Dealing with Big Data at CERN

Giuseppe Lo Presti
CERN IT Dep., Storage group

Agenda

- Setting the scene: CERN, the LHC, and its computing and storage challenges
- (Storage) Software Platforms for High Energy Physics
  - Usage patterns and comparison with Big Data industry
  - Our software stack “inventory”
- Future Outlook
  - High-Luminosity LHC & friends

SKA Signs Big Data Cooperation Agreement With CERN

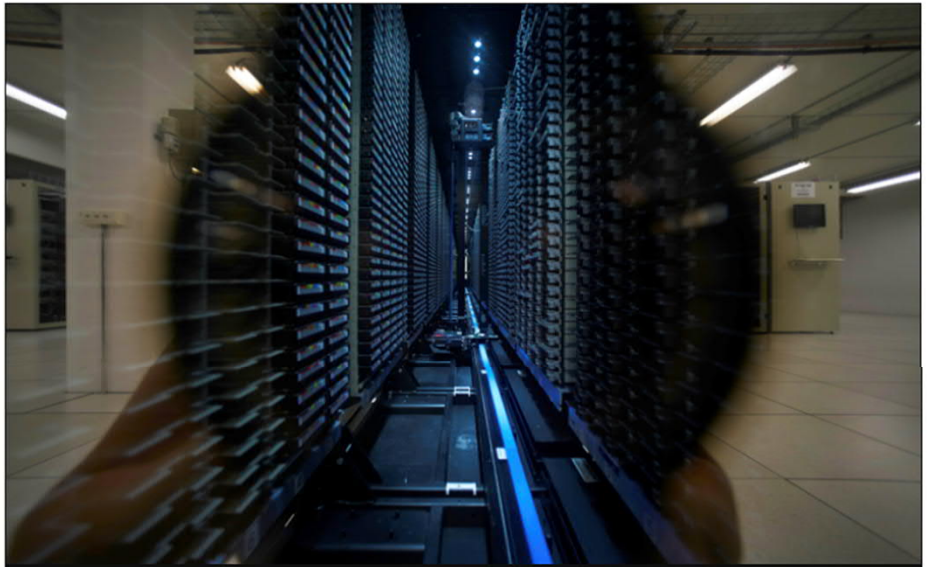
Dr. Fabiola Gianotti, CERN Director-General, and Prof. Philip Diamond, SKA Director-General, signing a cooperation agreement between the two organisations on Big Data. © 2017 CERN

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



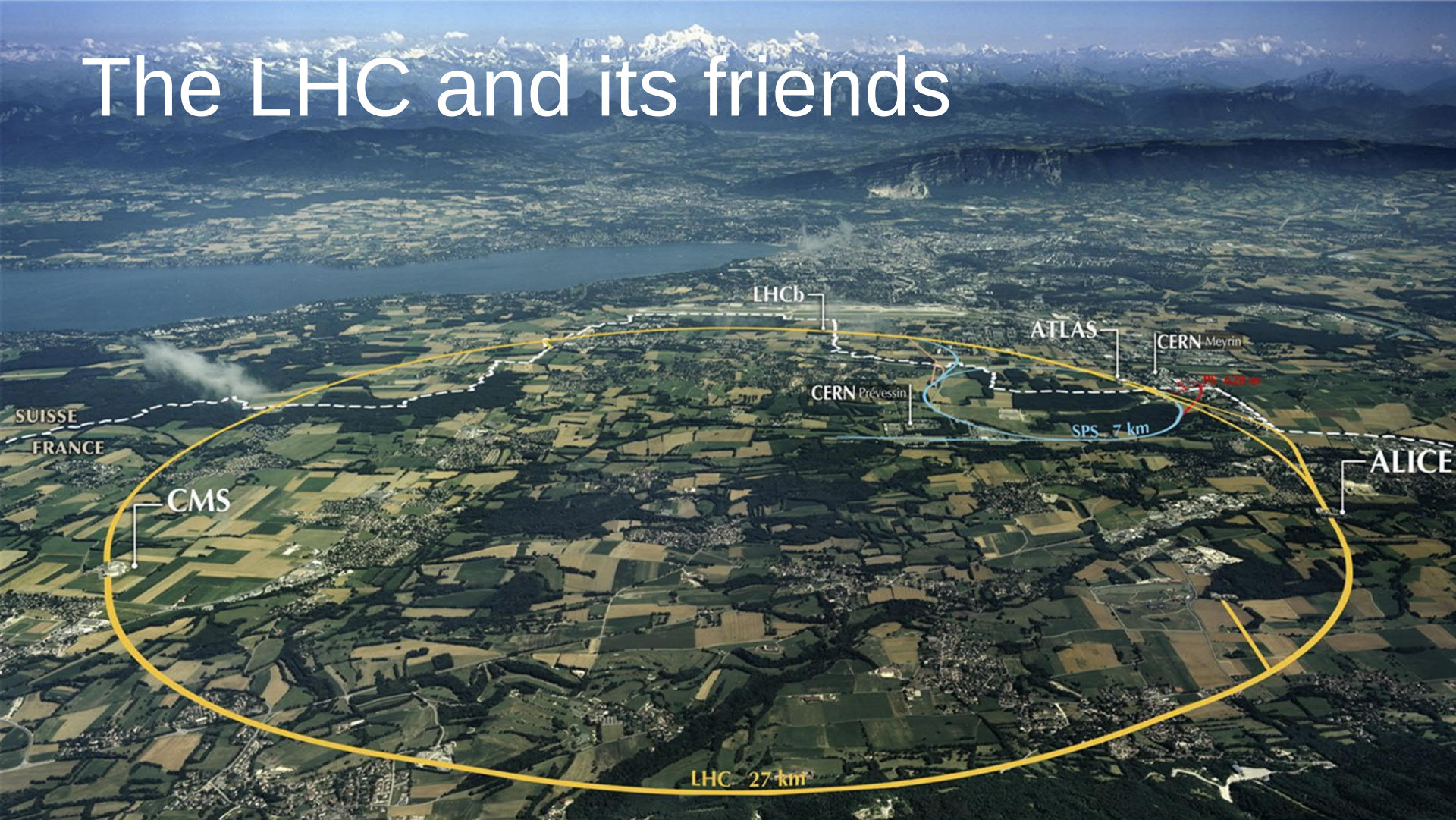
Magnetic tapes, retrieved by robotic arms, are used for long-term storage (Image: Julian Ordan/CERN)

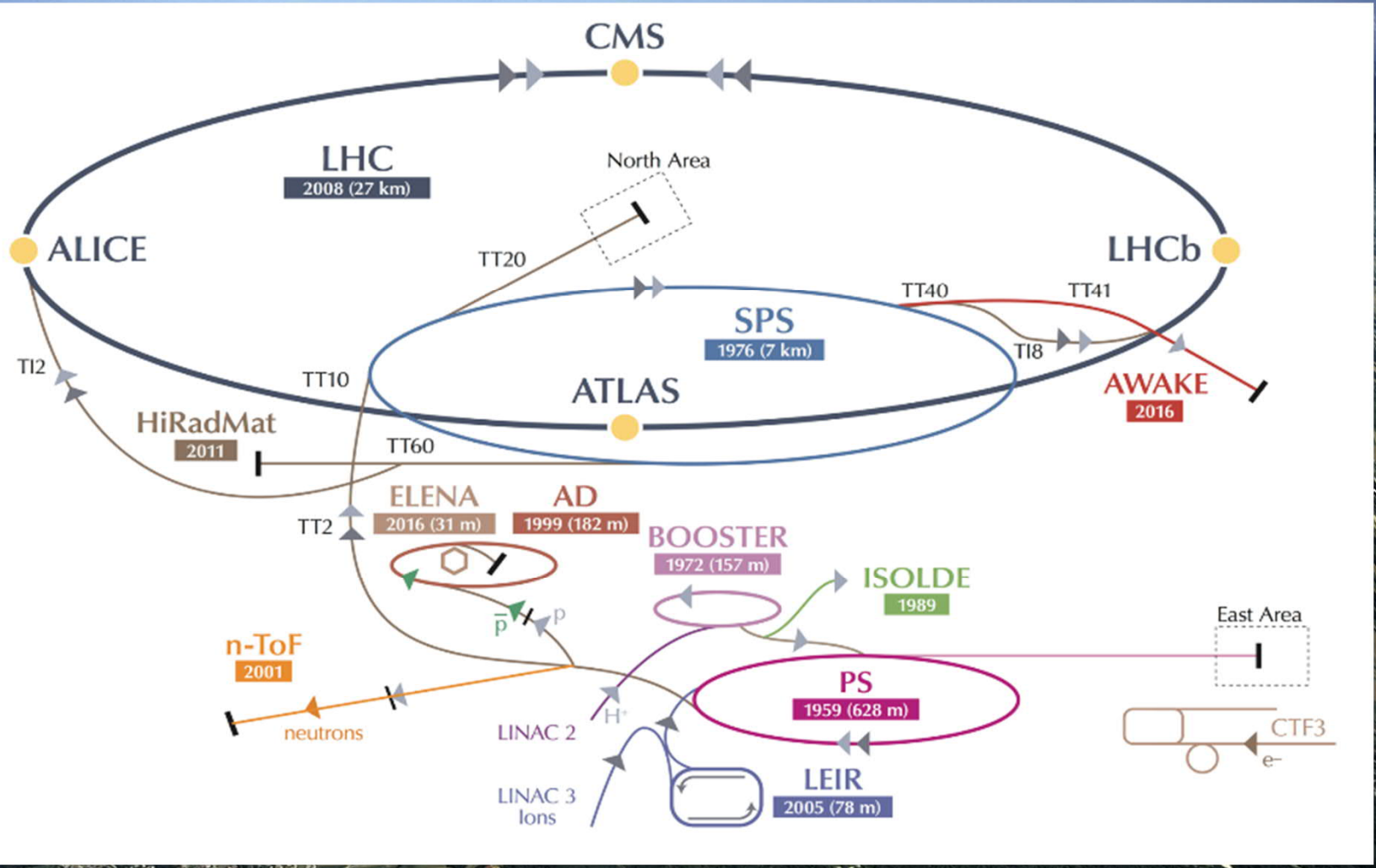
This year CERN's data centre broke its own record, when it collected more data than ever before.

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.

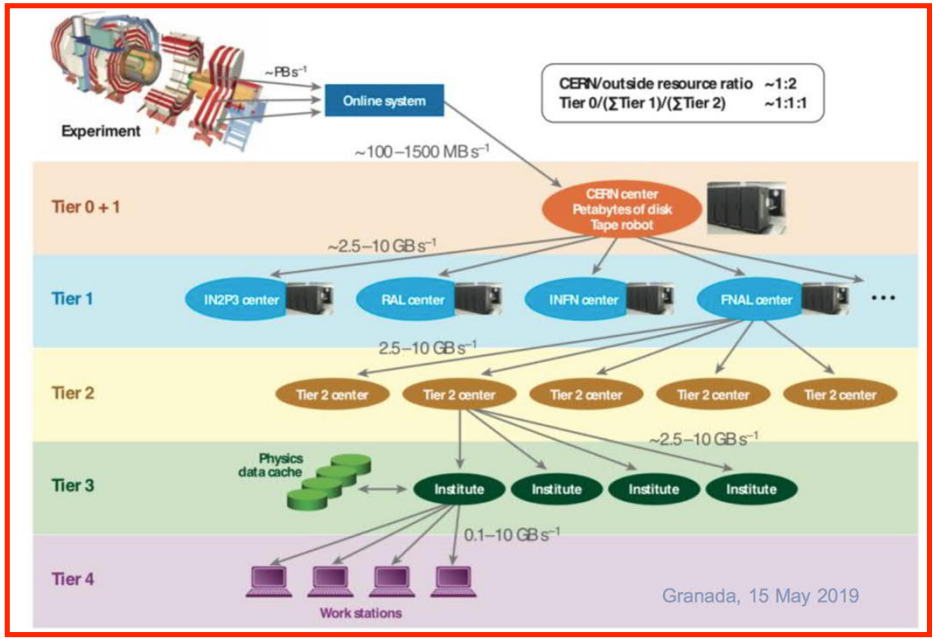
The LHC and its friends




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
ELENA
2016 (31 m)
AD
1999 (182 m)
TT2
BOOSTER
1972 (157 m)
ISOLDE
1989
East Area
n-ToF
2001
PS
1959 (628 m)
neutrons
LINAC 2
LINAC 3
Ions
LEIR
2005 (78 m)
CTF3

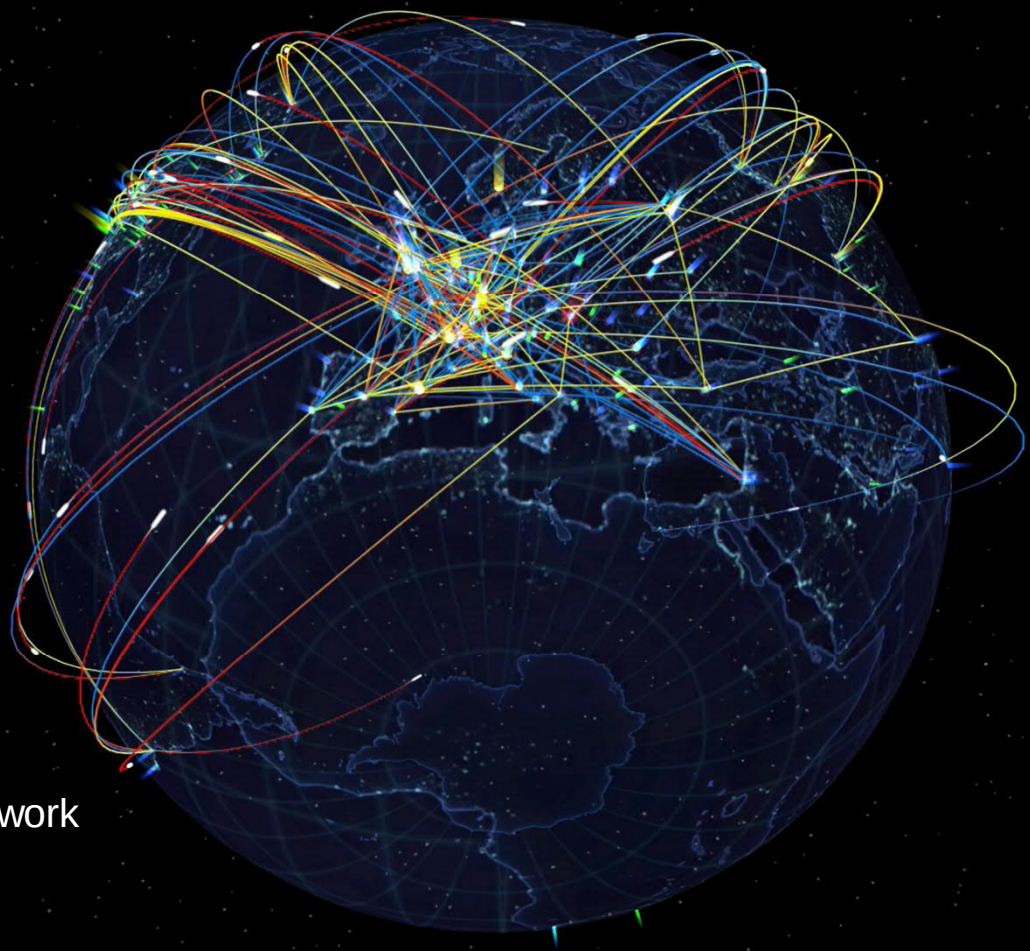
The Worldwide LHC Computing **Grid**

- 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



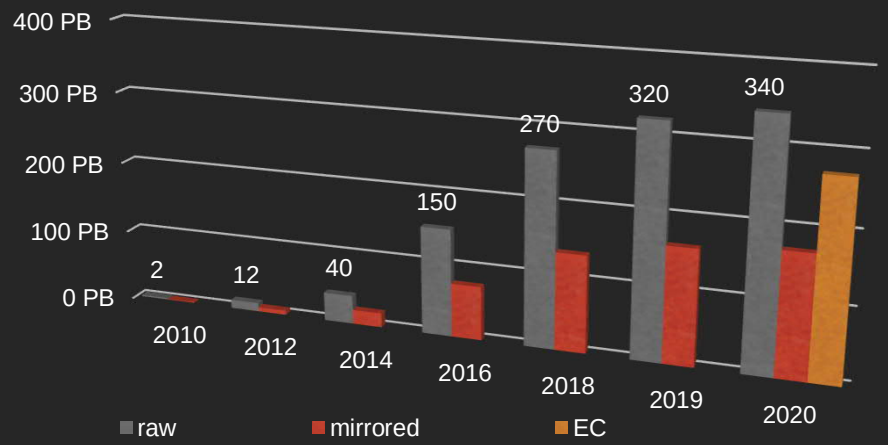
Software Platforms for HEP

- 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

Disk & Tape for Physics storage



EOS

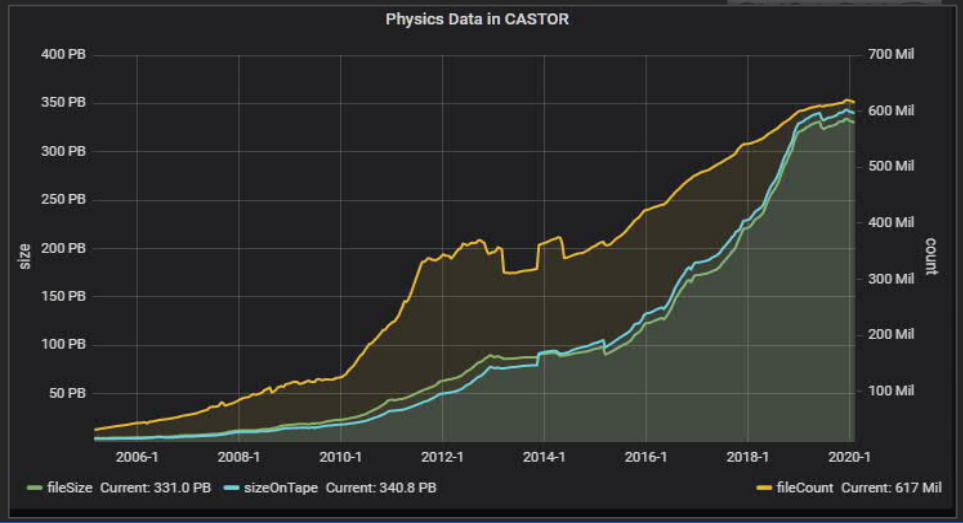
Files Stored 4.9 B

Storage Nodes 1500

IO Streams >100k

Hard Disks 60k

CASTOR
CERN Advanced STORage manager



Files Stored 617 M

Tape Cartridges 45k

Disk storage for IT infra

CERN Ceph Clusters		Size	Version
OpenStack Cinder/Glance	Prod.	6.4 PiB	mimic
	Hyperconv.	245 TiB	mimic
CephFS (HPC+Manila)	Prod.	1.09 PiB	luminous
	Pre-prod.	164 TiB	mimic
	Hyperconv.	356 TiB	mimic
CASTOR	Disk Buffer	5.5 PiB	nautilus
S3+SWIFT	Prod. (4+2EC)	1.92 PiB	luminous

Ceph

- Openstack RBD remain the biggest use-case
  - ~1PB/year
- Openstack Manila NFS-like shares (CephFS)
- Storage for Kopano

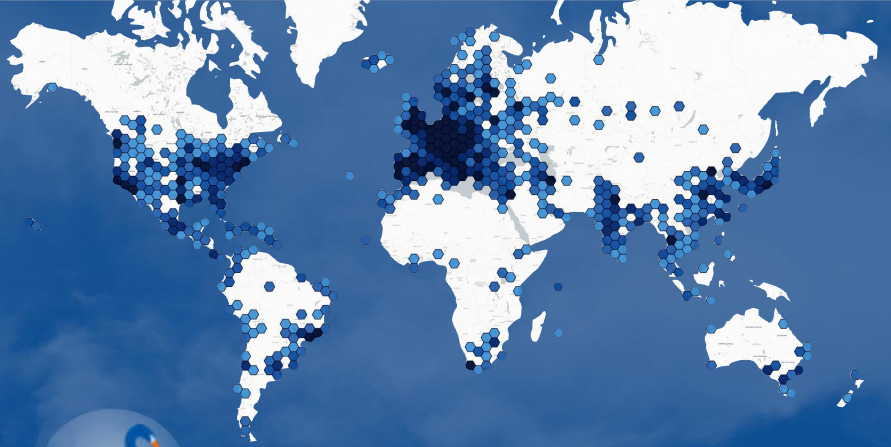
S3

- IT applications, ATLAS+LHC@home
- CVMFS repositories
- CERNBox backups using Restic

CVMFS

- 49 repositories with 868 M files and 52 TB
  - 43 using block storage 6 using S3 objects

Disk storage for the Community: **CERNBox** and **SWAN**



CERNBox

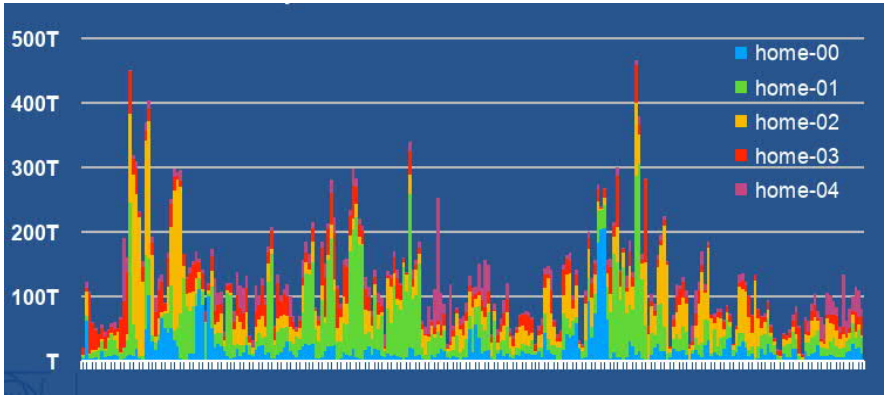
- 5 years of production
  - still growing service with 18k users
  - 4PB+ data, 1B+ files, 110k+ shares
- Consolidating “home dirs” into CERNBox
  - Migration out of Windows DFS ongoing
  - HA SAMBA gateway in production
- Central Hub for CERN data and apps
  - Draw.io, Onlyoffice, Collabora, Kopano, ...

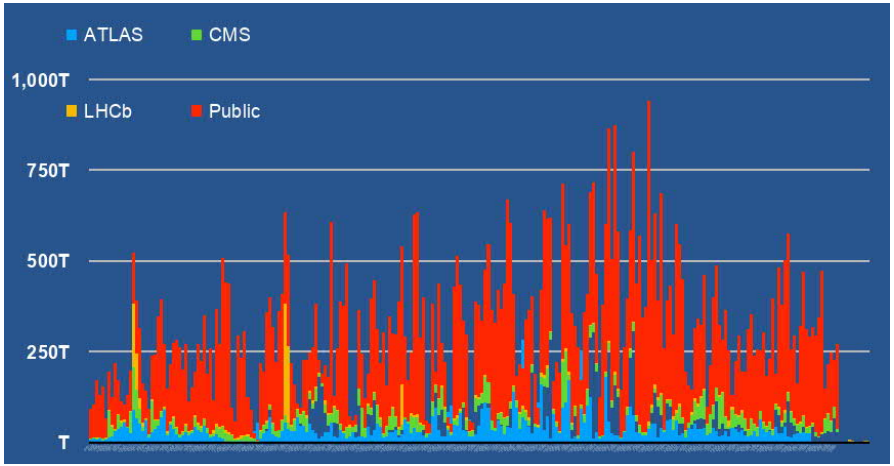
SWAN

- Turn-key data analysis platform
- Accessible from everywhere via a web browser
- Support for ROOT/C++, Python, R, Octave
- Fully integrated in CERN ecosystem
  - Storage on EOS, Sharing with CERNBox
  - Software provided by CVMFS
  - Massive computations on Spark

Disk for all uses: filesystem access

- FUSE (Filesystem) getting more and more popular, both for “homes” and for Physics





Storage hardware

Profiting from economy of scale

- Minimise price per TB
- Network may become the bottleneck!

Current generation of storage servers:

- 8 trays (24x disks) per system unit
  - ~2300 TB (12TB drives)
- 4 trays (24x disks) per system unit
  - ~1150 TB (12TB drives)
  - ~1340 TB (14TB drives)

Going towards high-density JBOD

- 2 trays (60x disks) per system unit
- ~1680 TB (14TB drives)

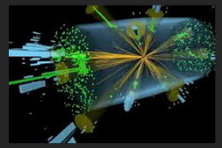
HEP & Big Data Technology

- If we would use “Big Data” analytics in Physics, we could profit from all existing Big Data storage technologies, protocols & analytics frameworks
  - Not mentioning that “Machine Learning” is known for decades to Physicists (used to be called MVA...)

- Why is that not yet mainstream?

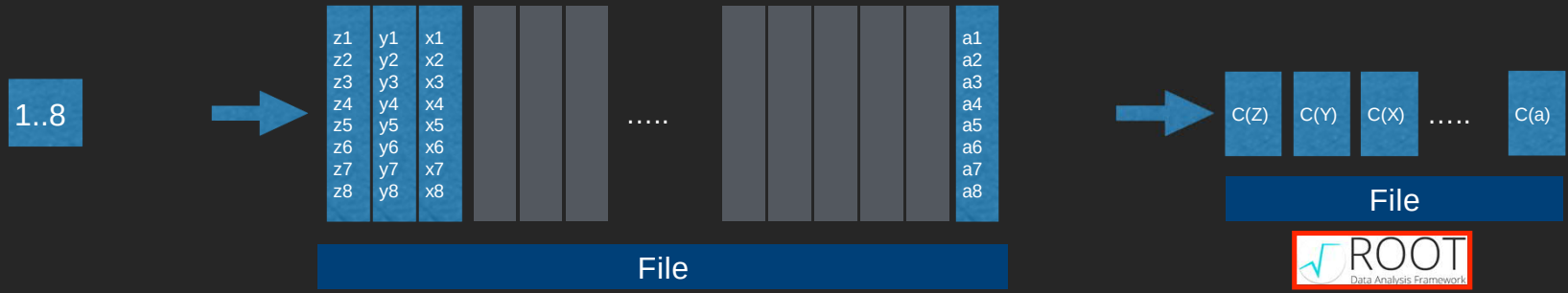
Physics Data Formats

unstructured raw data - each physics event is stored in a compound block - events are assembled during data taking from many detector systems



compressed raw data with individual events

structured data - data is stored optimised for volume and access patterns

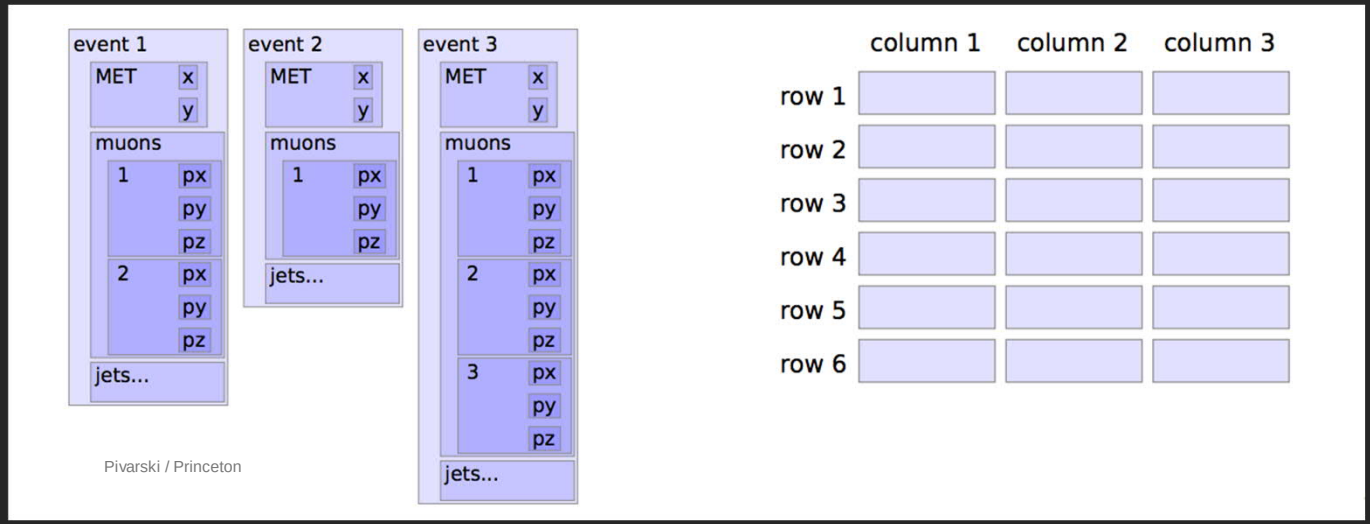


Physics Selection vs Big Data Analytics

HEP data are variable-length, nested data structures

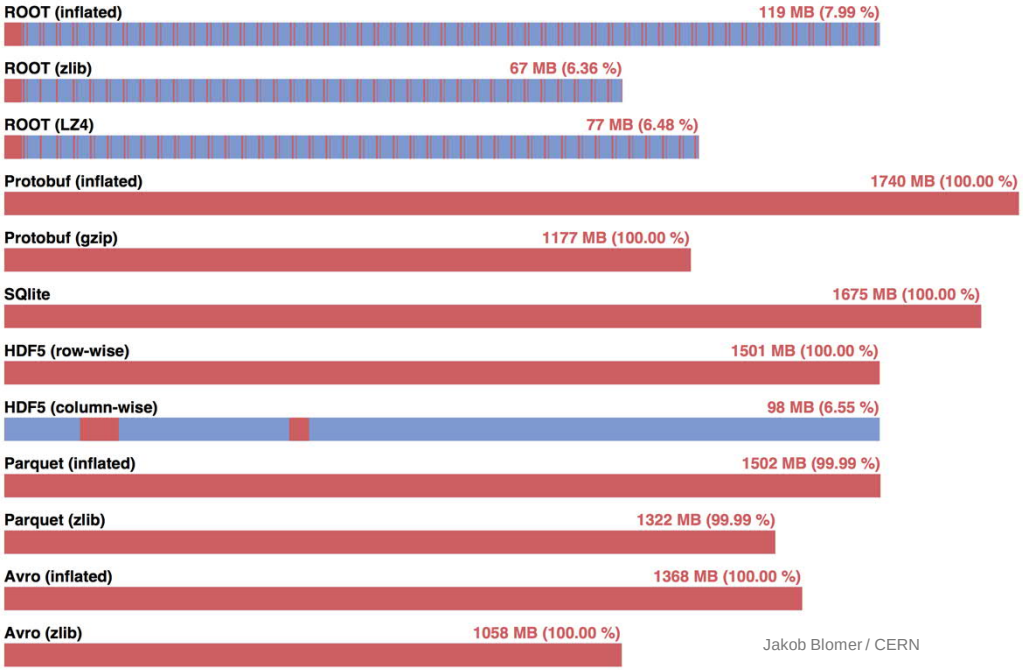
need to loop over combinations of particles

Data is typically in tabular form



Data Formats & Storage Access Patterns

Read pattern (red) in a selective physics analysis workflow



- ***sparse access*** pattern *cry for* certain access protocol capabilities in LAN & WAN environments

- predictable read patterns allows to use ***asynchronous multi-byte-range read*** requests to to compensate latencies

- good news: most of traffic in HEP is still mainly sequential forward-seeking IO

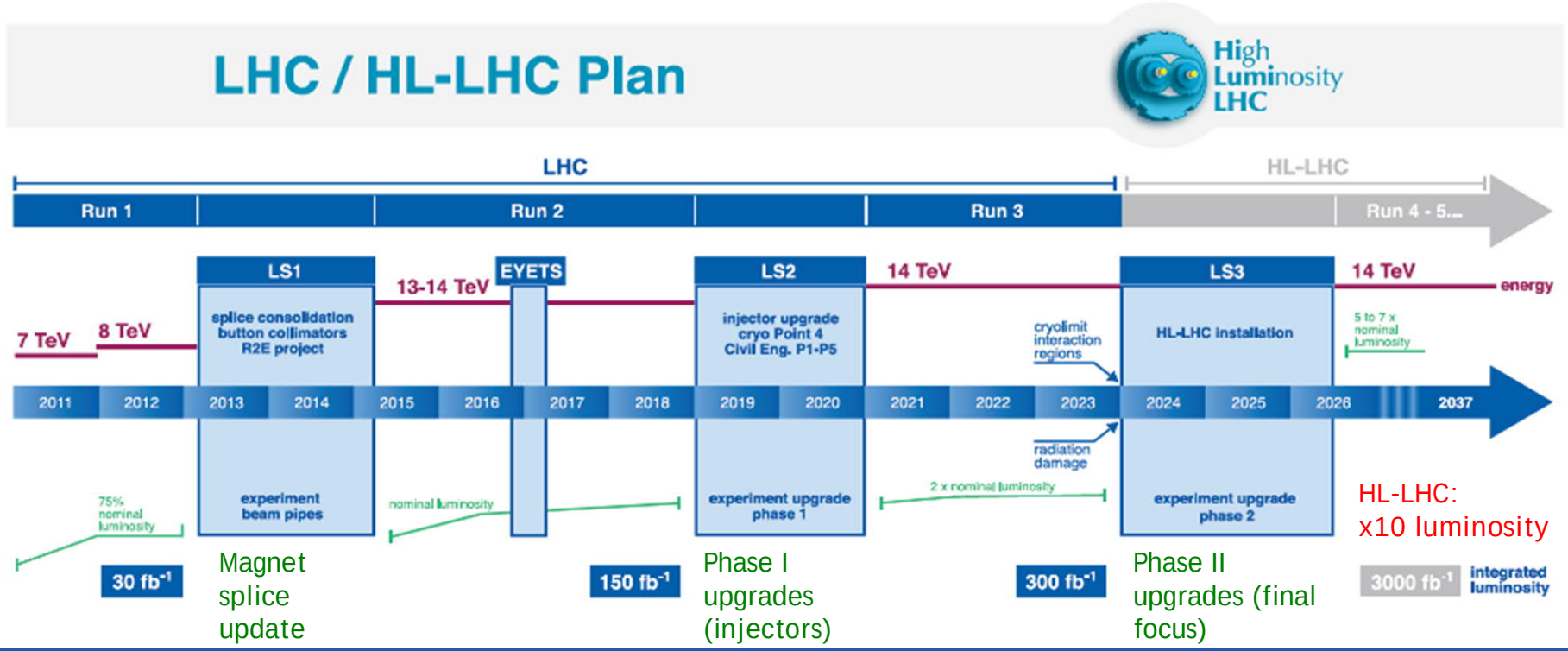
jobs@CERN like 100K people watching all a different movie with 1 MB/s streaming average

Authentication in HEP

- The Grid has established X509 proxy certificates to delegate identities between federated services across multiple sites
  - Extensions include granting dedicated roles to users
- The ensemble of certificates, CAs, Grid map files, proxy extensions and proxy certificate delegations is quite awkward…
  - Often a nightmare to implement and maintain => regular source for authentication problems
- Positive evolution: adoption of industry standard OpenID/OAuth2

Future Outlook

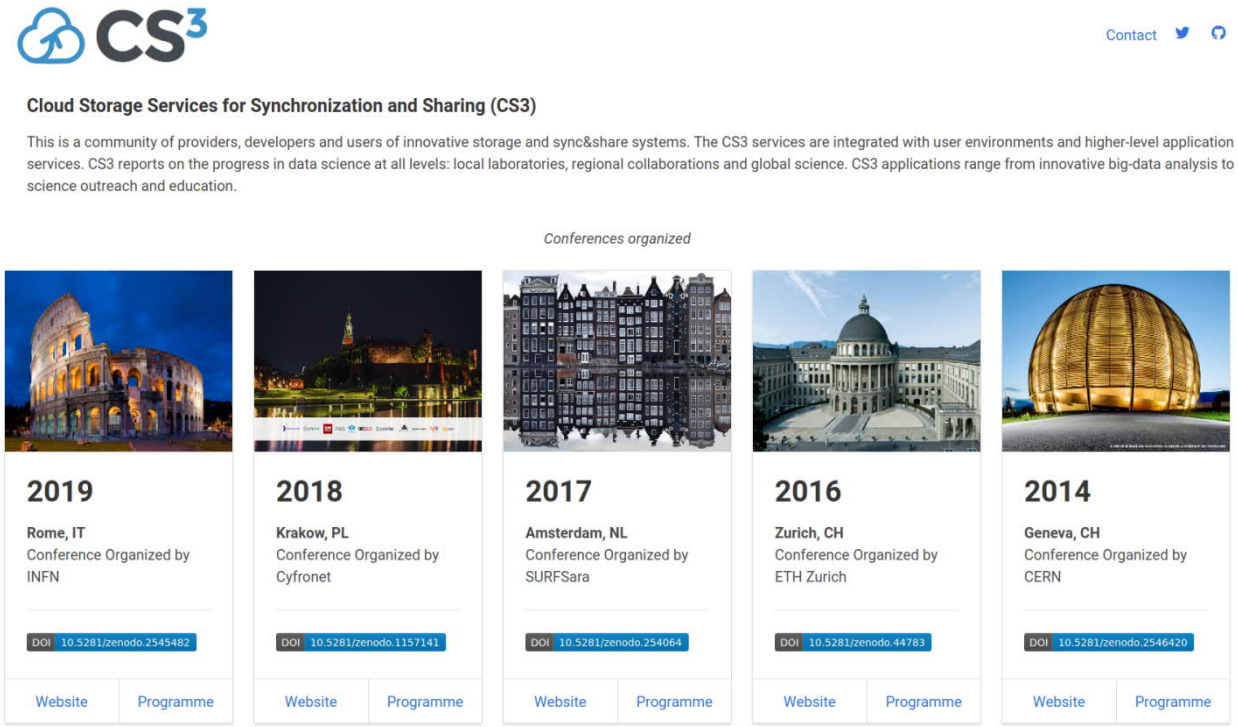
HL-LHC: a computing challenge



HL-LHC and friends

- High Luminosity LHC is not alone in the current arena of large scientific projects, several **Big Science collaborations** coming up:
  - Square Kilometer Array (**SKA**)
  - Cherenkov Telescope Array (**CTA**)
  - Deep Underground Neutrino Experiment (**DUNE**): prototype at CERN, full-sized experiment in USA
  - Low Frequency Array (**LOFAR**)
  - What about **Fusion**?
- Time for R&D, opportunity for new **synergies**

CS3: A new emerging community



6 workshop meet-ups since 2014

Last edition: Copenhagen, two weeks ago!
http://cs3.deic.dk
- 130+ participants
- 60+ contributions
- Launched a new EU project: the **CS3 MESH**

Industry participation:
- Start-Ups: Cubbit, pydio, …
- SMEs: OnlyOffice, ownCloud
- Big: AWS, Dropbox, …

Community website:
http://www.cs3community.org

Inventory

- The HEP community has built a modular stack of storage related software components
  - => allowed to integrate more or less any storage solution into the GRID
- In the context of the HEP Software Foundation and EU projects like XDC, ESCAPE, CS3MESH, etc. **many of these components are made available to other sciences**
- Changes in storage service implementations/technology tend to be slow, but happen
  - Stateful nature of the services and limited resources to adopt new technologies

Conclusions

- This was a very brief overview of a complex software ecosystem for a diverse user community
  - Two decades of development, fine-tuning, and production data taking/analysis
- Similarities and differences vs. industry
- More challenges ahead
  - And more science communities getting involved

Thanks for your attention! Questions?

Accélérateur de science

Credits to all CERN IT Storage colleagues