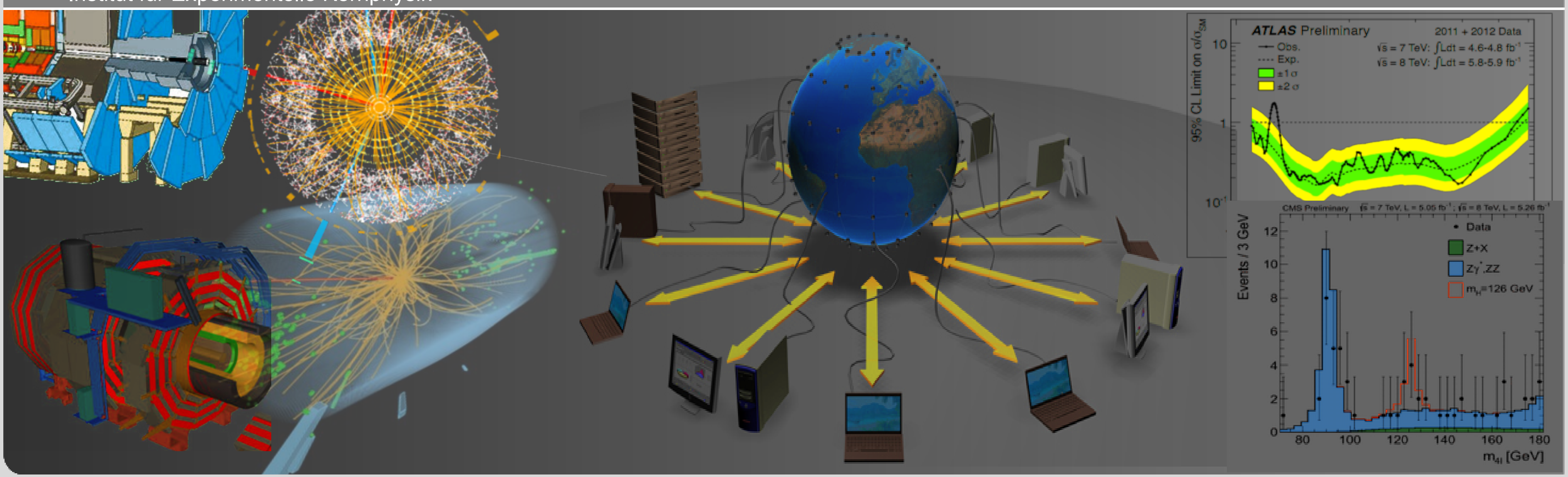
Karlsruher Institut für Technologie

Challenges for Software and Computing in HEP

Günter Quast

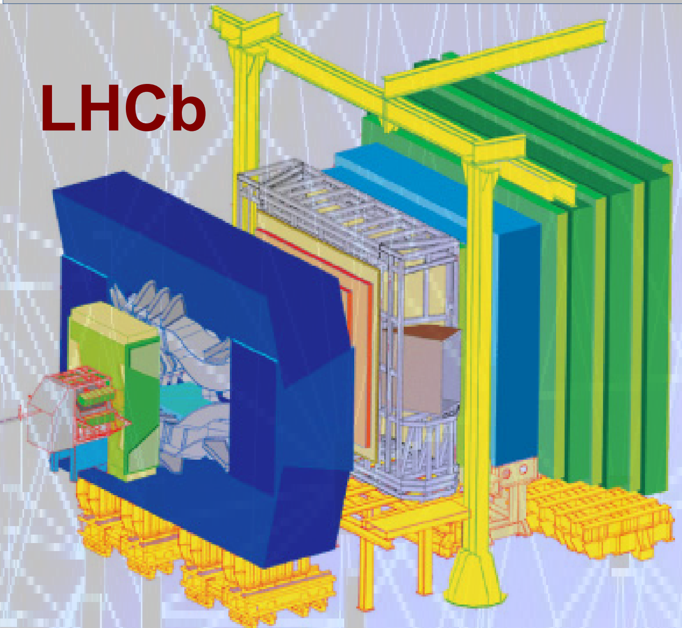
Fakultät für Physik
Institut für Experimentelle Kernphysik

HvdS Festkolloquium, 20 Dez. 2016



Scientific Data Sources

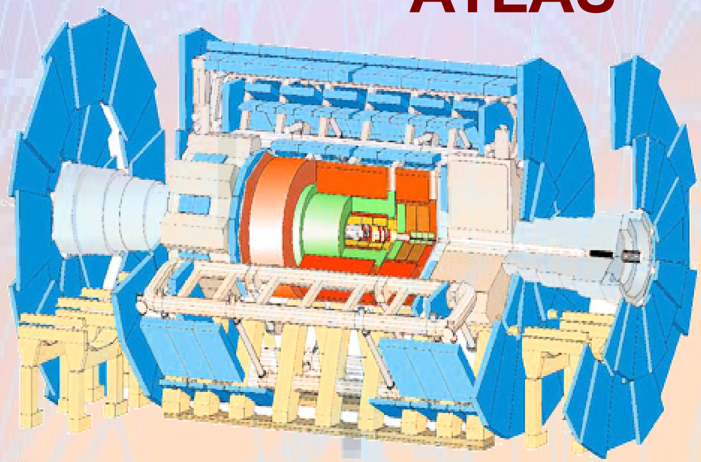
LHCb



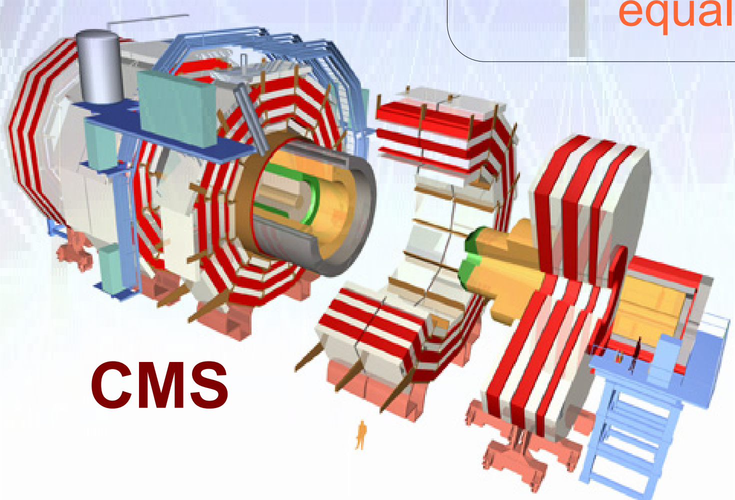
LHC detectors are large international projects

- Built by institutes from all over the world, ~10'000 scientists fom ~70 countries
- Detecors spcialised on different sicientific questions

ATLAS

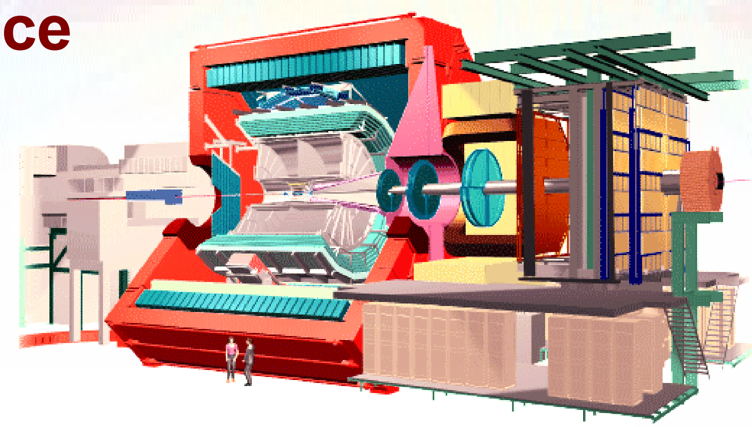


Important guiding principle:
equal access to all data for all participating scientists

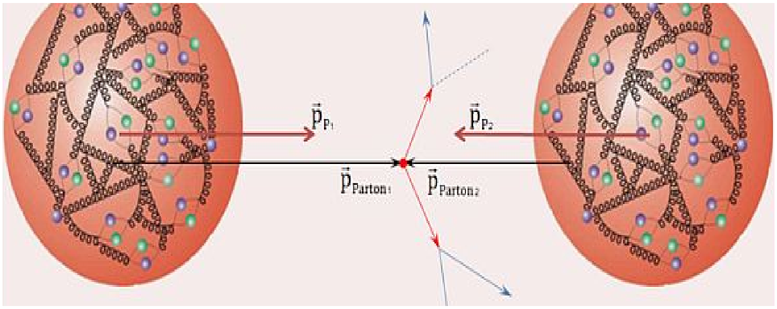


CMS

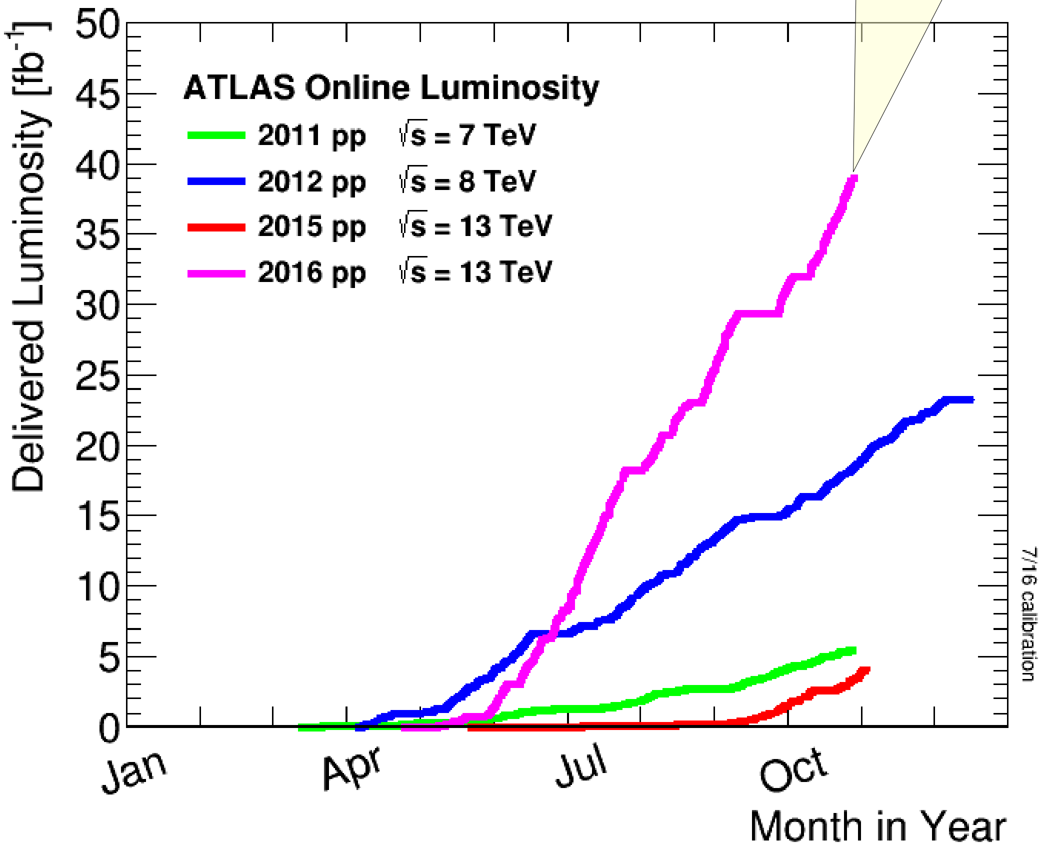
Alice

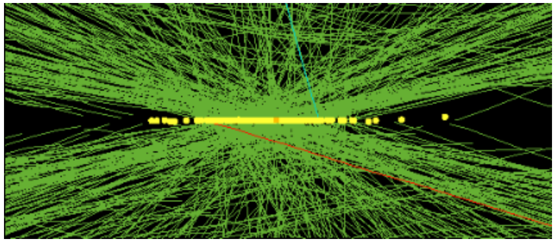


LHC Performance over the years



2016 run delivered more data than in in the 3 years before !





corresponds to 3.5 Trillion pp-collisions/experiment

Data Sources – example CMS

40 MHz (1000 TB/sec) equivalent

Level 1 - Hardware

100 kHz (100 GB/sec digitized)

Level 2 - Online Farm

10 kHz (10 GB/sec)

Level 3 - Online Farm

~1000 Hz (1 GB/sec)

Worldwide Community

- ~ 100 Million Detector cells
- LHC collision rate: 40 MHz
- 10-12 bit/ cell

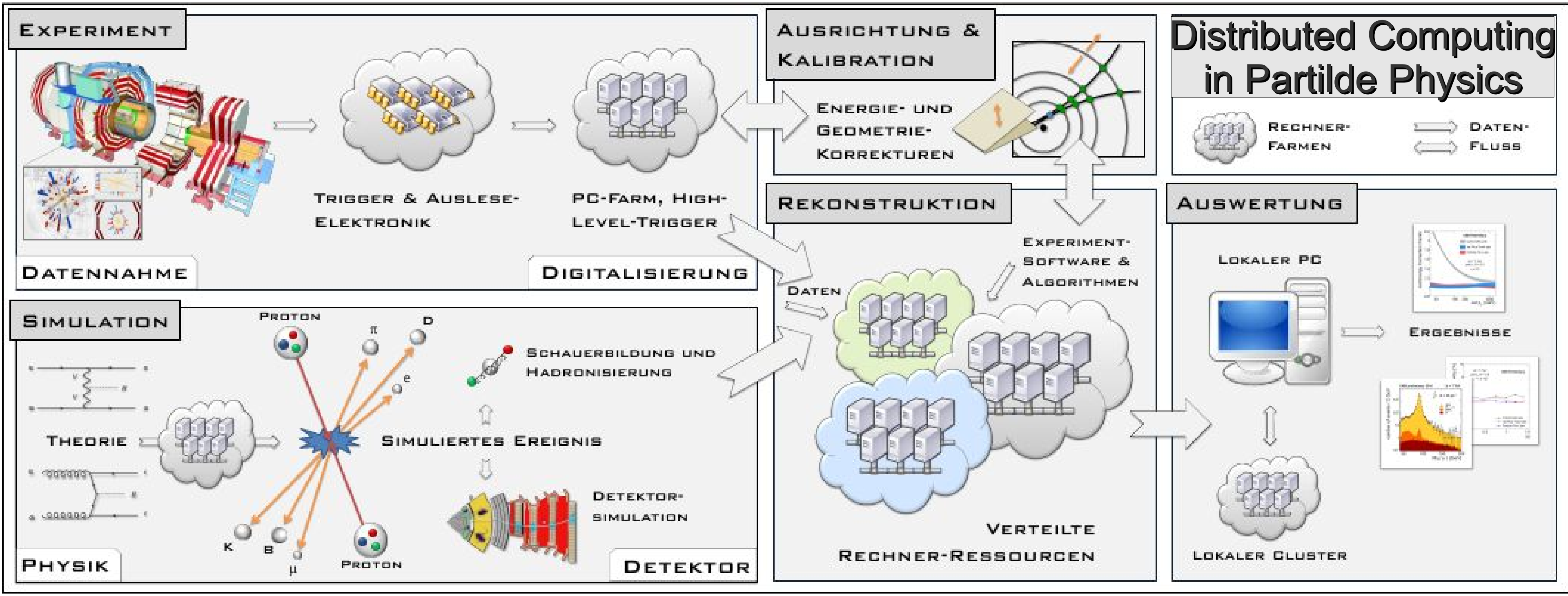
→ **~1000 Tbyte/s raw data**

- zero-suppression and trigger reduce this to

„only" ~1 Gbyte/s

i.e. ~1  */sec*

Heterogeneous Spectrum of Applications



Broad palette of applications
with different requests for CPU, Storage and I/O:

– centrally organised simulation and reconstruction of large data sets
  mainly need (much) compute power
– data selection and distribution
  need high I/O- and network bandwidth

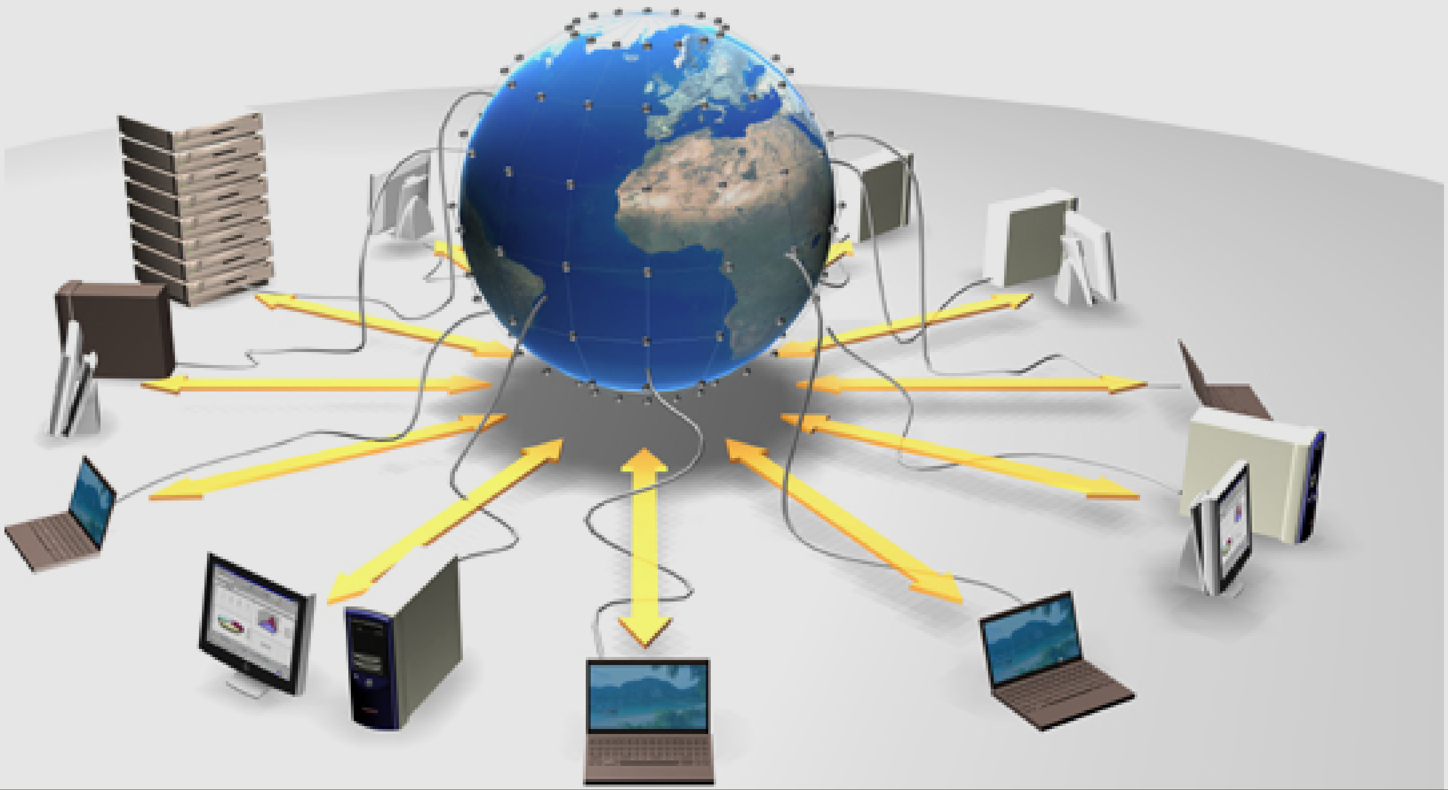
– physics analyis by many individuals with random access to data and resources with (often poorly desinged private) code → another challenge

1998: Idea of „Grid Computing“

came at the right time for the LHC:

“A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities.”

C. Kesselman, I. Foster, 1998

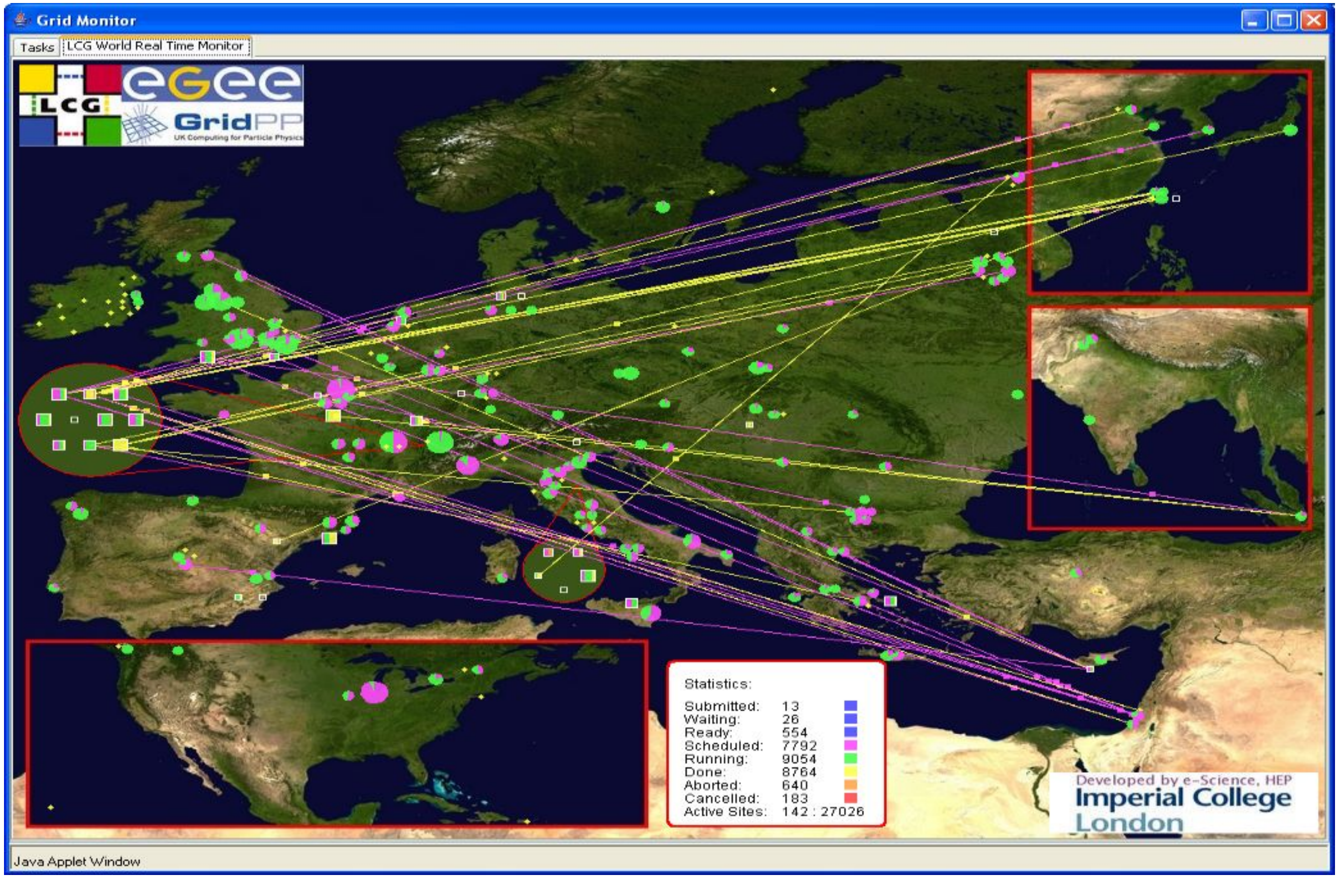


- *Coordinates resources that are not subject to central control ...*
- *… using standard, open, general-purpose protocols and interfaces …*
- *… to deliver non-trivial quality of services*

I. Foster, 2002

Data- and Information flow in the WLCG

The Worldwide LHC Computing Grid 2016



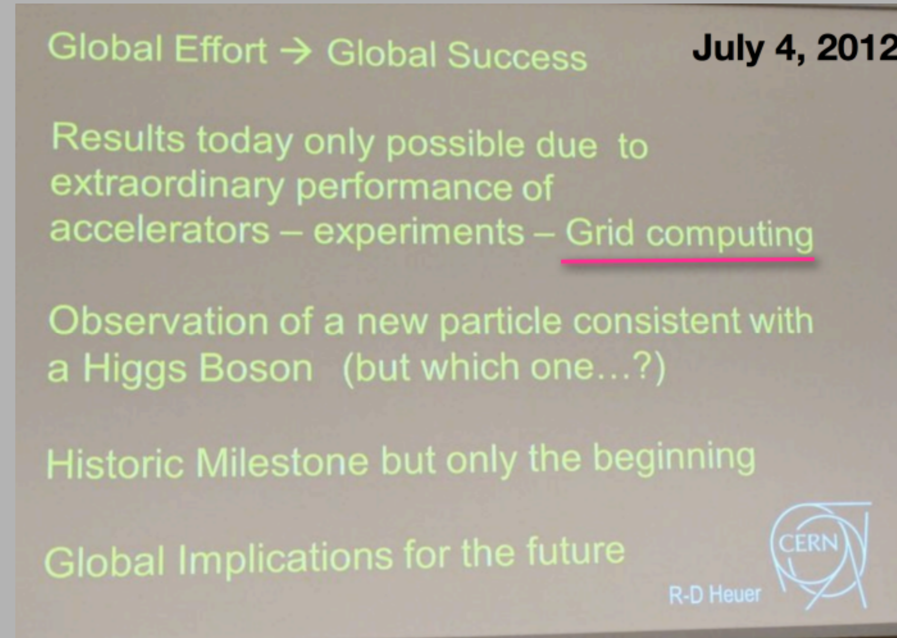
167 centres
in 42 countries:
11Tier-1
156 Tier-2

Ein Supercomputer mit
400'000 Prozessorkernen
310'000 TB disk storage
390'000 TB tape storage
~2 Million Jobs/day
~500 TB/day network transfers

Tier1 in D
GridKa at KIT

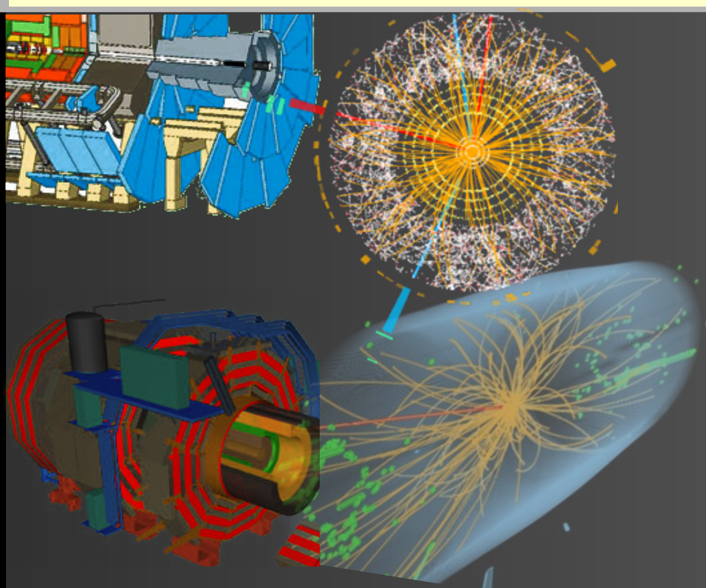
T2 in D
- DESY
- MPI München
- RWTH Aachen
- Uni Freiburg
- Uni Göttingen
- LMU München
- Uni Wuppertal

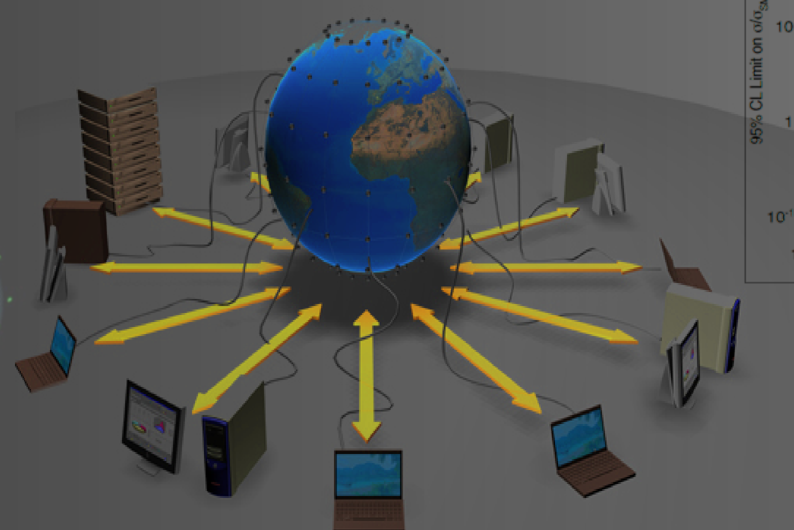
- German fraction WLCG
  ~15% Tier-1
  ~10% Tier-2

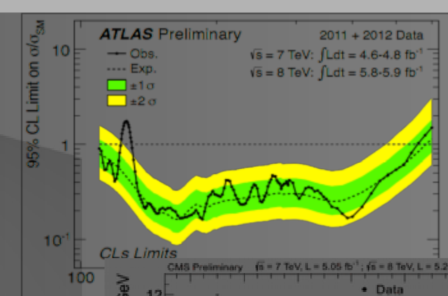


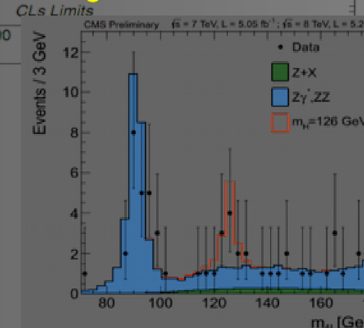
success of the LHC Phyics Program (also) thanks to the extraordinary perfomance of "Grid Computing"

Did the LHC Computing work ?

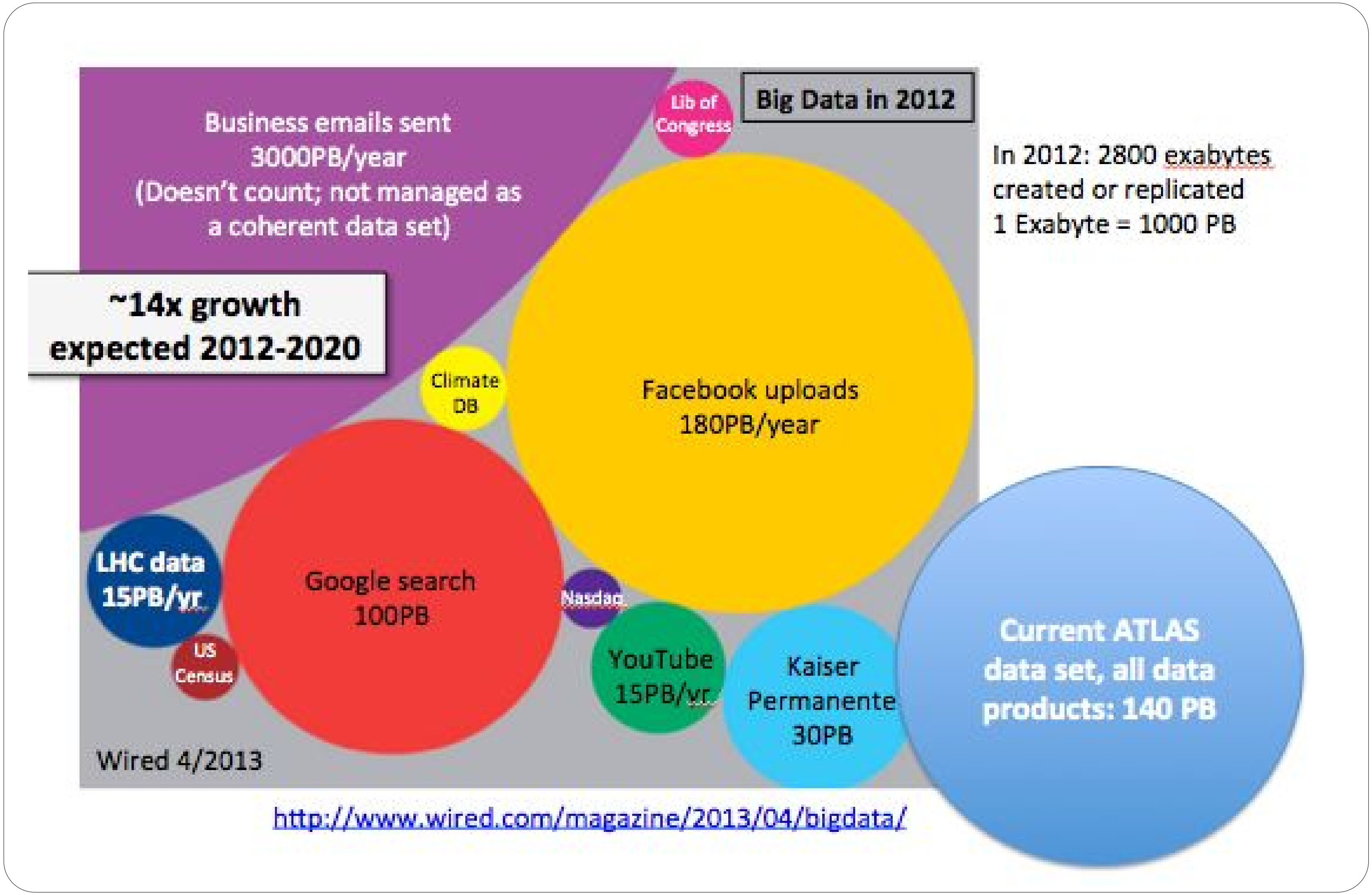








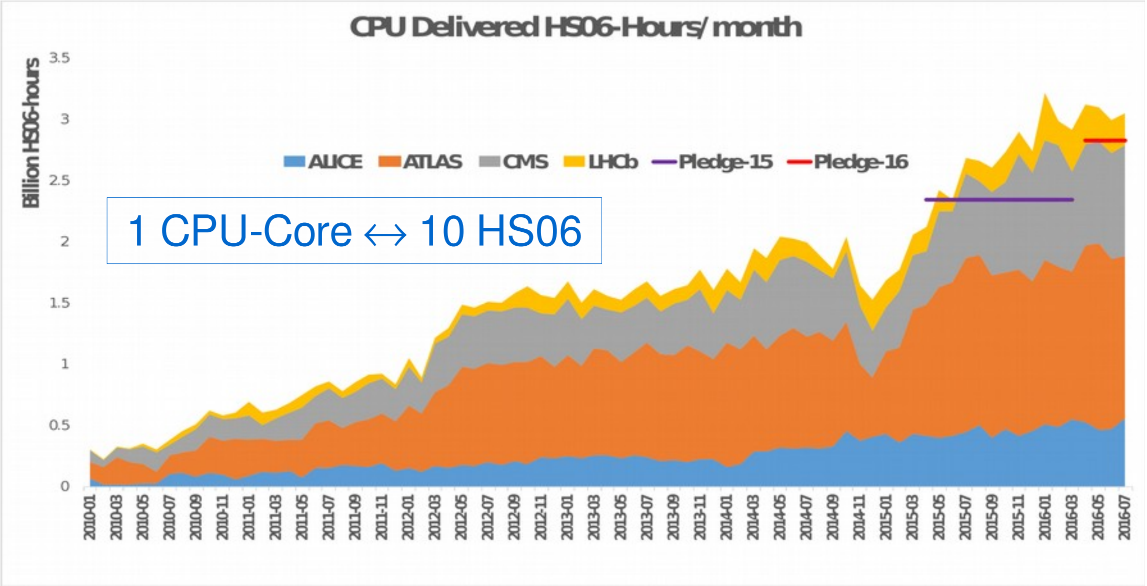
Ist HEP „Big Data“ ?



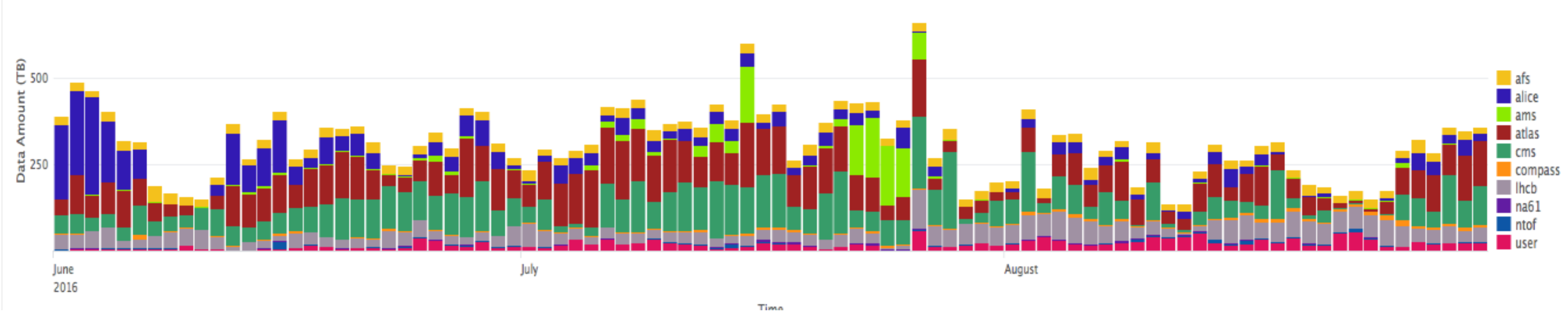
Yes !

New Records 2016

- 300 000 000 CPU hours / month delivered to experiments

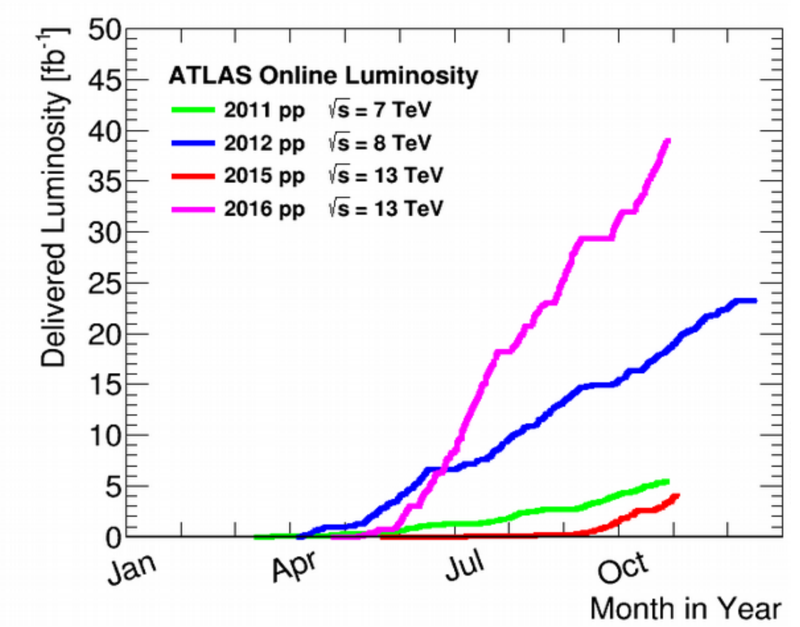


- 500 TB data transferred routinely per day (2x more than peak performance Run 1)



The year 2016: Run 2 computing needs revised

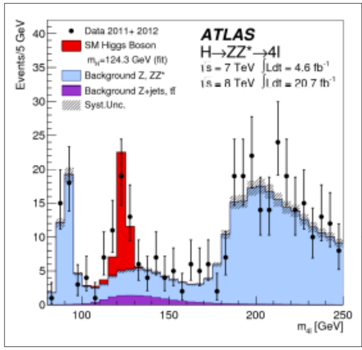
- LHC performance was above expectations,
  need for additional compute resources driven (mainly) by:
  - LHC live time (37% → > 60%)
  - Luminosity (1.0×10^{34} → 1.2×10^{34} or better)
  - Pile-up (CMS, ATLAS) (21 → 33 on average)



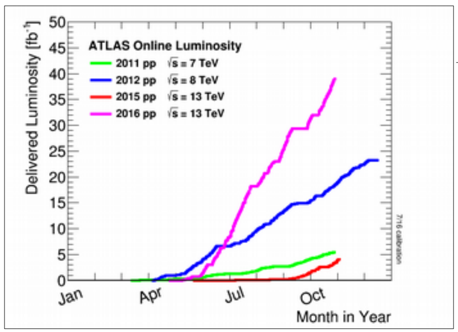
- For 2016, the available resources were barely sufficient

  But:
  future hopes in continued superb LHC performance
  already led to an increase of requirement estimates by ~25%

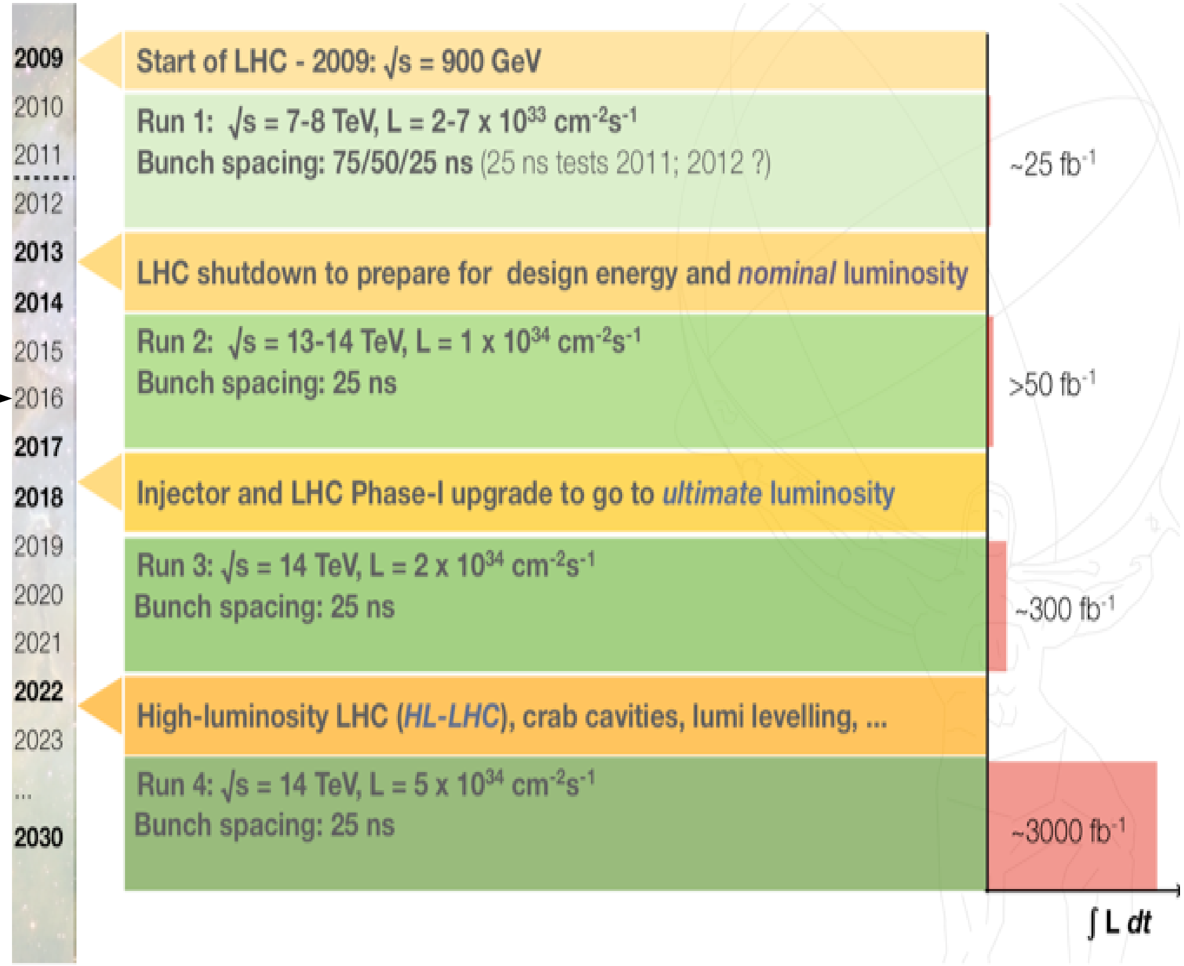
The long-term perspective: Future Plans for the LHC



Higgs Boson



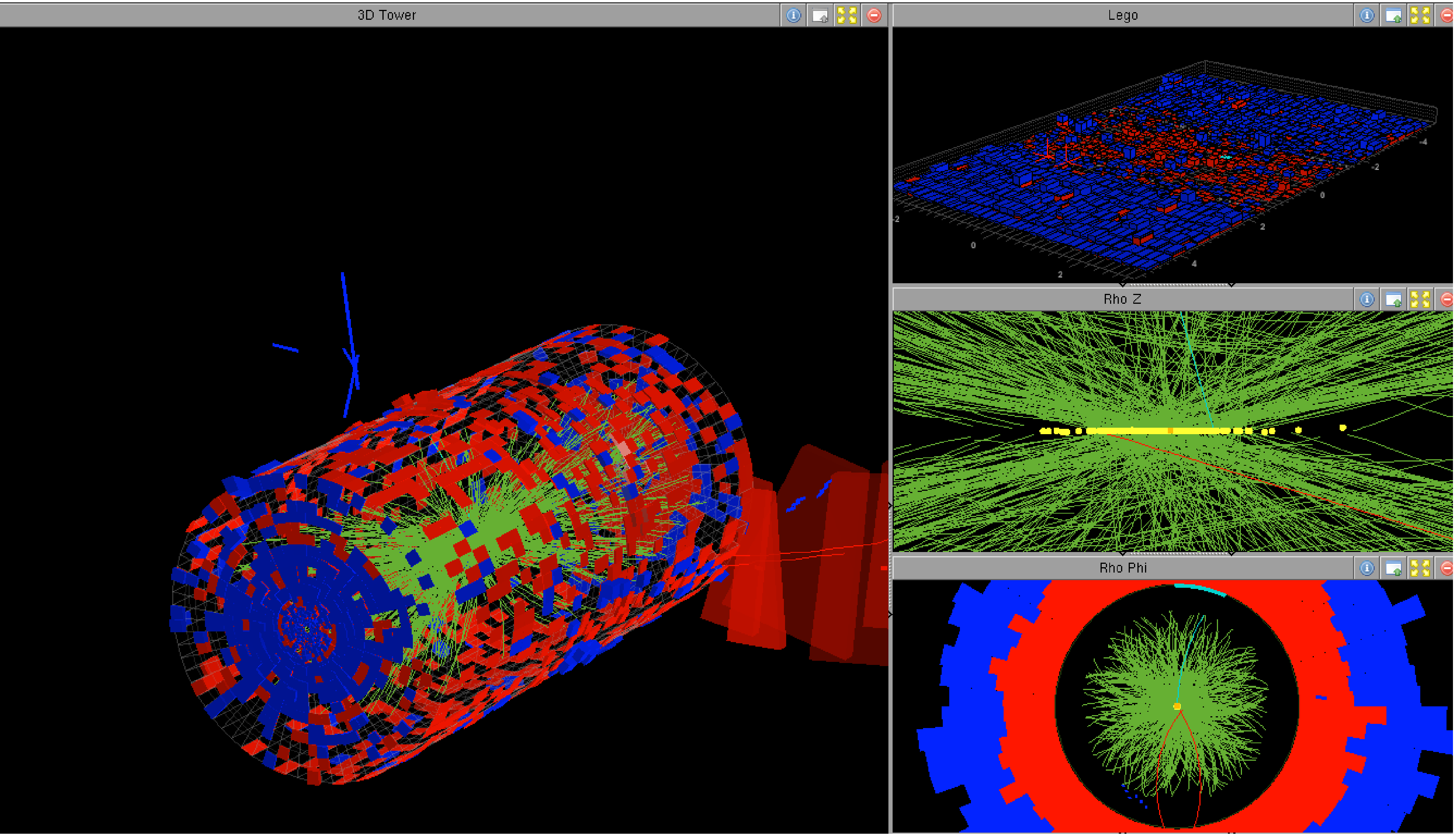
today



so far, only ~2% of total expected ∫ L recorded

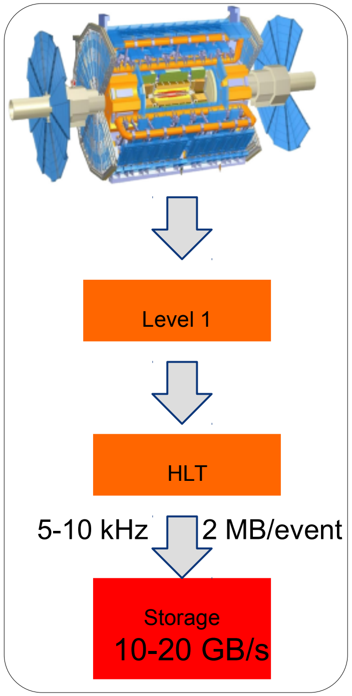
- still large physics discovery potential
- but there are enormous challenges ahead !!!

High-Luminosity LHC → high hit density in detectors

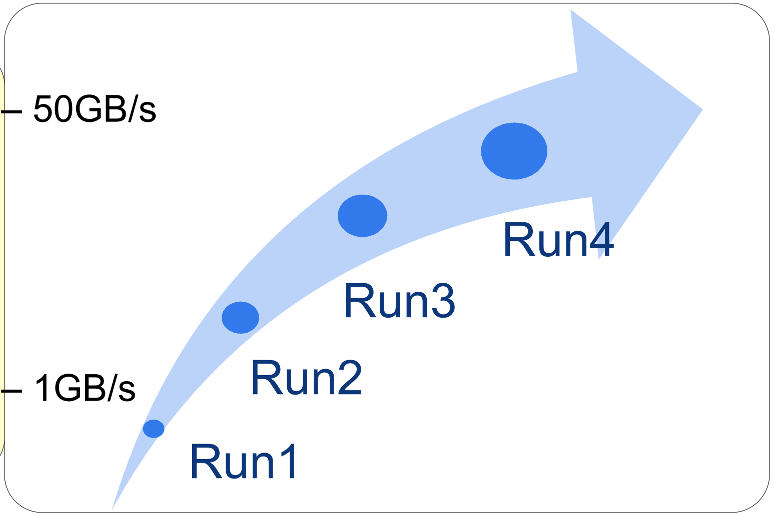


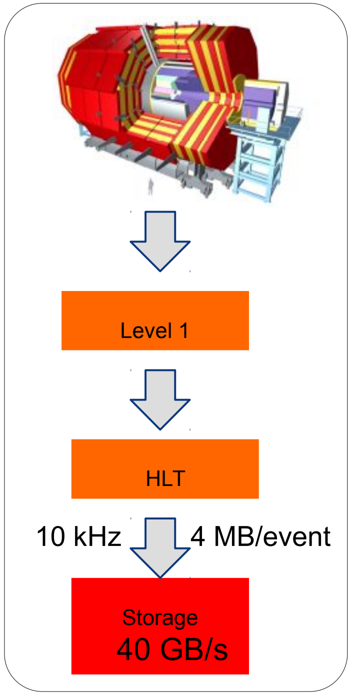
78 pp-collisions in one bunch crossing

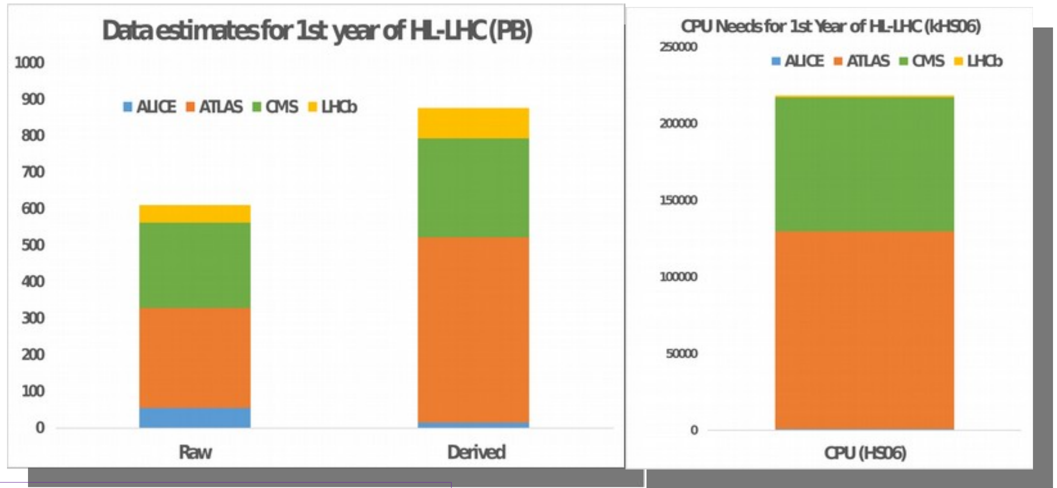
The long-term perspective: effects on SW & Computing



ATLAS
& CMS
expect
~10x higher
data rates
in run 4







Storage: 2016 2027
- raw 50 PB → 600 PB
- derived 80 PB → 900 PB

CPU:
- x60 from 2016

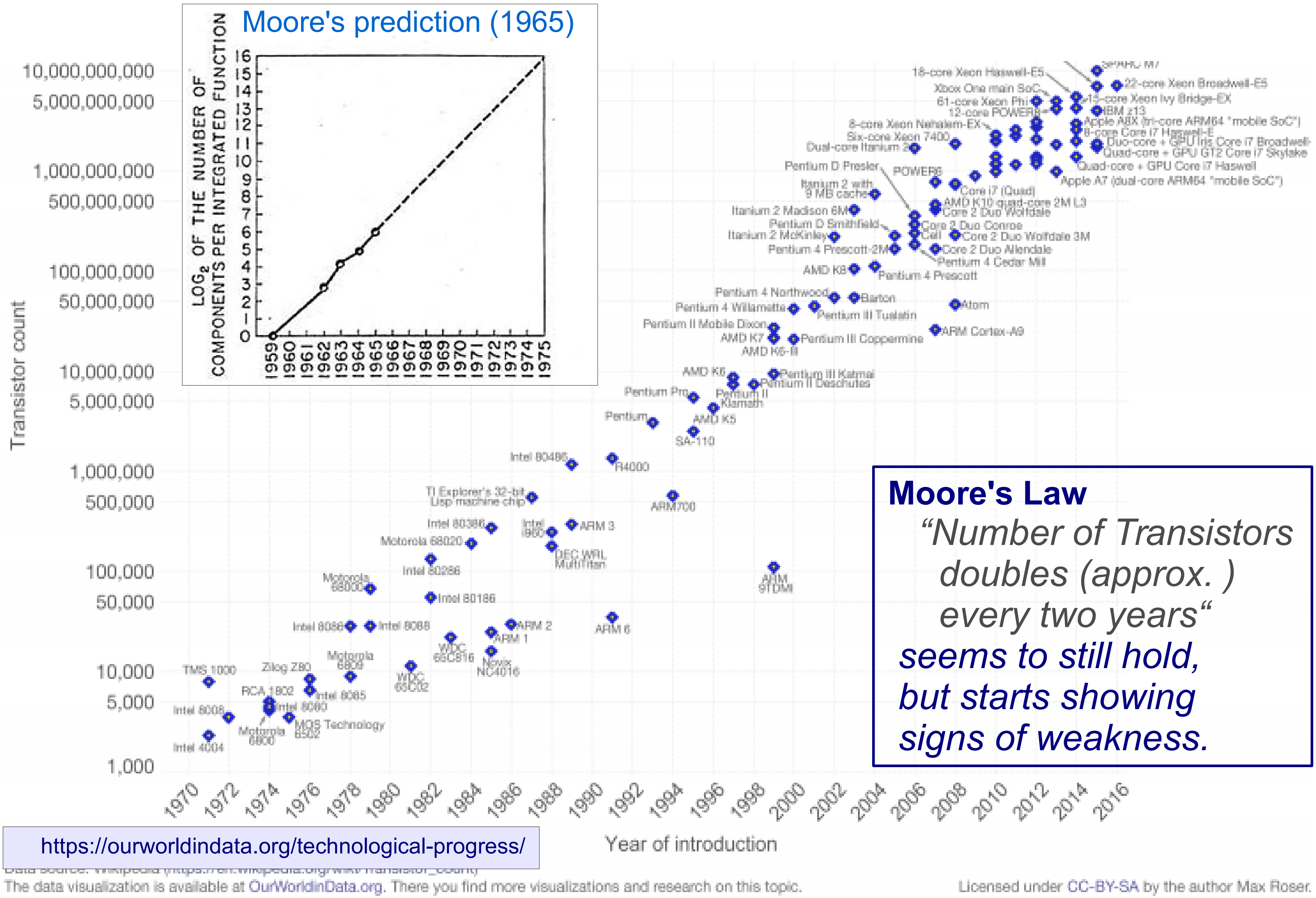
Source: Ian Bird
(CHEP '16)

- needs driven by new detectors, higher data rates, more complex events
  hardware technology expected to bing only factor 6-10 in 10-11 years

Part II

The Challenge(s)

The end of Moore's Law ?



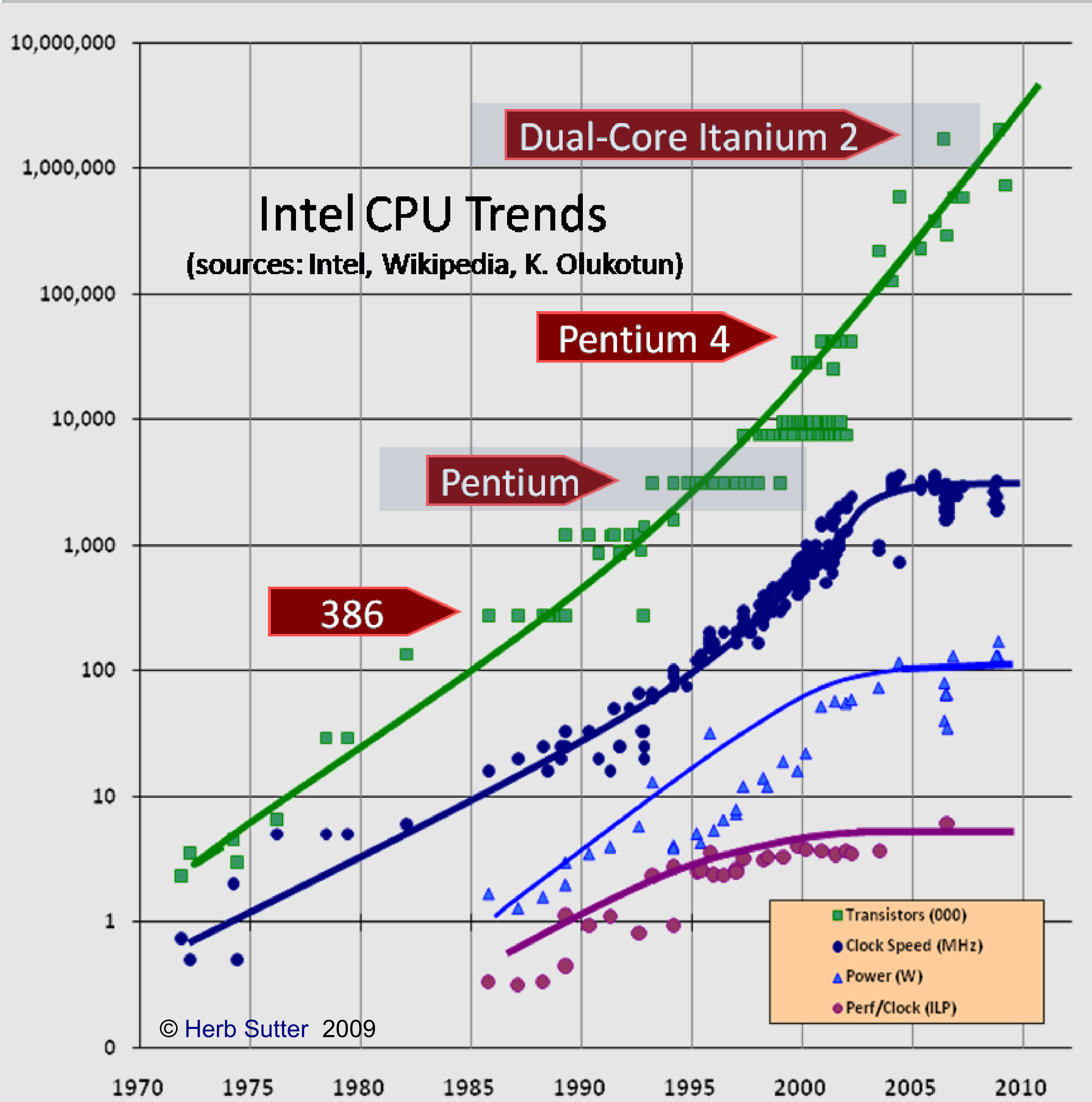
Moore's Law

“Number of Transistors doubles (approx.) every two years“

seems to still hold, but starts showing signs of weakness.

https://ourworldindata.org/technological-progress/

facing a new Situation since ~2005



past:
smaller structures led to higher clock rates and hence software performance

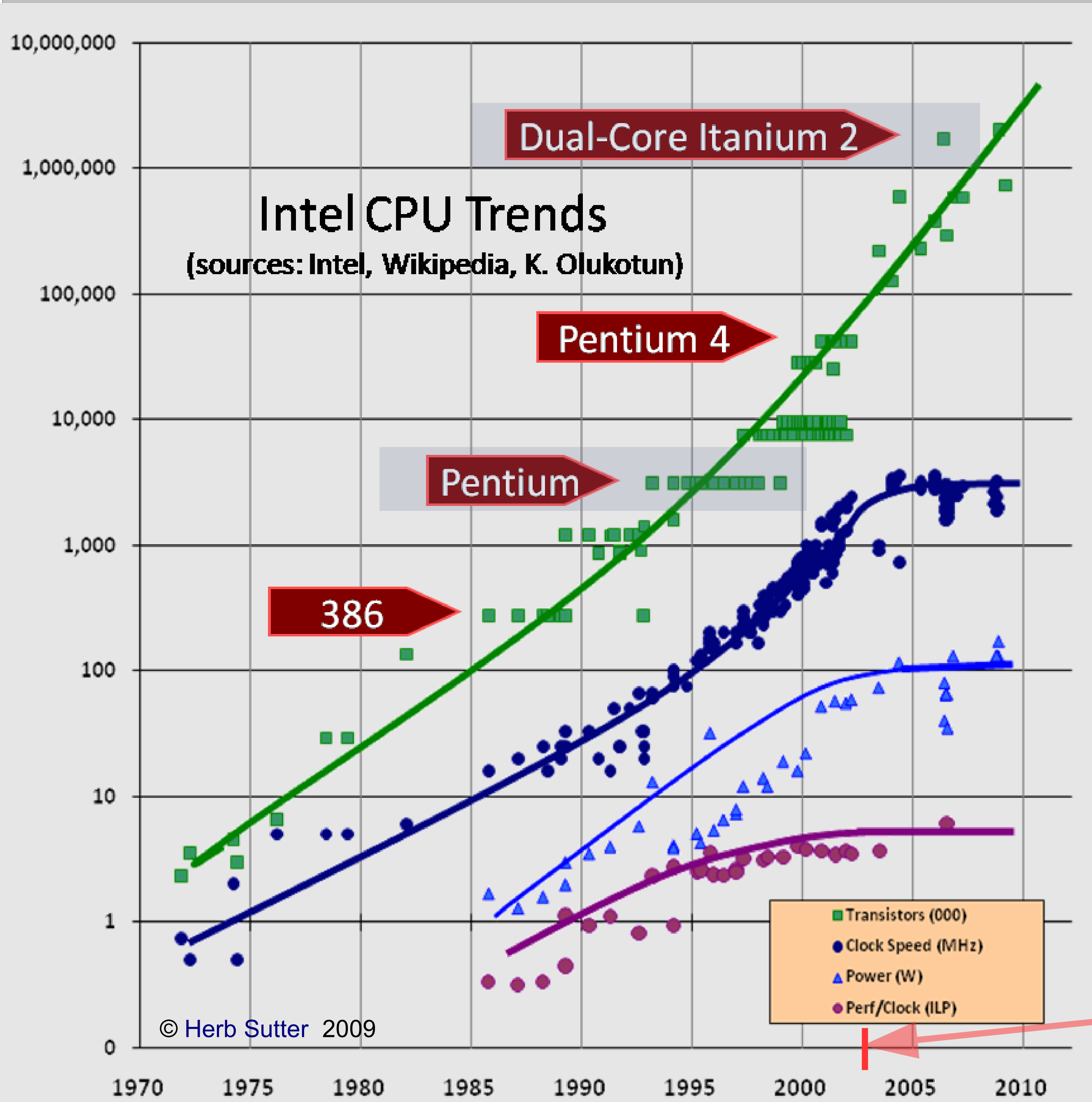
today:
limits of energy and thermal budgets reached

→ increase in complexty of CPU
- parallelism
- grapics cores

→ multi-core architectures

big challenge for software development

facing a new Situation since ~2005



past:
smaller structures led to higher clock rates and hence software performance

today:
limits of energy and thermal budgets reached

→ increase in complexty of CPU
- parallelism
- grapics cores

→ multi-core architectures

LHC code designed here

Recent Developments

more and more transistors / Chip, but no large increase in clock rates since 2005 – *„free Lunch is over“*

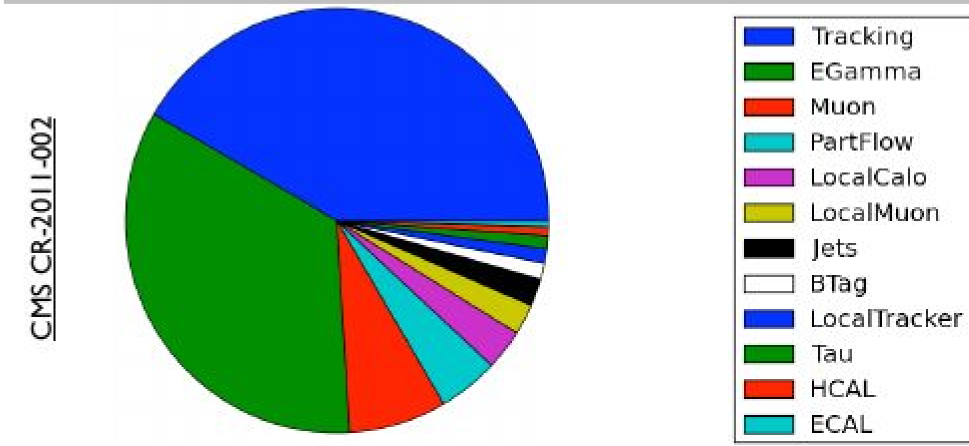
→ more complex processor architectures with more and larger registers

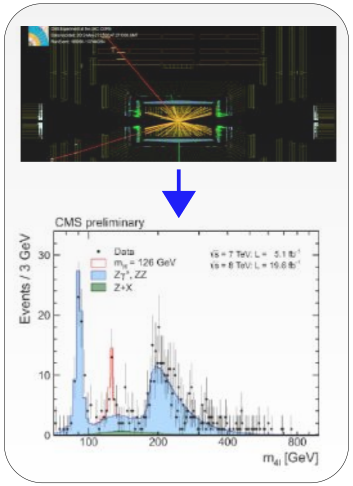
- larger, multi-stage cache storage
- vectorisization (**SIMD**=„**S**ingle **I**nstruction **M**ultiple **D**ata“)
  e.g. MMX, SSE, AVX
- multi-core architectures with 2/4/6/8/12 … CPU cores/chip
  „Hypterthreading“ on Intel Architecture
- pipelining (i.e. parallel execution) of instructions
- improved branch prediction
- integrated graphics processors (GPU)

Increasing parallelism and heterogeneity of architectures:
→ Challenge for the development of efficient program packages

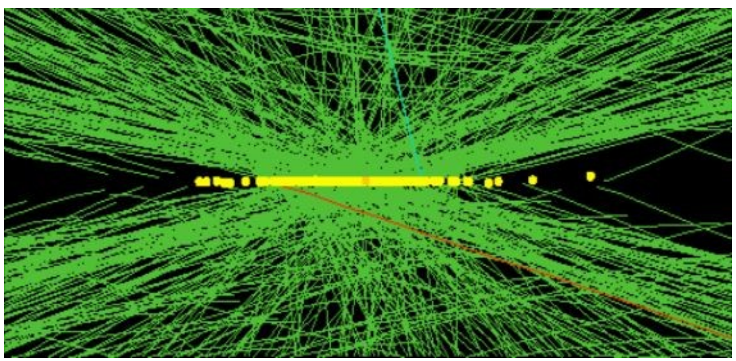
example: Event Reconstruction in CMS

fraction of CPU time for reconstruction steps
in top-pair events (CMS 2011)



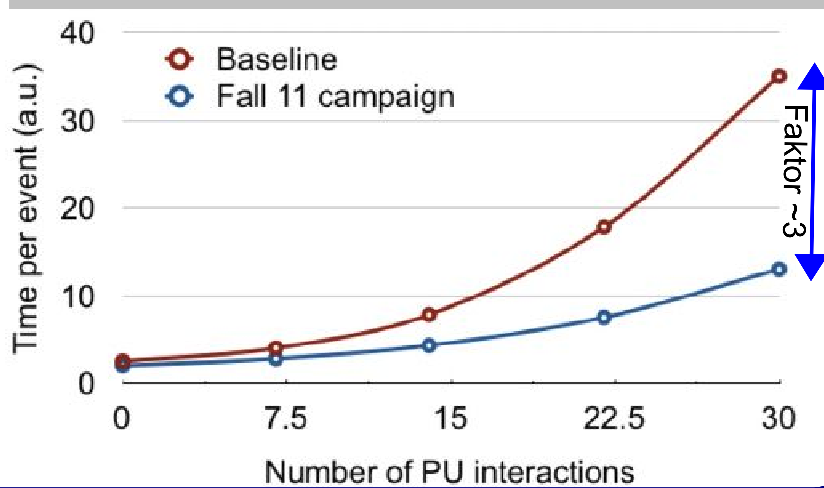


many concurrent pp collisions („Pile-up“)



drives up reconstruction time

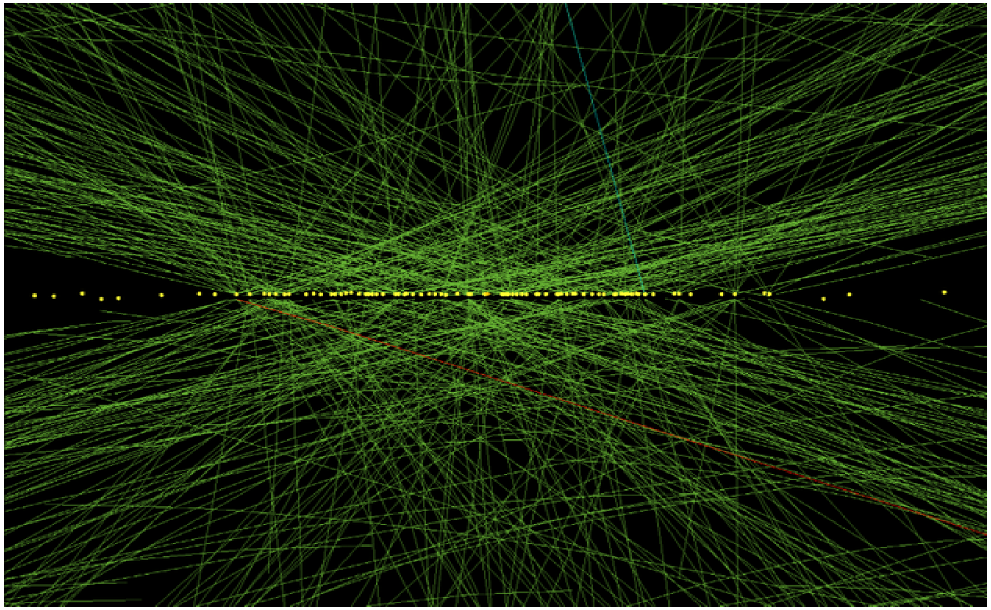
CPU time for QCD events **(CMS 2011 & 2012)**



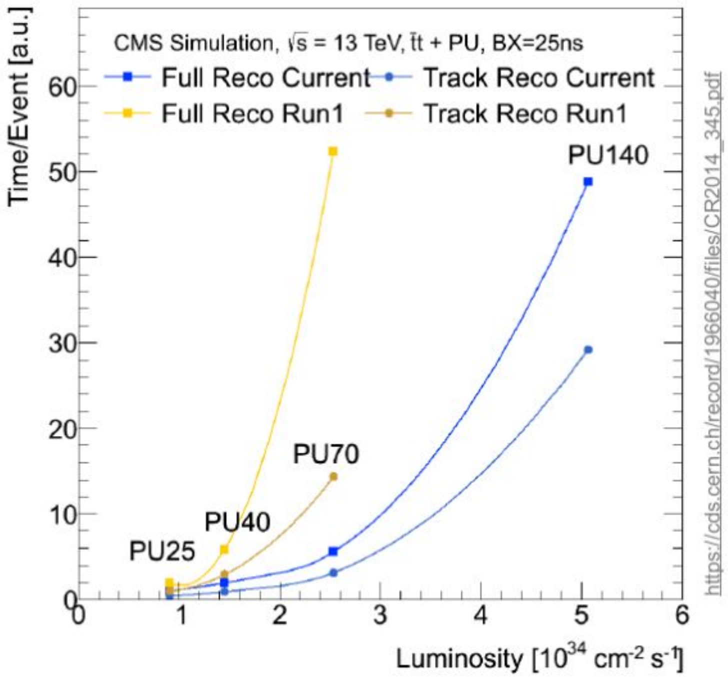
~3x faster reconstruction by optimisation and & new techniques

D. Piparo: *„Higgs in CMS was discovered with C++11“*

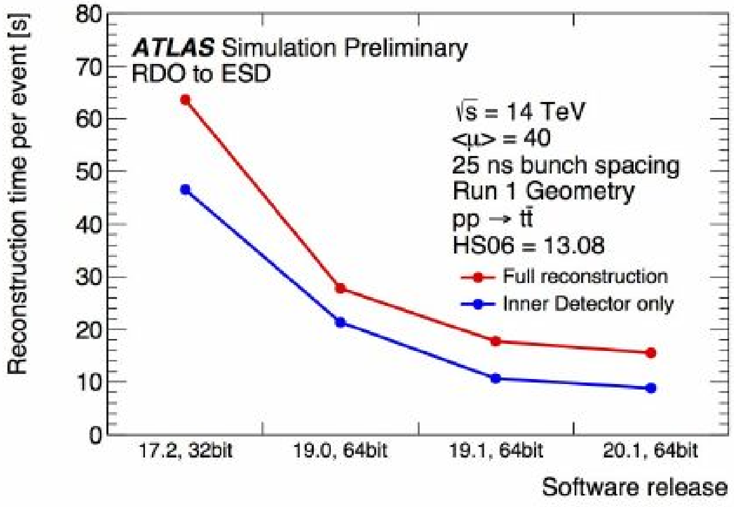
Run 2 achievements



vertex region in an event with 78 pp collisions in one bunch crossing



Pile-up drastically affects CPU time needed for reconstruction & simulation

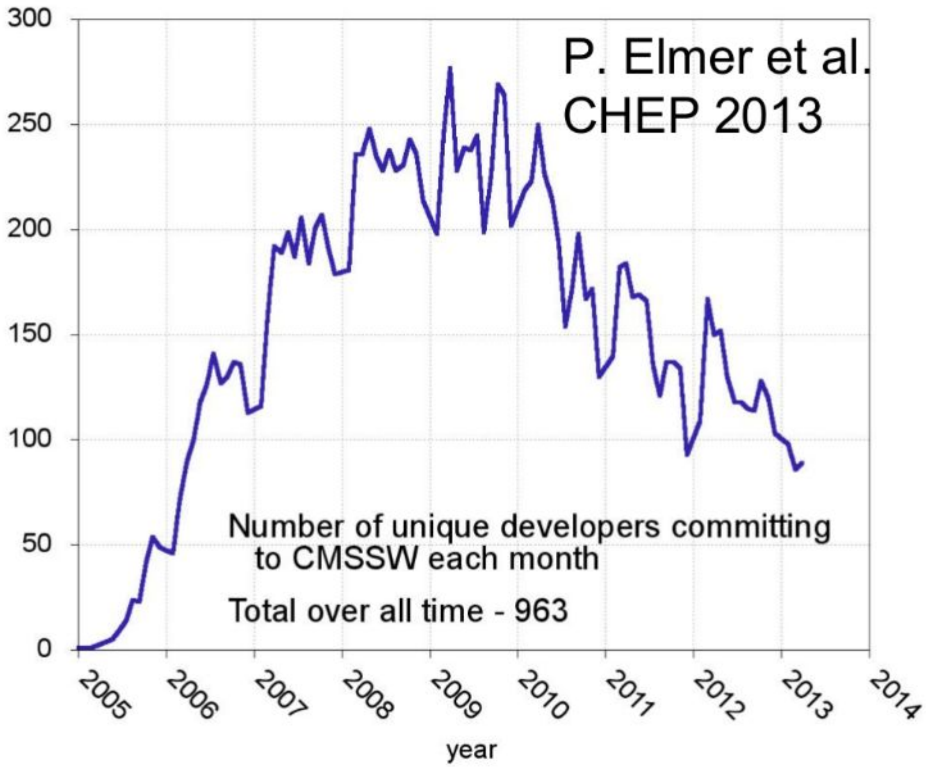


A lot has been achieved already, but still a long way to go to handle 10x more data

Sociological Challenge

Number of Software Developers depends on phases of an experiment

- large number of people during commissioning phase
- more and continued focus on physics analysis reduces the number of developers



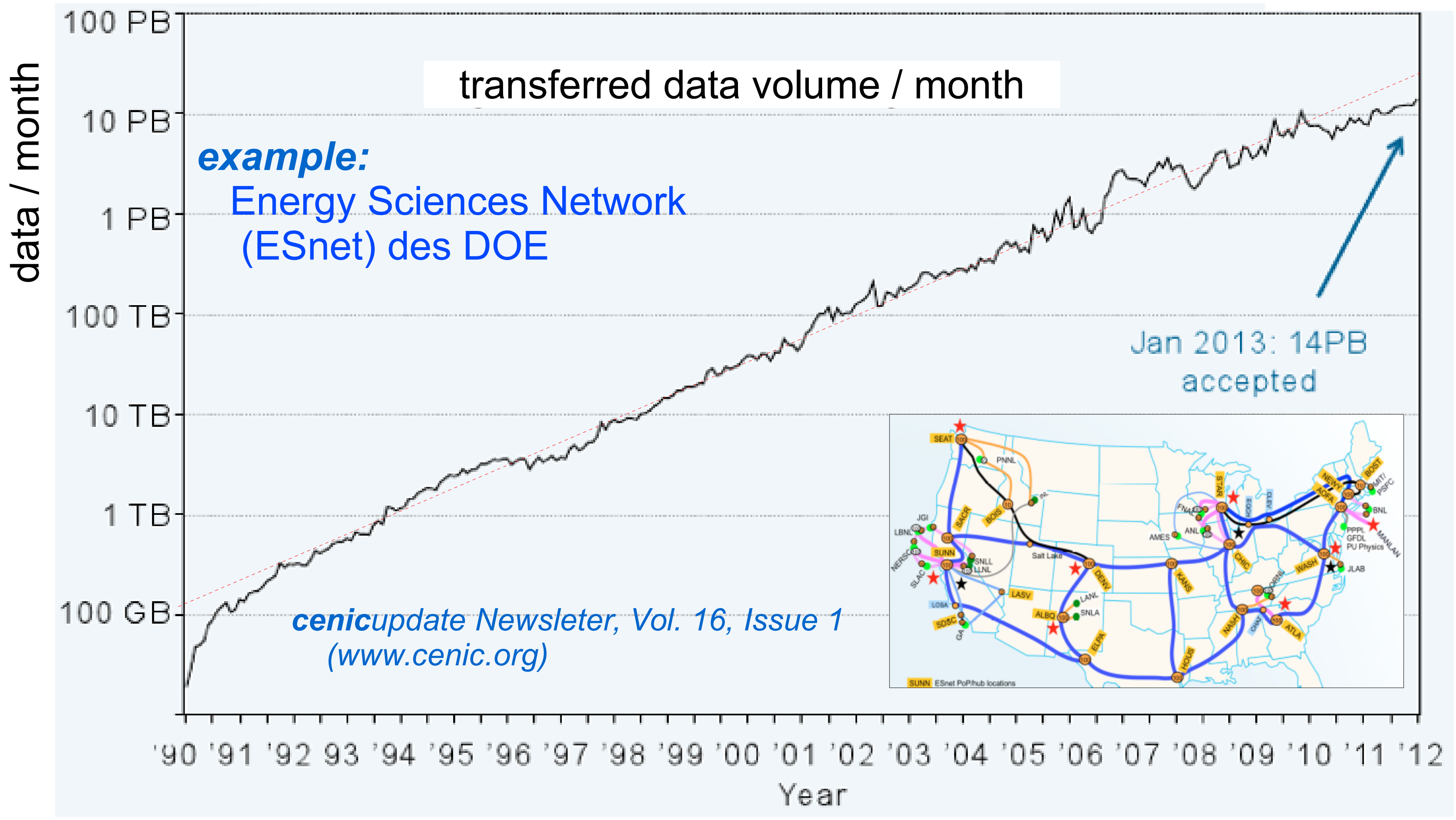
Must make work on SW & Computing
- more attractive and
- scientifically rewarding
  especially for young people !

How to ramp up again to have enough expertise for the SW&computing challenge ?

HSF
HEP Software Foundation

The HEP Software Foundation facilitates coordination and common efforts in high energy physics (HEP) software and computing internationally.

the good news: Network Bandwidth still increasing



- still unbroken exponetial growth, factor ~1,8 / a in last 20 years

Part III

Accepting the Challenge !

Network (2)

since long: **LHC Optical Private Network** for T0↔T1, T1↔ T1

relatively new: **LHCOne** for transfers T1 ↔ T2 & T2 ↔T2

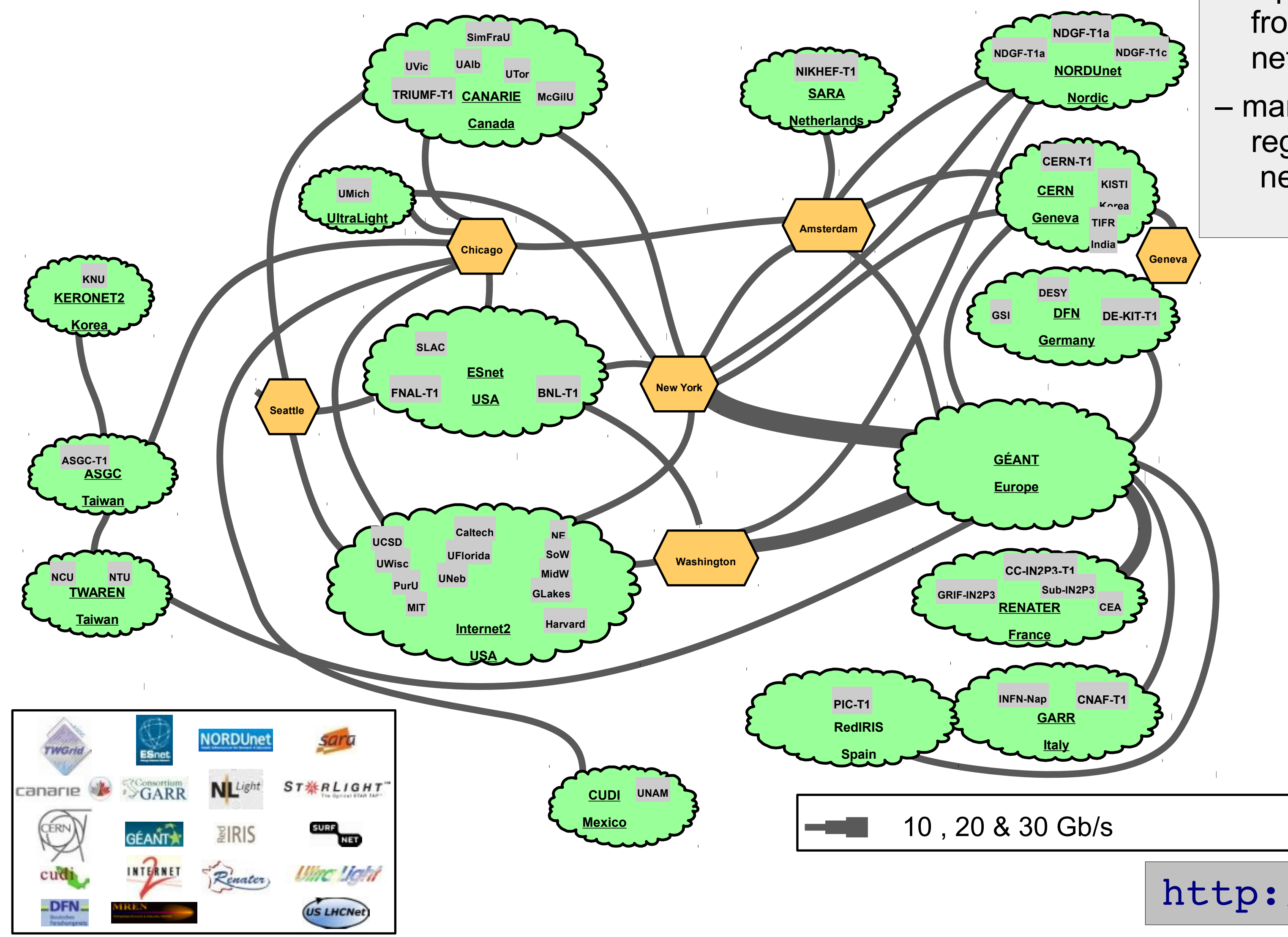
relevant for physics analysis

– **Private Network**

→ only trusted partners, no firewalls needed

– separates LHC traffic fromgeneral scientific networks

– management by regional / nationale network providers

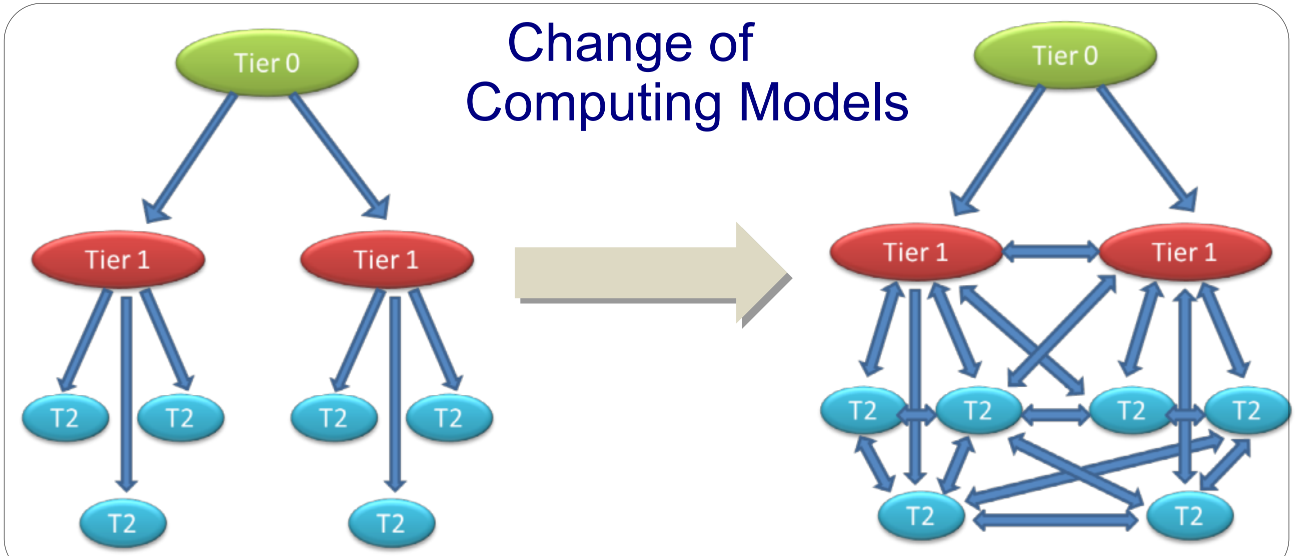


http://lhcone.net

Development of Computing Models

increasing network bandwidth

&



enable

Datea processing via network
„transferring is cheaper than storing"

„Data federation": **T1, T2 (& T3) as unified, distributed storage system**

based on (HEP-spezific) XRootD

ATLAS: „FAx" (**F**ederating **A**TLAS Storage Systems using **x**rootd)
CMS: „AAA" (**A**ny data **A**nytime **A**nywhere)

btw: concept since long exercised by ALICE !

100 GB/s WAN bandwith

2x100 GB/s Links
- CERN ↔ Wigner Data Centre (Budapest)
- as extension of CERN T0 in production

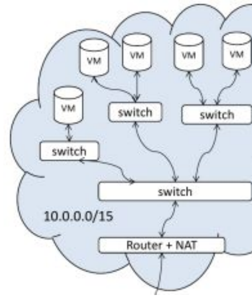
expect more 100 GB/s networks, also in Germany:
- connection between HGF Centres
- some "Ländernetze" (e.g. BelWü end 2017)

Resources external to WLCG - Heterogeneous Hardware

- Cloud Resources
  - private (e.g. institute clusters)
  - commercial (Amazon, Google, Telekom …)

left R&D Phase since long

→ „Grid of Clouds“ is a reality



z.B. Open Stack

Beispiel: Nutzung der ATLAS & CMS HLT-Farmen während des LS 1

- HPC – Cluster
  - many HEP applications run
  - MPI interfaces and SAN not needed for HEP

→ **an expensive way!**

e.g. Super-MUC at LRZ in Munich, SDCC at CMS or the new bwHPC Cluster in Freiburg (ATLAS/CMS/LHCb)

- no Grid services / authentication
- fast, but small and expensive disk
- often small WAN bandwidth
- different „Site Policies“

special solutions for every single case → personal !

- other processor architectures (ARM)

  Goal: **Optimisation of CPU cycles / Watt** for special applications

  Tests ongoing, many HEP-applications run on ARM

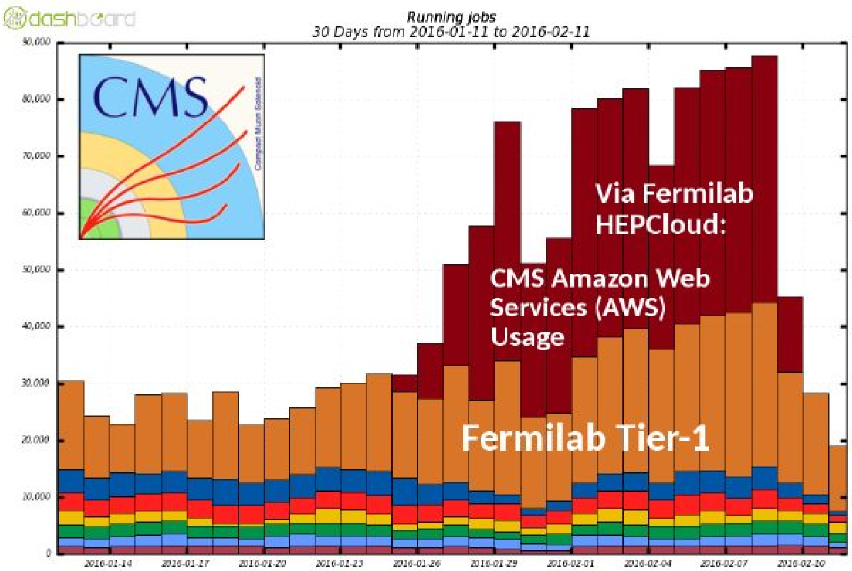
  Is this (part) of the future ?

test-cluster with 16 ARM processors and SATA Disks at KIT (SCC)

since 2012 more mobile than x86 processors sold

Massive usage of Cloud Resources

- Amazon Cloud (Amazon Web Services) provided to CMS via FNAL Tier1
  – more than doubled available CPU at Tier Ones
- – still sponsored by provider, but cost on spot market approaching reasonable levels: (FNAL: 0.9 Cent / CPU hour, Amazon: 1.4 Cent / CPU hour)

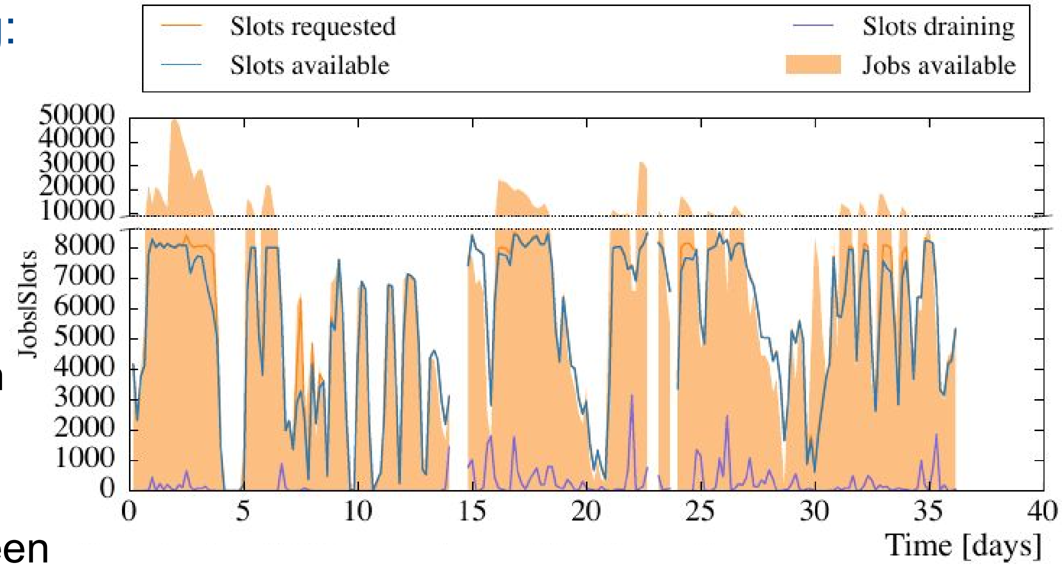


- bwFOR cluster NEMO in Freiburg:

  for Neuroscience, Elementary Particle Physics and Microsystems engineering

– fully virtualized set-up; controlled by ROCED (KIT) and HTCondor

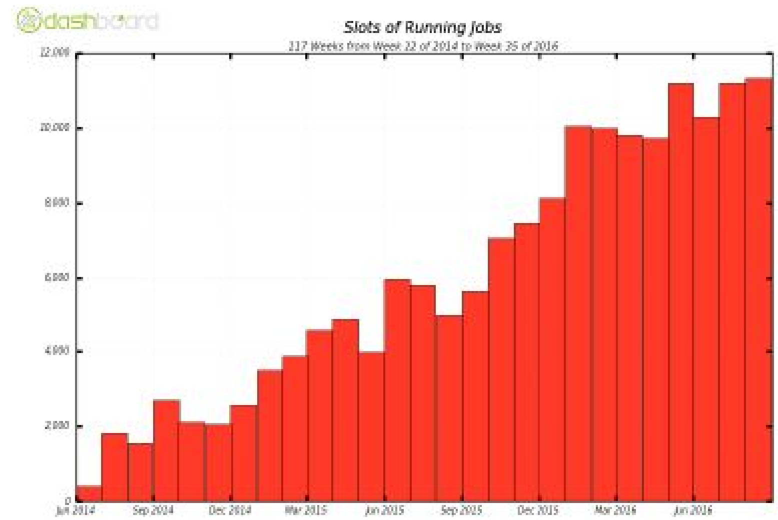
– Production system scaled up to 11k virtualized cores, more than 7 million CPU hours of user jobs processed in four months

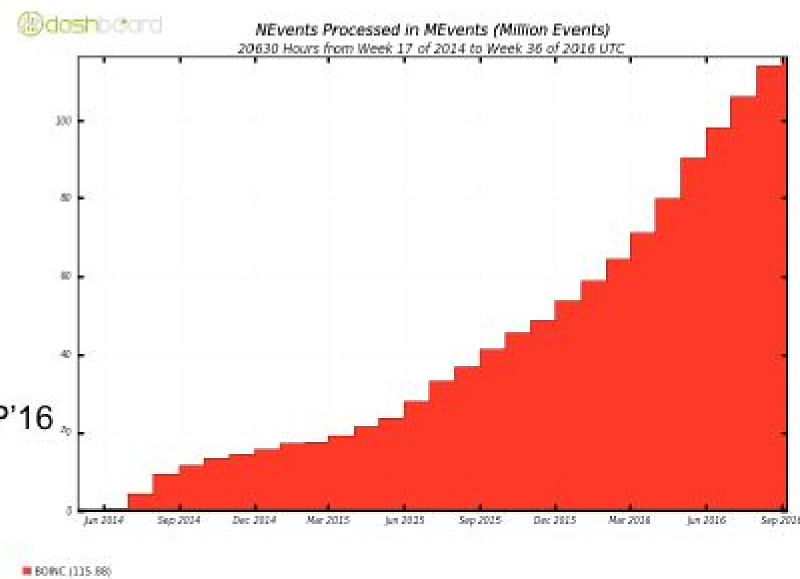


– saturating 10 GBit/s BelWü link between Karlsruhe - Freiburg and NRG Grid storage at GridKa

Volonteer Computing

- People volunteering their PC's spare CPU cycles for science
- most commonly used software is BOINC
- ATLAS MC simulation jobs inside a CernVM
- Jobs are taken from ATLAS job management system and submitted to BOINC server through ARC CE
- Steady growth of volunteers, 11-12k „running" job slots Providing 1-2 % of overall ATLAS CPU





D. Cameron, CHEP'16

Cloud-enabling Technologies in HEP

CernVM:

- Virtual machine based on Scientific Linux (maintained by CERN)
- Very lightweight, can be directly deployed on various cloud sites

CernVM-FS:

- On-demand HTTP based file system (Caching via HTTP Proxy)
- Many big experiments use it to deploy software to WLCG compute centres
- works excellently also on cloud sites

HTCondor:

- Free and open-source batch system commonly used in HEP
- Excellent with integrating dynamic worker nodes (even behind NATed networks)

xRootD and **data federations** for remote access

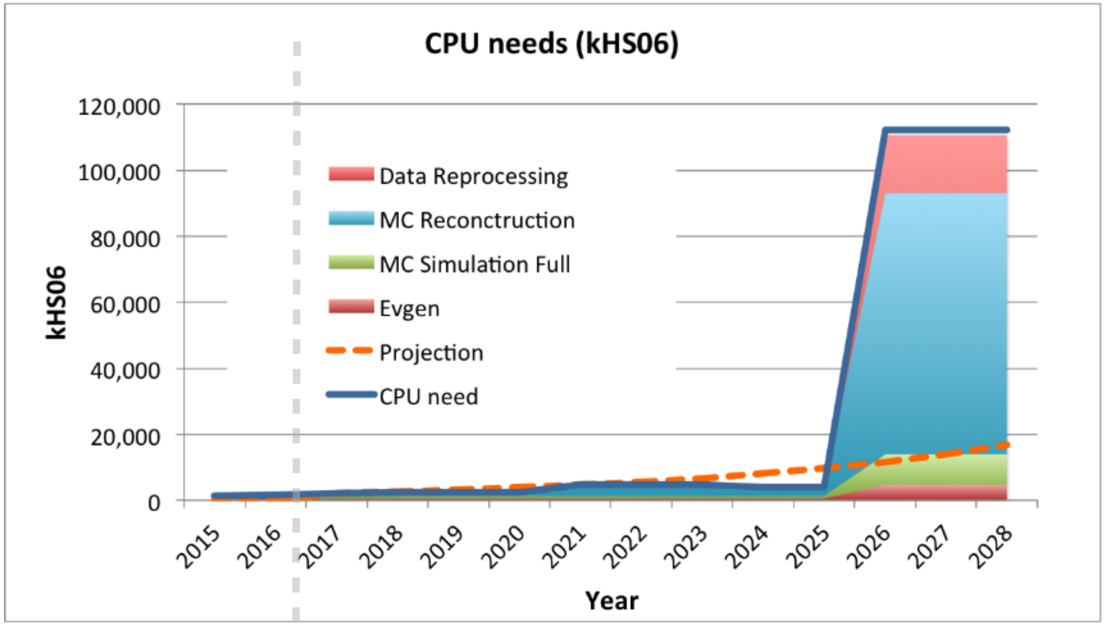
"Cloud manager", *e.g.* ROCED [KIT]:

- Cloud scheduler that supports multiple cloud APIs (OpenStack, Amazon EC2 and other commercial providers)
- Easily extendable thanks to modular design
- Parses HTCondor ClassAds and boots VMs on cloud sites depending on the number of queued jobs

Extrapolated future resource needs

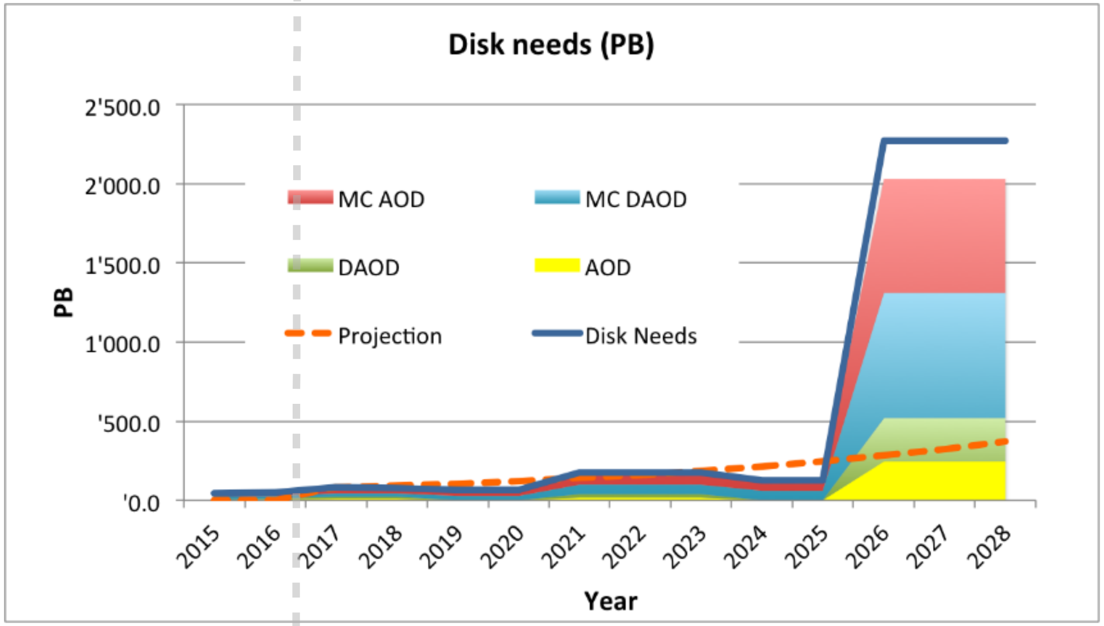
Large increase in resource needs for both CPU and storage driven by Monte Carlo (MC) simulation

attention:
"simple" extrapolation from 2016 ATLAS computing model
(S.Campane, CHEP 16)

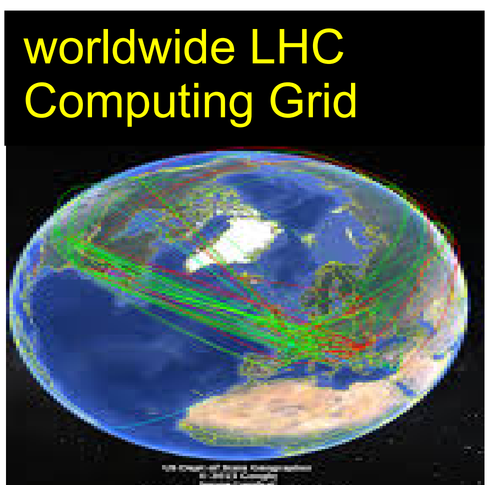


Need to follow multiple roads to face the challenge:

- code optimisation
- efficient use of new architectures
- additional resources, shared with other science communities



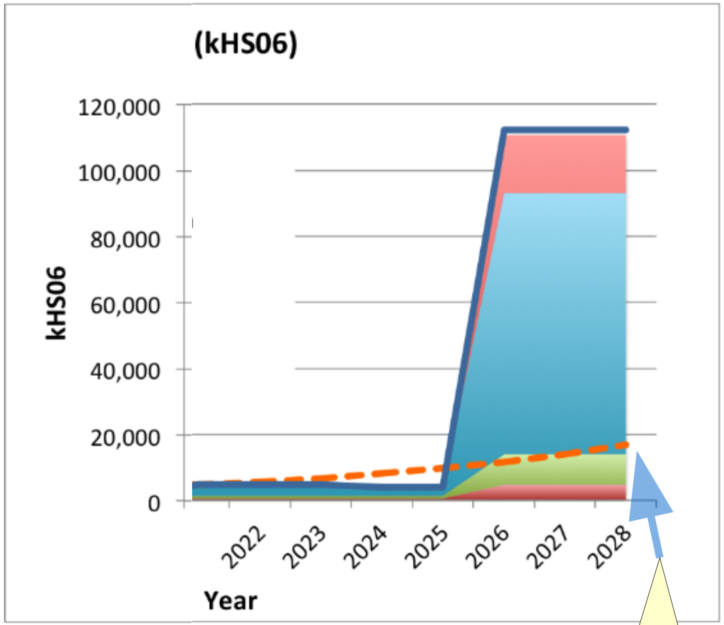
From Mainframe to the Cloud



Hans has seen and helped shaping these developments

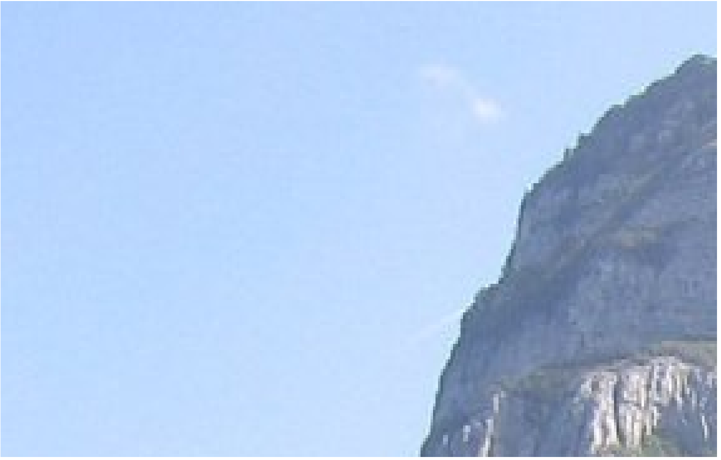
Helix Nebula Sciene Cloud

Mountains and surmountable problems



Factor ~9 missing
if we simply "sit and wait"

expected by technological progress



**But of course,
the challenge is accepted
and will hopefully be met !**

Need experienced and brave mountaineers

like you, Hans !

Dear Hans,

I wish you all the best,
enjoy your “Retirement” and
the new freedom it brings.

I'm sure it will be a very active time!