

Introduzione ai Big Data

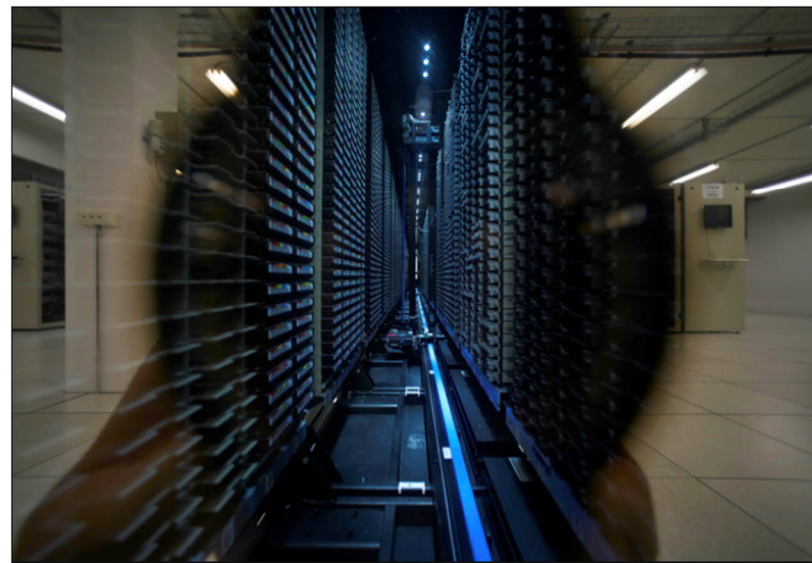
La gestione dell'informazione dal Data Taking al Cloud

Giuseppe Lo Presti
CERN IT Department

Italian Teachers Programme 2023 - Discovery

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.

SIGN IN / UP

The Register

STORAGE

CERN swells storage space beyond 1EB for LHC's latest ion-whacking experiments

A petabyte or more a day of readings? No problem, pal

Tobias Mann

Mon 2 Oct 2023 · 19:48 UTC

12



In preparation for its latest round of ion-smashing tests, CERN boosted its storage array for the experiments to more than one million terabytes in total size.

The facility's data store now exceeds an exabyte of raw capacity — with much of it on hard disk drives and an "increasing fraction of flash drives," the European super-lab's team explained in a [report](#).



CERN Courier April 2018

Software and computing

Time to adapt for big data

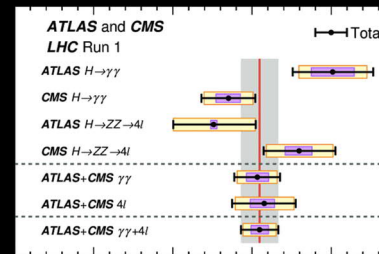
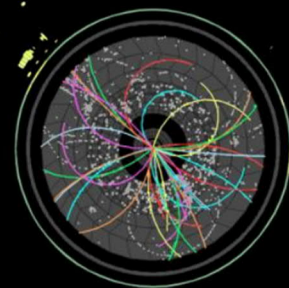
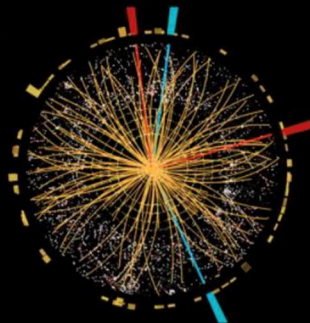
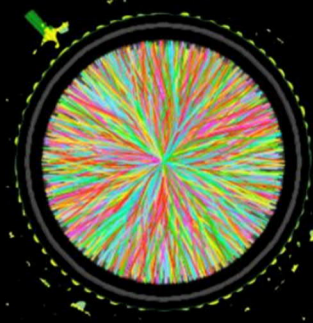
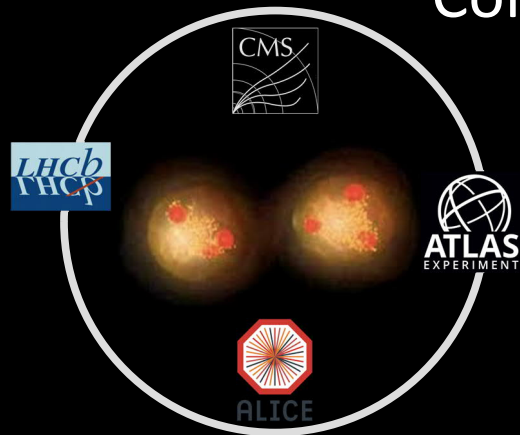
Radical changes in computing and software are required to ensure the success of the LHC and other high-energy physics experiments into the 2020s, argues a new report.

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



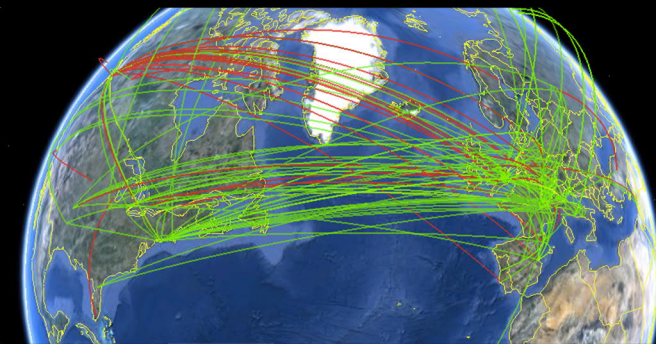
[...]Professor Bakker wrote that Mr Klein had been recommended by the director of the Zeeman laboratory in Amsterdam as a remarkable calculator[...] He needed no desk calculator and performed exceedingly well, exceeding in speed even my own desk calculator[...] I needed tables of combinations of so-called 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[...]**

Computing at CERN: The Big Picture



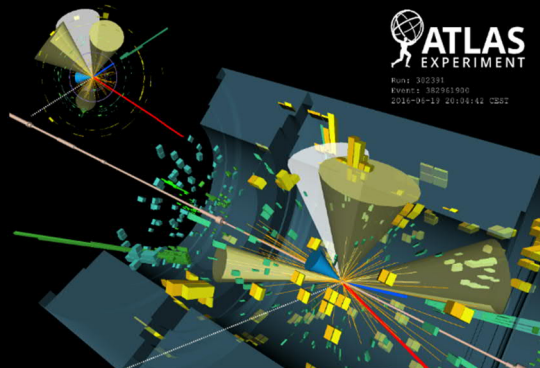
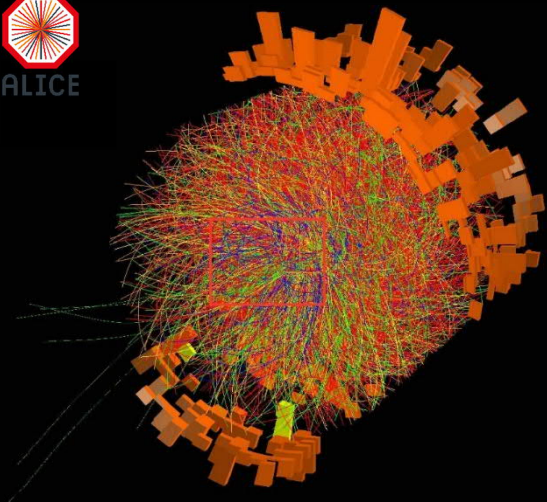
Data Storage - Data Processing - Event generation - Detector simulation - Event reconstruction - Resource accounting

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

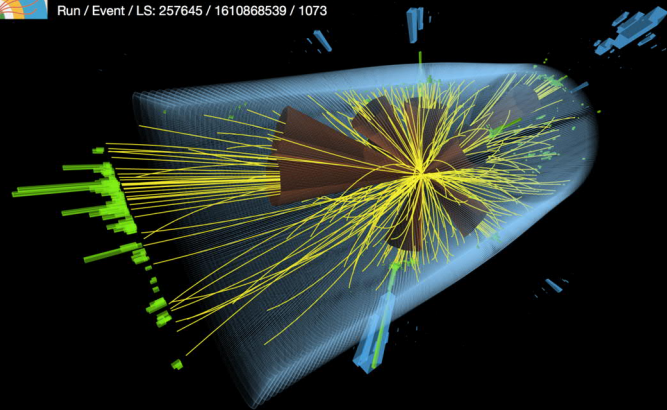


Lo Presti - Italian Teachers Programme 2023 - Discovery

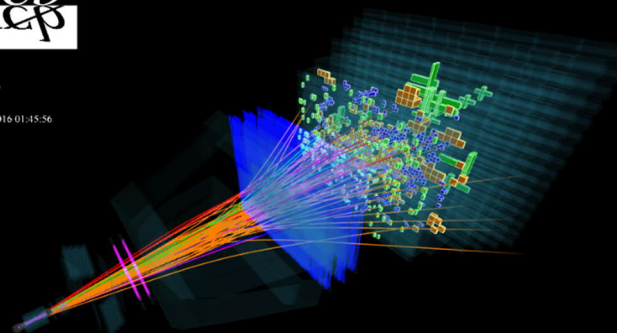
From the Hit to the Bit: Data Acquisition



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



Event 74374700
Run 173768
Min: 09 May 2016 01:45:56



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 bunch cross per second

Fast, simple information

Hardware trigger in a few micro seconds

L2: 100,000 events per second

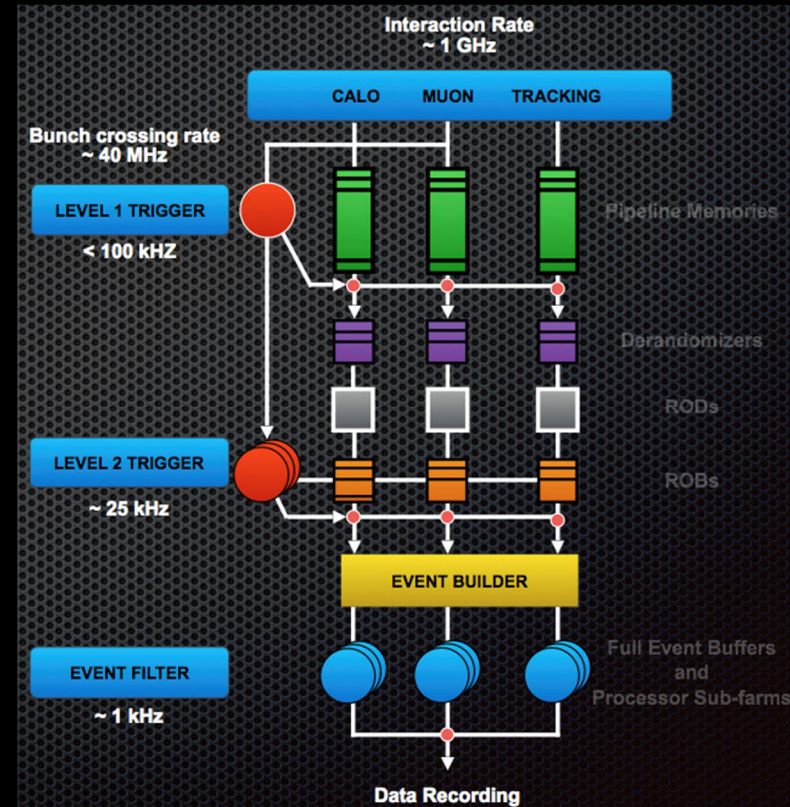
Fast algorithms in local computer farm

Software trigger in <1 second

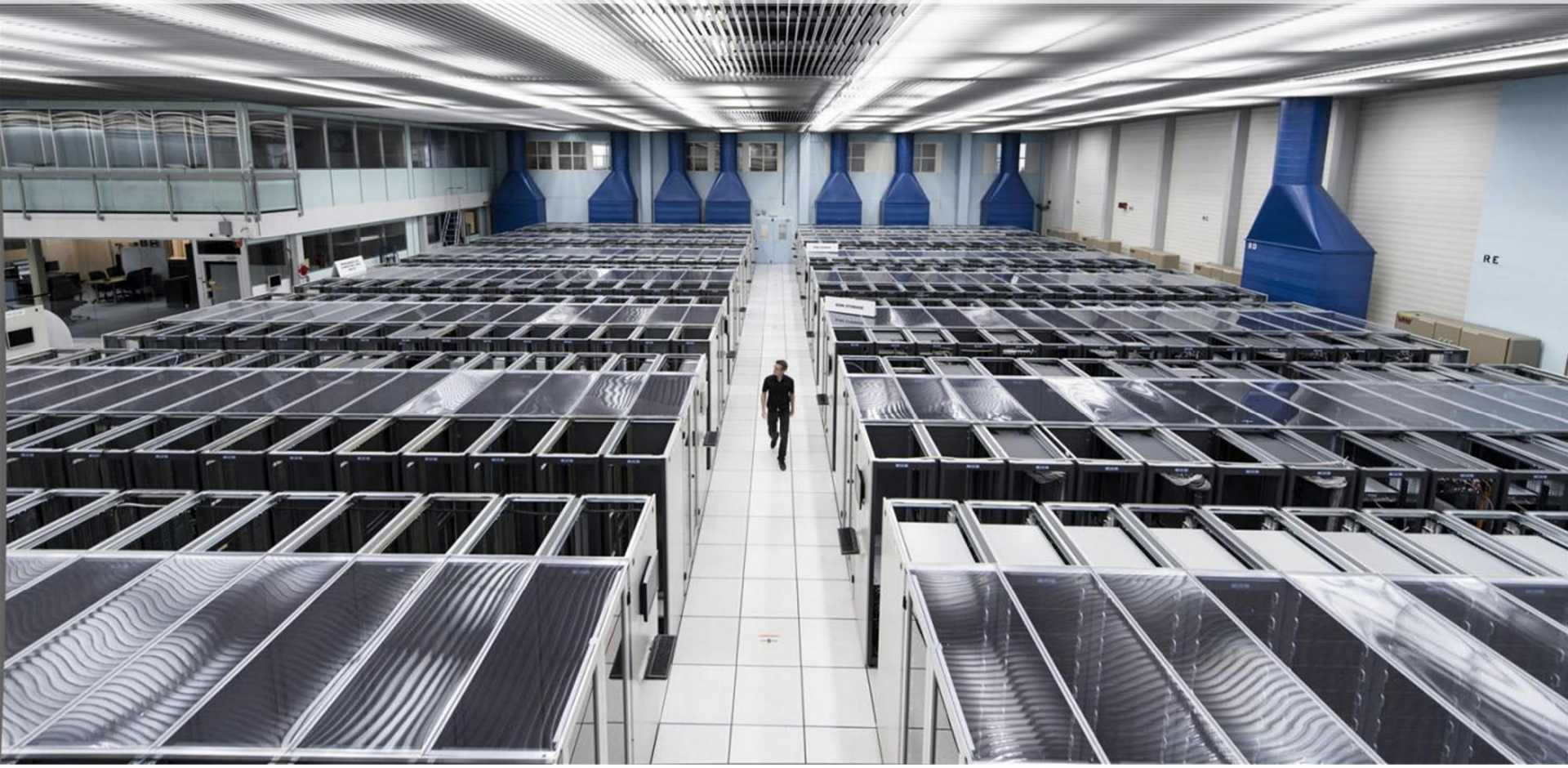
Which OS for such task?

EF: Few 1000s per second recorded
for offline analysis

By each experiment!



The CERN Data Centre



CERN DC: an ordinary week in numbers

Servers
9.4 K

Cores
442.6 K

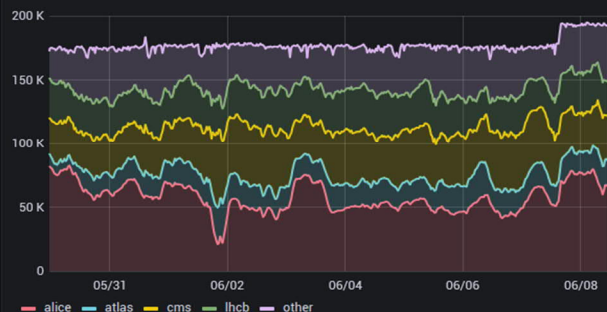
Disks
102.1 K

Tape Drives
163

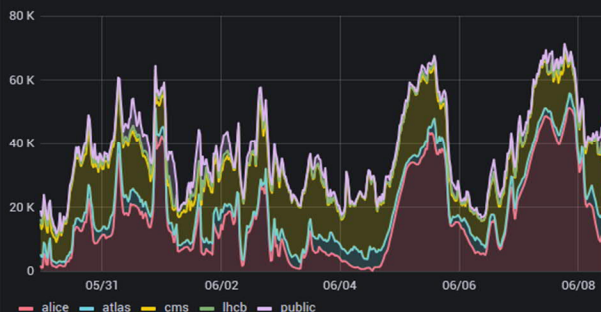
Routers
294

Wifi Points
4.9 K

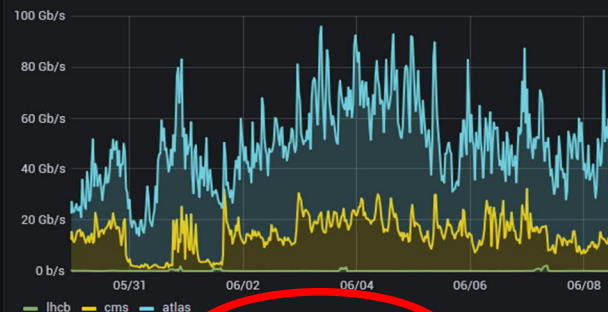
Batch Jobs Running



EOS Active Data Transfers



File Transfer Throughput



Cloud Virtual Machines Created



Databases Activity



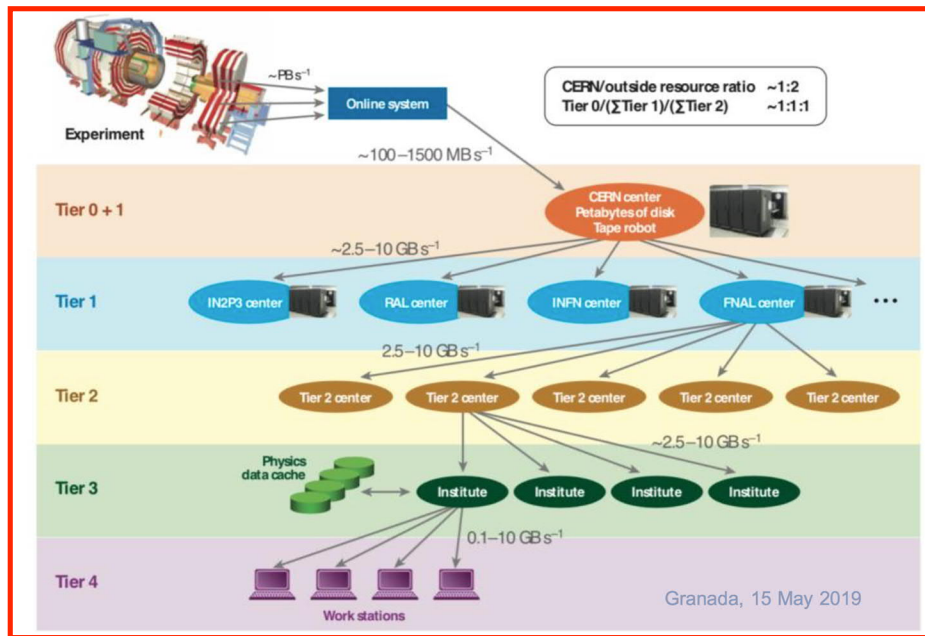
LHCOPN and LHCONE Total traffic

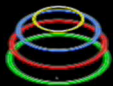


The Worldwide LHC Computing Grid



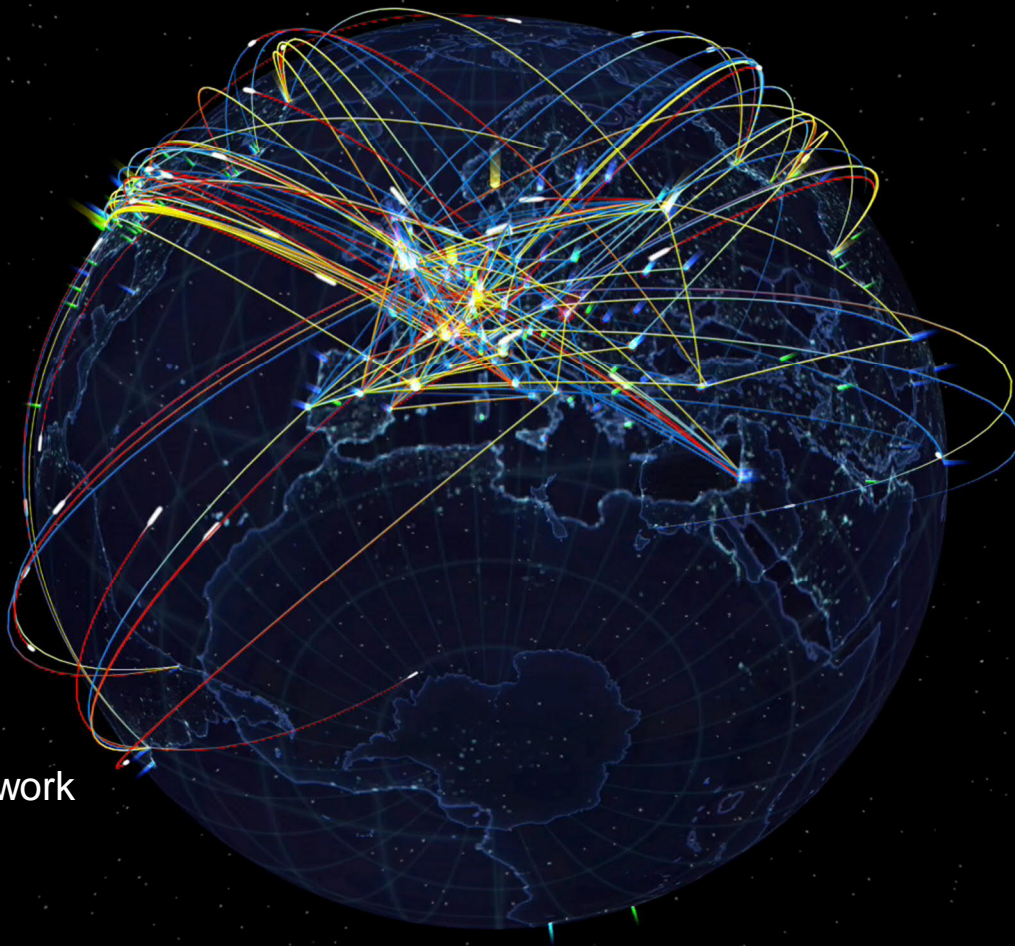
- The Worldwide LHC Computing Grid (WLCG) is a global collaboration of more than 170 data centres around the world, in 42 countries
- The CERN data centre (Tier-0) distributes the LHC data worldwide to the other WLCG sites (Tier-1 and Tier-2)
- WLCG provides global computing resources to store, distribute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





Data Distribution in WLCG

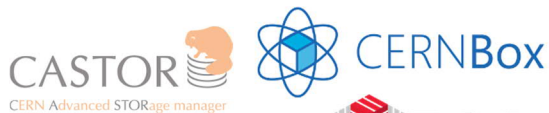
- Global transfer rates regularly exceeding **80 GB/s**
- **1+ EB** and 1.1B files transferred yearly in Run 3
- Main **challenge** is to have the **useful data close** to available computing resources
=> match storage/compute/network



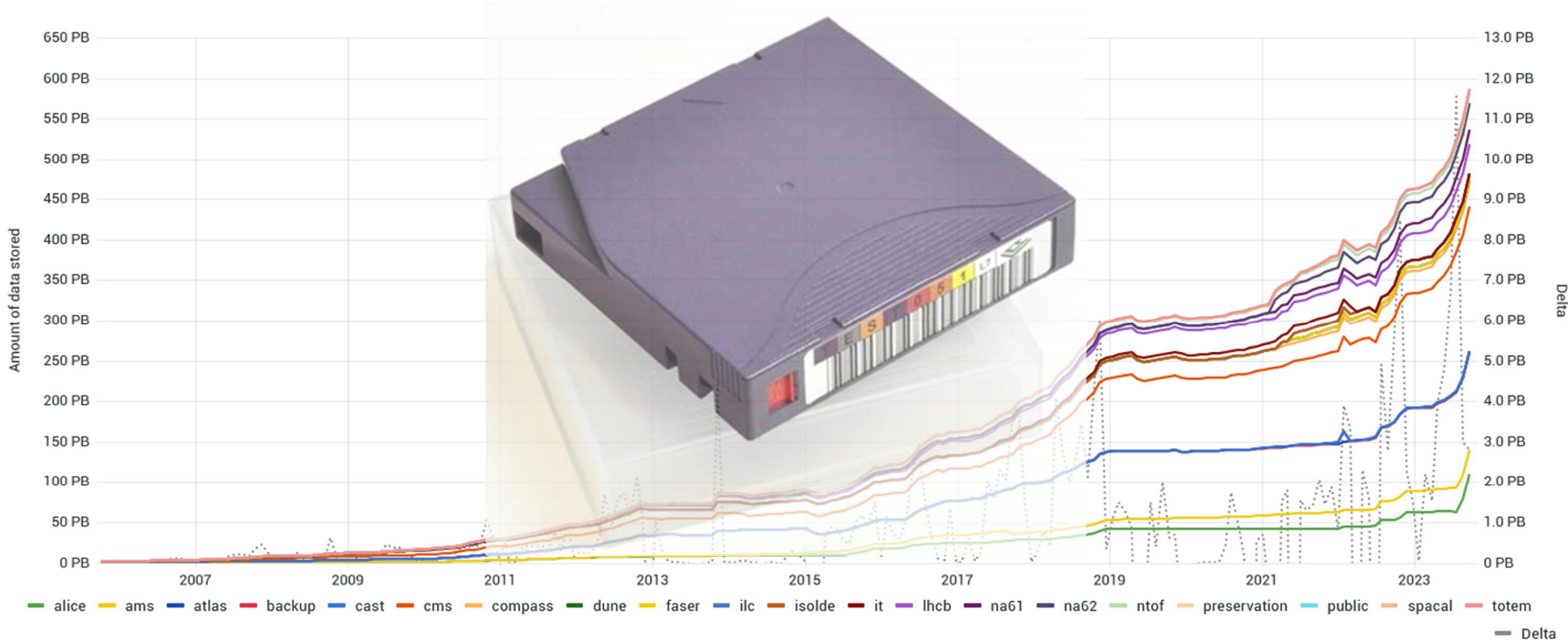
Running jobs: 365644
Active CPU cores: 807139
Transfer rate: 21.54 GiB/sec

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**

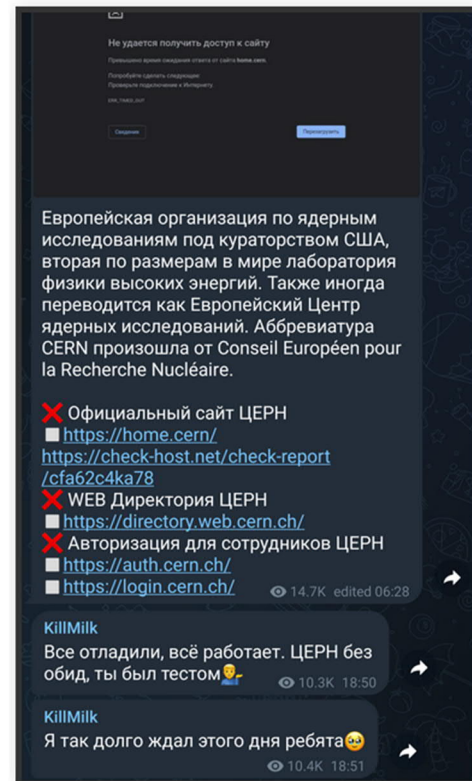


Largest scientific data repository



...And an appealing attack target

- CERN is permanently under cyber attack. Last attempt just happened
- Computer Security is a pillar of the whole IT infrastructure
 - Raising awareness at CERN and at partner institutes
 - It's not a matter of "if", but "when"!
 - Phishing campaigns, role games, presentations about real cases and mitigation measures, ...
 - Mandatory "Dual-Factor Authentication" (2FA) for IT operators
 - Continuous "white hat" penetration testing, in collaboration with the wider scientific community



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

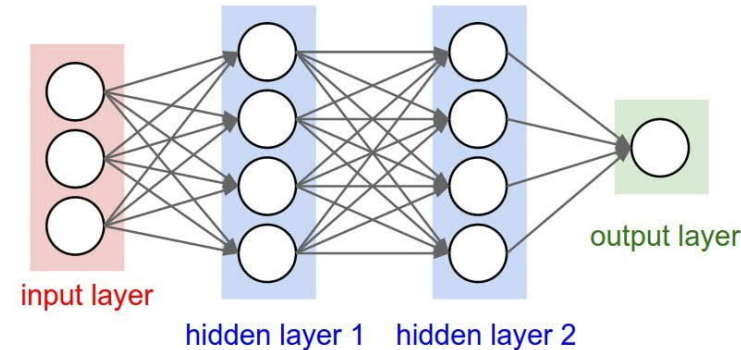


Big Data


- *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 1960s, 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



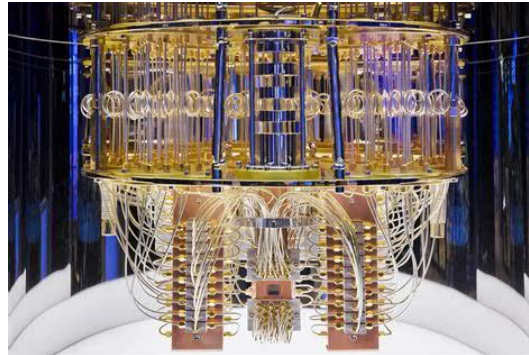
New frontiers: Heterogeneous Computing

- (Deep) Machine Learning is so **crucial** that industry has long invested into **hardware acceleration**
 - **GPUs** (Graphical Processing Units) for videogames (!) are being used on top of CPUs for faster matrix computations
 - **TPUs** (Tensor Processing Units), developed by Google, are offered in the Google Cloud Platform
- 
- A close-up photograph of a Raspberry Pi 4 single-board computer. It is equipped with a silver-colored aluminum heat sink and a black cooling fan, which is mounted on top of the board. The board itself is green and populated with various components, including a USB Type-C port, a USB-A port, and a micro-HDMI port. The fan has a CE mark and the model number '8C21A00030' printed on it.



New frontiers: Heterogeneous Computing

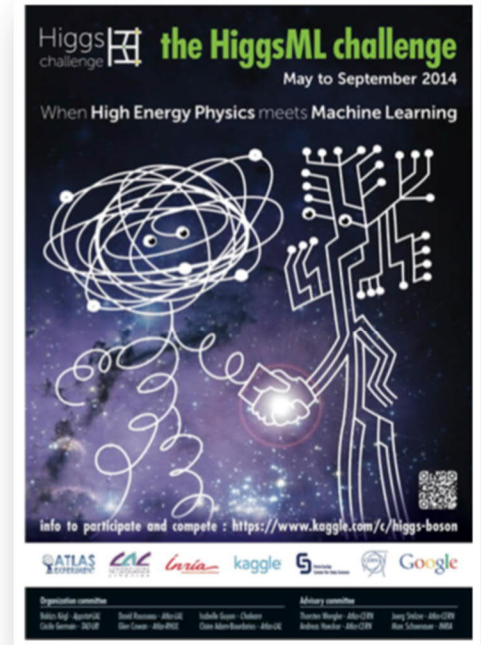
- A potential game changer: **Quantum Computing**
 - Quantum Computers can only execute a **very limited set of “programs”**, but with **exponential parallelism** (on paper)
 - *Quantum Machine Learning* is being demonstrated – at CERN – as one of those programs, which can be executed by such specialized hardware
 - Stay tuned...



G. Lo Presti - Italian Teachers Programme 2023 - Discovery

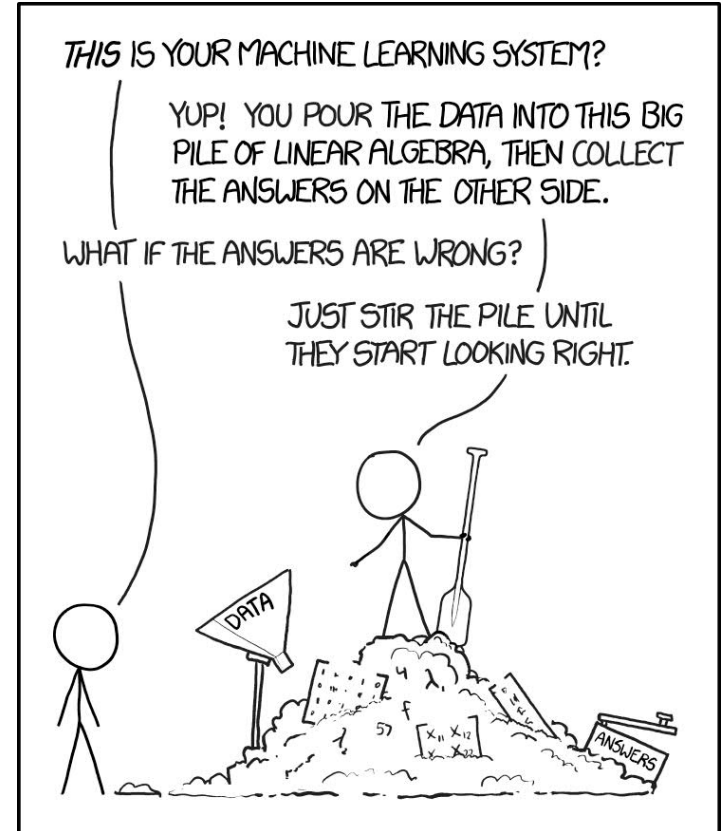
Big Data at CERN, history

- 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 ran in the past years
 - 2018 edition: best results obtained with pure parallel processing, without ML!



Big Data at CERN

- ...Quoted at the CERN Academic Training on Machine Learning

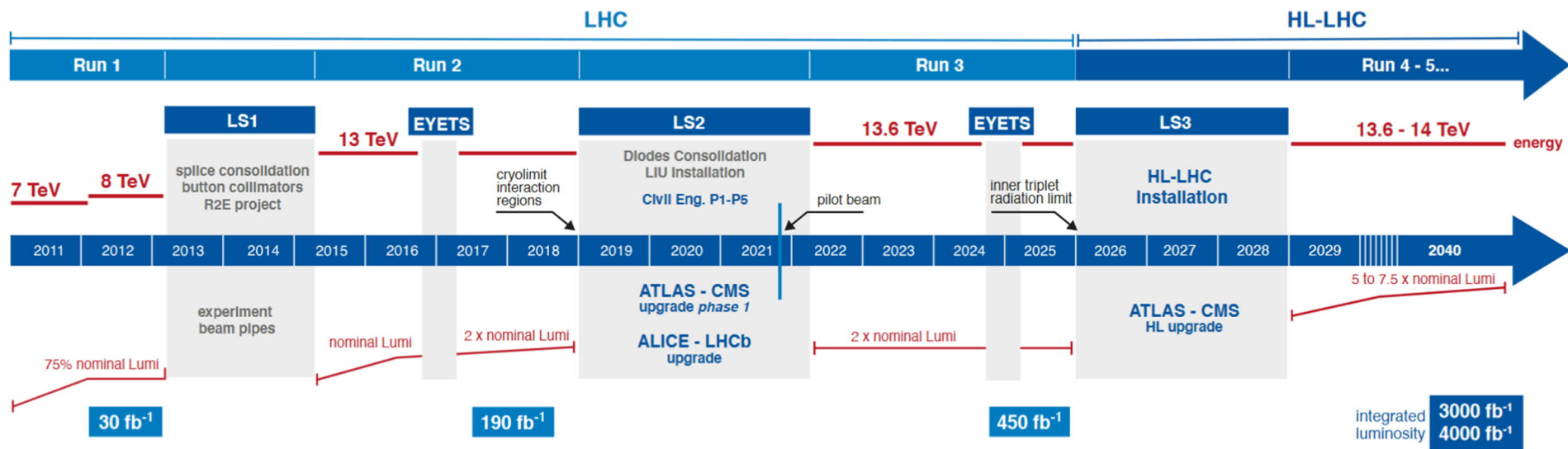


<https://xkcd.com/1838>, May 2017

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

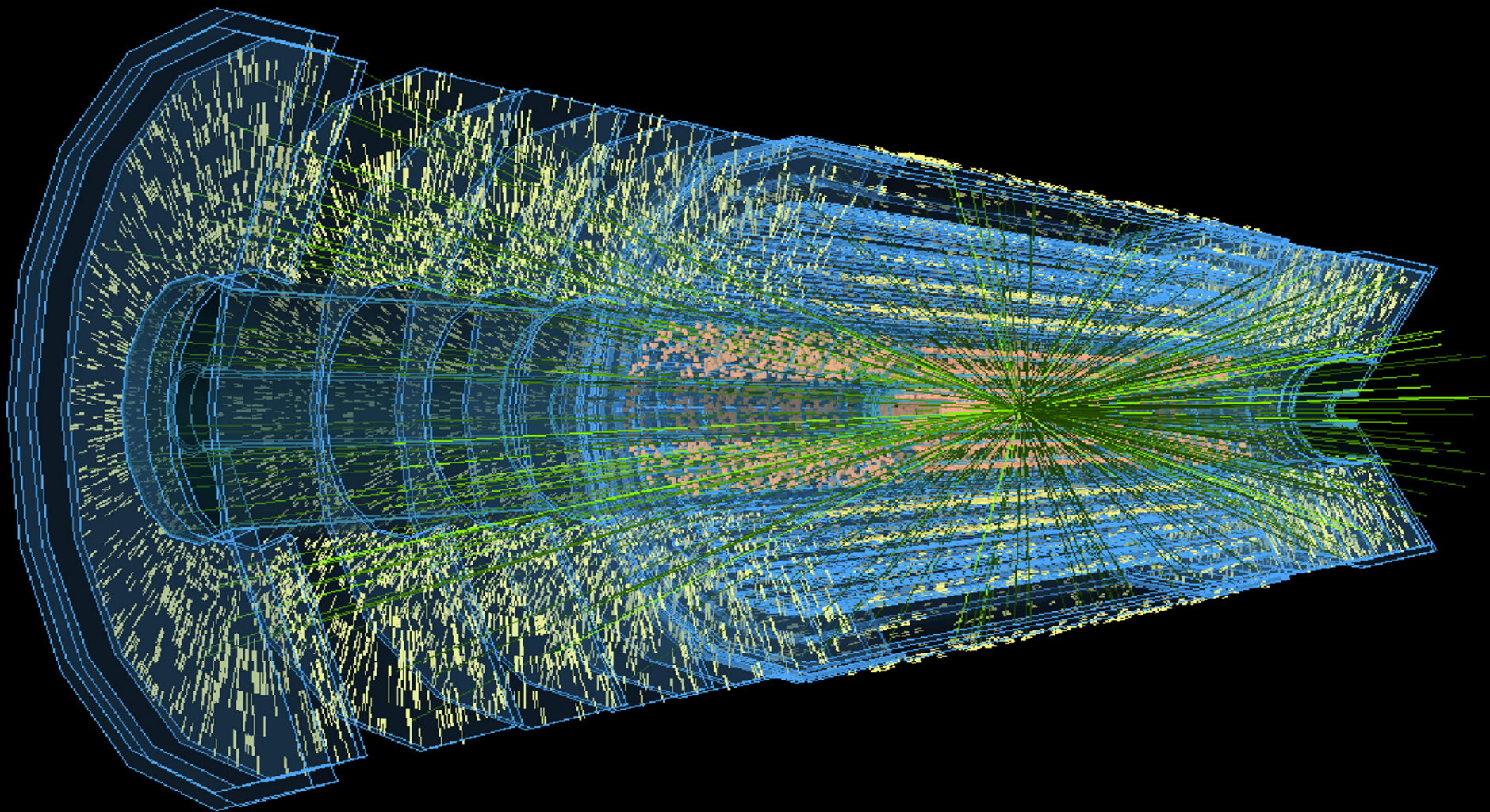


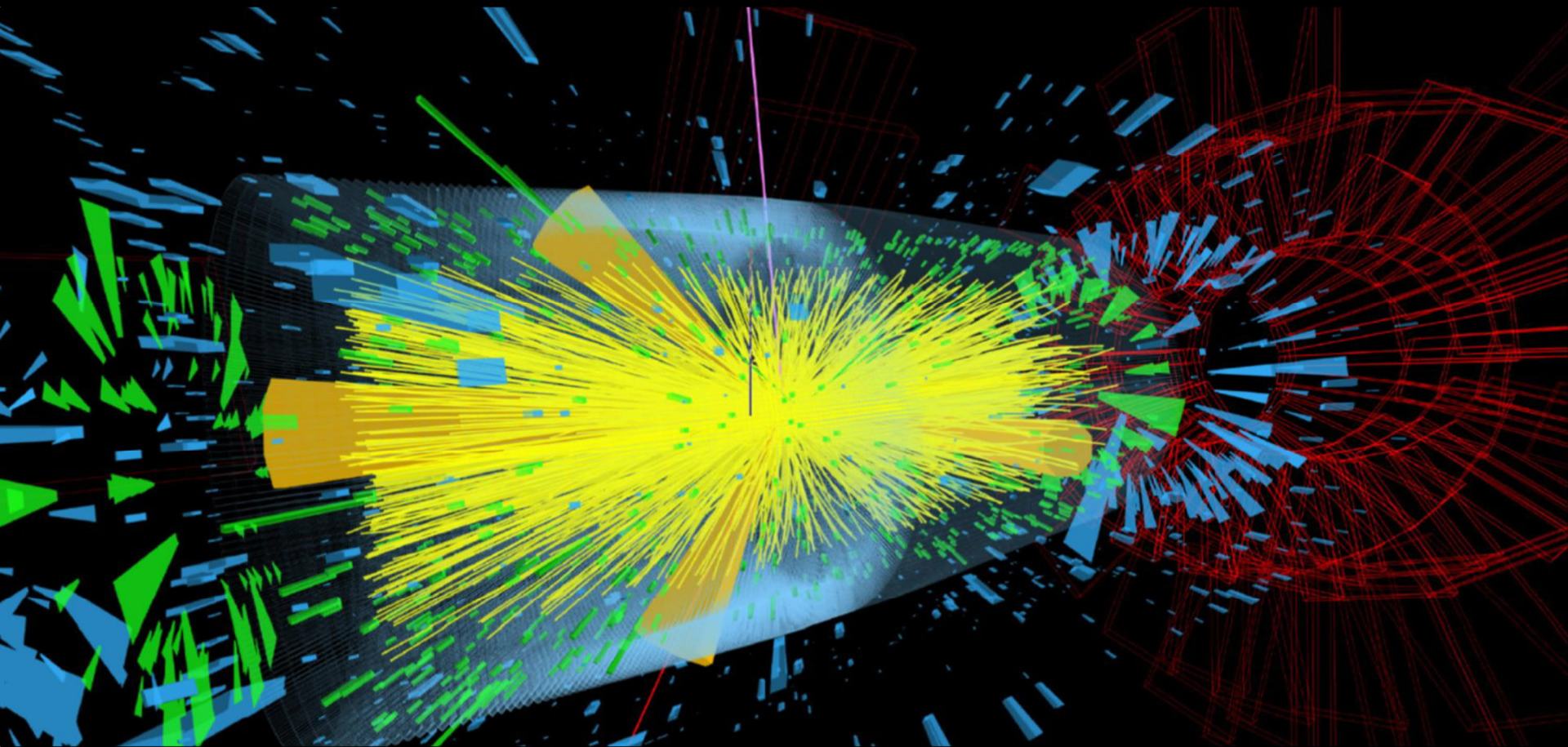
HL-LHC TECHNICAL EQUIPMENT:



HL-LHC CIVIL ENGINEERING:







HL-LHC and friends

- High Luminosity LHC is not alone in the current arena of large scientific collaborations – especially if we look into Astronomy

- Square Kilometer Array (**SKA**)
- Cherenkov Telescope Array (**CTA**)
- NASA James Webb Space Telescope
- ESA Euclid “3D” Telescope
- Etc...

Large discovery potential, beyond what particle colliders can do!

- Time for R&D, opportunity for new **synergies**
 - **Increasing role of ML techniques**
 - **LIGO/Virgo**: automatic GW signal detection and alerting

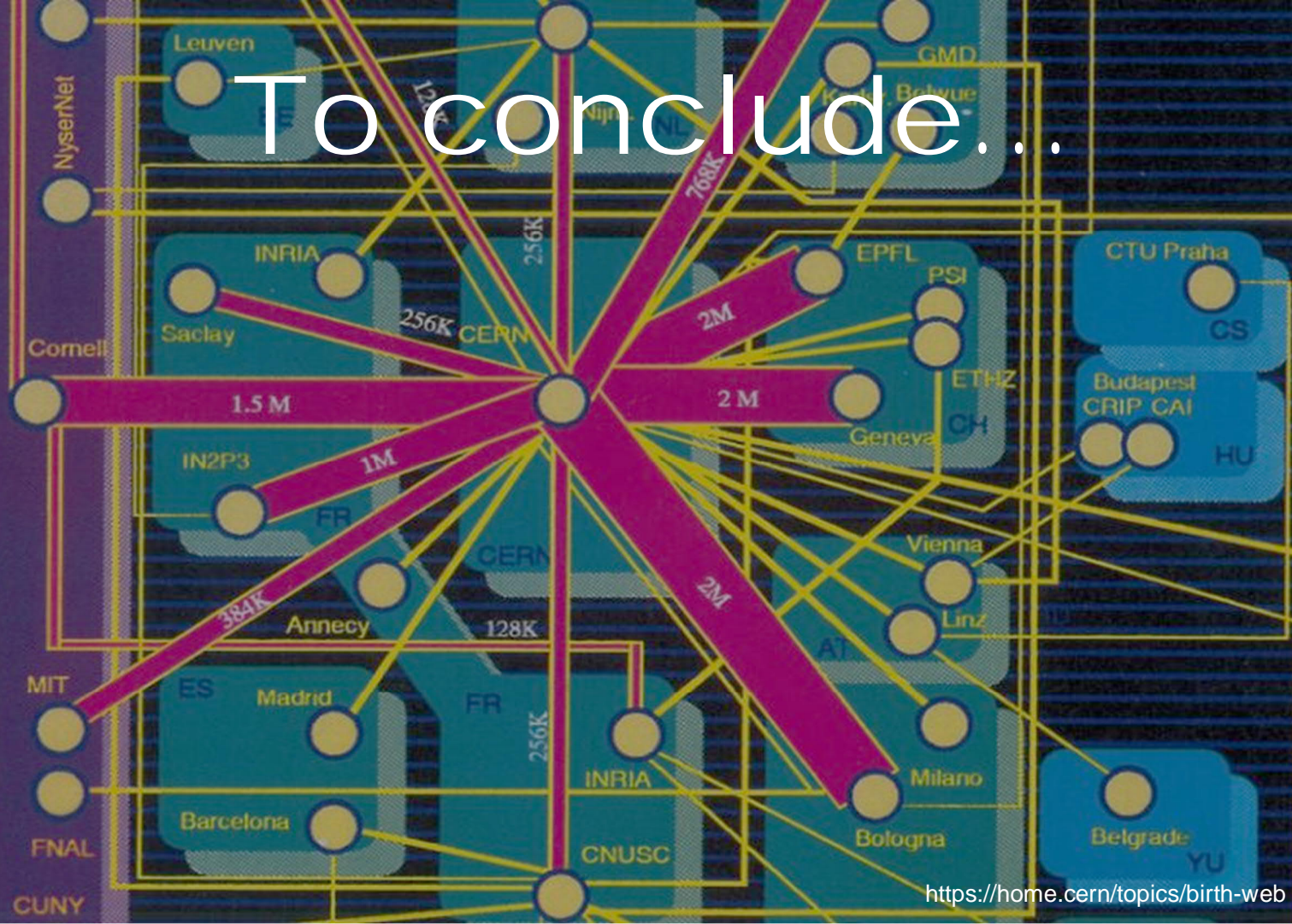


Opportunities and Risks...

- Big Data and **Data Science** are popular career paths, crossing the boundaries between **Computer Science**, **Physics** and **Statistics**
- Fundamental science and engineering remain the pillars to understand technology!
- Big Data and Machine Learning demonstrate **data's ever-growing value**, in particular when dealing with **personal data**
 - 2022: **7** out of the **top 10** world-largest companies by capitalization (including the GAMAM) are largely or entirely **based on the Data economy**
 - At 11 T\$, they compare with the GDP of **Germany + UK + France + Italy** !



To conclude...



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)
 - 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/hands?



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**, pioneering the concept of **Computing Clouds**
- Open access to science
 - Need for sharing the results had led CERN to pave the way to open access to documents and now data: **LHC@home** and **CERN Opendata Portal**
- **Openlab**
 - *Public-private partnership to accelerate 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, IBM, Intel, Microsoft, Oracle, ...**
 - Large student internship programme
 - Hosts the **CERN Quantum Technology Initiative**



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