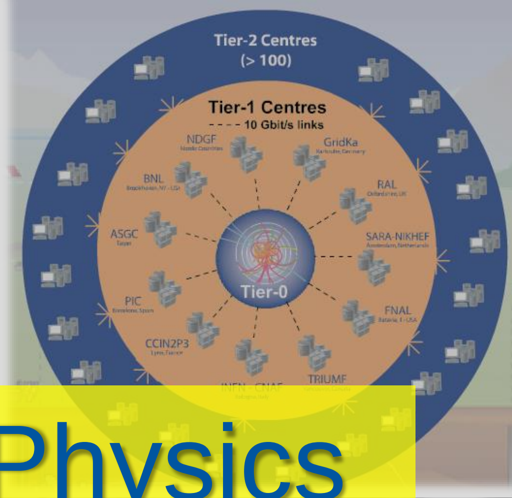


Computing in High-Energy Physics

Dr Helge Meinhard / CERN-IT
CERN openlab summer student lecture
(with additions/updates by D. Duellmann)
3 July 2018

CERN

"Science for peace"

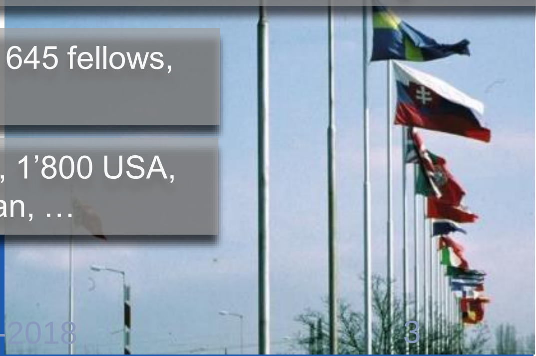
- International organisation close to Geneva, straddling Swiss-French border, founded 1954
- Facilities for fundamental research in particle physics
- 22 member states, 1.1 B CHF budget
- ~ 2'500 staff, +fellows, +apprentices, +you, …
- ~ 12'000 visiting scientists

1954: 12 Member States

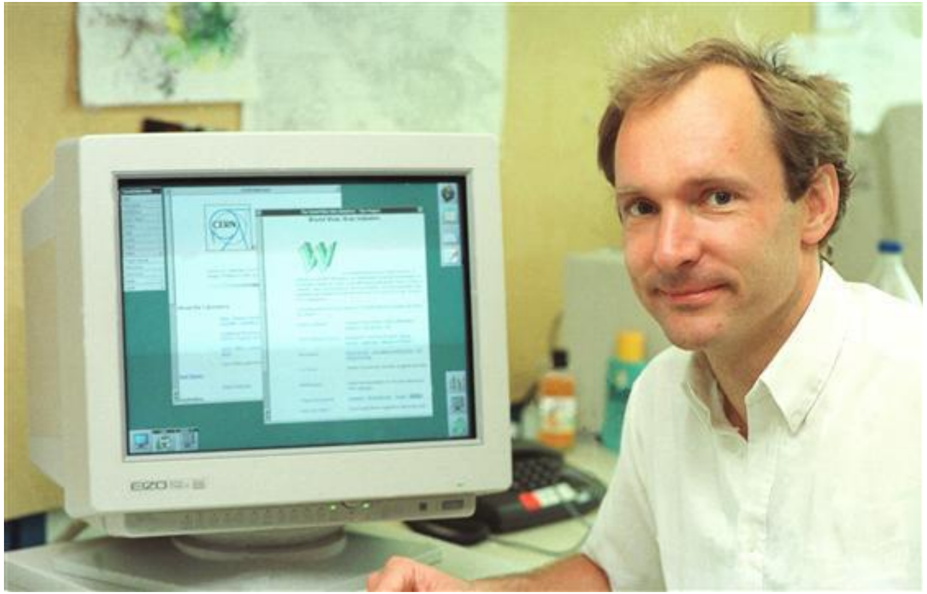
Members: Austria, Belgium, Bulgaria, Czech Republic, Denmar
Finland, France, Germany, Greece, Hungary, Israel, Italy, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Spain, Sweden, Switzerland, United Kingdom
Candidate for membership: Romania
Associate members: India, Lithuania, Pakistan, Turkey, Ukrain
Observers: European Commission, Japan, Russia, UNESCO, United States of America
Numerous **non-member states with collaboration agreement**

2'531 staff members, 645 fellows, 21 apprentices

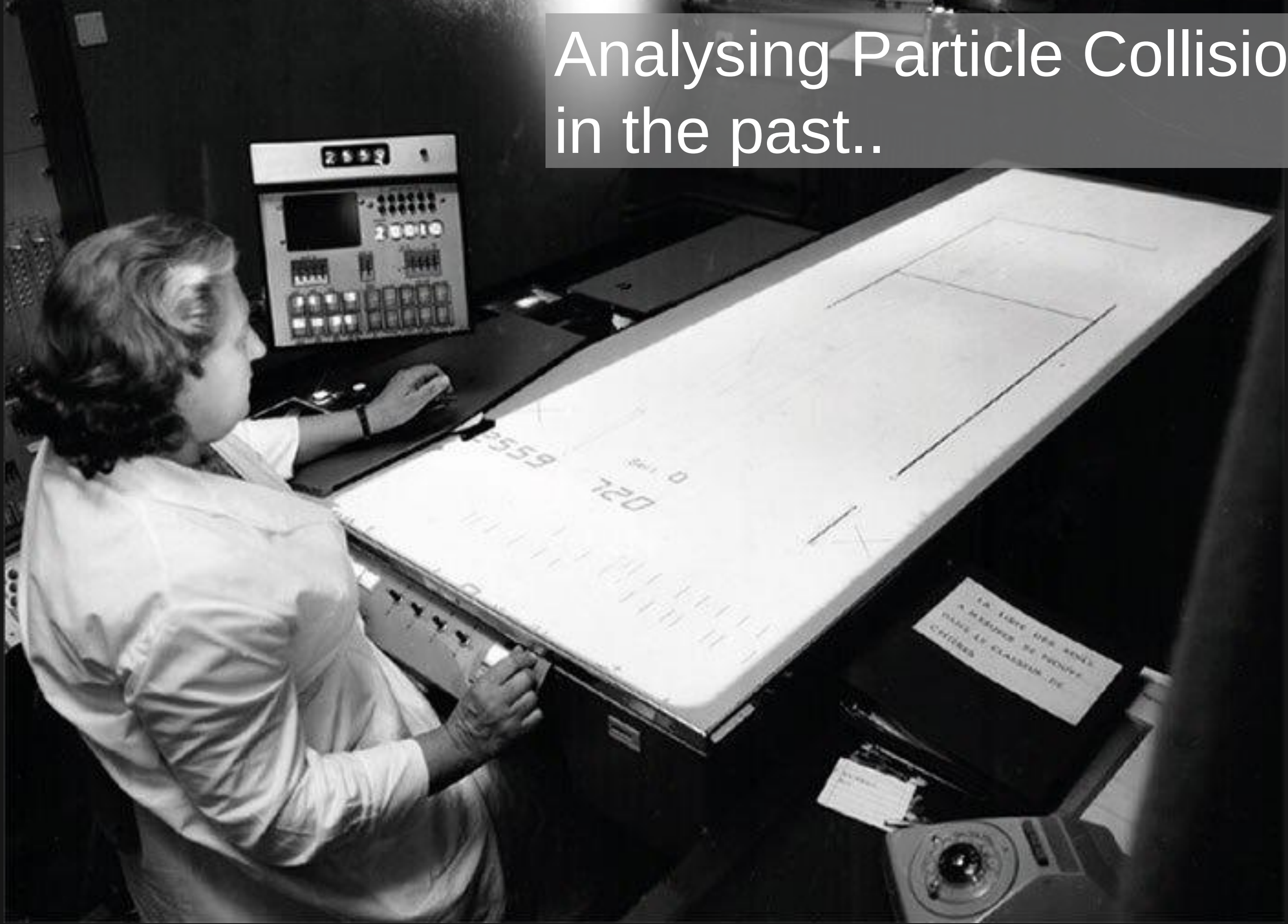
7'000 member states, 1'800 USA, 900 Russia, 270 Japan, …



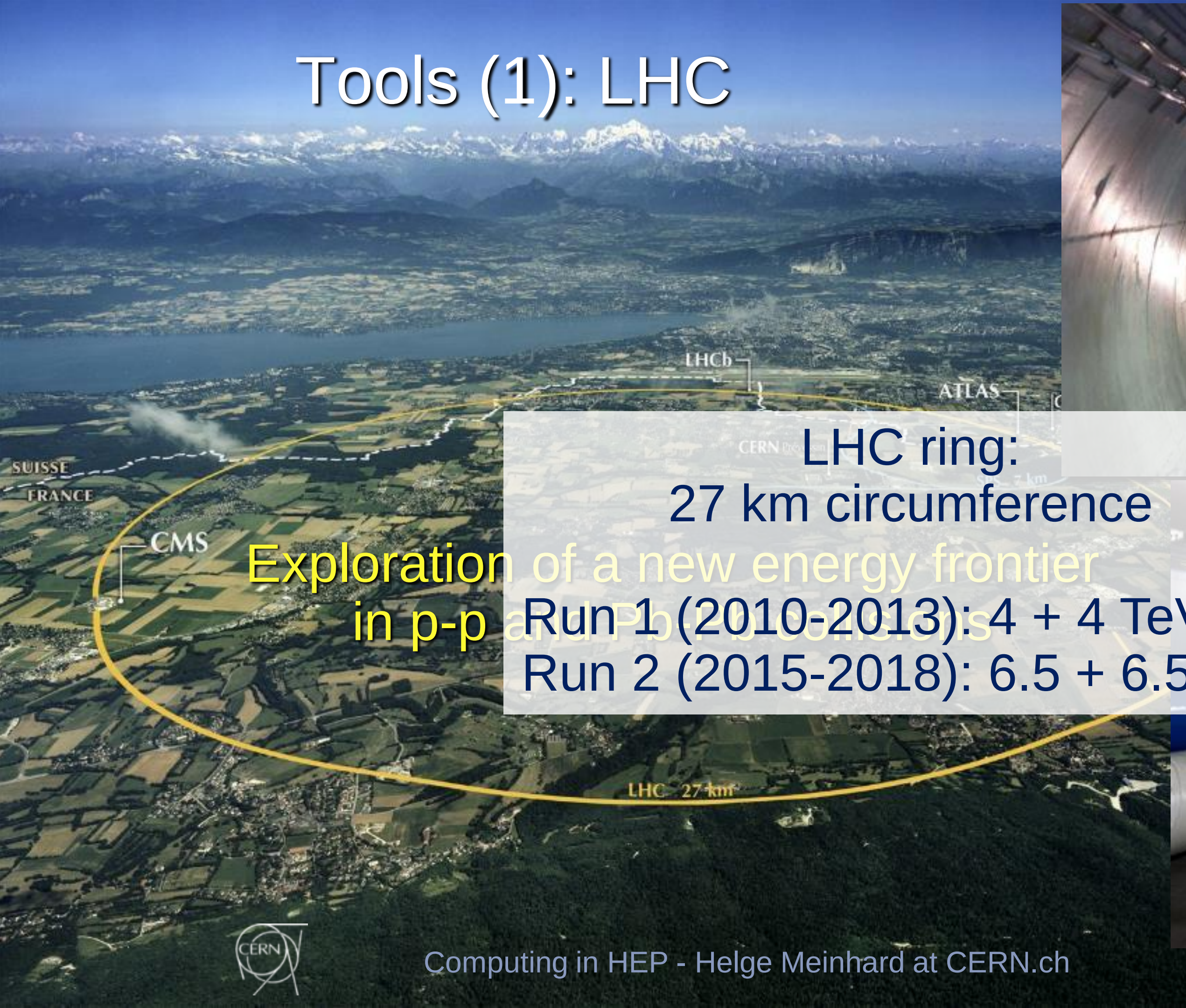
CERN – Where the Web was Born



Analysing Particle Collisions in the past..



Tools (1): LHC



LHC ring:
27 km circumference

Exploration of a new energy frontier
in p-p and Pb-Pb collisions

Run 1 (2010-2013): 4 + 4 TeV
Run 2 (2015-2018): 6.5 + 6.5 TeV

Tools (2): Detectors

pp, B-Physics, CP Violation
(matter-antimatter symmetry)

ATLAS (A Toroidal Lhc ApparatuS)

- 25 m diameter, 46 m length, 7'000 tons
- 3'000 scientists (including 1'000 grad students)
- 150 million channels
- 40 MHz collision rate
- Event rate after filtering: 300 Hz in Run 1; 1'000 Hz in Run 2

General Purpose,
proton-proton, heavy ions
Discovery of new physics:
Higgs, Supersymmetry

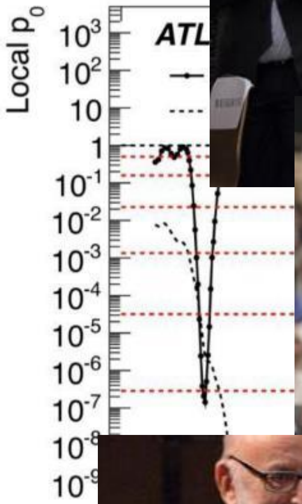
Exploration of new energy frontier
in p-p and Pb-Pb collisions

ALICE

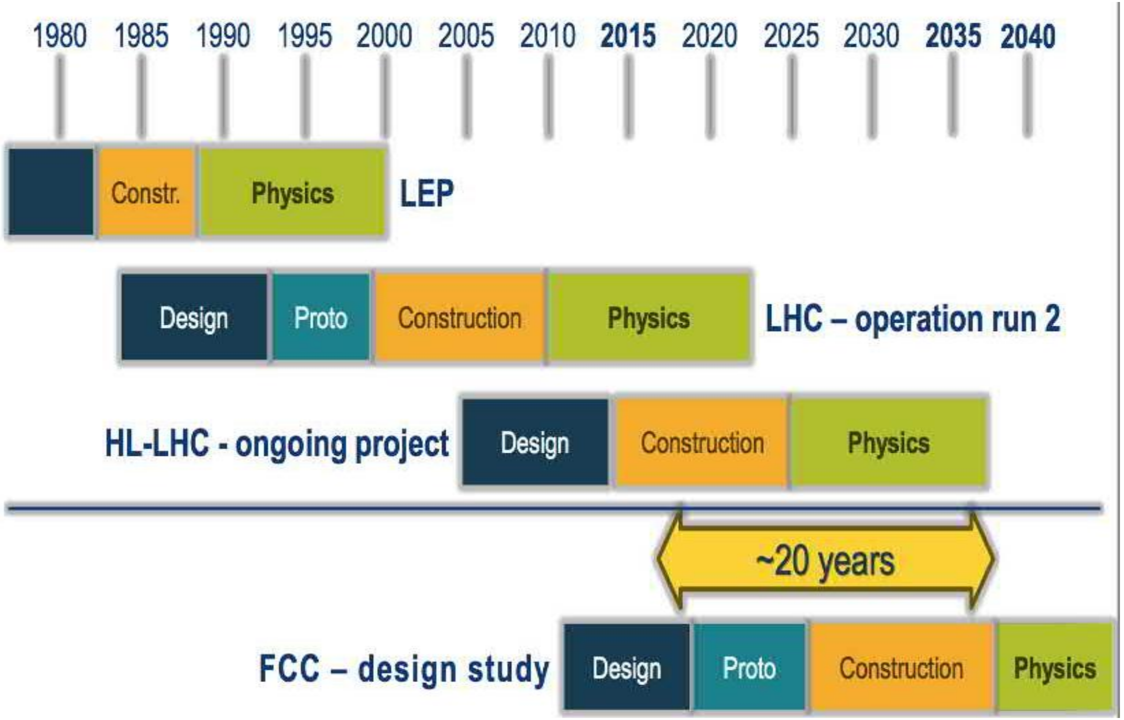
Heavy ions, pp
(state of matter of early universe)

Results so far

- Many… the most spectacular one being
- 04 July 2012: Discovery of a “Higgs-like particle”
- March 2013: The particle is indeed a Higgs boson
- 08 Oct 2013 / 10 Dec 2013: Nobel price to Peter Higgs and François Englert
  - CERN, ATLAS and CMS explicitly mentioned

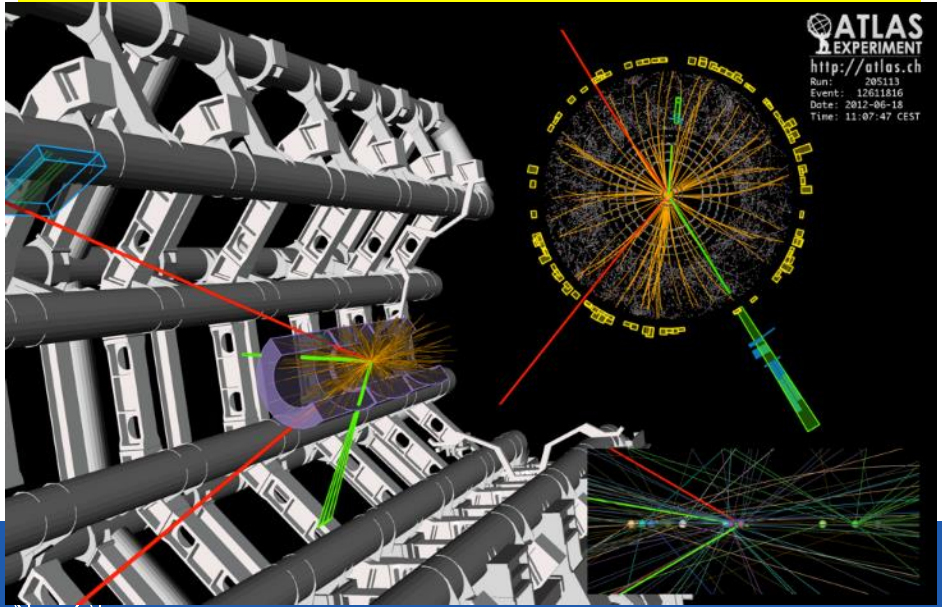


HEP Timescales



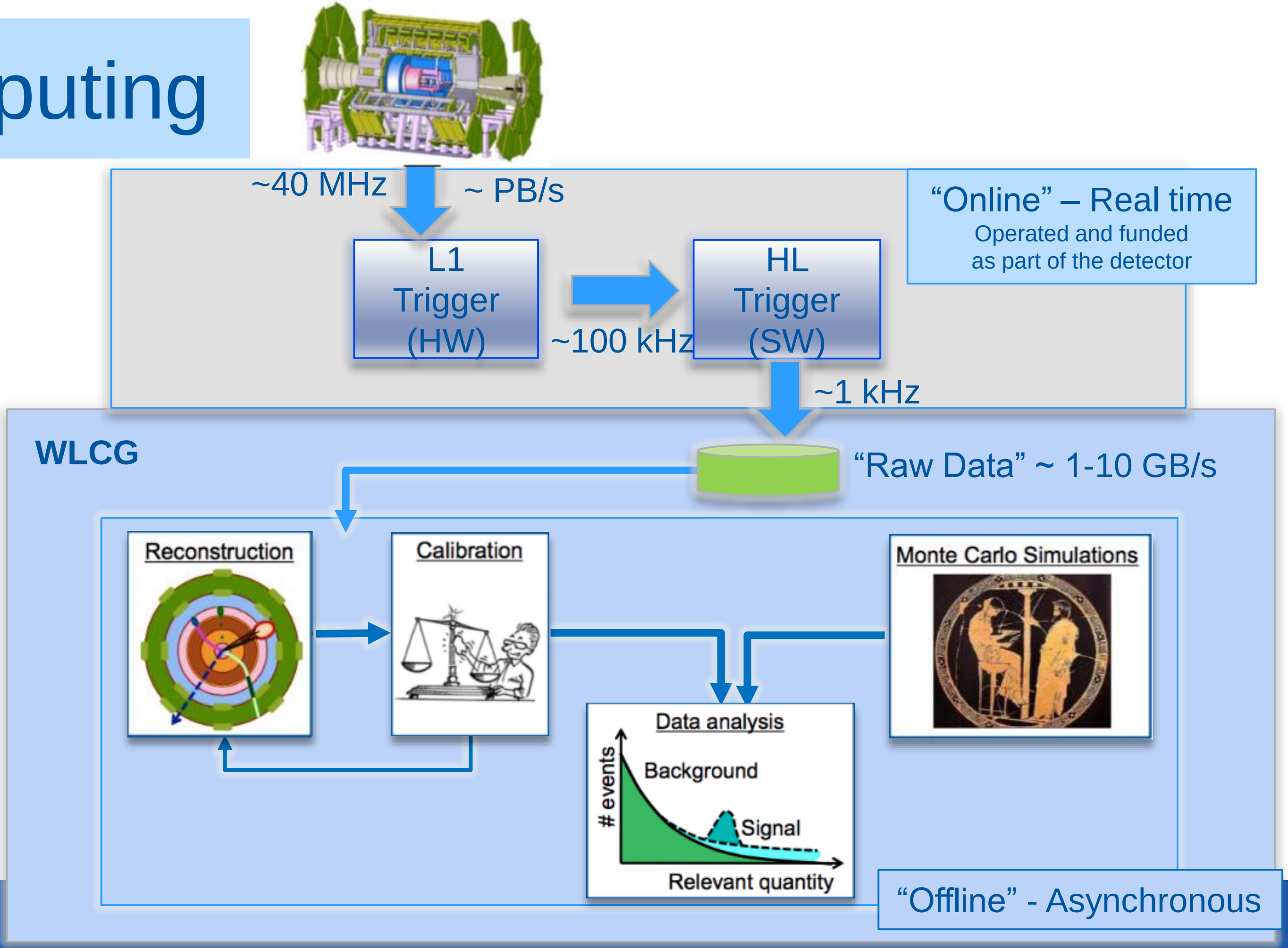
LHC Data – what does it consist of?

- 150 million sensors deliver data … 40 million times per second
- Generates ~ 1 PB per second

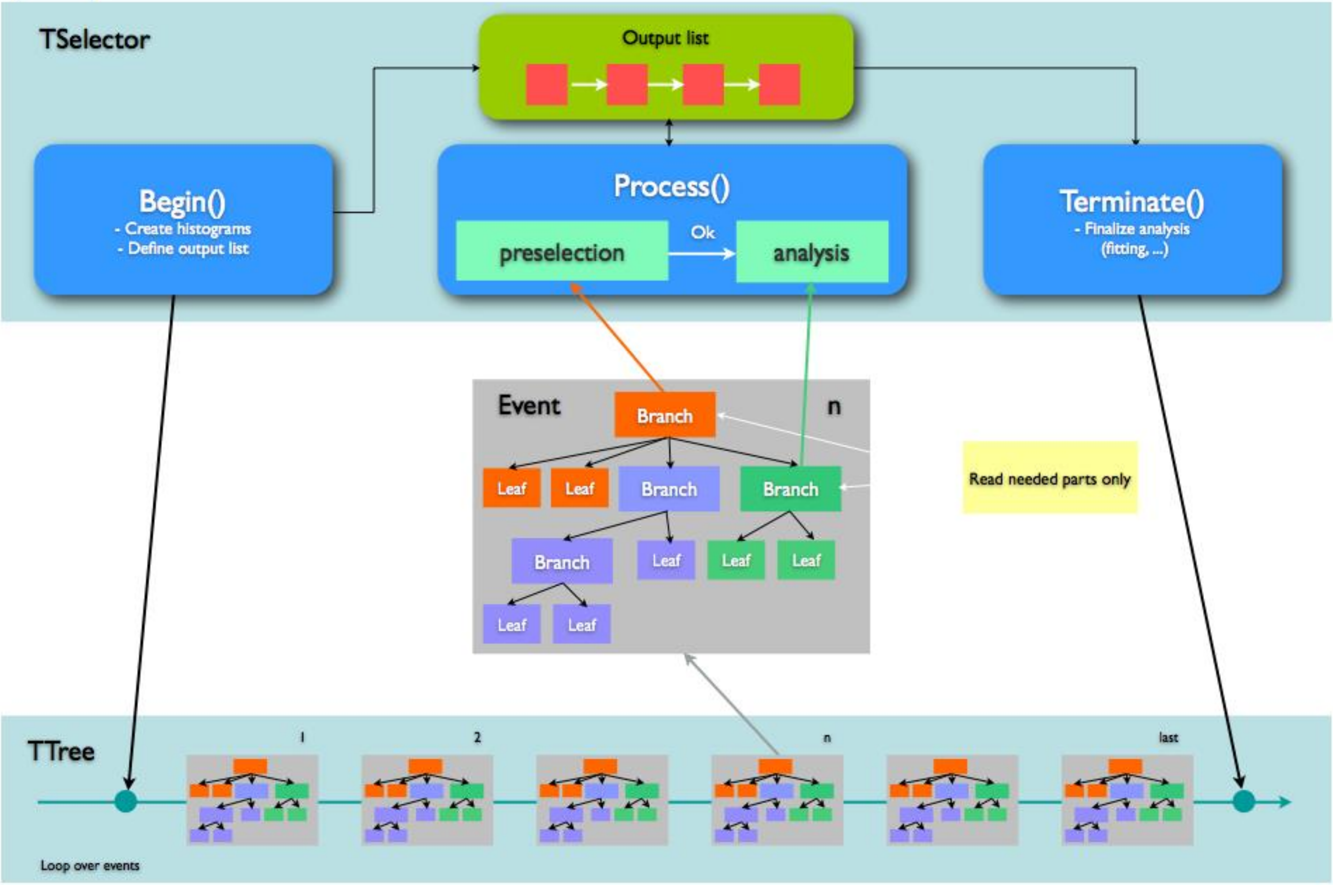


- Raw data:
  - Was a sensor hit?
  - How much energy deposit?
  - What time?
- Reconstructed data:
  - Momentum of tracks (4-vectors)
  - Origin
  - Energy in clusters (jets)
  - Particle type
  - Calibration information
  - …

HEP Computing



A Typical Physics Analysis Flow



Nature of the Computing Task

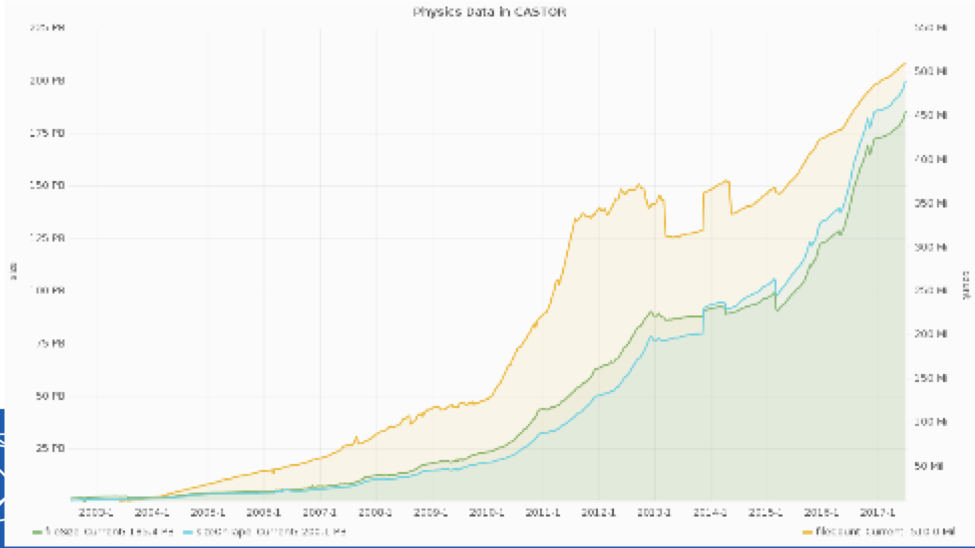
- Enormous number of proton or heavy ion collisions
  - Data from each collision are small (for protons: order 1…10 MB)
  - Each collision is “independent” of other collisions
- No supercomputers needed
  - Most cost-effective solution is standard PC architecture (x86) servers with 2 sockets, SATA drives (spinning or SSD), Ethernet network
  - Linux (RHEL variants: Scientific Linux, CentOS) used everywhere
- Calculations are mostly combinatorics
  - Rather integer than floating-point intensive

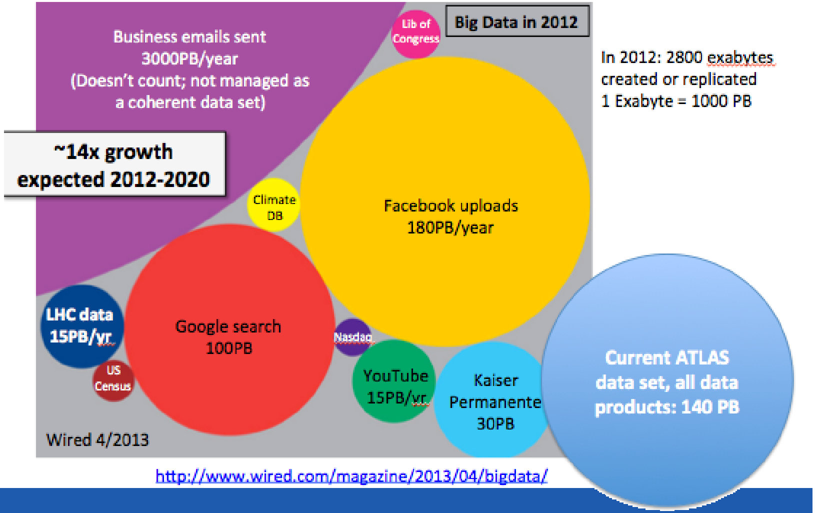
Scale of the Computing Problem

- Raw data: order 1…10 MB per collision event
  - 1 kHz, for ~7.10^6 live seconds / year
  - ➢ 7 PB/year …. per detector
- Several copies, derived data sets, replicated many times for performance, accessibility, etc

➢ ATLAS (for example) has a managed data set of ~ 285 PB

➢ CERN data archive on tape is ~200 PB



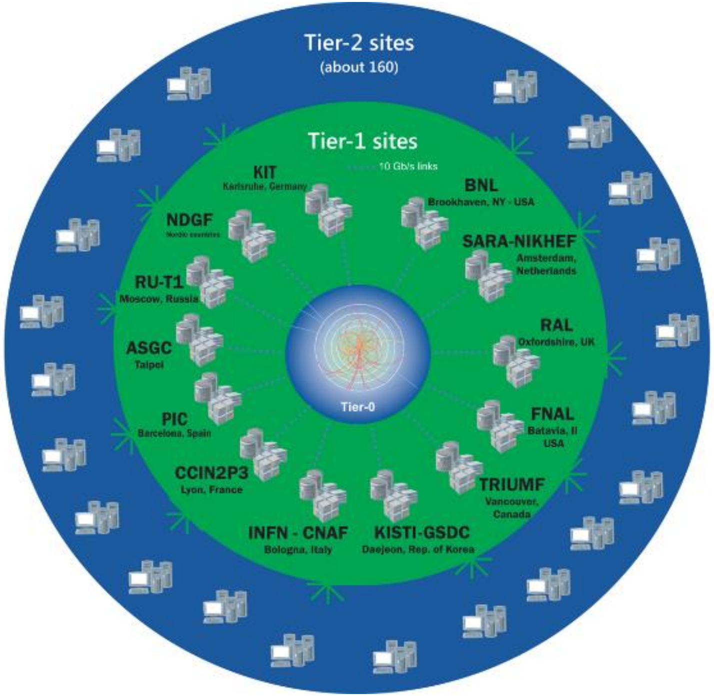


The Worldwide LHC Computing Grid

Tier-0 (CERN):
data recording, reconstruction and distribution

Tier-1:
permanent storage, re-processing, analysis

Tier-2:
Simulation, end-user analysis



~170 sites, 42 countries

~750'000 cores

~1'000 PB of storage

> 2 million jobs/day

10-100 Gb links

WLCG: An international collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource to a single infrastructure accessible by all LHC physicists

WLCG – a World-wide Infrastructure

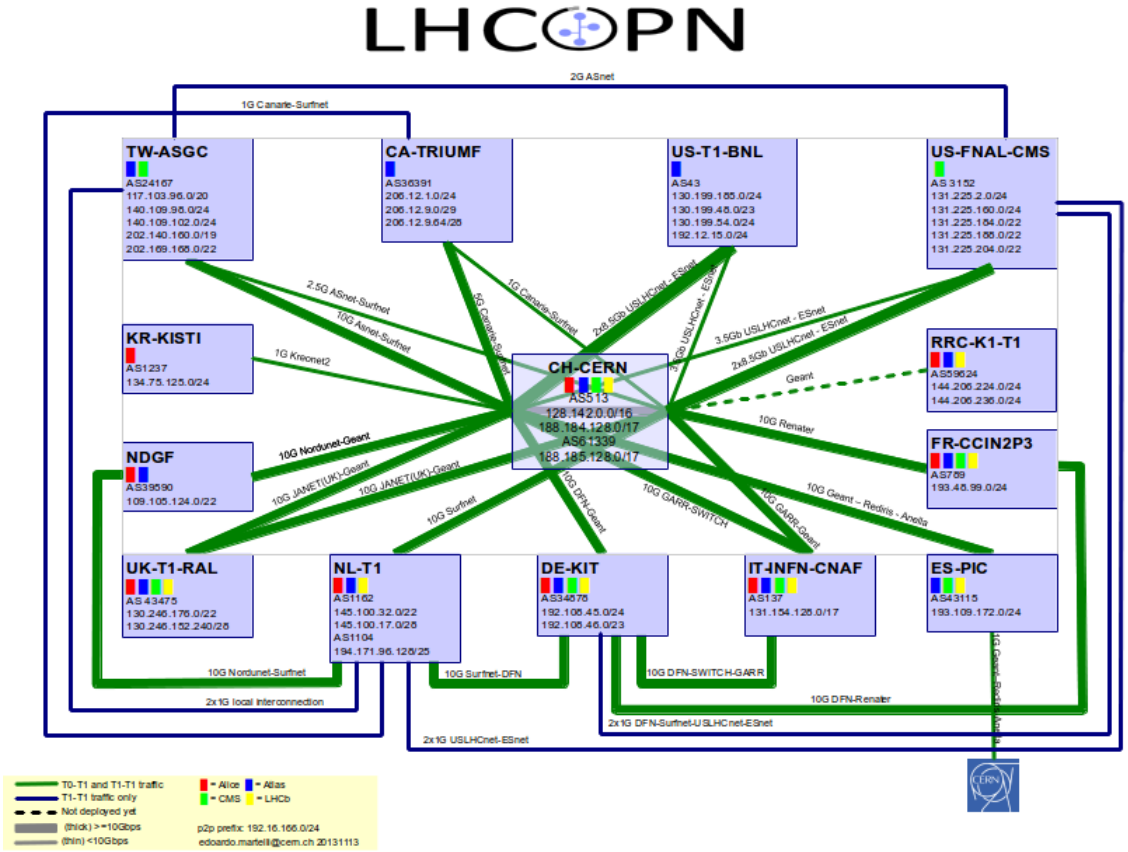
A distributed Tier-0

CERN data centre
3.5 MW - full

Tier-0 extension:
Wigner Research Centre,
Budapest/Hungary

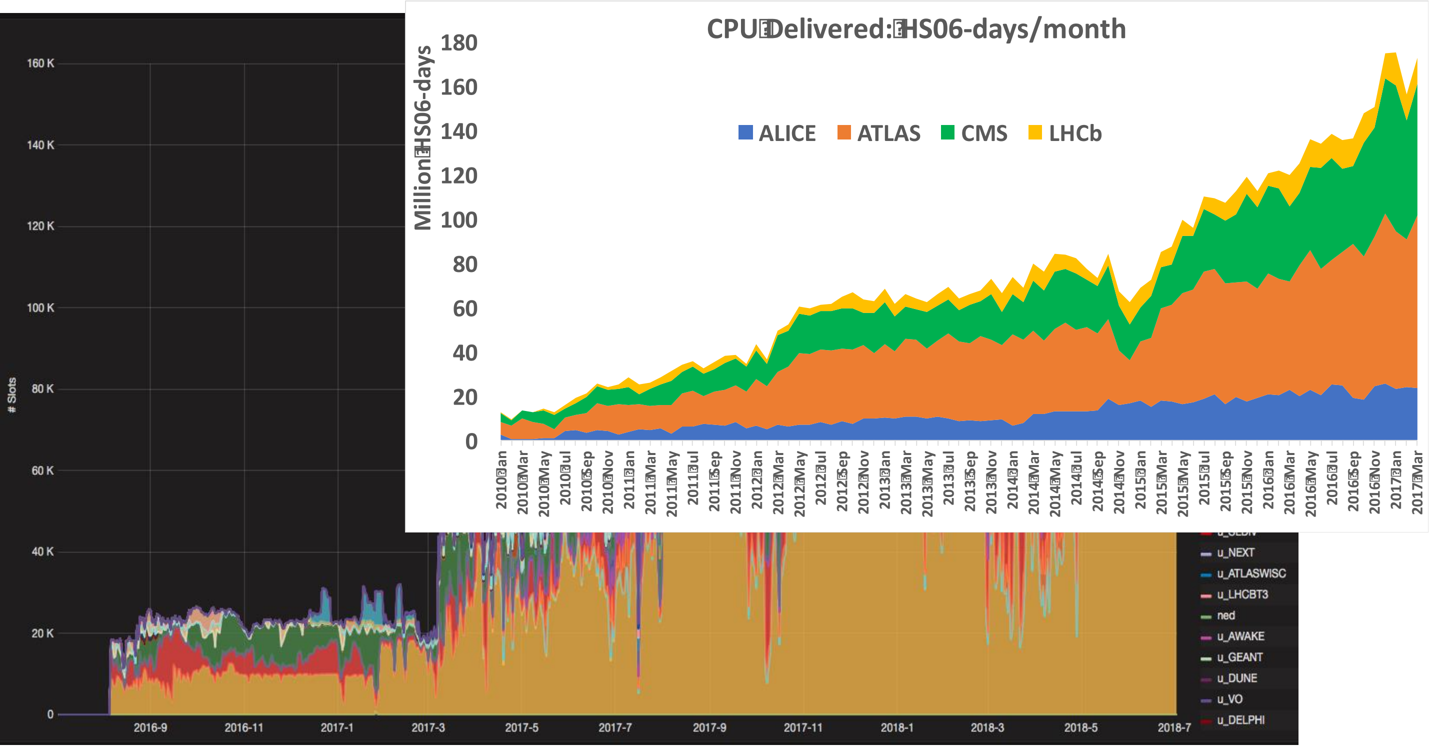
Three dedicated
100GE connections,
~25 ms RTT

COMPUTING		STORAGE	
Servers (Meyrin)	Cores (Meyrin)	Disks (Meyrin)	Tape Drives
11.5 K	174.3 K	61.9 K	104
Servers (Wigner)	Cores (Wigner)	Disks (Wigner)	Tape Cartridges
3.5 K	56.0 K	29.7 K	32.2 K

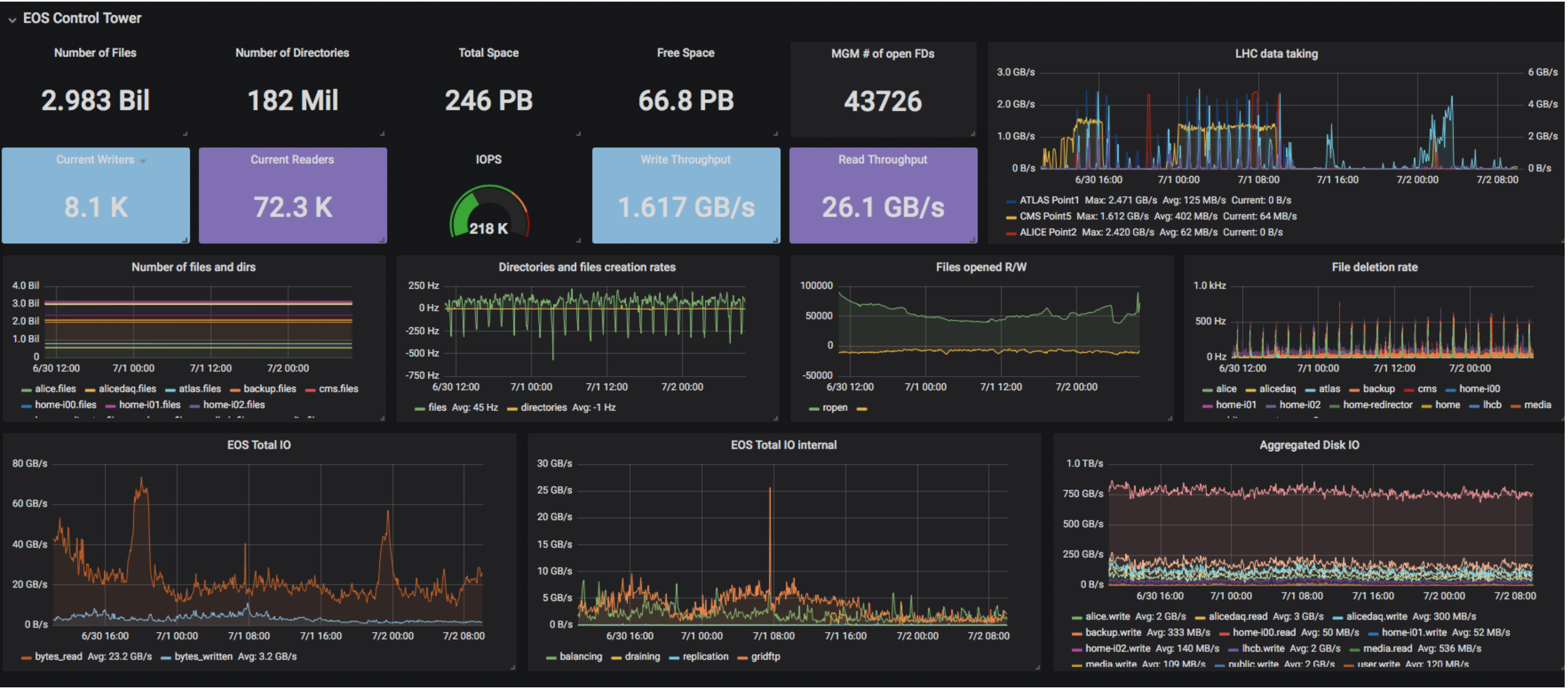


- Optical Private Network
- Support T0 – T1 transfers
- Some T1 – T1 traffic
- Managed by LHC Tier 0 and Tier 1 sites

Processing Scale



WLCG:
> 2 M jobs/day
on ~1M CPU cores

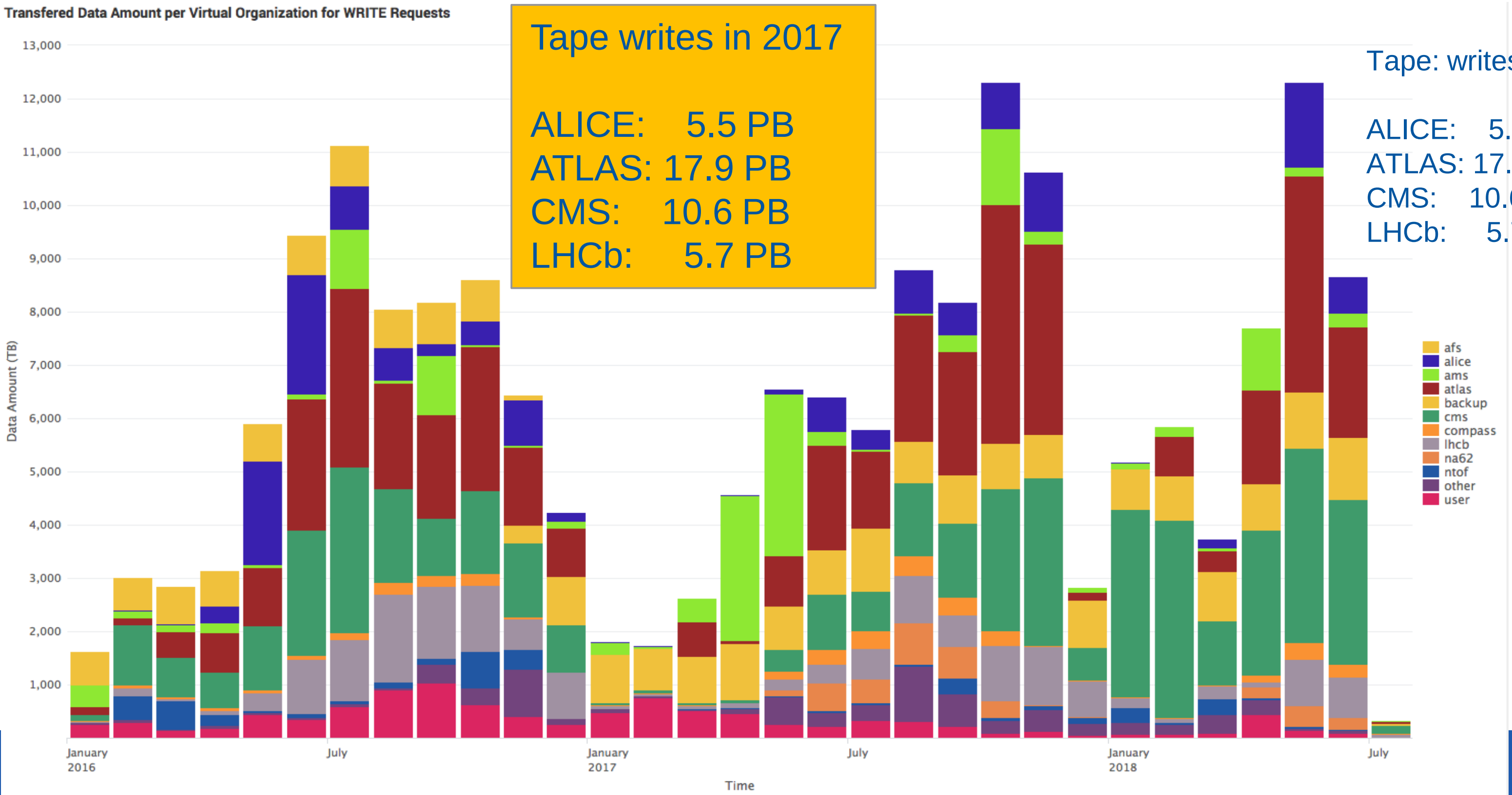

EOS Control Tower
Number of Files
2.983 Bil
Number of Directories
182 Mil
Total Space
246 PB
Free Space
66.8 PB
MGM # of open FDs
43726
LHC data taking
ATLAS Point1 Max: 2.471 GB/s Avg: 125 MB/s Current: 0 B/s
CMS Point5 Max: 1.612 GB/s Avg: 402 MB/s Current: 64 MB/s
ALICE Point2 Max: 2.420 GB/s Avg: 62 MB/s Current: 0 B/s
Current Writers
8.1 K
Current Readers
72.3 K
IOPS
218 K
Write Throughput
1.617 GB/s
Read Throughput
26.1 GB/s
Number of files and dirs
Directories and files creation rates
files Avg: 45 Hz
directories Avg: -1 Hz
Files opened R/W
File deletion rate
EOS Total IO
bytes_read Avg: 23.2 GB/s
bytes_written Avg: 3.2 GB/s
EOS Total IO internal
balancing
draining
replication
gridftp
Aggregated Disk IO

Media hierarchy

- We still use tape! Why?
  - $/PB (TCO incl. power)
  - separate physical copy with high “destruction” latency
- We stopped trying “automatic” HSM (Hierarchical Storage Management)
  for large experiment users
  - file based HSM interface did not allow to specify user priorities
- Disk content is stable (until the experiment decides to replace active data)
  - thousands of job streams at relatively low rate (cpu bound)
- Tape access enabled only for a few production activities

CERN

Scale Examples: Tape Archive



Tape writes in 2017

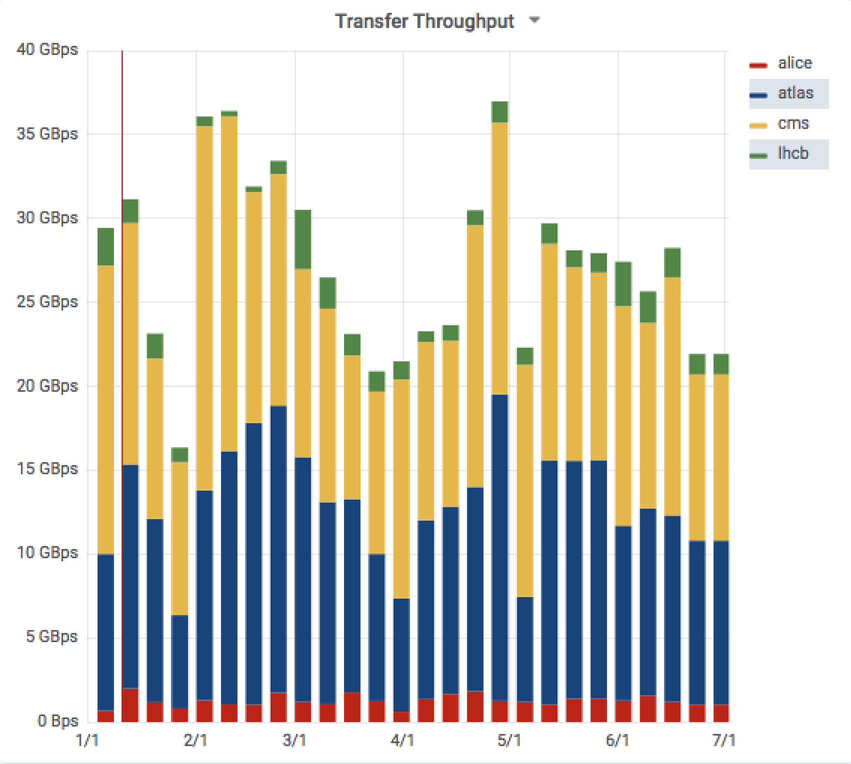
ALICE: 5.5 PB
ATLAS: 17.9 PB
CMS: 10.6 PB
LHCb: 5.7 PB

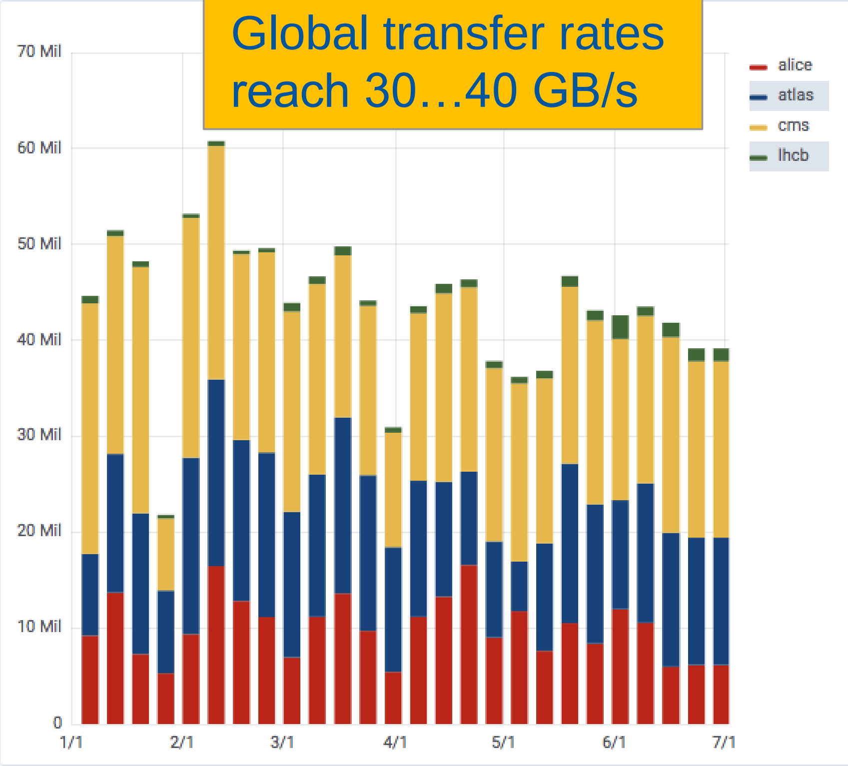
Tape: writes in 2017

ALICE: 5.5 PB
ATLAS: 17.9 PB
CMS: 10.6 PB
LHCb: 5.7 PB

Scale Example: Data Transfer

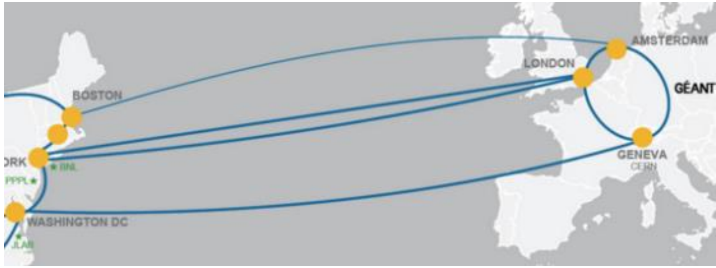
WLCG:
Global transfer rates reach 30…40 GB/s



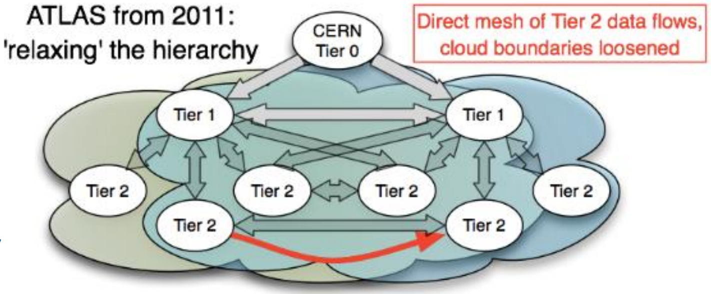


Distributed model

- Performance & reliability of the networks has *exceeded earlier expectations*
  - 10 Gb/s → 100 Gb/s at large centres
  - >100 Gb/s transatlantic links in place
  - Many Tier 2s connected at 10 Gb/s or better
  - NB. Still concern over connectivity at sites in less-well connected countries



- Strict hierarchical model of Tiers evolved during Run 1 to optimize the use of available resources
  - Move away from the strict roles of the Tiers to more functional and service quality based
  - Better use of the overall distributed system
- Focus on use of resources/capabilities rather than “Tier roles”
  - Data access peer-peer: removal of hierarchical structure



Transforming In-House Resources (1)

Before Wigner deployment:

- Physical servers only
  - Inefficient resource usage
  - Strong coupling of services with HW life-cycle
- Vertical view
  - Service managers responsible for entire stack
- Home-made tools of 10 years ago
  - Successful at the time, but Increasingly brittle
  - Lack of support for dynamic host creation/deletion
  - Limited scalability
- Person-power: (at best) constant
  - … despite many more machines

Transforming In-House Resources (2)

Current situation:

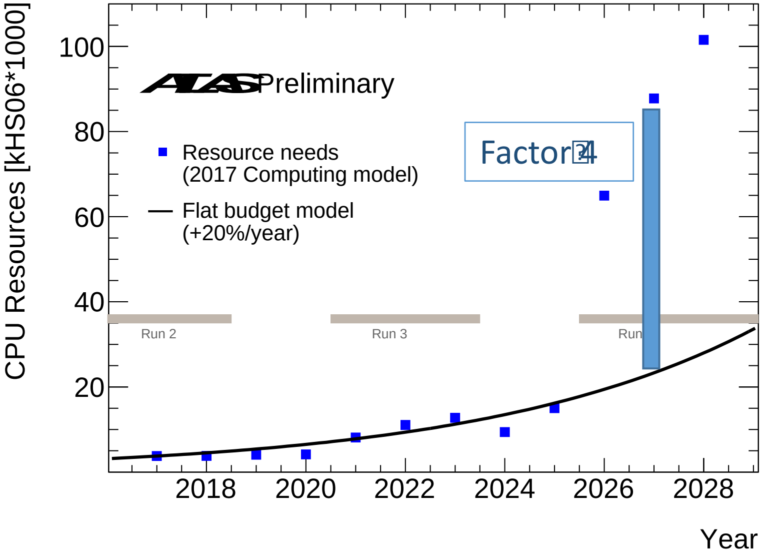
- Full support for physical and virtual servers
- Full support for remote machines
- Horizontal view
  - Responsibilities by layers of service deployment
- Large fraction of resources run as private cloud under OpenStack
- Scaling to large numbers
  (> 15’000 physical, several 100’000s virtual)
- Support for dynamic host creation/deletion
  - Deploy new services/servers in hours rather than weeks/months
  - Optimise operational and resource efficiency

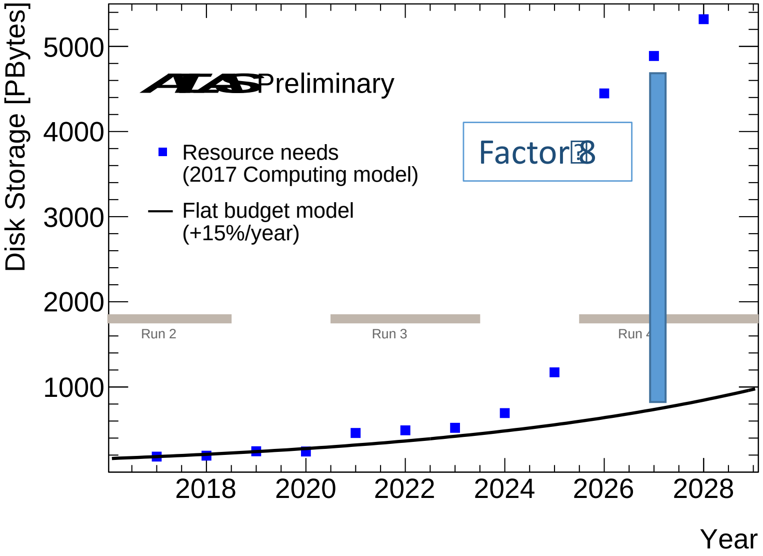
Future Challenges for LHC

- Assuming 20% per year from technology, still factors missing in terms of cores, storage etc.
- Moore's law coming to an end for business and financial reasons
- Large effort spent to improve software efficiency
- Exploit multi-threading, new instruction sets, …

Data:
- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:
- x60 from 2016


ATLAS Preliminary
Resource needs
(2017 Computing model)
Flat budget model
(+20%/year)
Factor 4
Run 2
Run 3
Run
CPU Resources [kHS06*1000]
100
80
60
40
20
2018
2020
2022
2024
2026
2028
Year

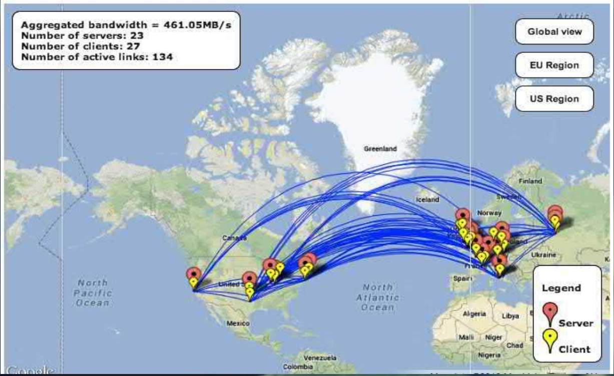

ATLAS Preliminary
Resource needs
(2017 Computing model)
Flat budget model
(+15%/year)
Factor 8
Run 2
Run 3
Run 4
Disk Storage [PBytes]
5000
4000
3000
2000
1000
2018
2020
2022
2024
2026
2028
Year

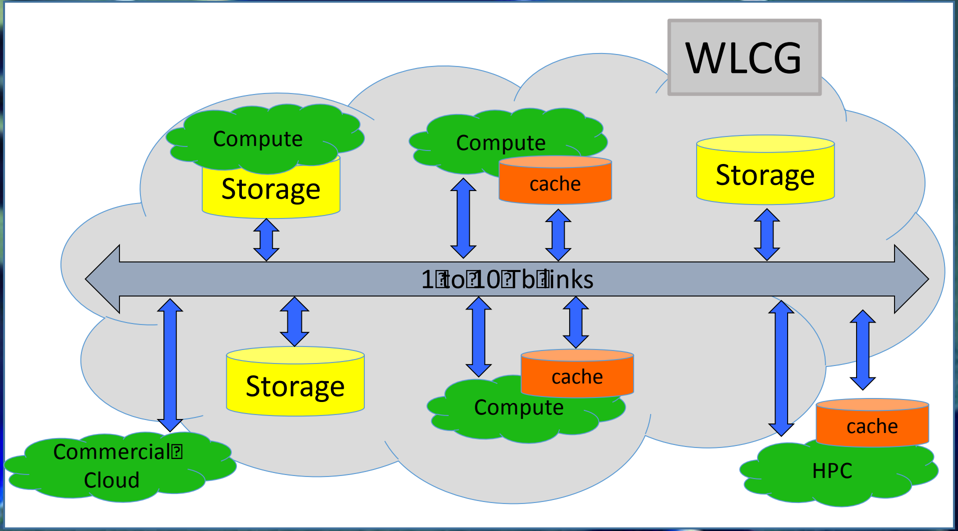
Trends – Software

- Recognizing the need to re-engineer HEP software
  - New architectures, parallelism everywhere, vectorisation, data structures, …

- HEP Software Foundation (HSF) set up (http://hepsoftwarefoundation.org/)
  - Community wide – buy-in from major labs, experiments, projects
  - Goals:
    - Address rapidly growing needs for simulation, reconstruction and analysis of current and future HEP experiments
    - Promote the maintenance and development of common software projects and components for use in current and future HEP experiments
    - Enable the emergence of new projects that aim to adapt to new technologies, improve the performance, provide innovative capabilities or reduce the maintenance effort
    - Enable potential new collaborators to become involved
    - Identify priorities and roadmaps
    - Promote collaboration with other scientific and software domains

Making hundreds of petabytes of data accessible globally to scientists is one of the biggest challenges of WLCG





Data Organization, Management and Access in WLCG

Maria Girone
CERN openlab CTO

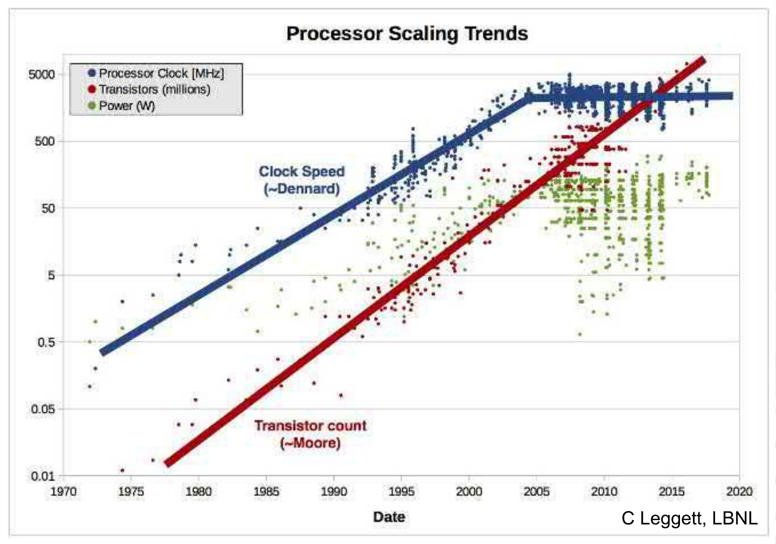
HEP has a vast investment in software

- Significant effort to make efficient multi-threaded and vectorized CPU code

Accelerated computing devices (GPUs, FPGAs) offer a different model

- Complexity of heterogeneous architectures

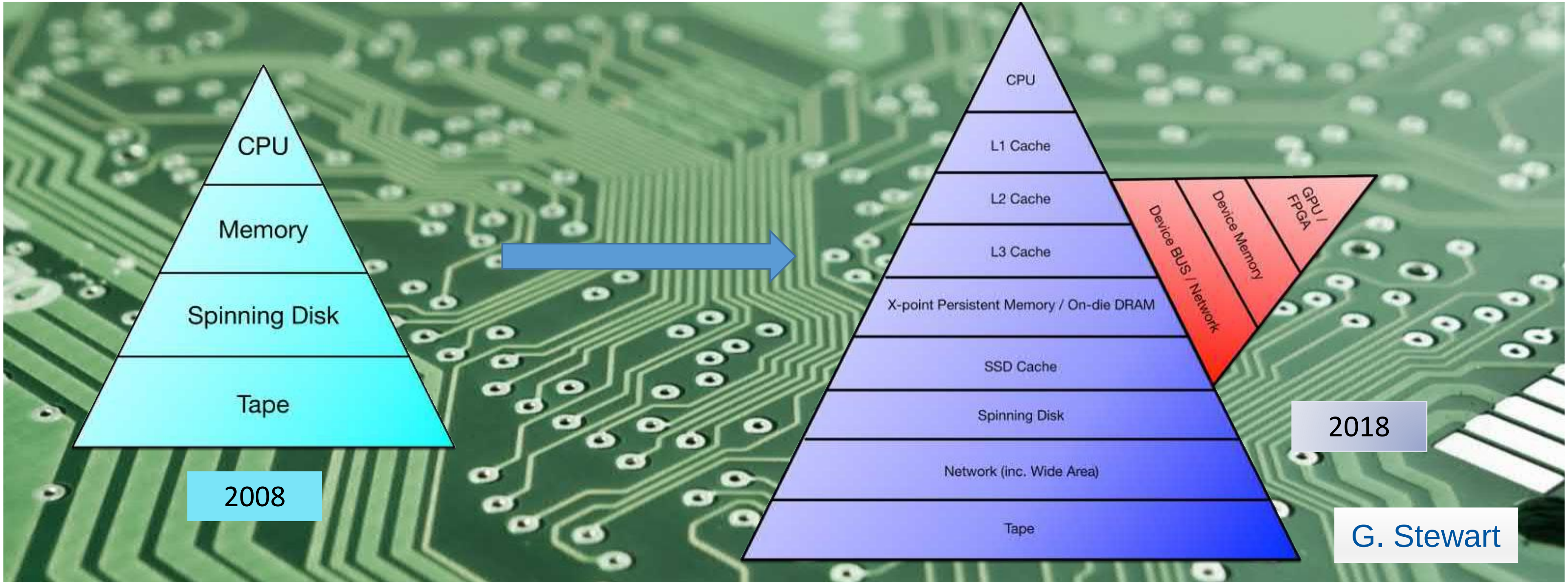
Simultaneously exploring lower performance but lower power alternatives like ARM



Software optimization can gain factors in performance.

Maria Girone
CERN openlab CTO

Computing Hierarchy Changes



Opportunistic resources

- Today this has become more important
  - Opportunistic use of:
    - HPC facilities
    - Large cloud providers
    - Other offers for “off-peak” or short periods
    - ...
    - All at very low or no cost (for hardware)
  - But scale and cost are unpredictable

- Also growing in importance:
  - Volunteer computing (citizen science)
    - BOINC-like (LHC@home, ATLAS/CMS/LHCb@home, etc)
    - Now can be used for many workloads – as well as the outreach opportunities

Drivers of Change

- Must reduce the (distributed) provisioning layer of compute to something simple, we need a hybrid and be able to use:
  - Our own resources
  - Commercial resources
  - Opportunistic use of clouds, grids, HPC, volunteer resources, etc.
- Move towards simpler site management
  - Reduce operational costs at grid sites
  - Reduce "special" grid middleware support cost
- Today (2015) it is cheaper for us to operate our own data centres
  - We use 100% of our resources 24x365
- We also get a large synergistic set of resources in many Tier 2s – essentially for "free" – over and above the pledged resources
- However, commercial pricing is now getting more competitive
  - Large scale hosting contracts, commercial cloud provisioning

Scaling up Further: New On-Premise Resources

- Option of new data centre at CERN explored
- On CERN Prévessin site for power reasons
- Multi-stage up to 12 MW
- Attractive solution feasible
  - Investments compensated by power and network cost savings over 10 years
- Project options and schedules being discussed
- Possible integration of experiment on-line needs later

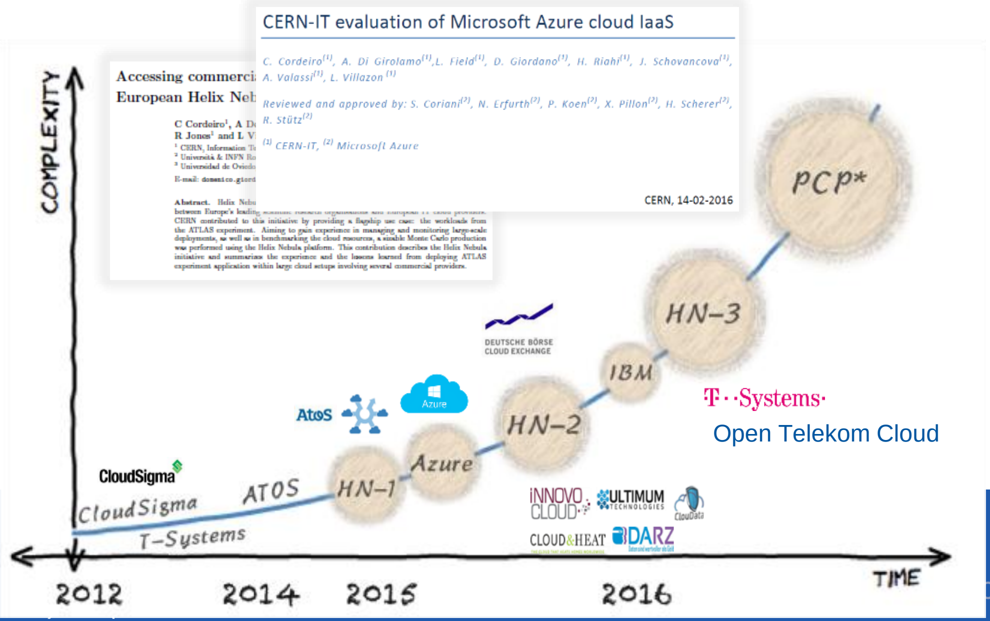
Scaling up Further: Commercial Clouds (1)

- Additional resources
  - Later to complement or replace on-premise capacity
- Potential benefits
  - Economy of scale
  - More elastic, adapts to changing demands
  - Somebody else worries about machines and infrastructure

- Potential issues
  - Cloud provider's business models not well adapted to procurement rules and procedures of public organisations
  - Lack of skills for and experience with procurements
  - Market largely not targeting compute-heavy tasks
    - Performance metrics/benchmarks not established
  - Legal impediments
  - Not integrated with on-premise resources and/or publicly funded e-infrastructures

Scaling up Further: Commercial Clouds (2)

- CERN
  - Series of short procurement projects of increasing size and complexity



- WLCG
  - Private cloud infrastructures at many sites
  - Use of AWS, Google, Rackspace etc. by FNAL, BNL, CERN, experiments, others
  - Helix Nebula The Science Cloud PCP project in Europe (together with other sciences)
  - Also testing real commercial procurements to understand cost
  - So far most use has been simulation, only now looking at data-intensive use cases

Conclusions

- LHC computing has successfully managed to collect and analyze in science unprecedented data volumes
- Initially used purpose-built tools, some of which of general utility for data-intensive sciences
  - Helping with adaptation / generalisation were needed
  - Focus on “core-business” and risk protection: eg ROOT, EOS
- Additional open-source tools and new technologies are being adopted/tested
  - Hadoop, Spark, Machine Learning, GPU based Deep Learning
  - This time some adaptation/generalization may be required on the side of HEP computing!
- Future expectations for data volume require further innovations,
  - Eg software optimisation and hardware investments beyond Moore/Kryder laws
- Integration between commercial clouds, scalable on-premise deployments and public e-infrastructures enables additional strategies
- The strength of HEP computing was always to seize the opportunities of the changing ITC market by exploiting the expertise in both computer engineering and scientific computing!

Thank you for your attention

Accelerating Science and Innovation