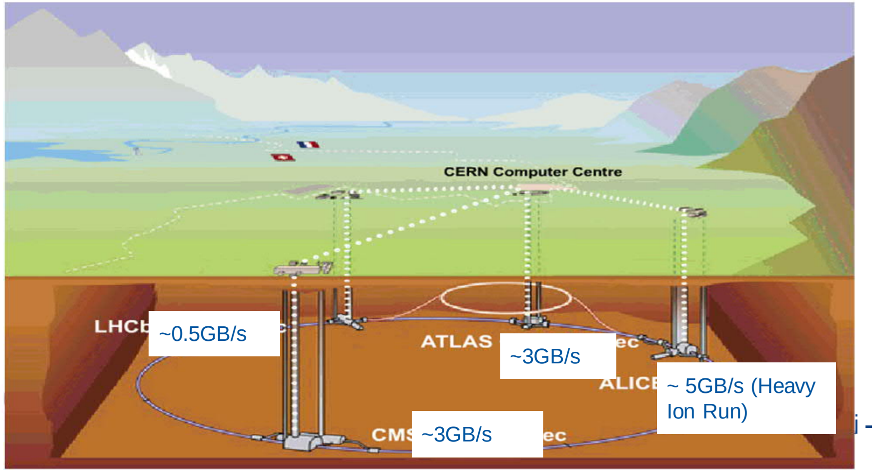
EF: Few 1000s per second recorded for offline analysis

- By each experiment!



Data Processing

- Experiments send over 10 PB of data per r
  - 100PB from all experiments in 2018
- The LHC data is aggregated at the CERN be stored, processed, and distributed

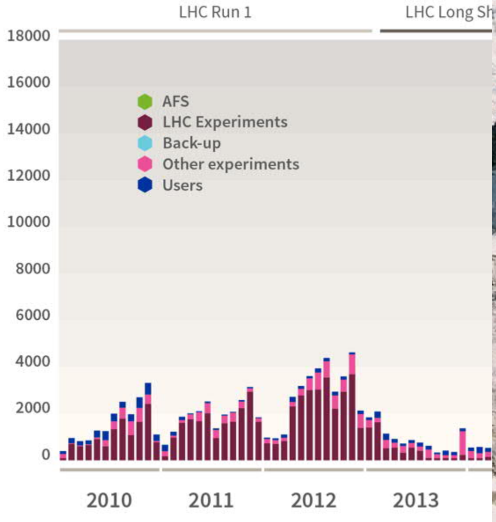


~0.5GB/s

~3GB/s

~ 5GB/s (Heavy Ion Run)

~3GB/s

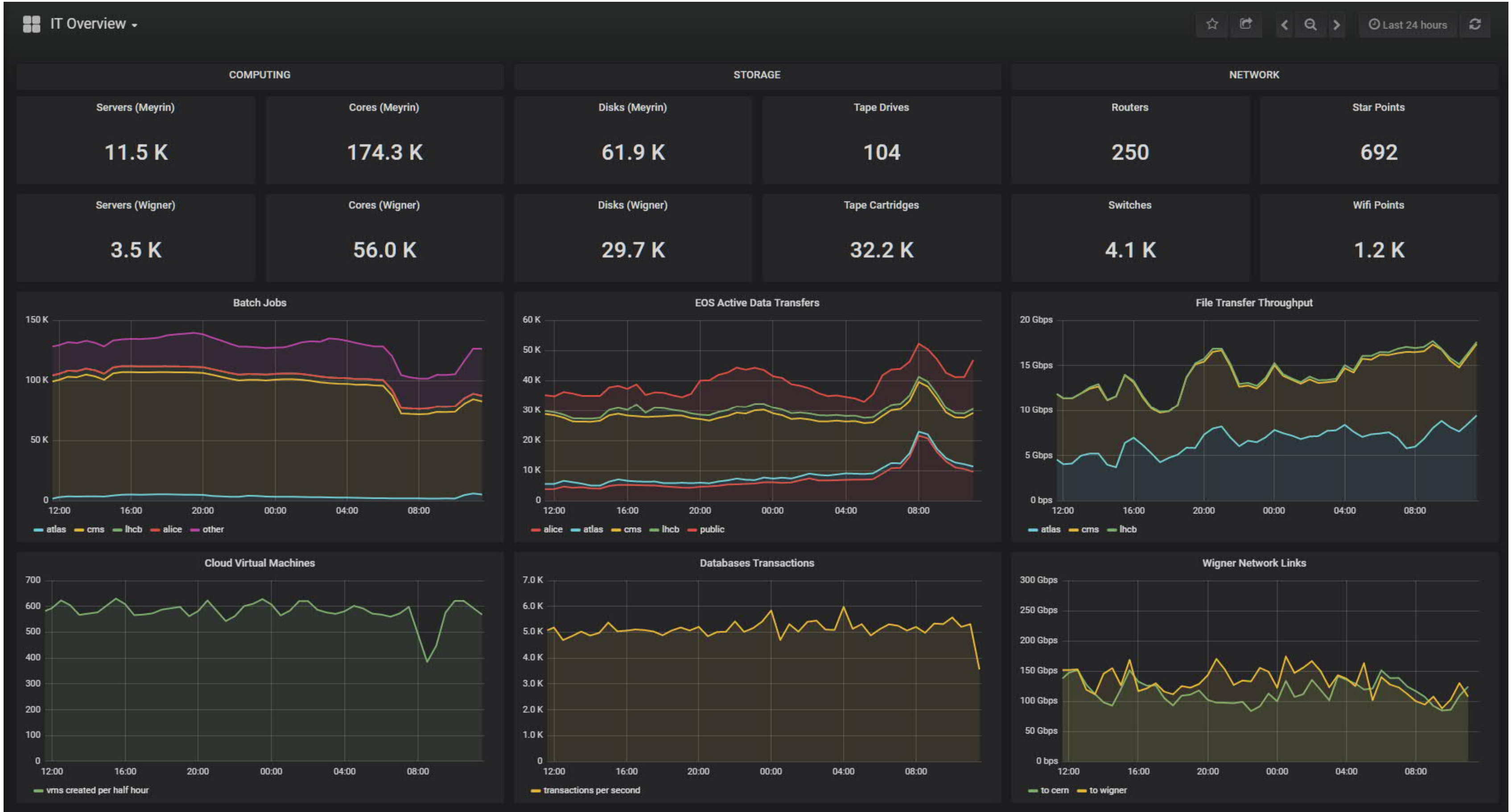


The CERN Data Centre

The CERN Data Centre

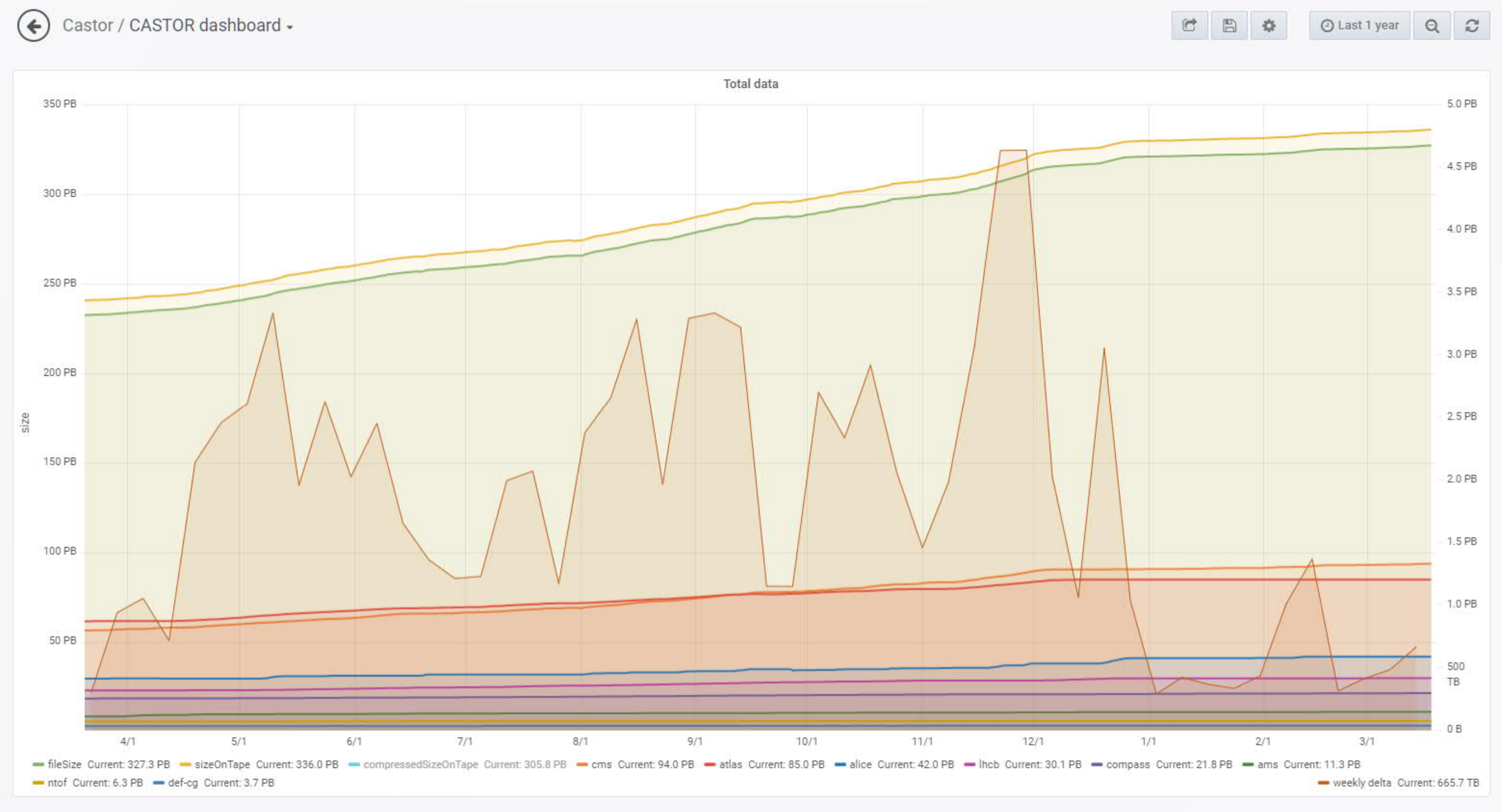
- Built in the 70s on the main CERN site
  - 3.5 MW for equipment
- Extension located at Wigner (Budapest) as of 2013
  - 2.7 MW for equipment
  - Connected to the Geneva CC with 3x100Gbps links (21 and 24 ms RTT)
- Nowadays, hardware generally based on commodity
  - ~**15,000** servers, providing **230,000** processor cores
  - ~**90,000** disk drives providing **280PB** disk space
  - ~**30,000** tapes, providing **0.5EB** capacity
- Power and Heat management: **PUE** (Power Usage Effectiveness) and Green-IT

CERN CC: an ordinary week in numbers



COMPUTING		STORAGE		NETWORK	
Servers (Meyrin)	Cores (Meyrin)	Disks (Meyrin)	Tape Drives	Routers	Star Points
11.5 K	174.3 K	61.9 K	104	250	692
Servers (Wigner)	Cores (Wigner)	Disks (Wigner)	Tape Cartridges	Switches	Wifi Points
3.5 K	56.0 K	29.7 K	32.2 K	4.1 K	1.2 K

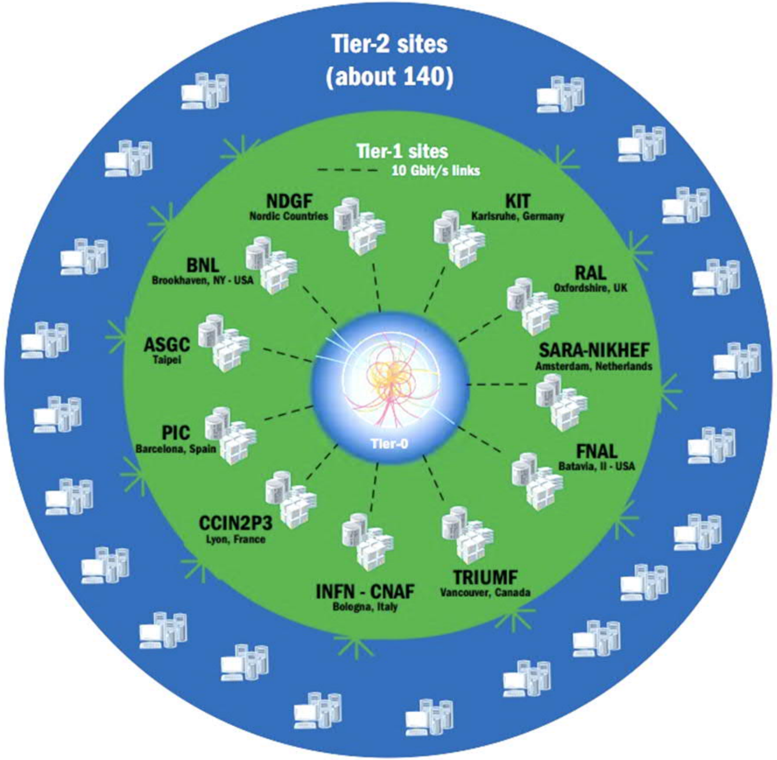
CERN CC: largest scientific data repository



The Worldwide LHC Computing **Grid**

- The Worldwide LHC Computing Grid (WLCG) is a global collaboration of more than 170 data centres around the world, in 42 countries
- The CERN data centre (Tier-0) distributes the LHC data worldwide to the other WLCG sites (Tier-1 and Tier-2)
- WLCG provides global computing resources to store, distribute and analyse the LHC data
  - CERN = only 15% of CPU resources
- The resources are distributed – for funding and sociological reasons

WLCG
Worldwide LHC Computing Grid



Take-away #1

- LHC data rates range from the PB/sec at the detector to the GB/sec after filtering
- Scientific data towards Exabyte scale
    - +100% of LHC data in 2018 vs 2017
- Data centres run on commodity hardware
- Commercial providers are (much) larger
    - CERN remains the world-largest scientific repository
- …Is this really “Big Data”?

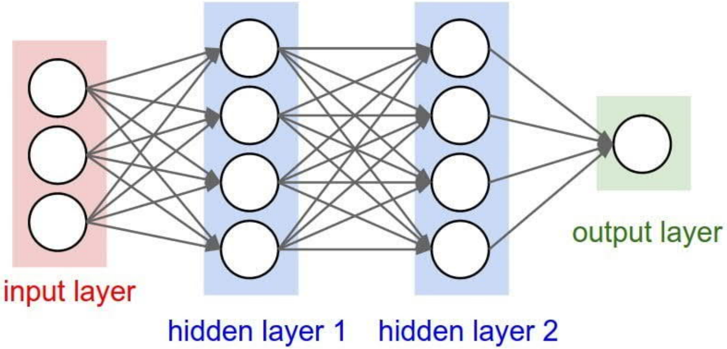
Big Data
and what’s coming next

Big Data

- *Big data* is a field that treats of ways to analyse […] or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software (*Wikipedia*)
  - **Moving target** by definition!
- From structured data, relational DBs, centralized processing…
- To ***unstructured*** data and decentralized (i.e. parallel and loosely-coupled) processing, more adapted to the Cloud
  - E.g. trend analysis, pattern recognition, image segmentation, natural language interpretation/translation, …

Big Data out there

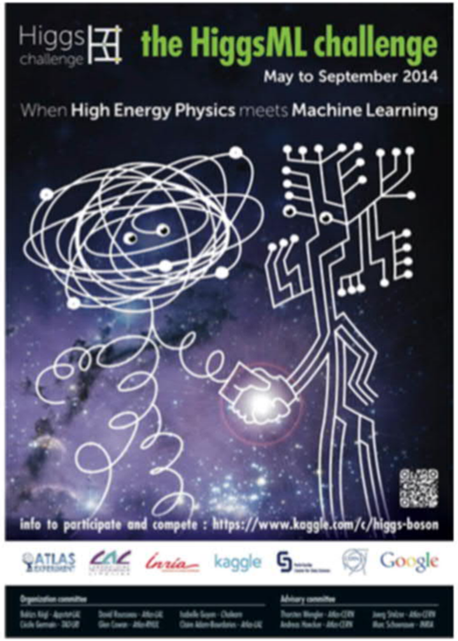
- Increasing interest in Big Data analysis
  - **The Power of Data:** Neural Networks are well known since the 1990s, but it’s only now with **very large** and **easily accessible** data sets that they become effective!
  - Lots of software frameworks for *Deep Machine Learning* with NNs coming up
    - TensorFlow PYTORCH Deep Learning with PyTorch



Big Data at CERN

- Experiments have long used Machine Learning (once called Multi-Variate Analysis) techniques
    - Track reconstruction ~ pattern matching
    - Deep Neural Networks coming to help?

- HiggsML and TrackML Challenges
    - 2018 edition: best results obtained with pure parallel processing, without ML!

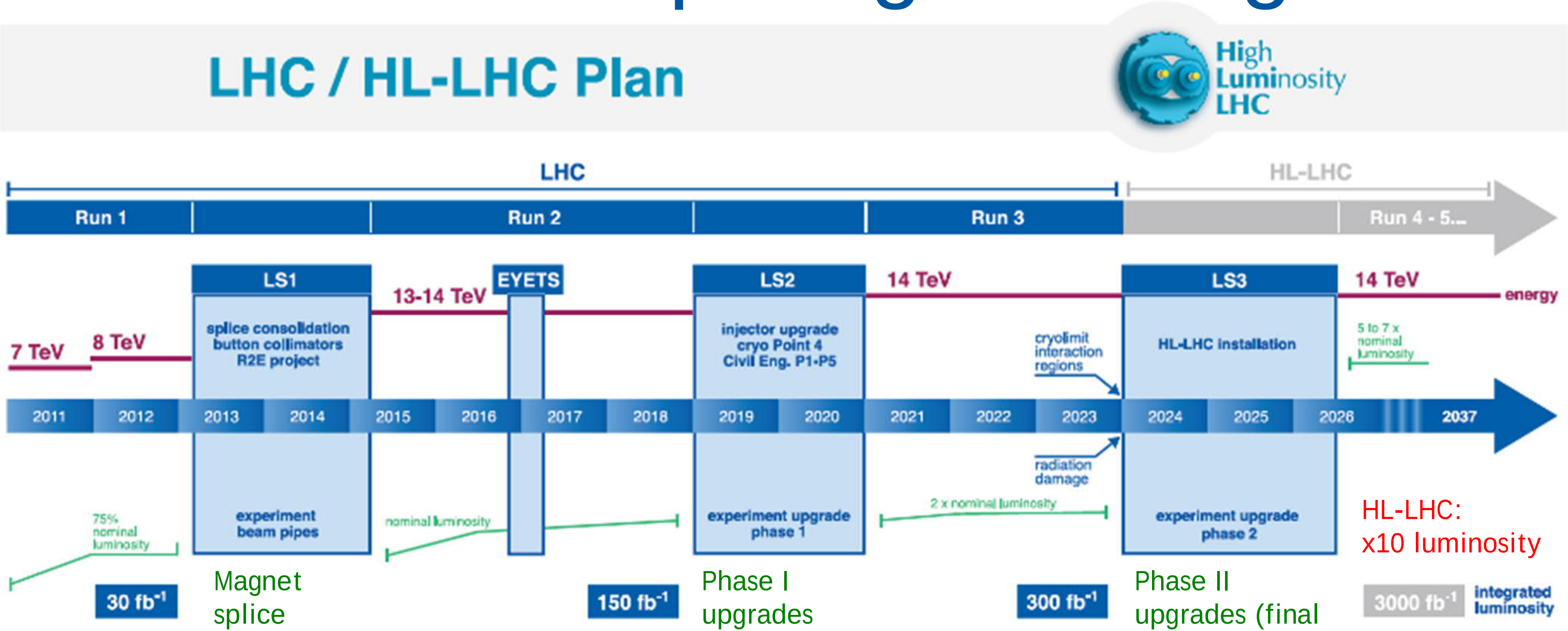

Higgs challenge
the HiggsML challenge
May to September 2014
When High Energy Physics meets Machine Learning
info to participate and compete : https://www.kaggle.com/c/higgs-boson
Organization committee
Advisory committee

Big Data at CERN

- More recently, LHC Beams Control Logging
  - Data migrated from Oracle DB to Hadoop
    - Explosion of data from (connected) sensors
  - Extract trends and detect/predict failures

- In general, ML techniques are getting attention in contexts where analytical approaches are inapplicable/unpractical
  - Security forensics, system analysis/profiling, etc.
    - Typically boiling down to **log analysis**

HL-LHC: a computing challenge

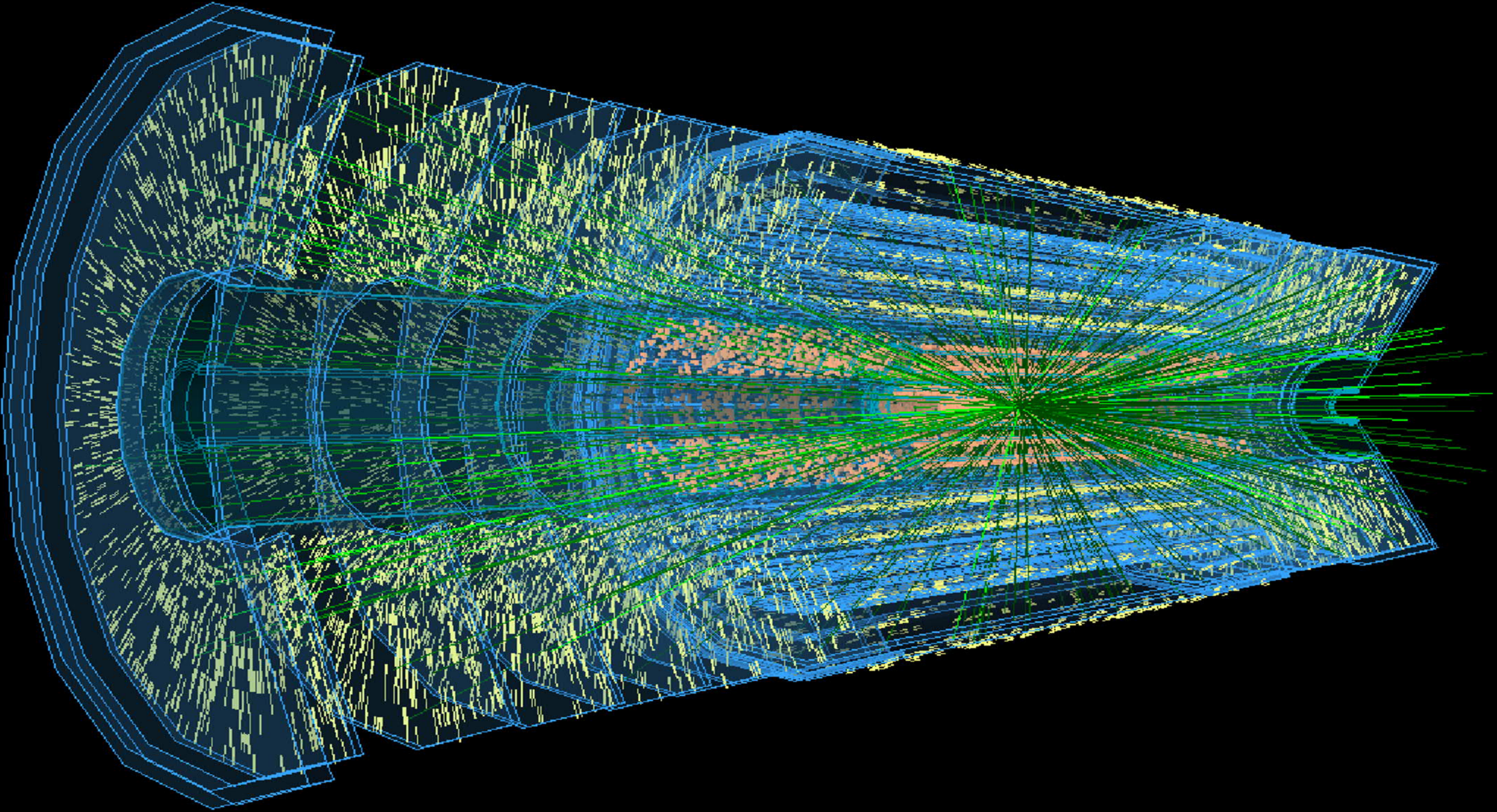


Magnet splice update

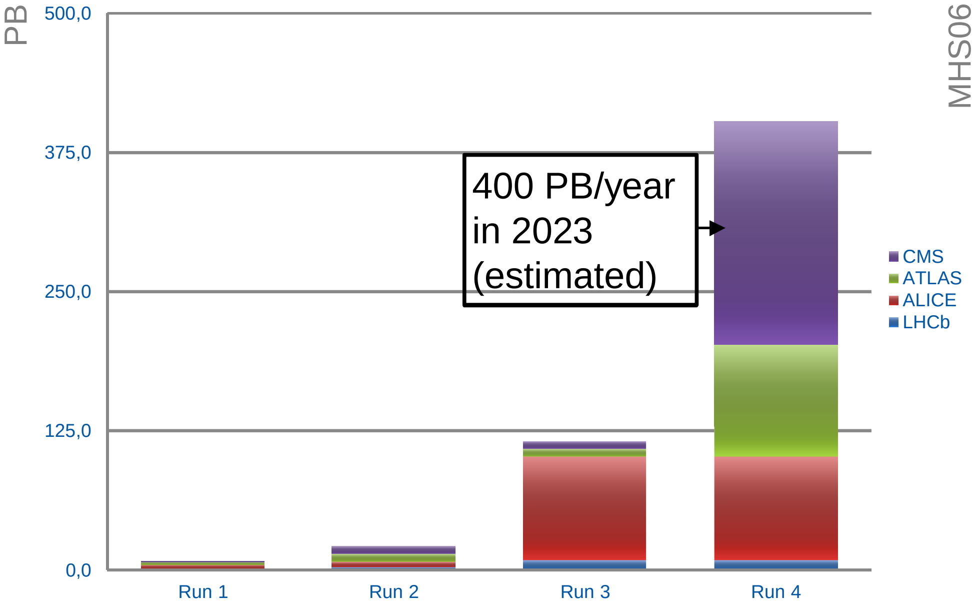
Phase I upgrades (injectors)

Phase II upgrades (final focus)

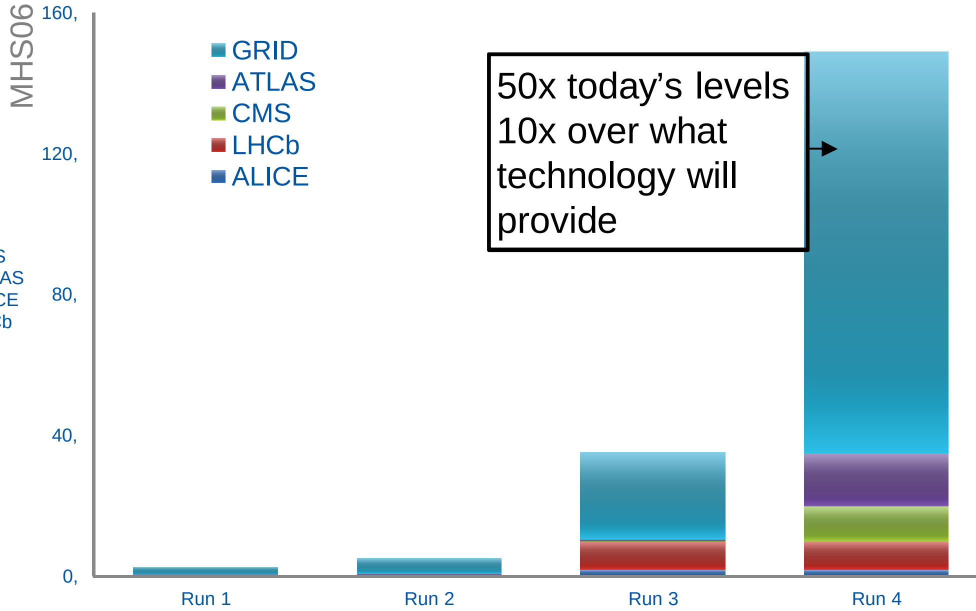
HL-LHC:
x10 luminosity



HL-LHC: a computing challenge



400 PB/year in 2023 (estimated)



50x today’s levels
10x over what technology will provide

HL-LHC and friends

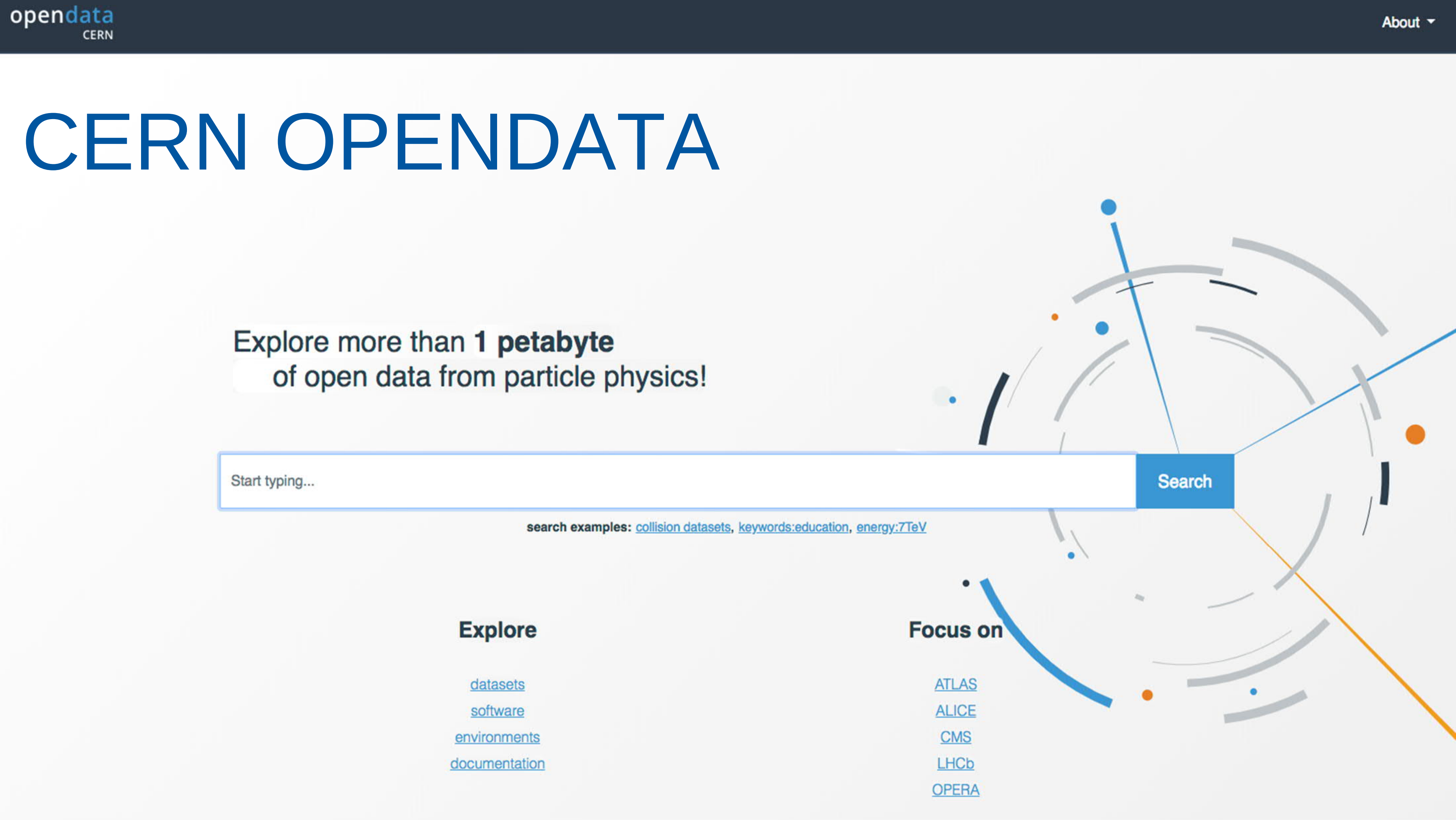
- High Luminosity LHC is not alone in the current arena of large scientific collaborations
- **New Big Science experiments** coming up:
  - Square Kilometer Array (**SKA**)
  - Cherenkov Telescope Array (**CTA**)
  - Deep Underground Neutrino Experiment (**DUNE**): prototype at CERN, full sized experiment in USA

- Time for R&D, opportunity for new **synergies**
  - Typical trend: migrating the 1st Level Trigger from FPGAs to GPUs
  - Increasing role of ML techniques, in particular in other sciences
    - LIGO: GW signal detection

CERN-IT: pushing boundaries

- CERN-IT impact on society through computing:
  - Need for collaboration of computing resources for the Global LHC led to adopt **Grid Computing** and first concept of **Computing Clouds**
- Open access to science
  - Need for sharing the results had led CERN to pave to way to open access to documents and now data: **LHC@home** and **CERN Opendata Portal**
- Openlab
  - "CERN openlab is a unique public-private partnership that accelerates the development of cutting-edge solutions for the worldwide LHC community and wider scientific research"
    - Testing software and hardware
    - **Important student internship program**
- EU projects:
  - HNSciCloud (cloud computing resources), EOSC (data infrastructures)

CERN OPENDATA



Explore more than **1 petabyte** of open data from particle physics!

search examples: collision datasets, keywords:education, energy:7TeV

Explore

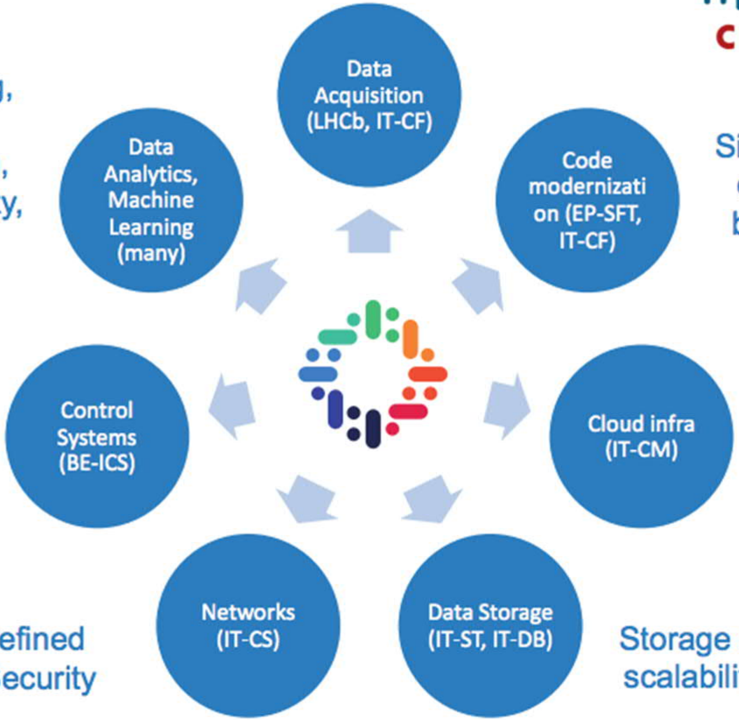
datasets
software
environments
documentation

Focus on

ATLAS
ALICE
CMS
LHCb
OPERA

PHASE V R&D PROJECTS

CERN openlab



Data Acquisition (LHCb, IT-CF)

Code modernization (EP-SFT, IT-CF)

Cloud infra (IT-CM)

Data Storage (IT-ST, IT-DB)

Networks (IT-CS)

Control Systems (BE-ICS)

Data Analytics, Machine Learning (many)

High-bandwidth fabrics, accelerated platforms

intel IDT CISCO

Simulation, HPC on the Cloud, benchmarking

intel HUAWEI Newcastle University

Cloud federations, containers, scalability

HUAWEI rackspace

Storage architectures, scalability, monitoring

ORACLE SEAGATE COMTRADE

Software Defined Networks, Security

BROCADE

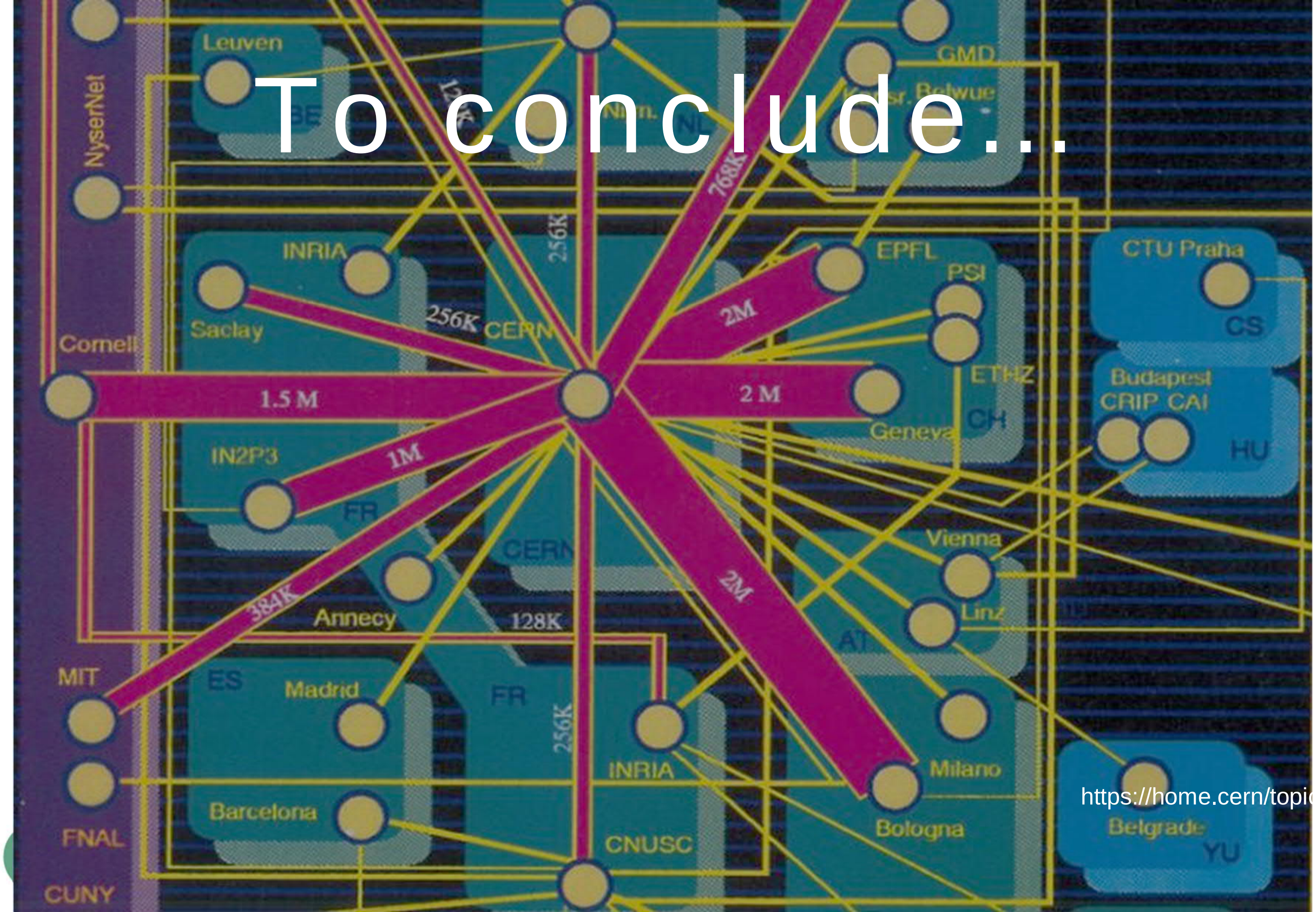
Predictive/proactive maintenance and operations

SIEMENS

Data quality monitoring, anomaly detection, physics data reduction, benchmarking/scalability, systems biology and large-scale multi-disciplinary platforms

intel ORACLE Yandex Scimpulse foundation KING'S College LONDON

To conclude...



https://home.cern/topi

From CERN to the world

- Fundamental Science always pushed technology boundaries, with large returns on investments
- For computing, CERN R&D led for instance to:
  - Invention of the Web (1989, cf. #Web30)
    - Key contribution to the Internet infrastructure
    - **80% of the total European** Internet traffic going through CERN in the late 1980s
  - Touch screens (1972)
    - Super Proton Synchrotron control system required complex controls and developed capacitive touch screen
    - It was based on open standards and moved into industry

…mmm… web + touch-screen: what do you have in your pocket?

Take-away #2

- **Fundamental** Science continues to be main inspiration for **revolutionary** ideas, due to revolutionary needs
    - Industry has well defined offer and demand. We do not. This is the key for **innovation.**

- IT industry has **globally** evolved **beyond our scale**
    - Big Data analysis techniques gaining more and more momentum
        - But there’s no silver bullet !
    - Hard to invent the ‘next Web’, but plenty of room for collaboration and – again – innovation
        - Openlab, EU projects, etc.

Thanks! (More) Questions?