Big Data

La gestione dell'informazione dal Data Taking al Cloud

Giuseppe Lo Presti
CERN IT Department

Italian Teachers Programme 2022 - Discovery

Ground-breaking ceremony for the High-Luminosity LHC

Posted by Corinne Pralavorio on 26 Jun 2018. Last updated 26 Jun 2018, 16.21.

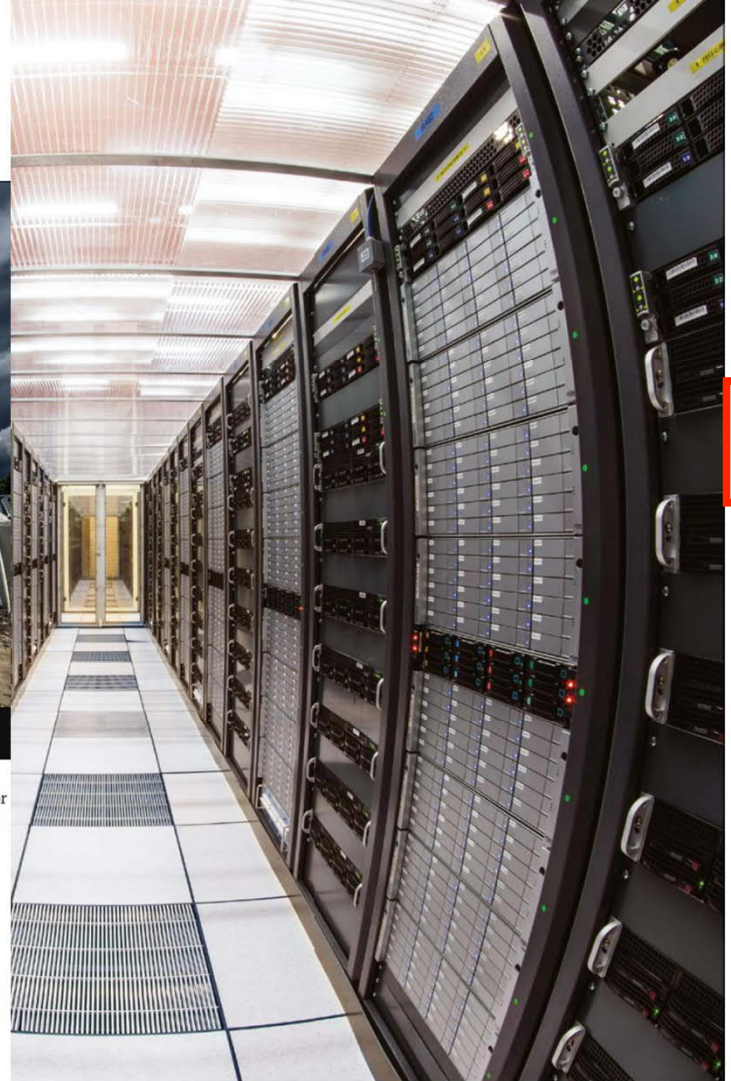
by Corinne Pralavorio

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.



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.

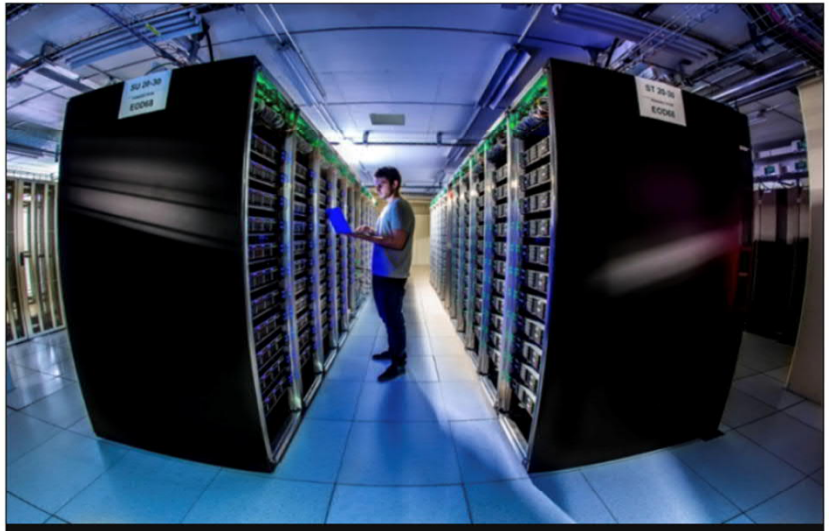
The high-luminosity upgrade of the LHC (HL-LHC), due to enter operation in the mid-2020s, will push the frontiers of accelerator and detector technology, bringing enormous challenges to software and computing (*CERN Courier* October 2017 p5). The scale of the HL-LHC data challenge is staggering: the 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.)

CERN Data Centre passes the 200-petabyte milestone

by *Mélissa Gaillard*

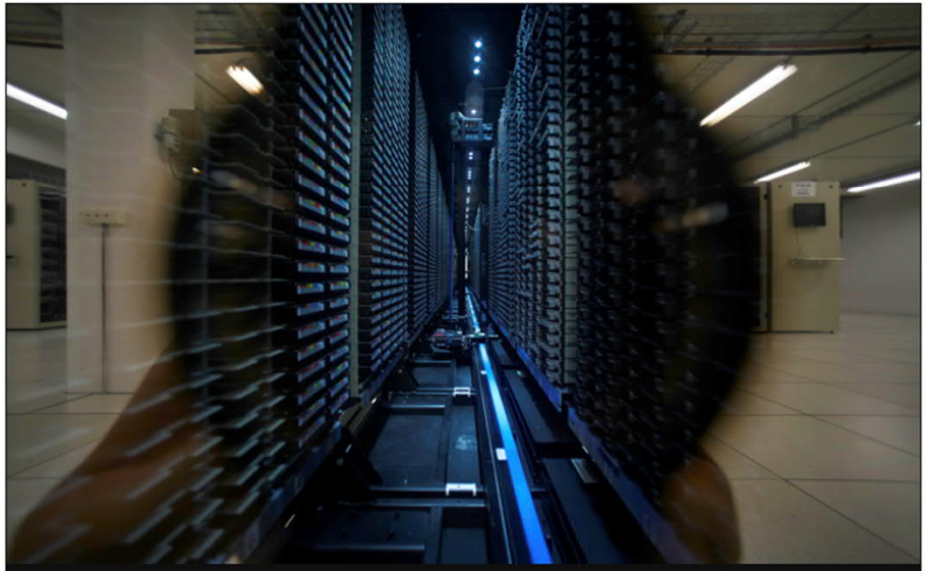


CERN's Data Centre (Image: Robert Hradil, Monika Majer/ProStudio22.ch)

On 29 June 2017, the CERN DC passed the milestone of 200 petabytes of data permanently archived in its tape libraries. Where do these data come from? Particles collide in the Large Hadron Collider (LHC) detectors approximately 1 billion times per second, generating about one petabyte of collision data per second. However, such quantities of data are impossible for current computing systems to record and they are hence filtered by the experiments, keeping only the most "interesting" ones. The filtered LHC data are then aggregated in the CERN Data Centre (DC), where initial data reconstruction is performed, and where a copy is archived to long-term tape storage. Even after the drastic data reduction performed by the experiments, the CERN DC processes on average one petabyte of data per day. This is how the the milestone of 200 petabytes of data permanently archived in its tape libraries was reached on 29 June.

Breaking data records bit by bit

by *Harriet Jarlett*

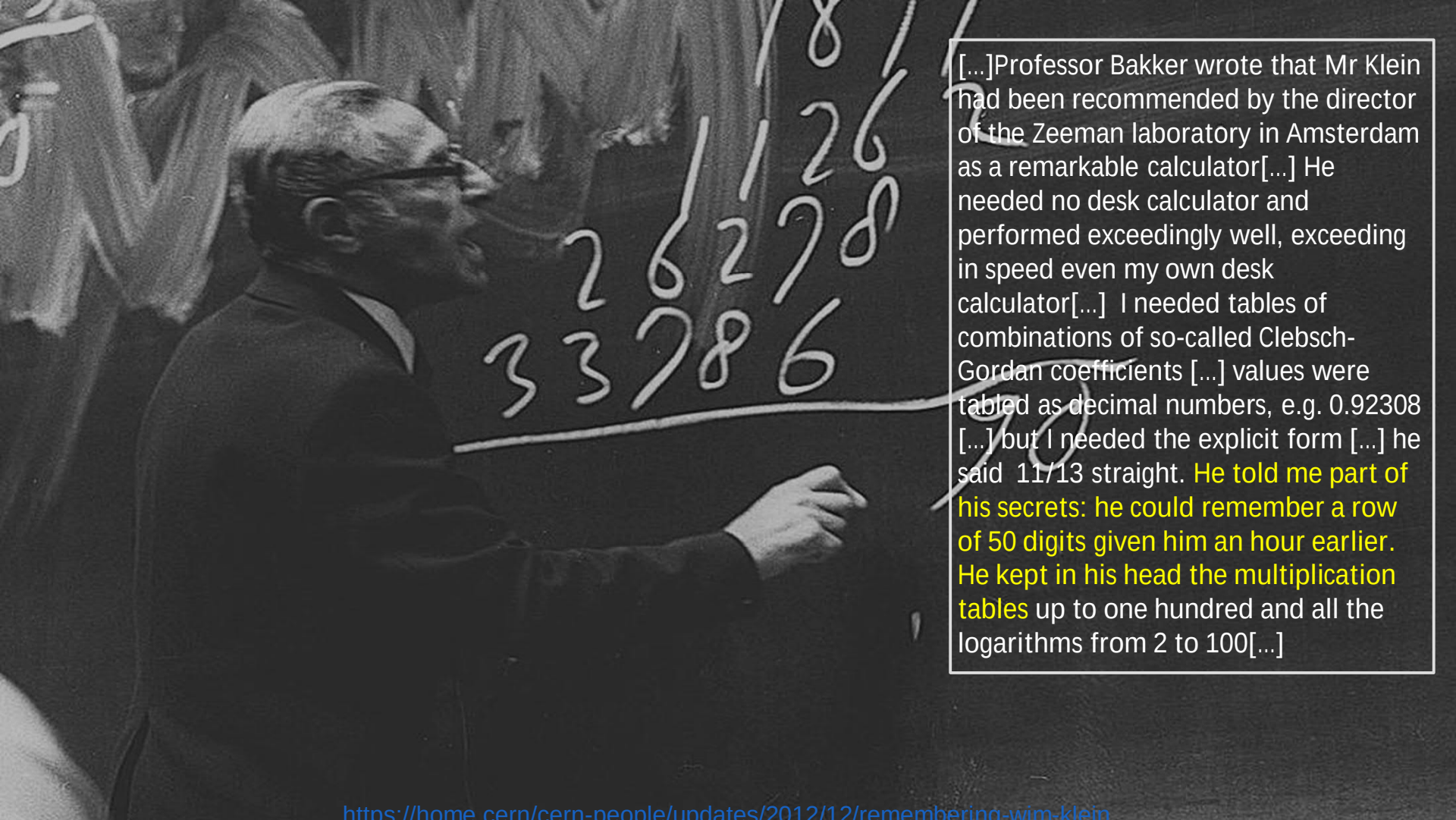


Magnetic tapes, retrieved by robotic arms, are used for long-term storage (Image: Julian Ordan/CERN)

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.

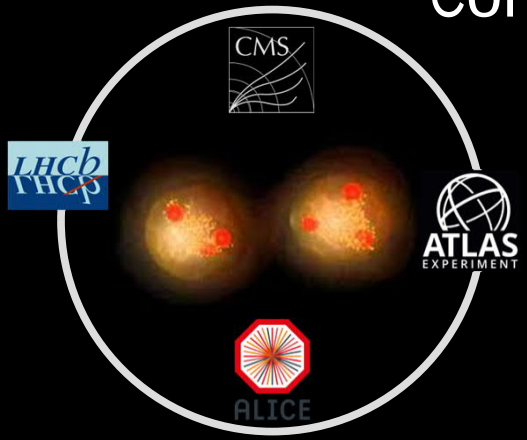
"For the last ten years, the data volume stored on tape at CERN has been growing at an almost exponential rate. By the end of June we had already passed a data storage milestone, with a total of 200 petabytes of data permanently archived on tape," explains German Cancio, who leads the tape, archive &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

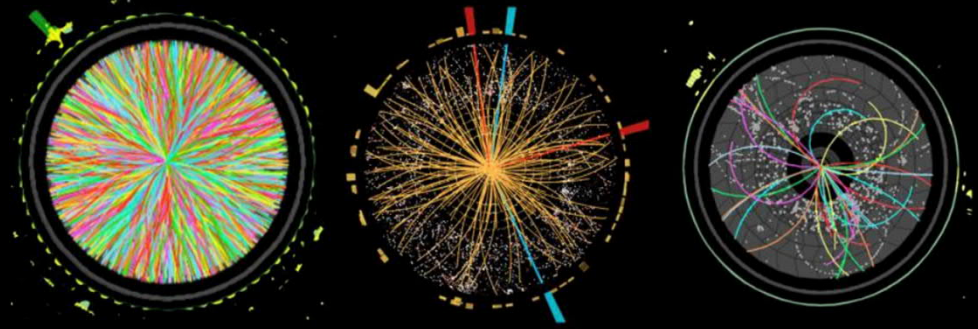


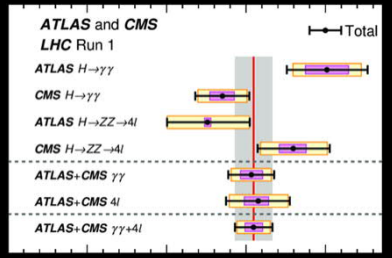
[...]Professor Bakker wrote that Mr Klein had been recommended by the director of the Zeeman laboratory in Amsterdam as a remarkable calculator[...] He needed no desk calculator and performed exceedingly well, exceeding in speed even my own desk calculator[...] I needed tables of combinations of so-called Clebsch-Gordan coefficients [...] values were tabled as decimal numbers, e.g. 0.92308 [...] but I needed the explicit form [...] he said 11/13 straight. He told me part of his secrets: he could remember a row of 50 digits given him an hour earlier. He kept in his head the multiplication tables up to one hundred and all the logarithms from 2 to 100[...]

https://home.cern/cern-people/updates/2012/12/remembering-wim-klein

Computing at CERN: The Big Picture

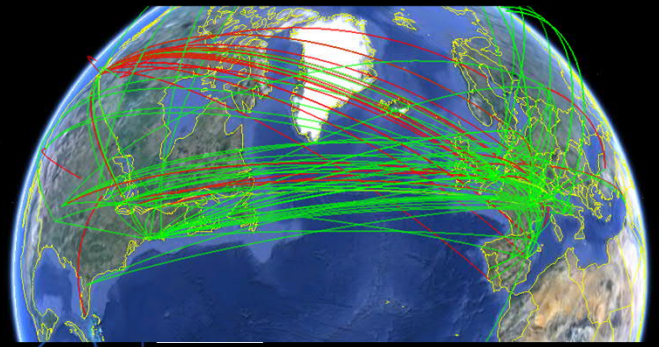






Data Storage - Data Processing - Event generation - Detector simulation - Event reconstruction - Resource accounting

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



GAUDI-LHCb

ATHENA-ATLAS

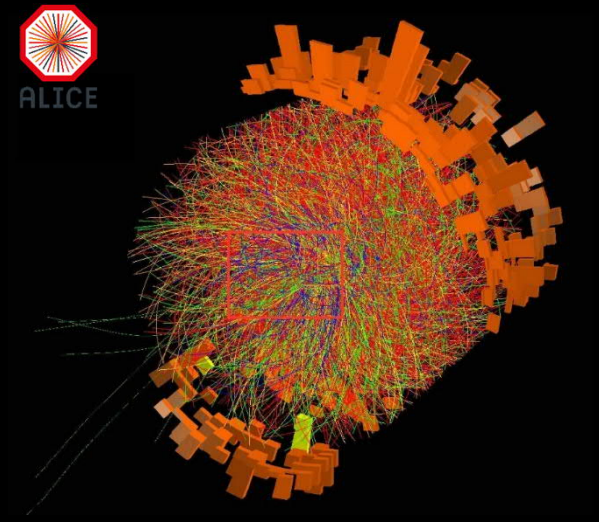
CMSSW-CMS

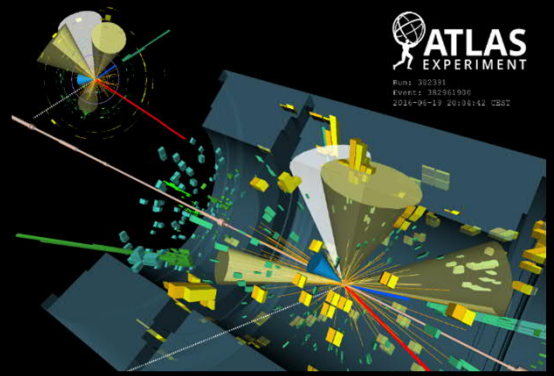
GEANT4
A SIMULATION TOOLKIT

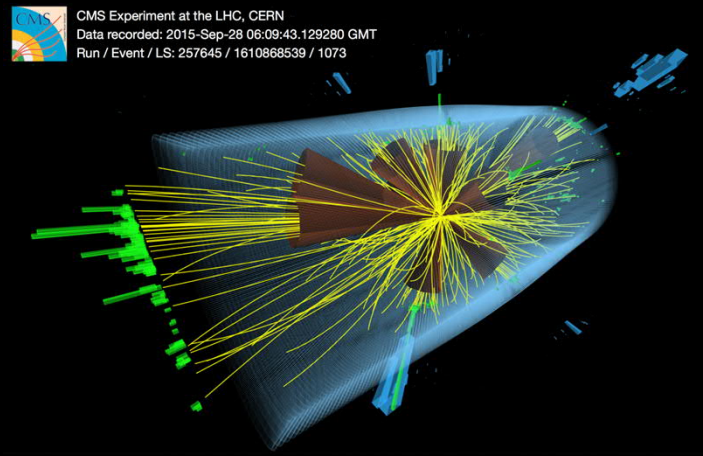
SHERPA

PYTHIA

From the Hit to the Bit: Data Acquisition







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

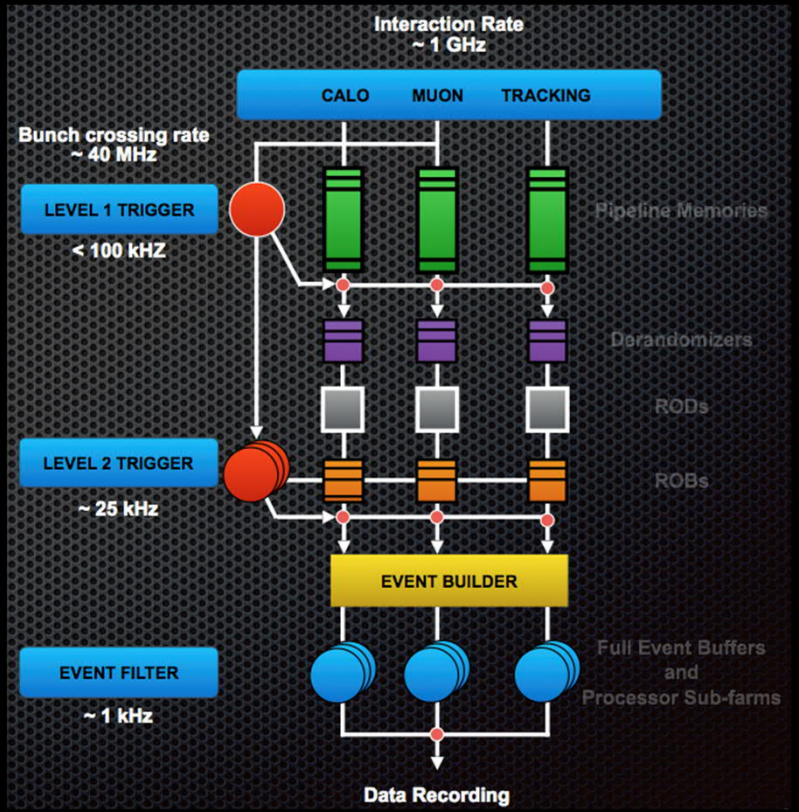
- Fast, simple information
- **Hardware** trigger in a few micro seconds

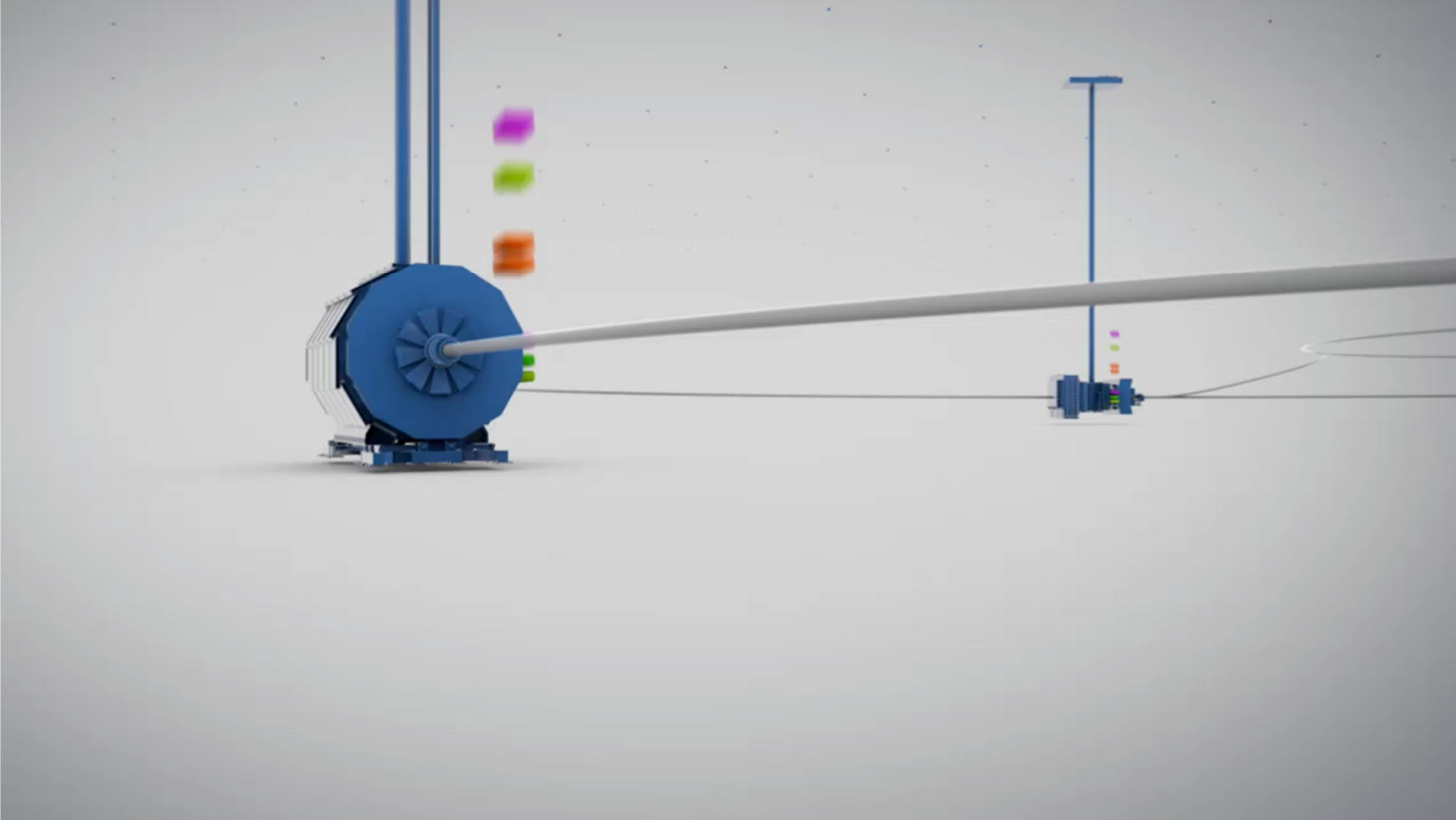
L2: 100,000 events per second

- Fast algorithms in local computer farm
- **Software** trigger in <1 second
- ***Which OS for such task?***

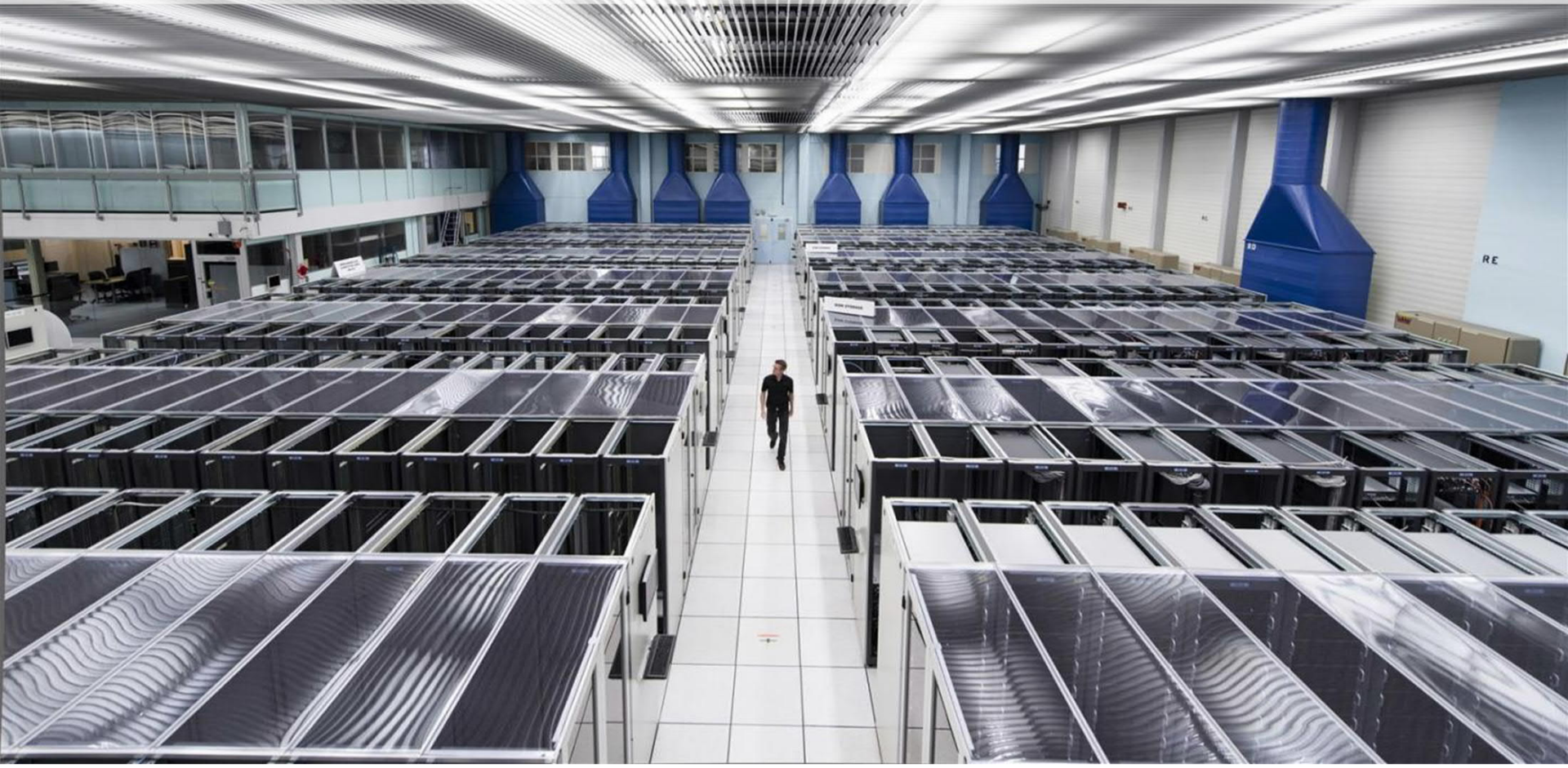
EF: Few 1000s per second recorded for offline analysis

- By each experiment!

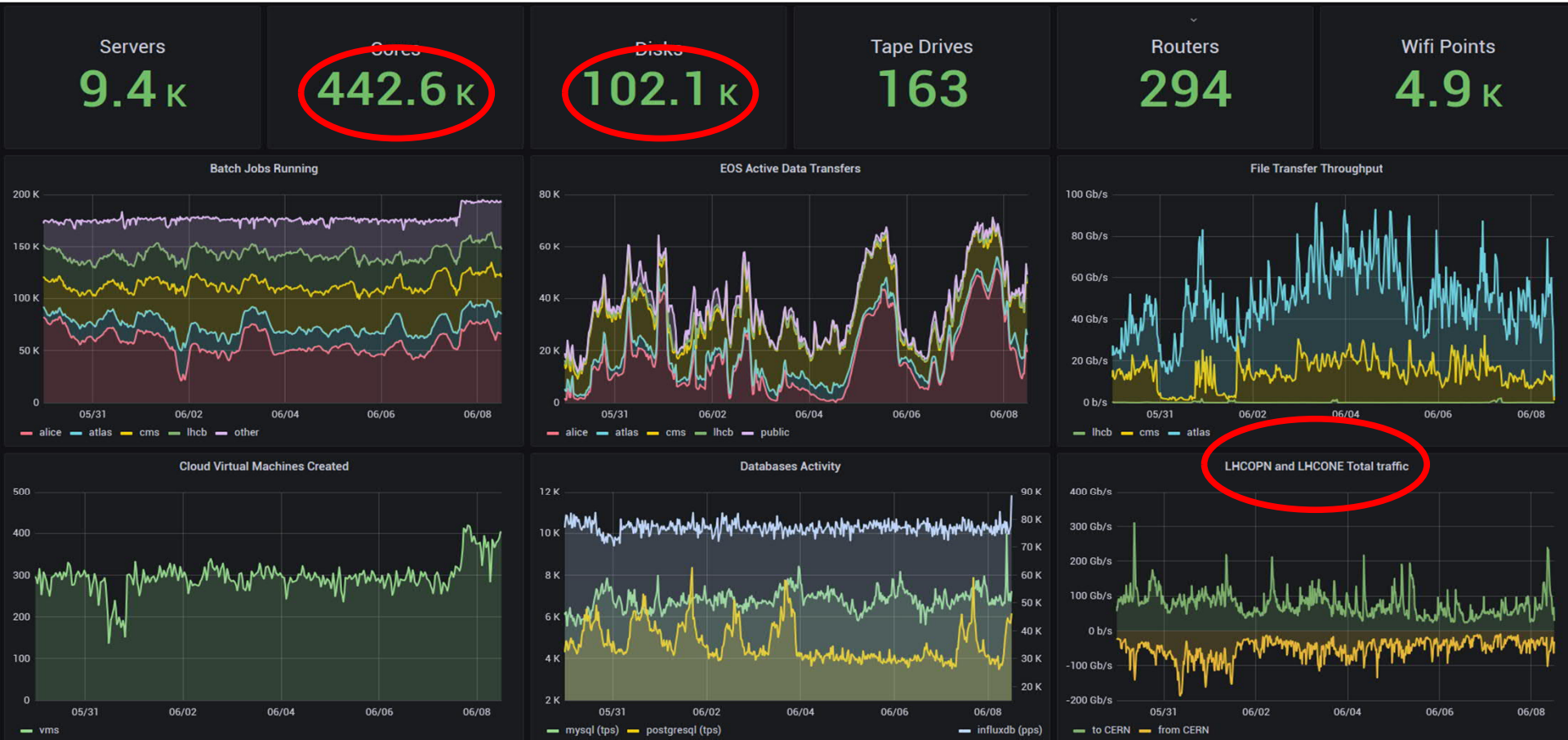




The CERN Data Centre

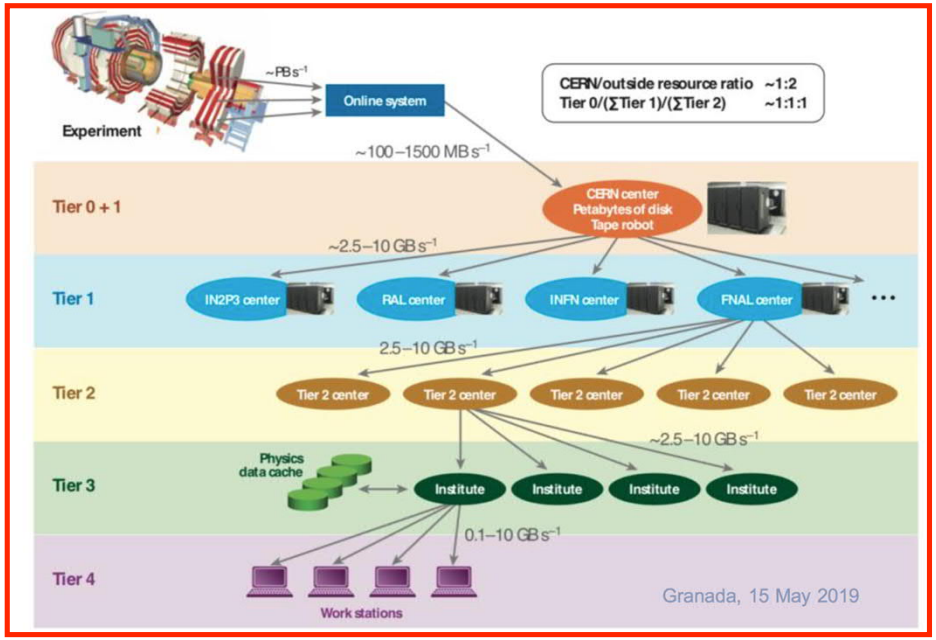


CERN DC: an ordinary week in numbers



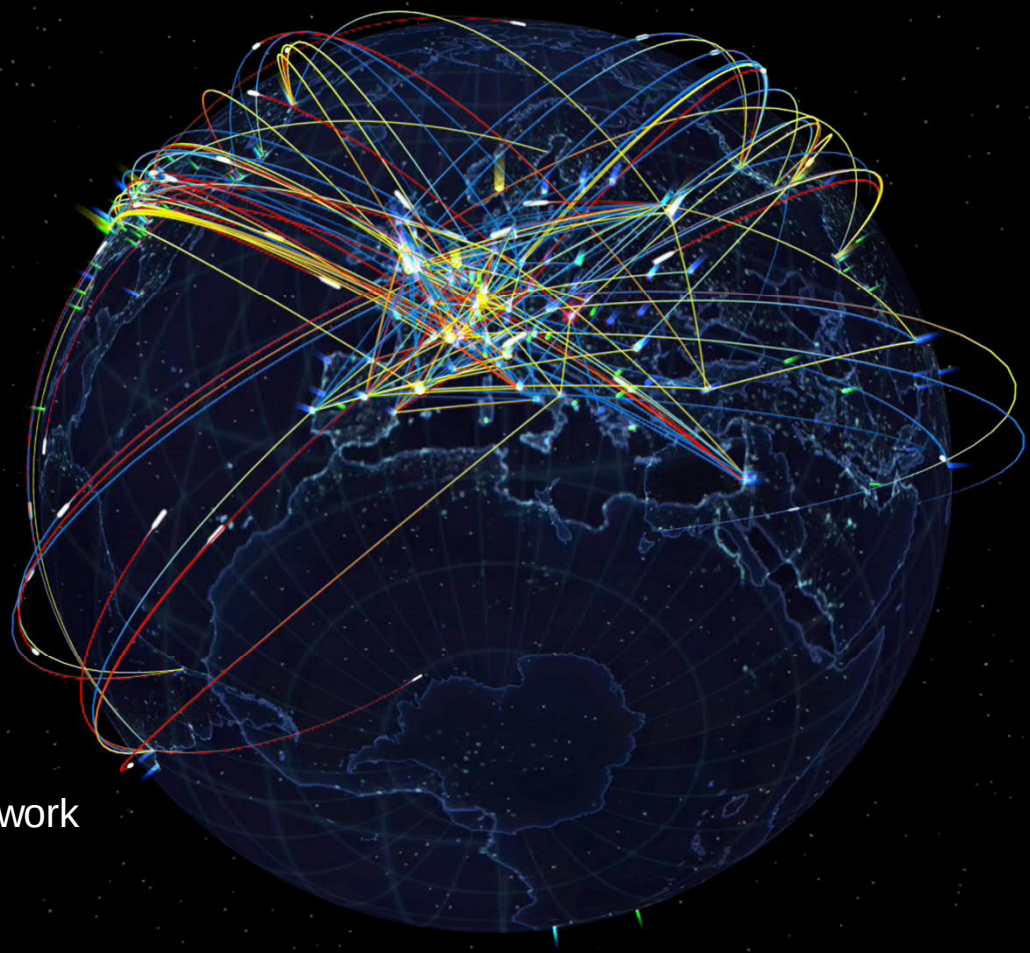
The Worldwide LHC Computing **Grid**

- 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
  - Distributed funding
  - “Sociological” reasons



Data Distribution in WLCG

- Global transfer rates regularly exceeding **60 GB/s**
- **830 PB** and 1.1B files transferred until end of LHC Run 2 (2010-2018)
- Main **challenge** is to have the **useful data close** to available computing resources
  => match storage/compute/network



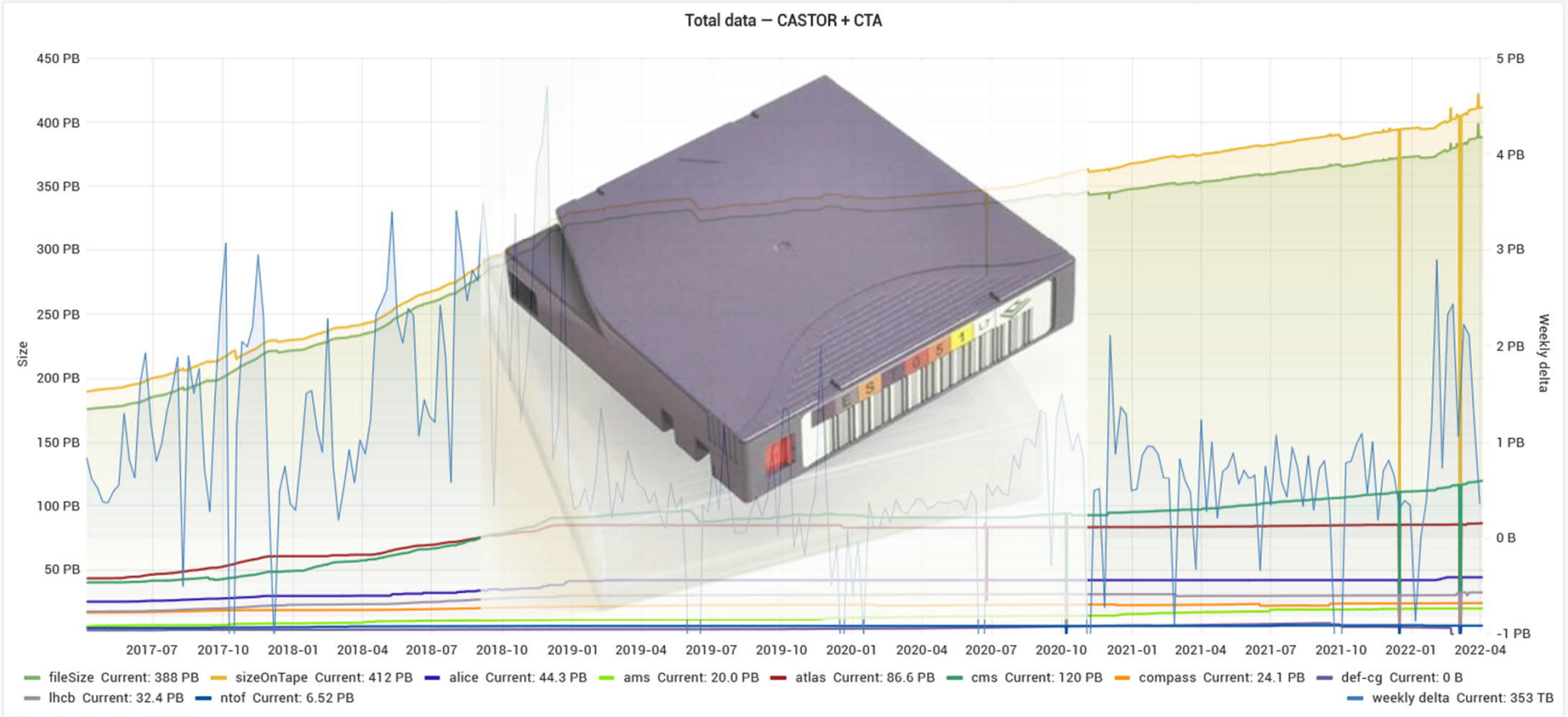
Software Platforms for HEP

- Home made solutions vs. integrating software platforms from the (open source) market
  - Infrastructure moving towards the latter as industry grew in front of us!

Software Platforms for HEP

- Home made solutions vs. integrating software platforms from the (open source) market
  - Infrastructure moving towards the latter as industry grew in front of us!
  - Yet, **high-level storage software customized** for our **specific** access patterns

Largest scientific data repository



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
- Data centres run on commodity hardware and open-source OSes
- Commercial providers are (much) larger
  - CERN remains the world-largest scientific repository
- …Is this really “Big Data”?

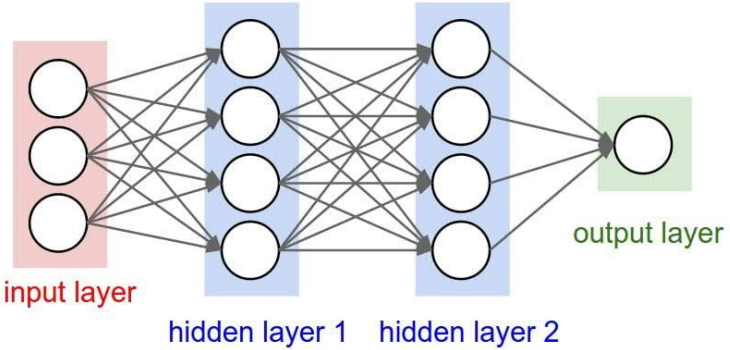
Big Data and what's coming next

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

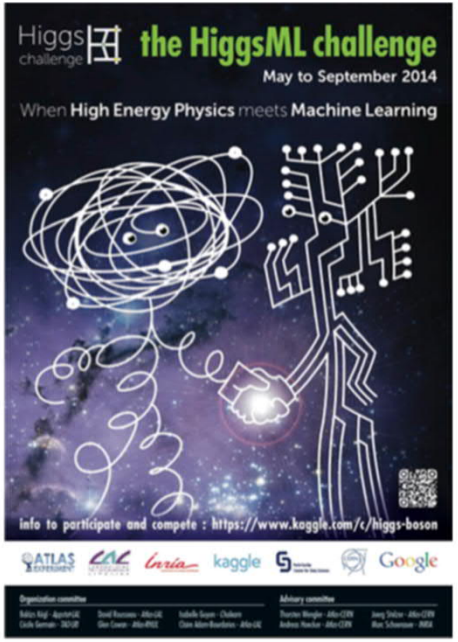
- Increasing interest in Big Data analysis
  - **The Power of Data:** Neural Networks are well known since the 1990s, but it's only now with **very large** and **easily accessible** data sets that they become effective!
  - Lots of software frameworks for *Deep Machine Learning* with NNs coming up
 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



Big Data at CERN, history

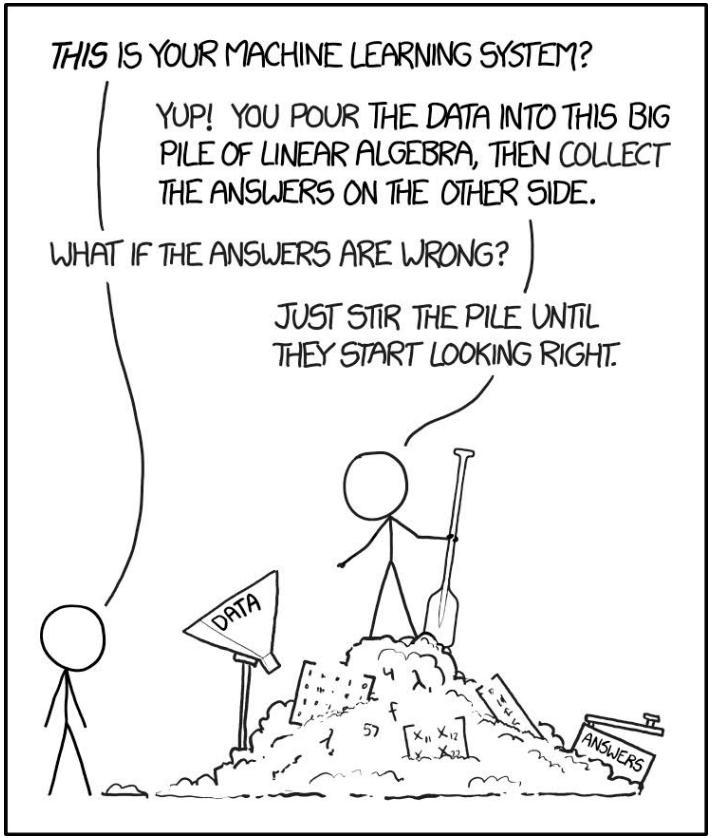
- Experiments have long used *Machine Learning* (once called *Multi-Variate Analysis*) techniques
  - Track reconstruction ~ pattern matching
  - Deep Neural Networks coming to help?

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



Big Data at CERN

- …Quoted at the CERN Academic Training on Machine Learning

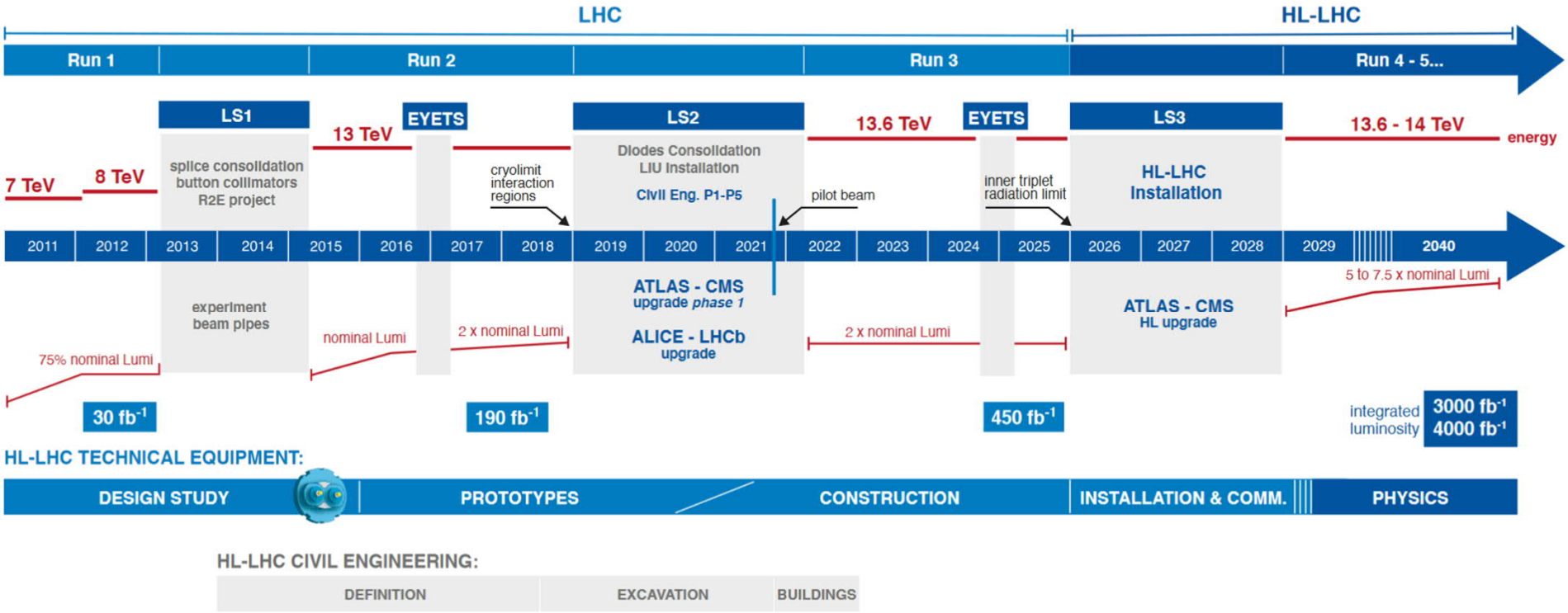


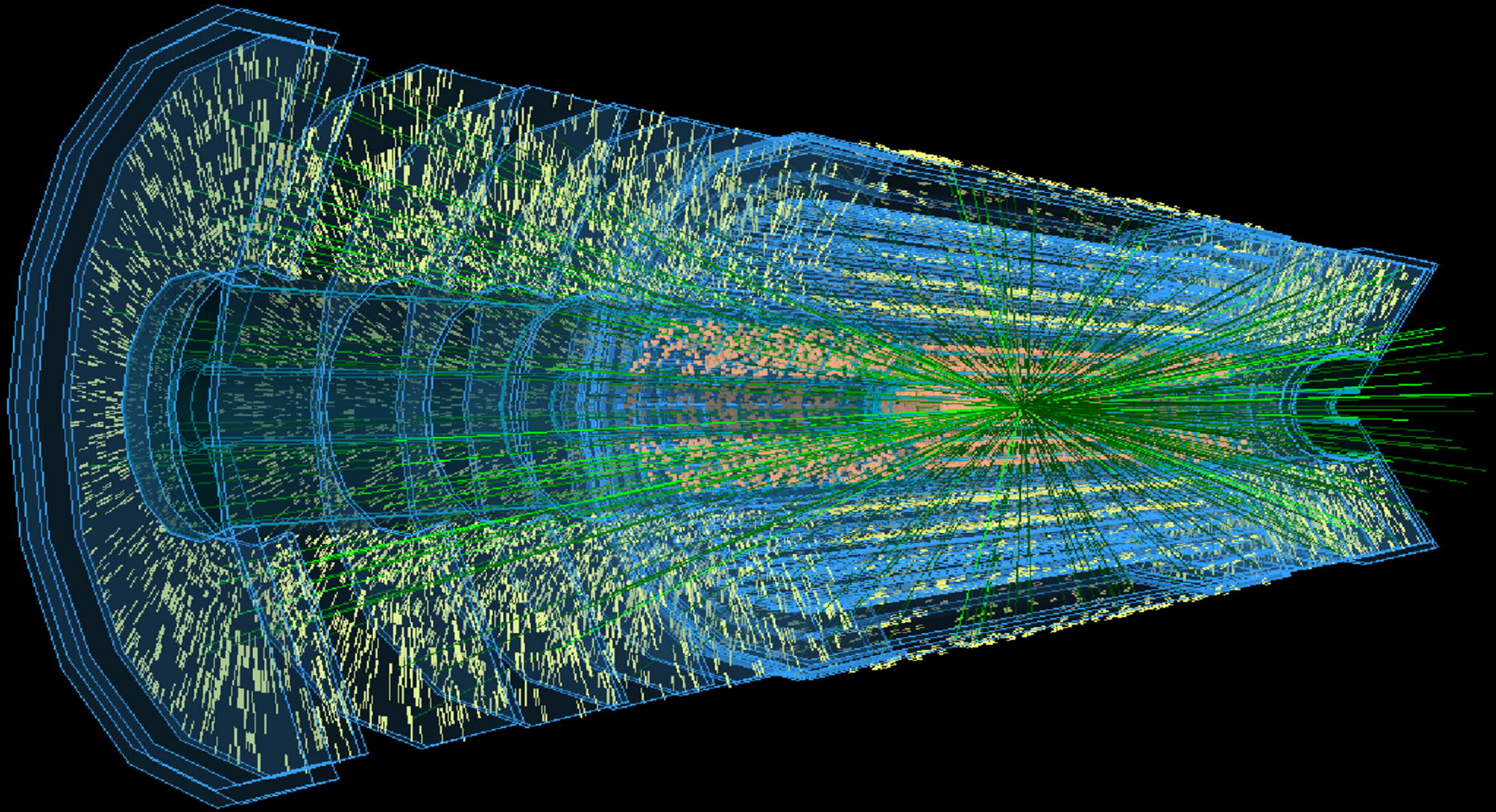
https://xkcd.com/1838/

Big Data at CERN

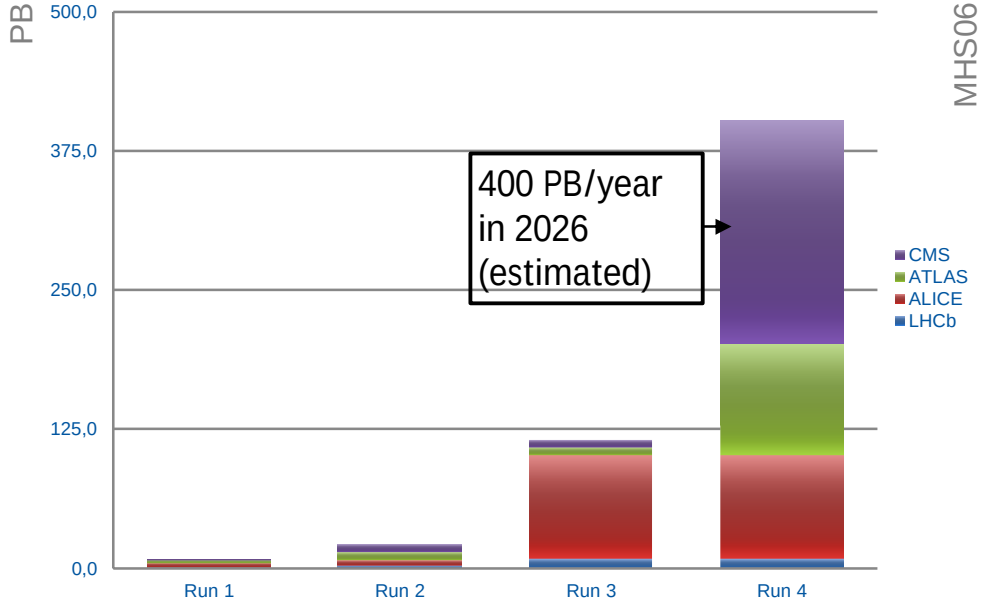
- More recently, LHC Beams Control Logging
  - Extract trends and detect/predict failures
- In general, ML techniques implemented where analytical approaches are inapplicable/unpractical
  - Security forensics, system analysis/profiling, etc.
    - Typically boiling down to **log analysis**
- Novel trends in data acquisition systems: use ML on GPUs to “learn” how to best select/discard events

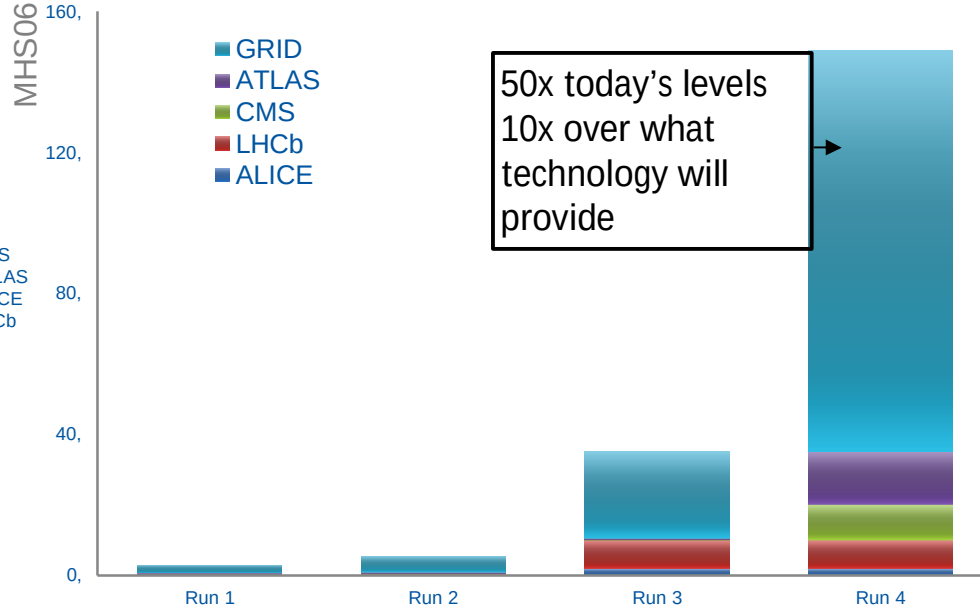
Hi-Lumi LHC: a computing challenge





HL-LHC: a computing challenge



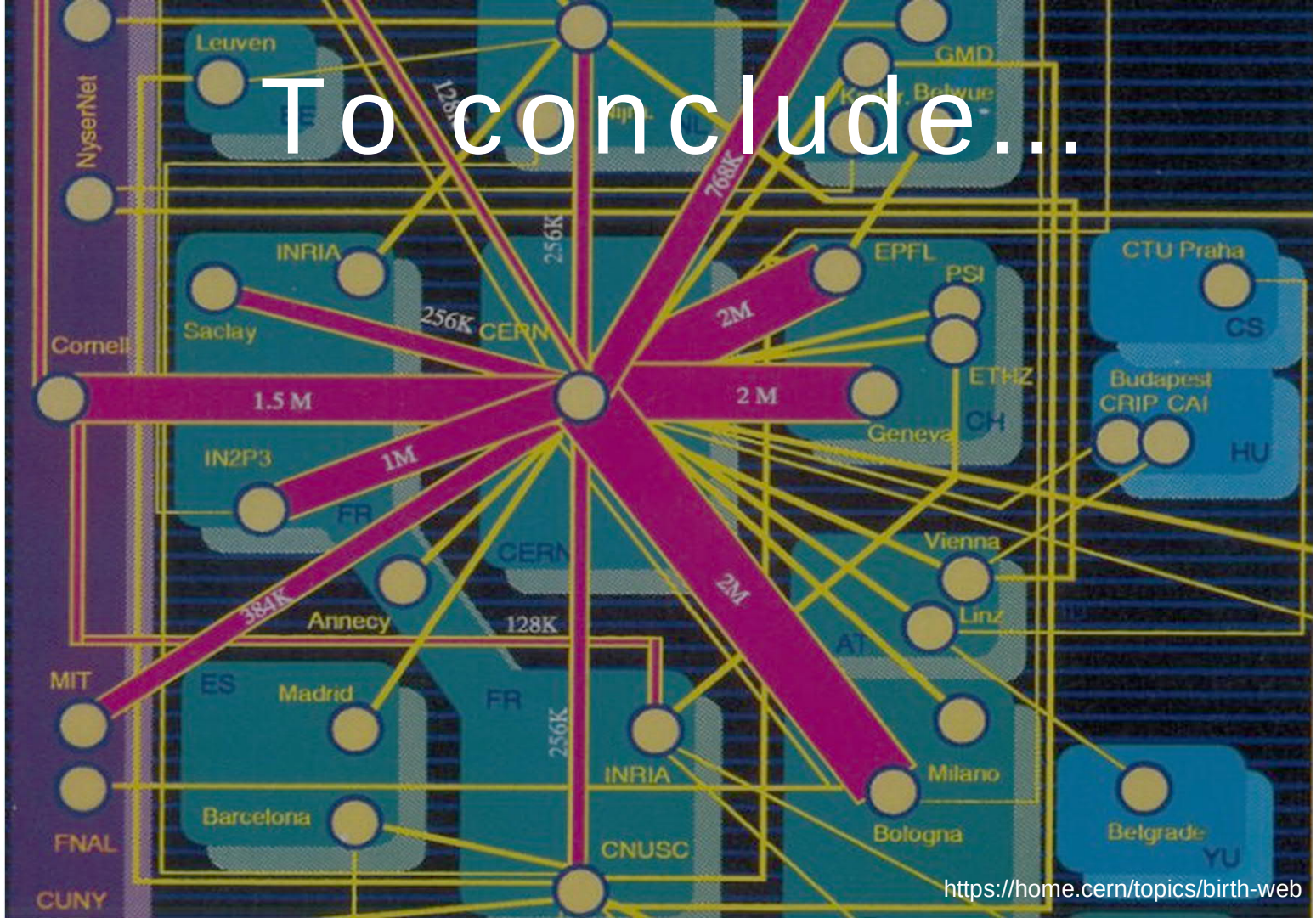


HL-LHC and friends

- 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**):
  - **Nuclear Fusion** Energy (**ITER**)

- Time for R&D, opportunity for new **synergies**
  - Increasing role of ML techniques, in particular in other sciences
    - **LIGO/Virgo**: GW signal detection

To conclude...



https://home.cern/topics/birth-web

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 team 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?

CERN-IT: pushing boundaries

- CERN-IT impact on society through computing:
  - Need for collaboration tools for Global Science led to invent the **World Wide Web**
  - 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"
    - Many big IT players involved, including (in alphabetic order) Google, Intel, Microsoft, Oracle, …
    - Testing software and hardware
    - Large student internship programme

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 !
  - The role of **Open Source** in software development is more and more crucial as scientific collaborations get larger

Thanks for your attention! Questions?

Accélérateur de science

Giuseppe.LoPresti@cern.ch
www.linkedin.com/in/giuseppelopresti

Credits to all CERN IT Storage colleagues