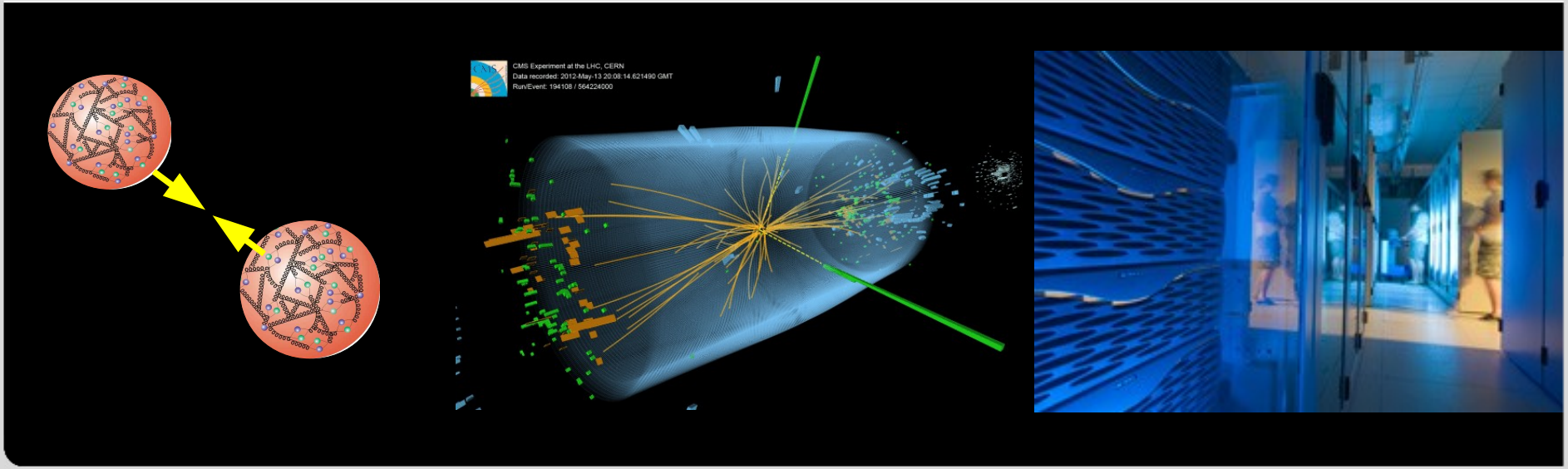
Karlsruhe Institute of Technology

From raw data to Pbytes on disk

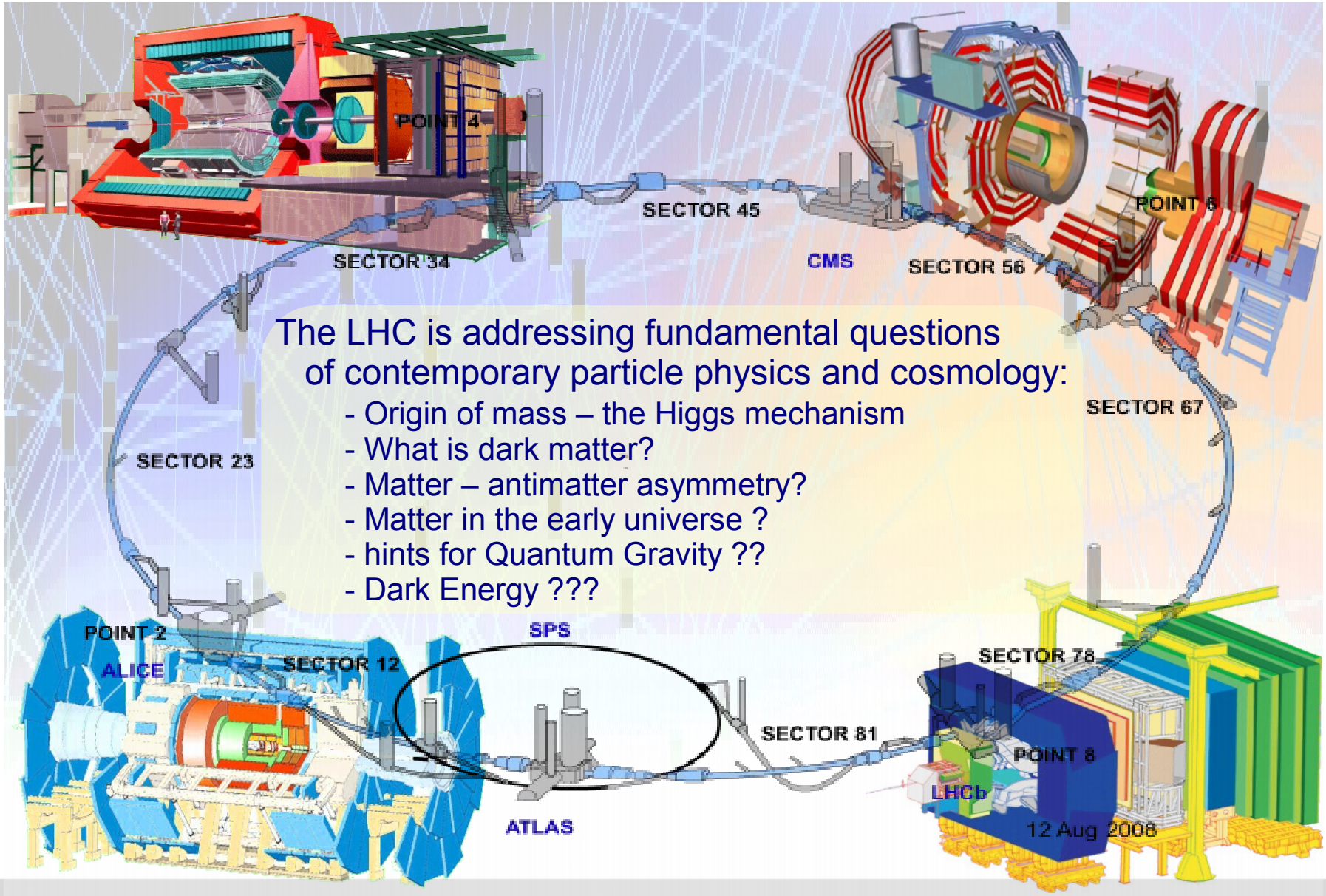
The world wide LHC Computing Grid

Günter Quast **HAP Workshop Bad Liebenzell, Dark Universe** Nov. 22nd 2012

Institut für Experimentelle Kernphysik



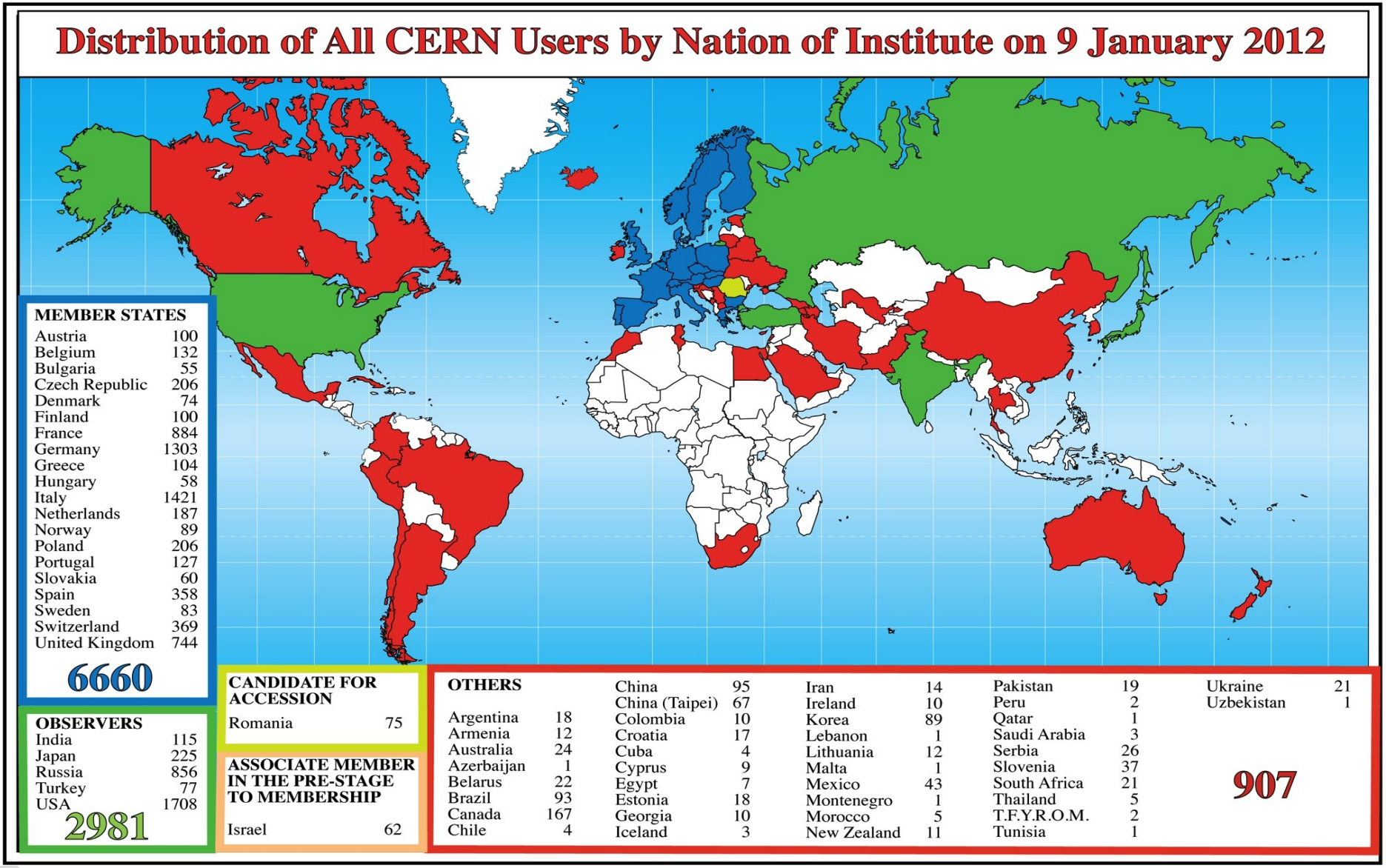
Large Hadron Collider and Experiments



Particle Physics is international teamwork

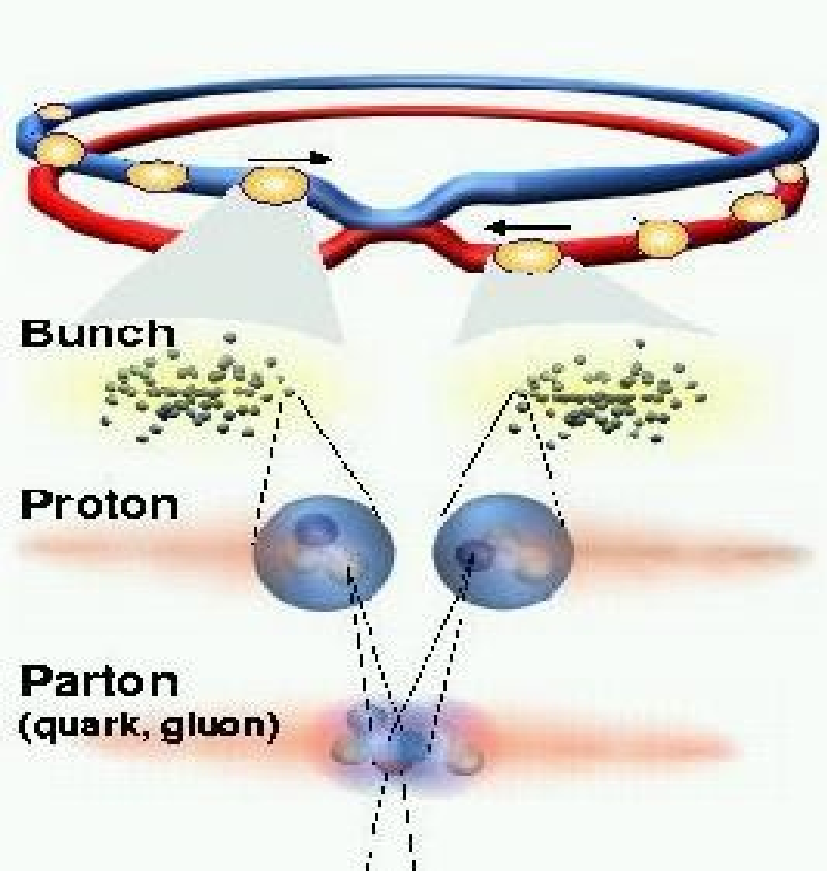
Working @Cern: ~300 institutes fromEurope, ~ 7000 Users
~300 institutes elsewhere, ~ 4000 Users

Particle Physics international



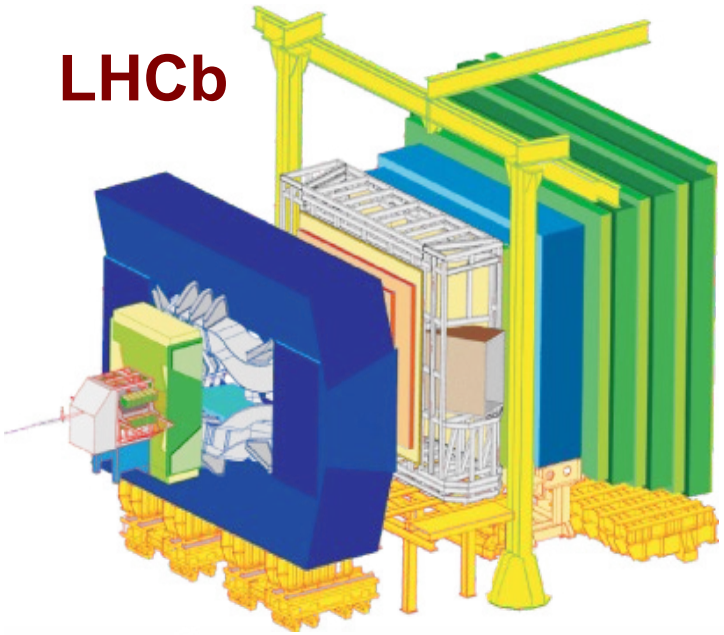
LHC Parameter

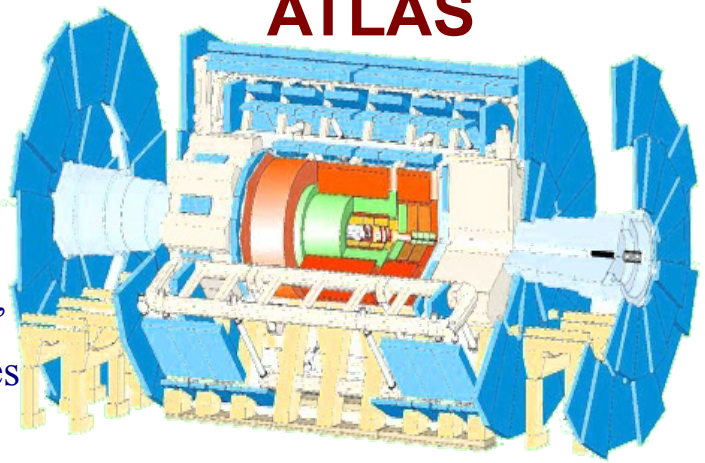
Luminosity: ~10^{34} $cm^{-2}s^{-1}$ design, already exceeded by now



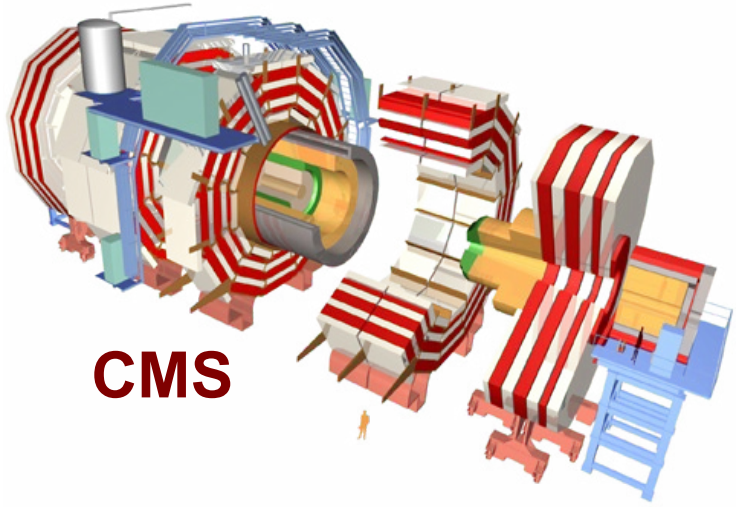
- max. 2835x2835 Proton-Proton bunches
- 10^{11} protons/bunch
- proton energie: 7 TeV
- collision rate: 40 Mhz
- up to 10^{9} pp-interactions/sec
- 4 detektors
  - ATLAS mulit-purpose, pp
  - CMS multi-purpose, pp
  - LHCb asymmetric for b physics
  - ALICE heavy ion physics (Pb ↔ Pb and p ↔ Pb)

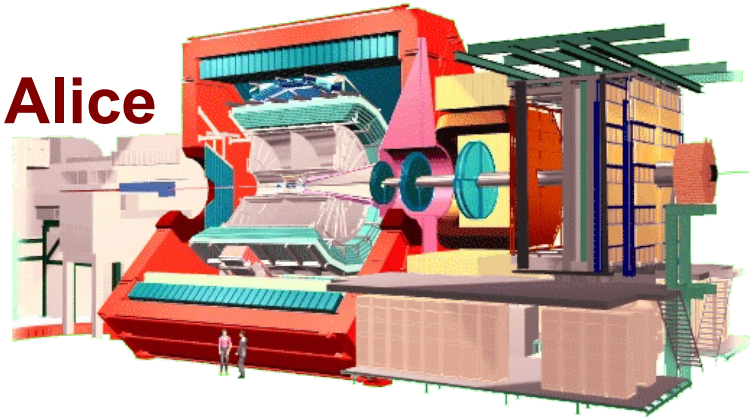
Data sources





- Built by institutes all over the world, ~10'000 physicists from ~70 countries
- Each detector has more than 100 million sensors
- 40 million pp-collisions recorded per second
- detetors specialized on different questions



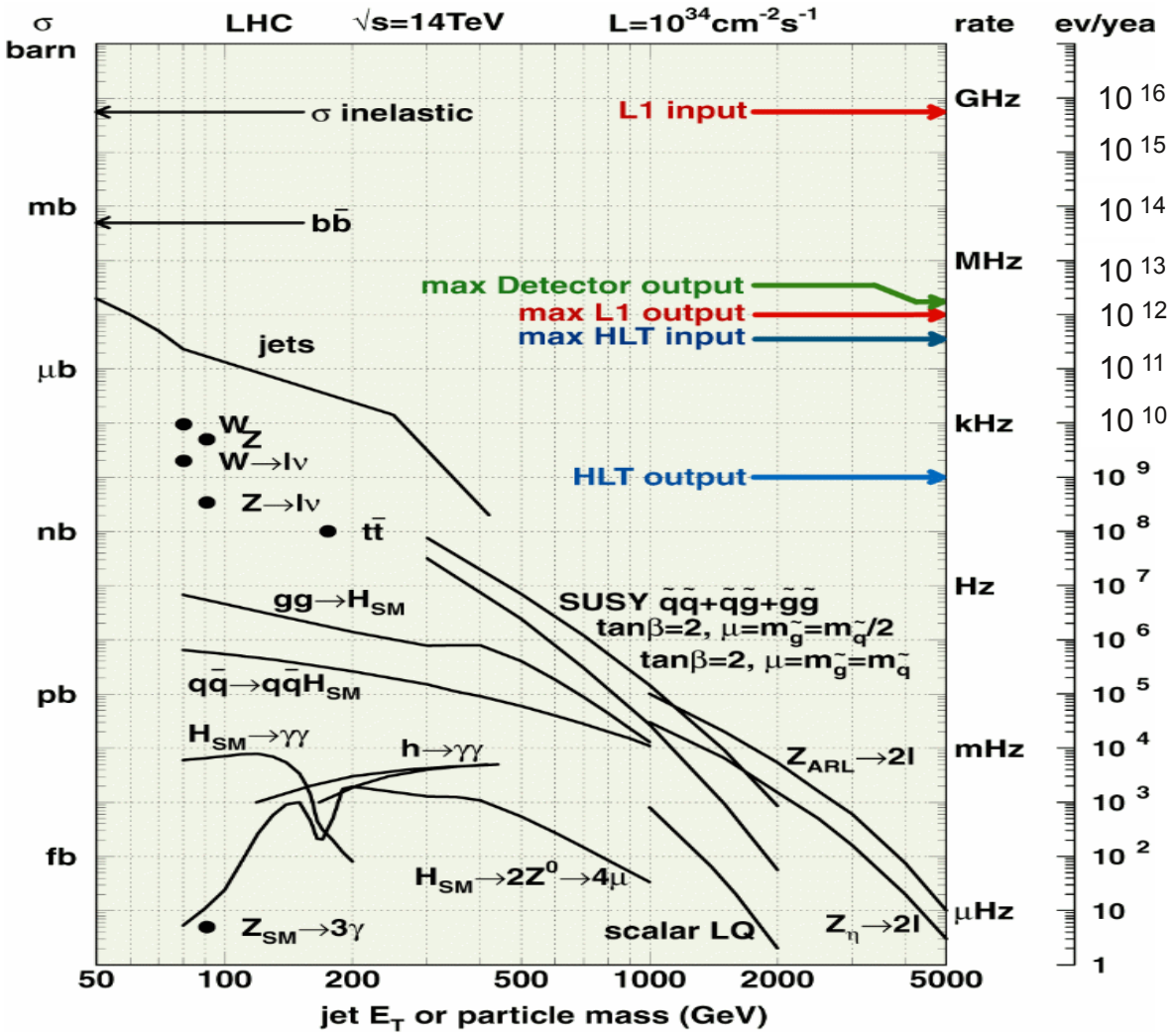


The Challenge

Rates of
"interesting physics"

~10^{10} times smaller than

inelastic pp cross section



Data Sources – example CMS

- ~ 100 Millionen detector cells
- LHC collision rate: 40 MHz
- 10-12 bit/cell

→ **~1000 Tbyte/s raw data**

- Zero-suppression and Trigger reduce this to „only" some 100 Mbyte/s

i.e. 1  /sec

world-wide physics community

Distributed Computing is ideal to do this job !

CMS Trigger & DAQ

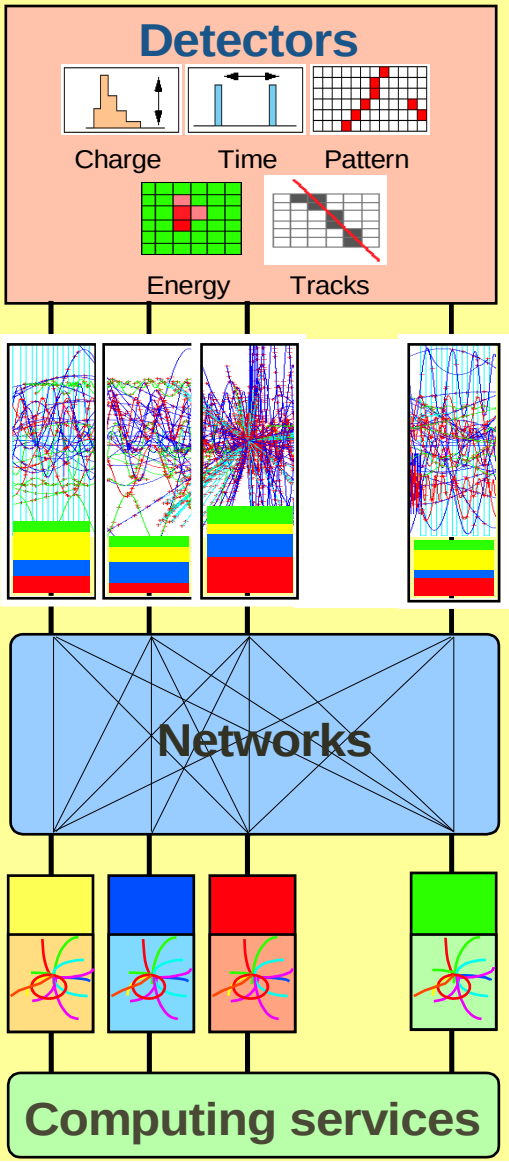
40 MHz
COLLISION RATE

100 kHz
LEVEL-1 TRIGGER

1 Terabit/s
(50000 DATA CHANNELS)

500 Gigabit/s

Gigabit/s SERVICE LAN



16 Million channels
3 Gigacell buffers

1 Megabyte EVENT DATA

200 Gigabyte BUFFERS
500 Readout memories

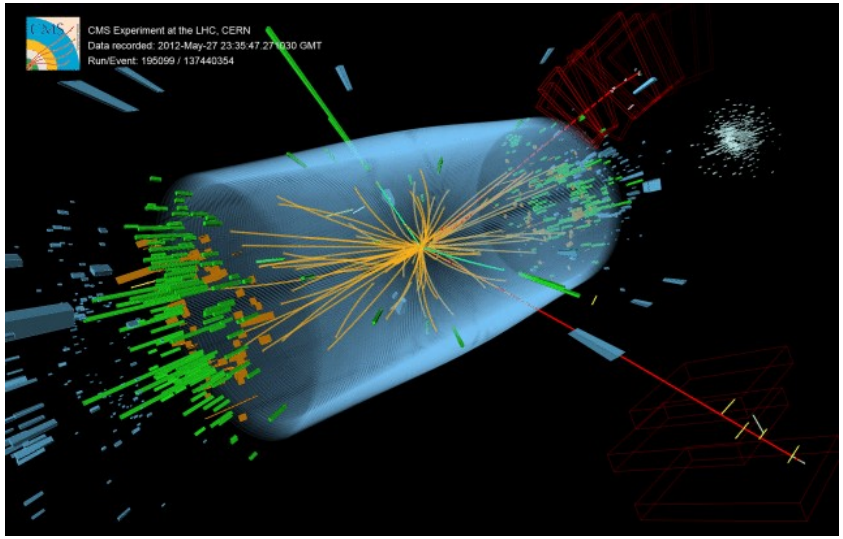
EVENT BUILDER. A large switching network (512+512 ports) with a total throughput of approximately 500 Gbit/s forms the interconnection between the sources (Readout Dual Port Memory) and the destinations (switch to Farm Interface). The Event Manager collects the status and request of event filters and distributes event building commands (read/clear) to RDPMs

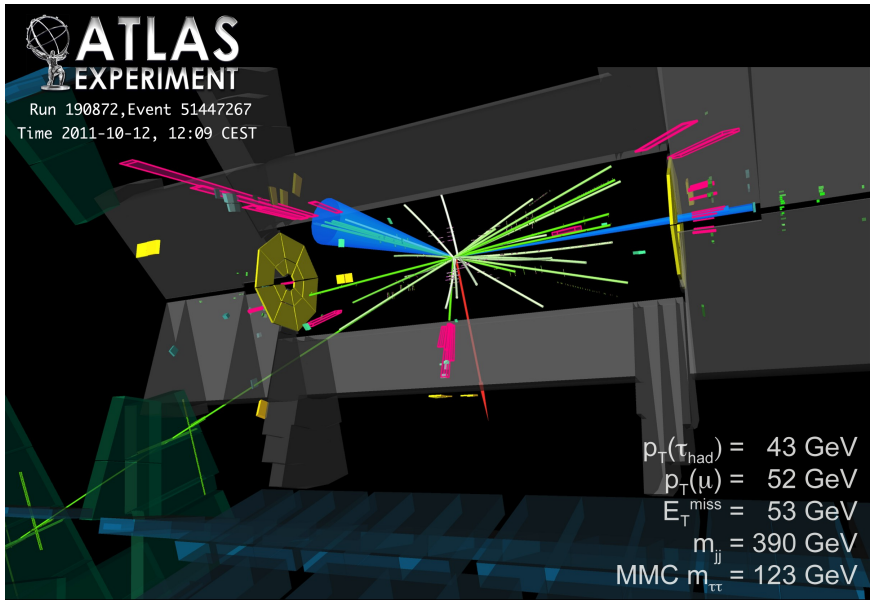
5 TeraIPS

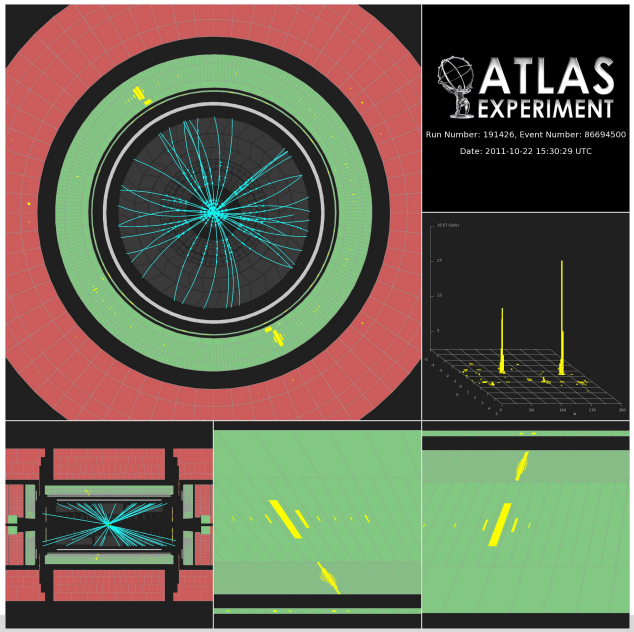
EVENT FILTER. It consists of a set of high performance commercial processors organized into many farms convenient for on-line and off-line applications. The farm architecture is such that a single CPU processes one event

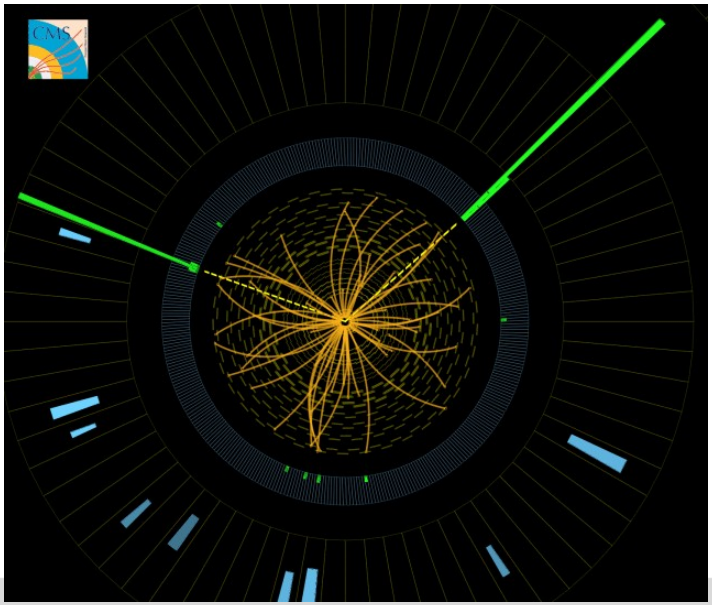
Petabyte ARCHIVE

The output: reconstructed events

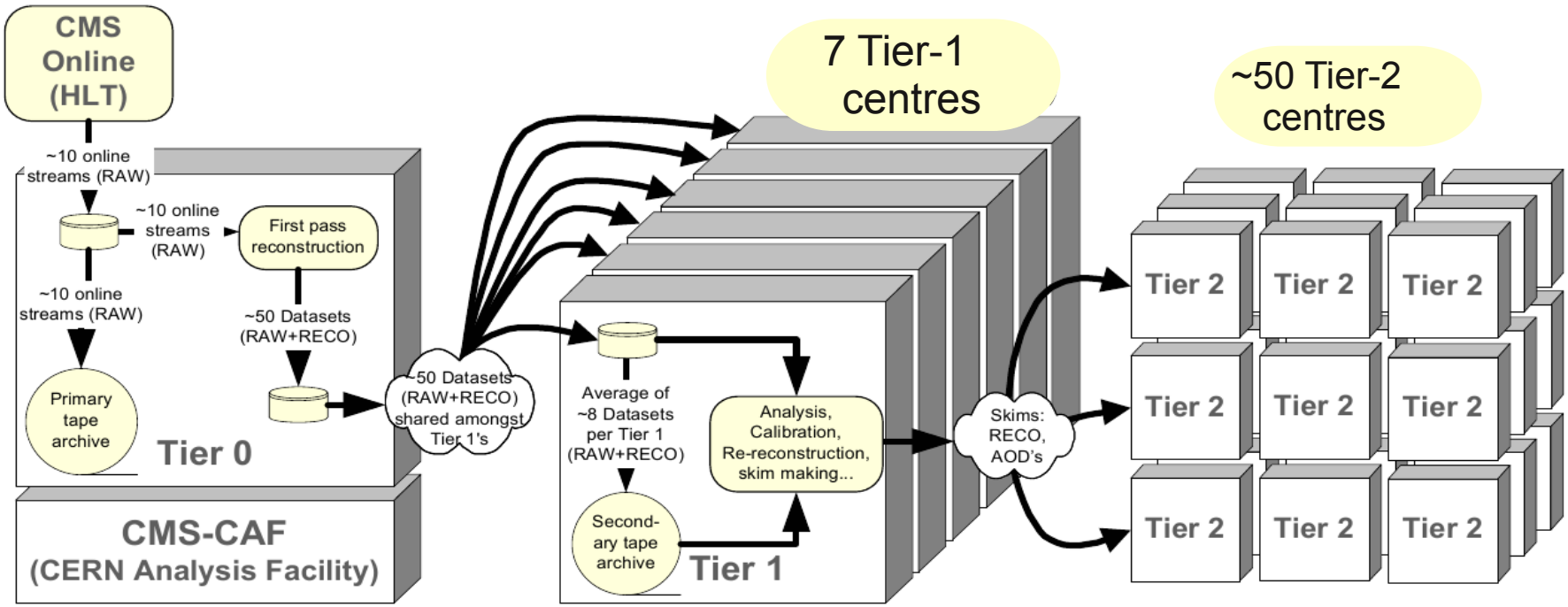








Computing Models – a hierarciacal structure example: C MS



CMS Computing TDR, 2005

LHC-Experiments

- typically **share big Tier-1**s, take responsibility for experiment-specific services
- have a **large number of Tier2s**, usually supporting only one experiment
- have an **even larger number of Tier-3s** without any obligations towards WLCG

The Grid Paradigm ...

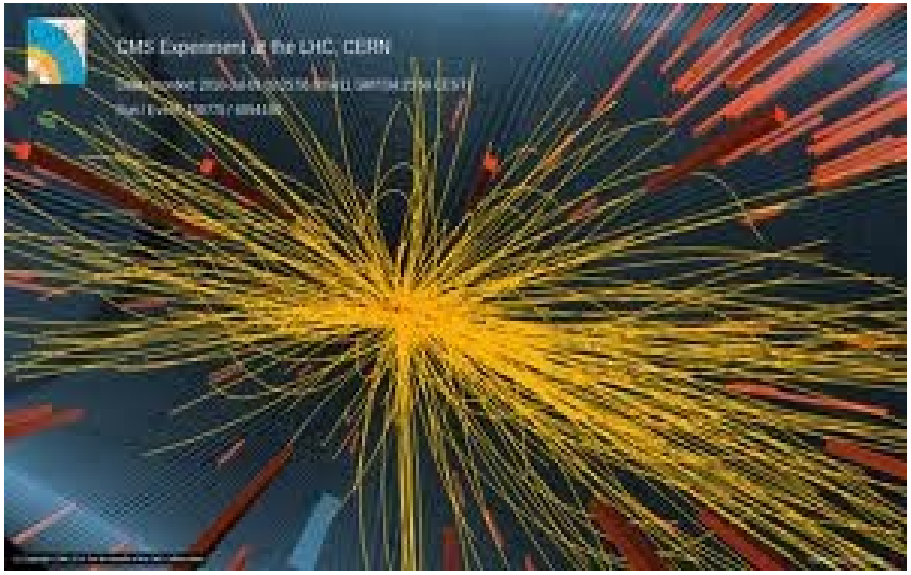
… came at the right time to provide a model for distributed computing in HEP

> **"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 nontrivial quality of services*

I. Foster, 2002

Why Grid is well suited for HEP



Experimental HEP codes - key characteristics:

- modest memory requirement (~2GB) & modest floating point
  → **perform well on PCs**
- independent events
  → **easy parallelism**
- large data collections (TBytes)
- shared by very large collaborations

Grid for LHC

- Given the international and collaborative nature of HEP computing must be distributed

  → harvest intellectual contributions from all partners, also funding issues

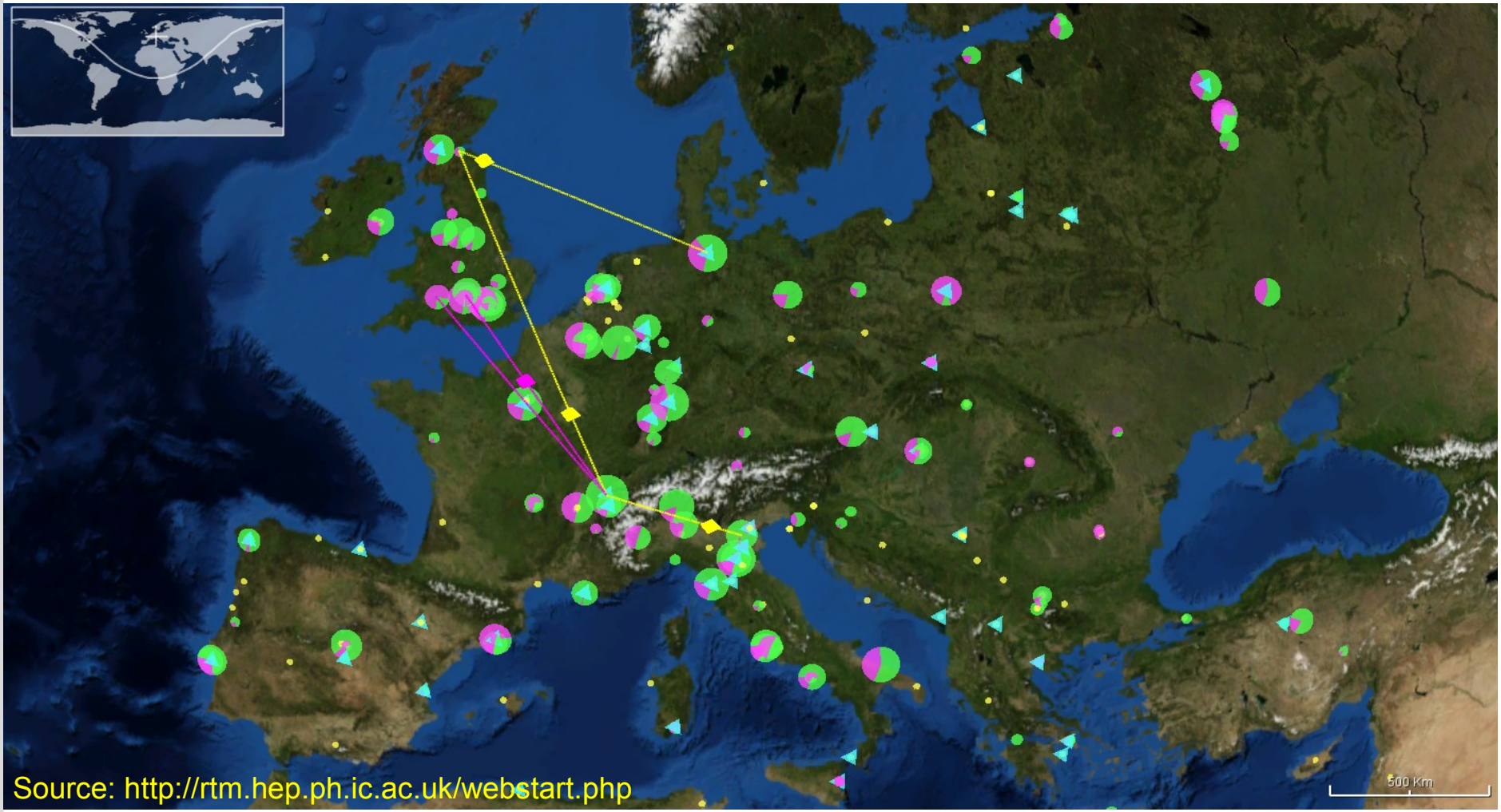
- Early studies in 1999 (Monarc Study group) suggested a hierarcical approach, following the typical data reduction schemes usually adopted in data analysis in high enery physics

- Grid paradigm came at the right time and was adopted by LHC physicists as the base line for distributed computing

- Major contributions by physicists to developments in Grid computing

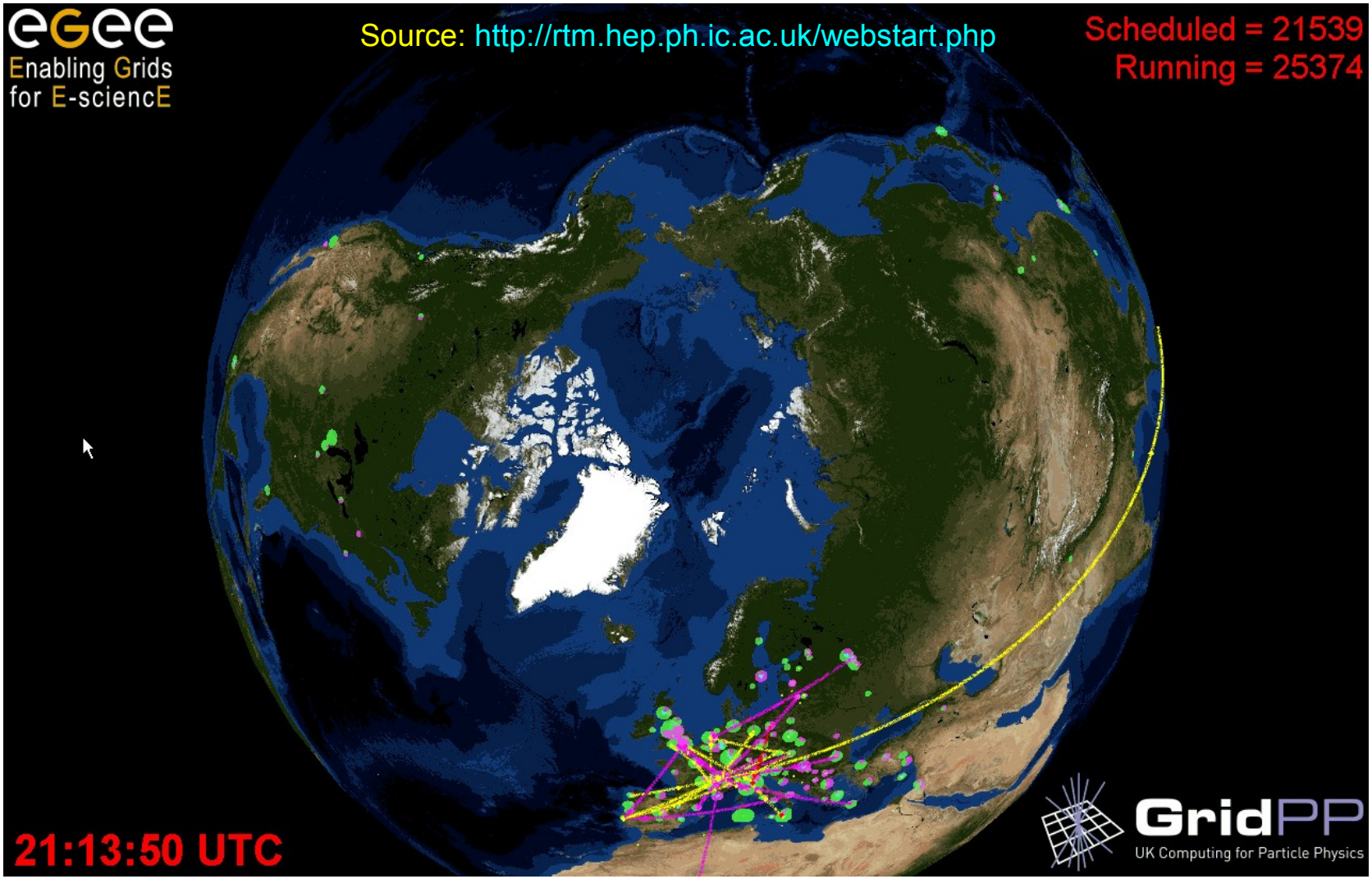
- Other HEP communities also benefit and contributed

WLHC Computing Grid in action



A truly international, world-spanning Grid for LHC data processing, simulation and analysis

WLHC Computing Grid in action 365d x 24h



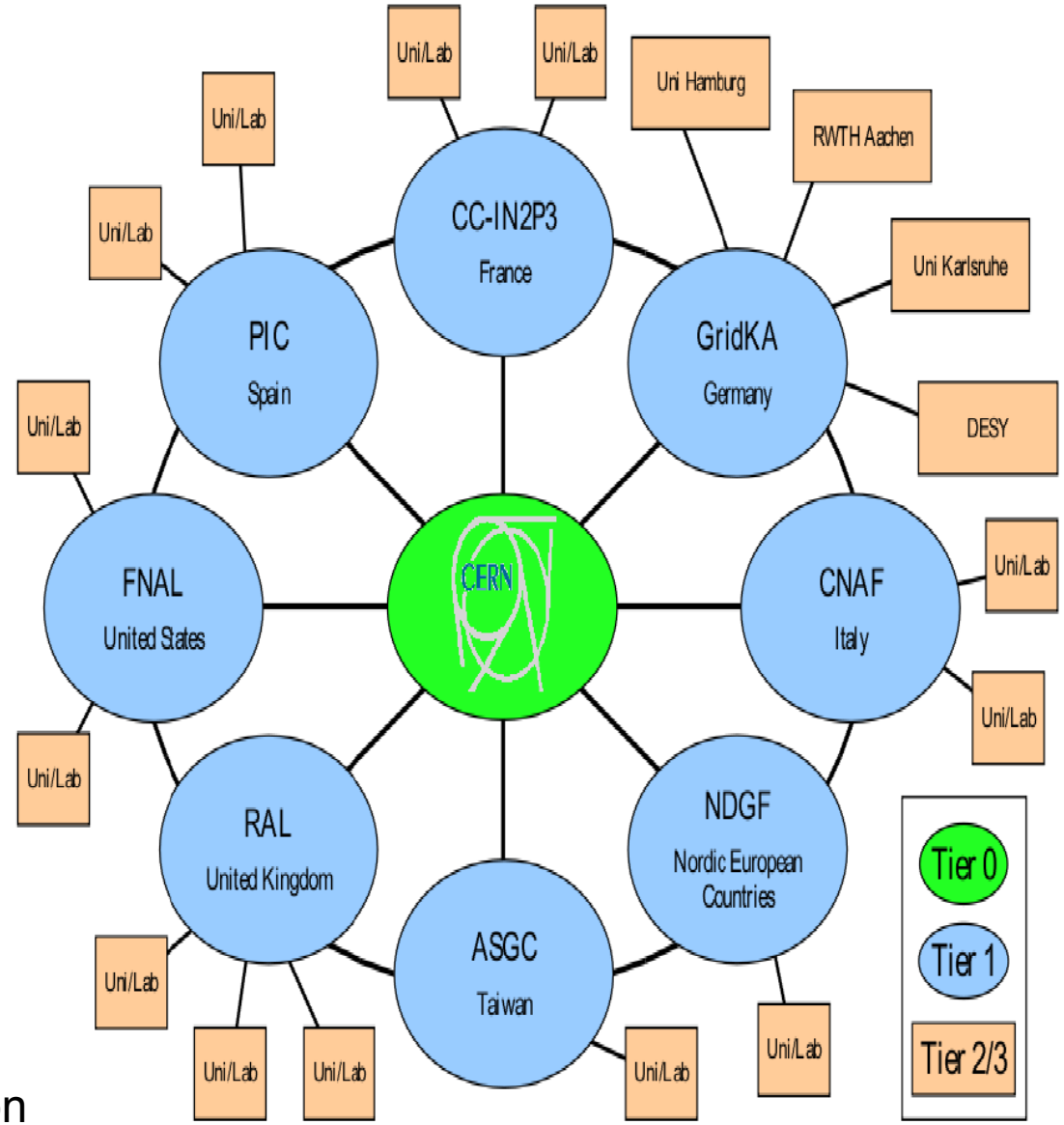
A truly international, world-spanning Grid for LHC data processing, simulation and analysis

Structure of the LHC Grid

A grid with hierarcies and different tasks at different levels

In addition, it is a **"Grid of Grids"** with interoperability between different middlewares:

- *gLite* middleware in most of Europe
- *Open Science Grid* in USA
- *NorduGrid* in Northern Europe
- *Alien* by the Alice collaboration



WLCG: Grid with hierarcical structure

Tier-0 **the accelerator centre**

- Data acquisition & initial processing
- Long-term data curation
- Distribution of data to T1/T2

Tier-1

11 Tier-1 Centres

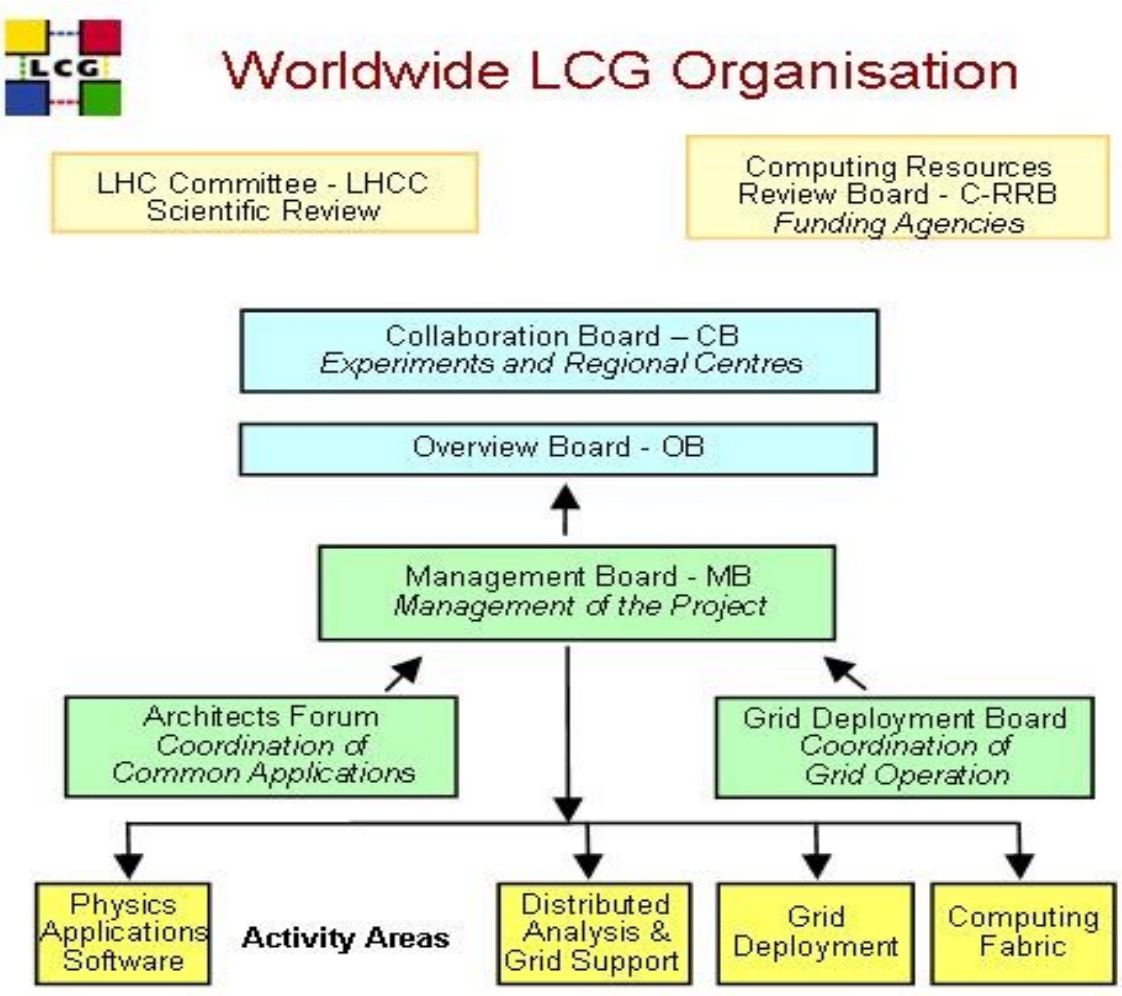
- "online" to the data acquisition process
  → high availability
- Managed Mass Storage
  → grid-enabled data service
- Data-intensive analysis
- National, regional support

Tier-2 **150 Centres in 60 Federations in 35 countries**

- **End-user (physicist, research group) analysis** & collaboration with T3 (= institute resources) **– where the discoveries are made**
- Monte Carlo Simulation

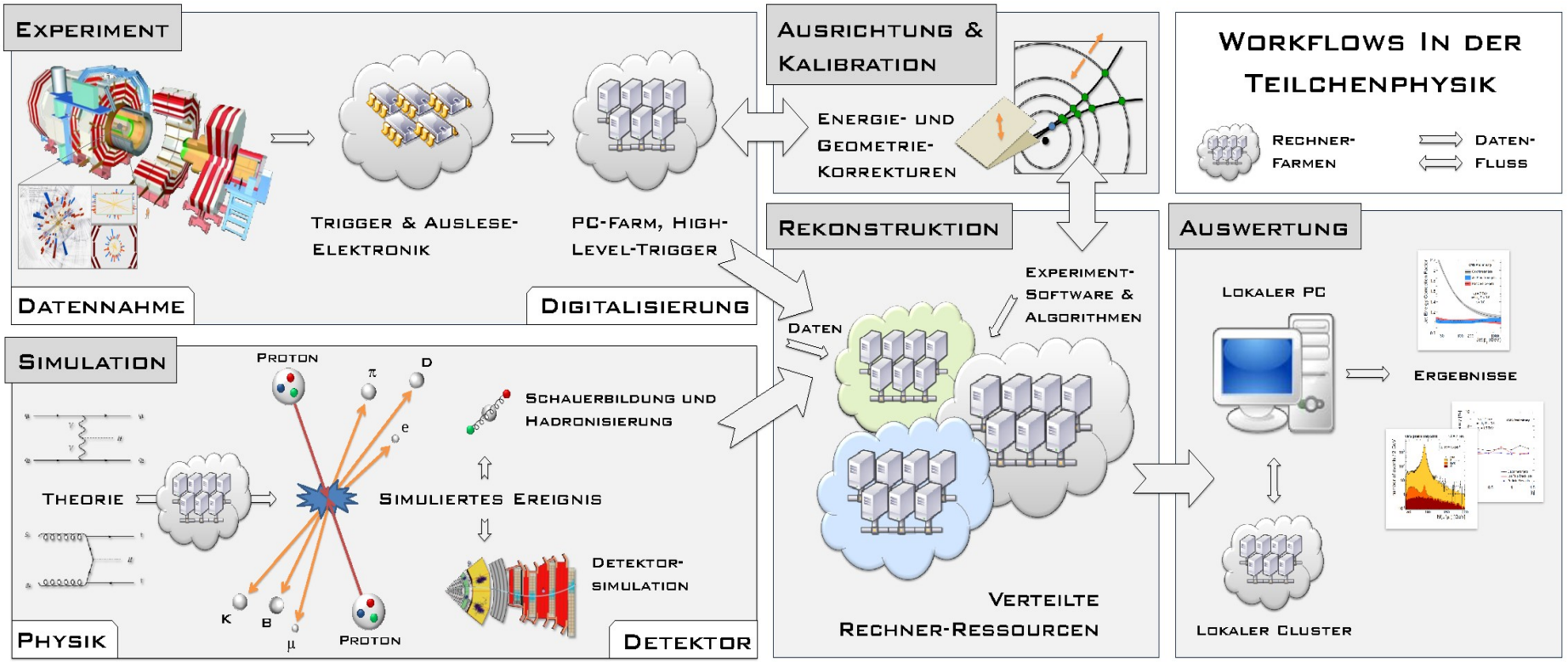
Tier-3 several 100 grid-enabled PC clusters @ institutes

Organisation of the World-wide LHC computing Grid



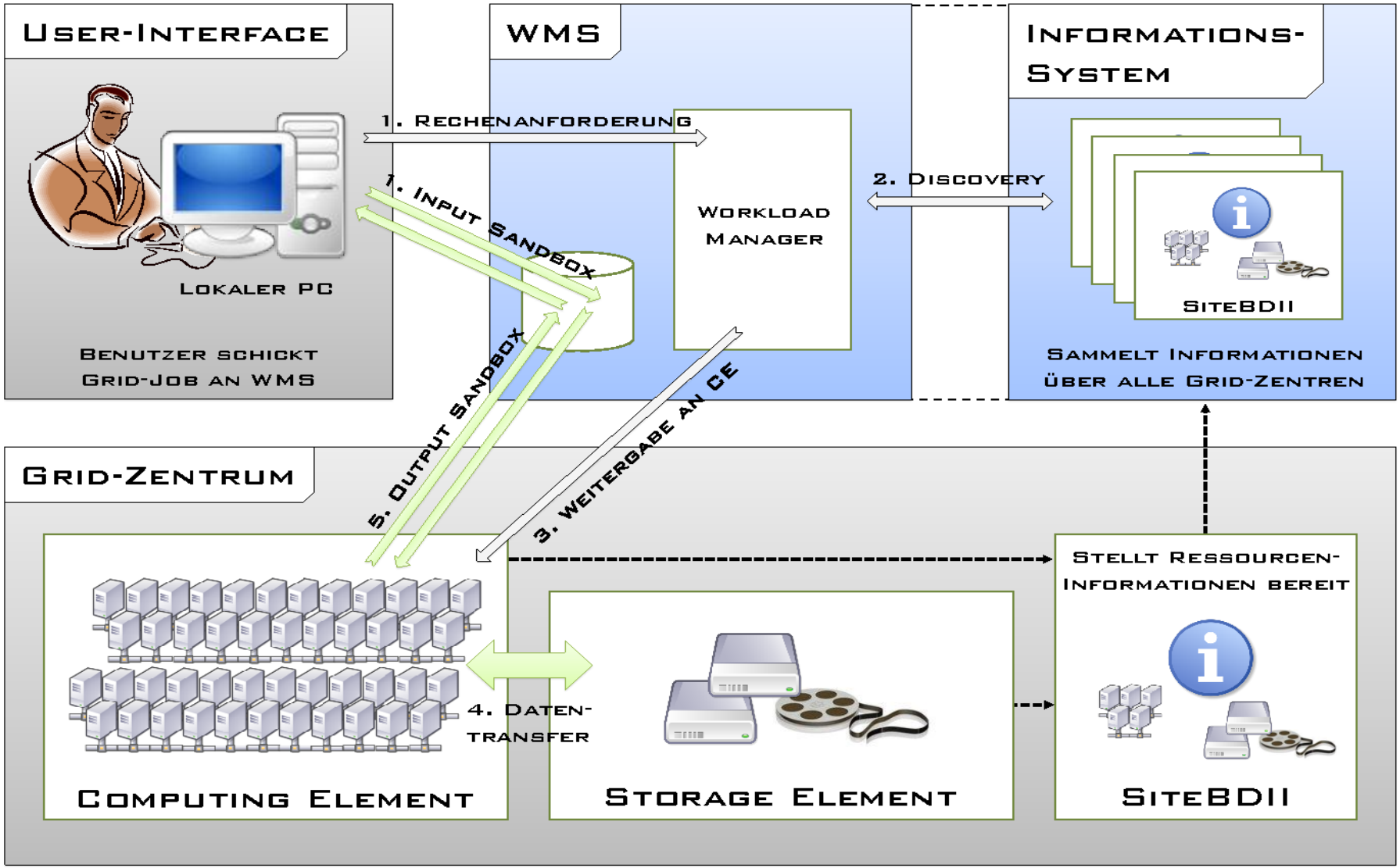
Grids can`t work without an organisational structure representing all parties involved

What Particle Physicists do on the Grid

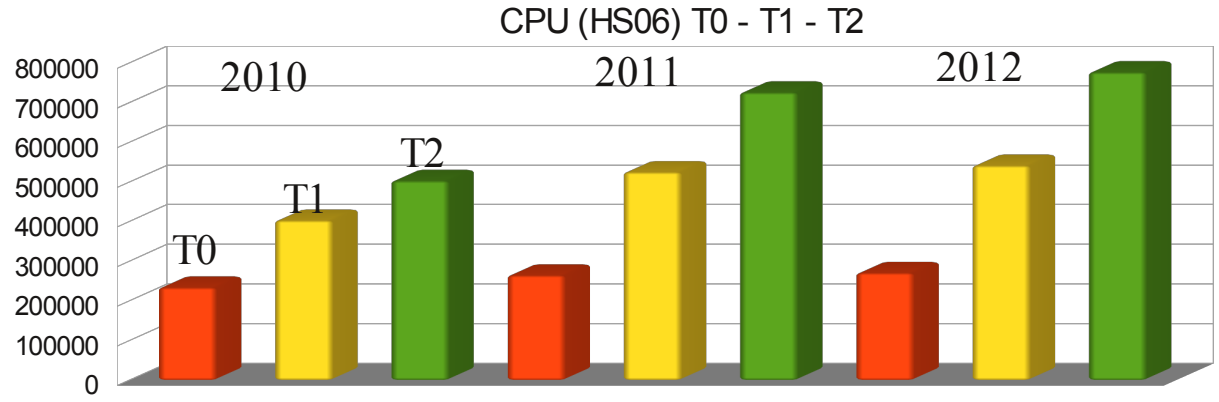


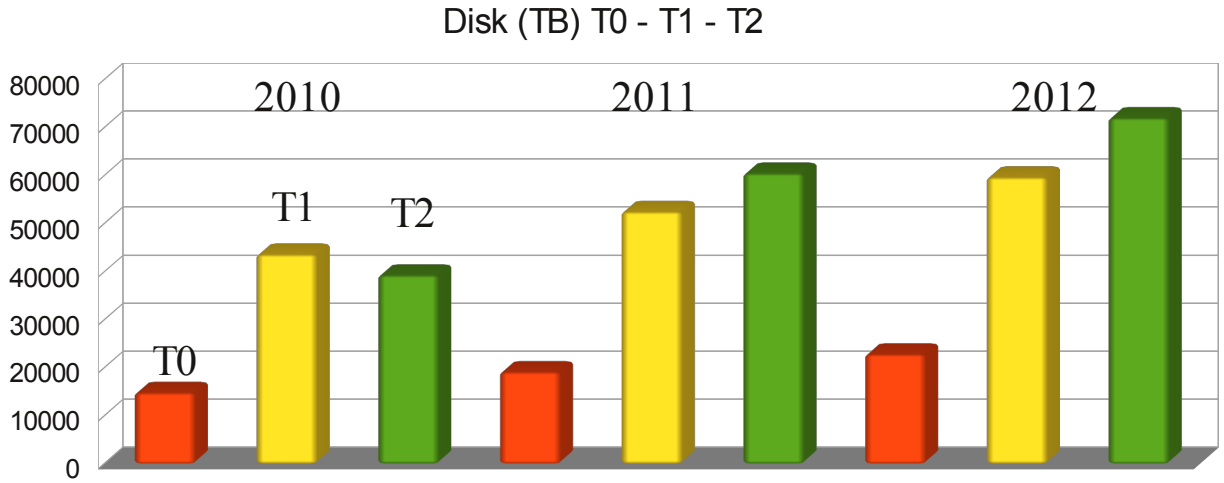
- **CPU-intense simulation** of particle physics reactions and detector response
- **processing** (=reconstruction) of **large data volumes**
- **I/O-intense filtering and distribution** of data
- **transfer** to local clusters and workstations for final physics interpretation

Typical workflows on the Grid



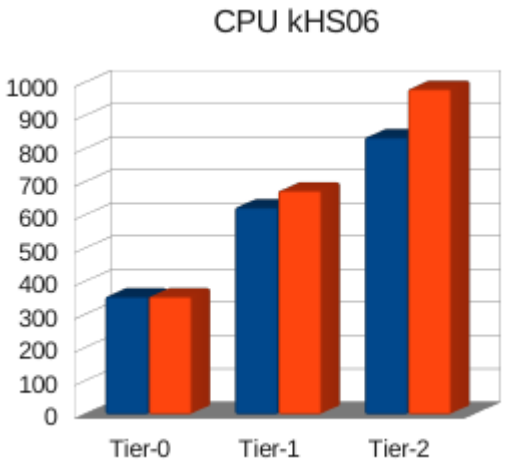
How big is WLCG ?



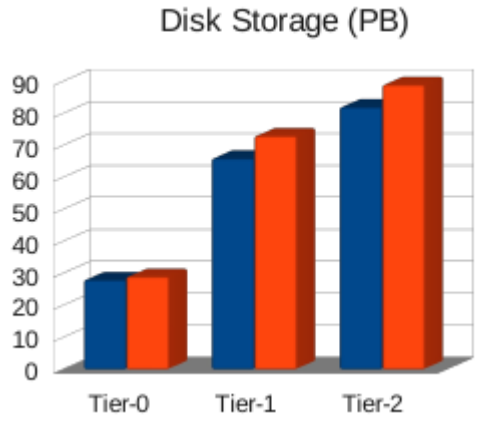


The largest Science Grid in the World

How big is WLCG today and compared with others ?



Total CPU 2013:
2010 kHS06
approx. equiv.
200'000 CPU cores



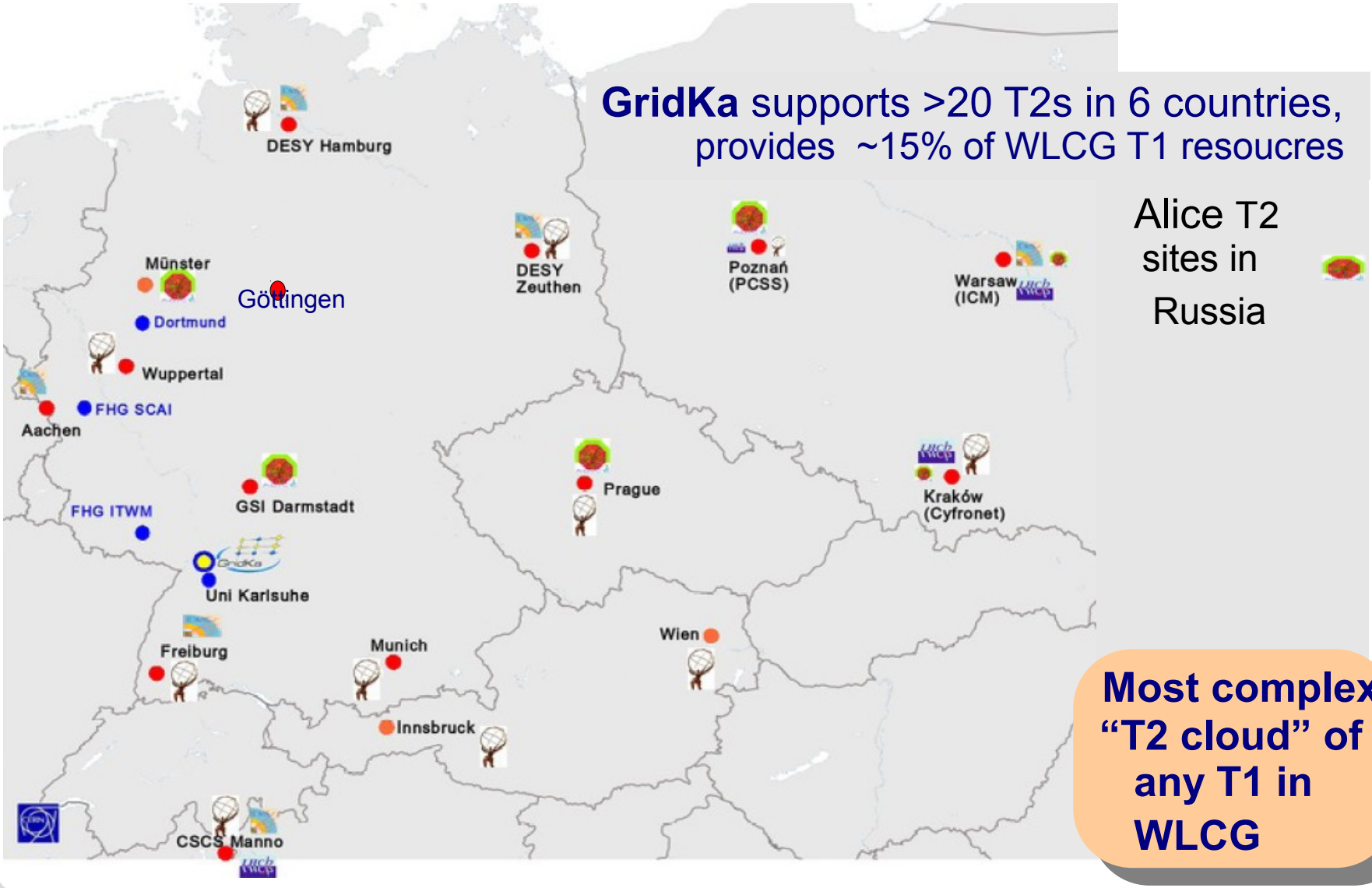
Total Disk 2013:
190 PB,

same amount as
Tape Storage
190 PB

Other "big players" ?

- new HPC Cluster in Munich
  (SuperMuc, #64 on Top 500 list)
  155 k Cores, 10 PB disk
- 1&1 Rechenzentrum Karlsruhe
  30'000 Server, ~2 PB/Monat I/O
- Amazon Cloud:
  ~ 500 PB
- Facebook
  ~30 PB
- Google Data Center
  power 200 MW (WLCG: ~6MW)

A closer look to the surroundings of GridKa



GridKa supports >20 T2s in 6 countries, provides ~15% of WLCG T1 resoucres

Alice T2 sites in Russia

Most complex "T2 cloud" of any T1 in WLCG

After thee years of experience with LHC operation:

How well did it work ?

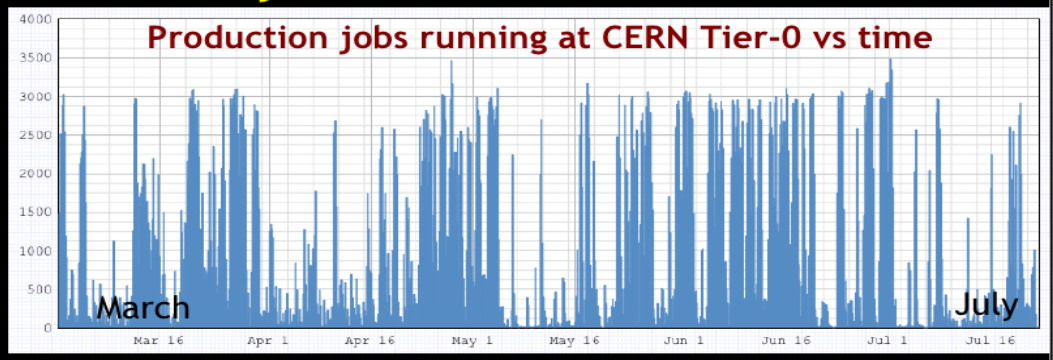
After 2 years of experience - Did it work ?

Up to the users to give feedback:

D. Charlton,ATLAS, EPS HEP 2011

Computing Grid Delivers Physics

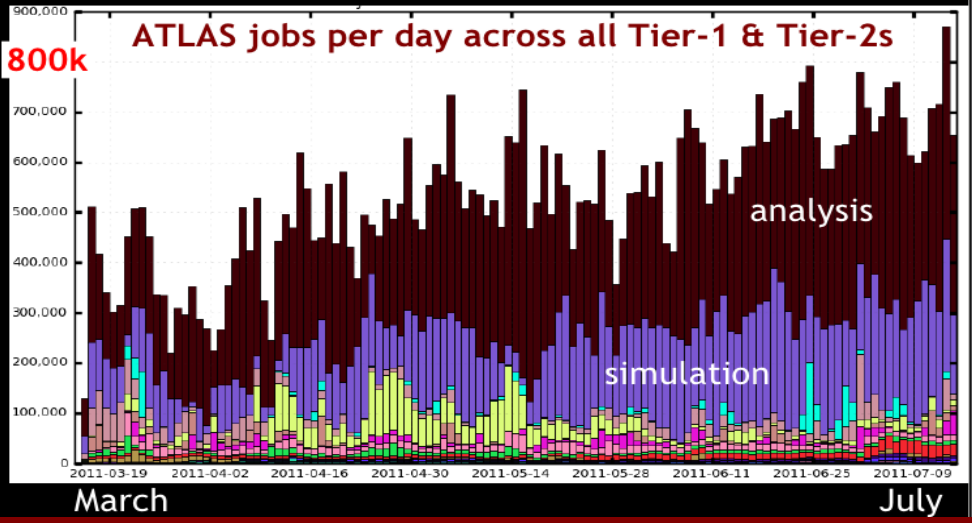
Data preparation:
- First-pass reco. at Tier-0 within ~2 days
- Calibration/DQ good for physics analysis
- Data analysable on Grid within ~1 week



Tier-1 and Tier-2's process ~⅔ M jobs per day
- simulation
- re-reconstruction (campaigns)
- group production (ntuples...)
- physics analysis

The high quality computing system allows us to show results on data taken until the end of June

Payback for the years of investment and hard work

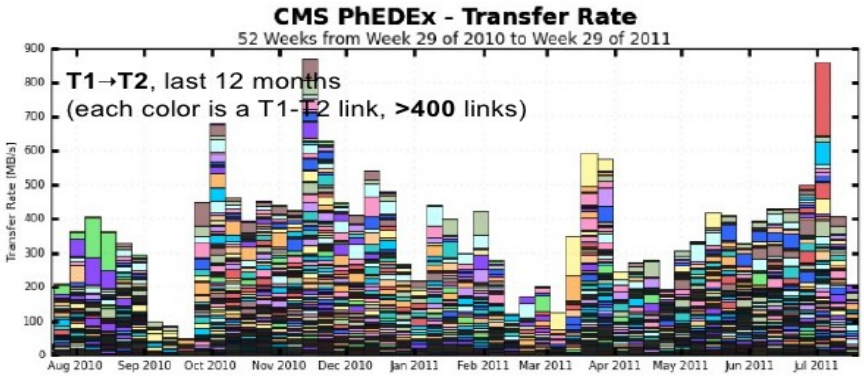


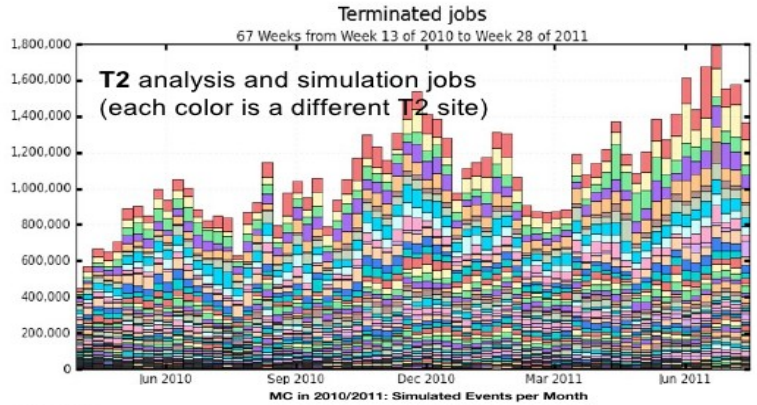
Did it work (2)

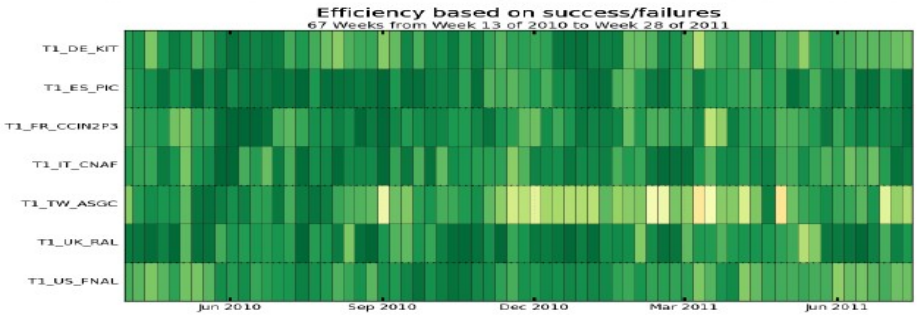
G. Tonelli, CMS, EPS HEP 2011

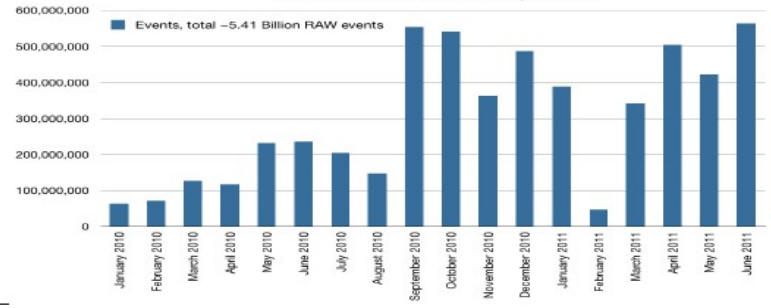
Offline and Computing running smoothly

•Smooth **Tier-0** operation, keeping up with the data taking. Increase in **Tier-1** utilization, for reprocessing and skimming jobs; High usage of **Tier-2** for analysis, **>400 (800)** individual users per week (month). More than 5.4 Billions MC events.









G. Tonelli, CERN/INFN/UNIPI *HEP_2011_GRENOBLE* *July 25 2011* 36

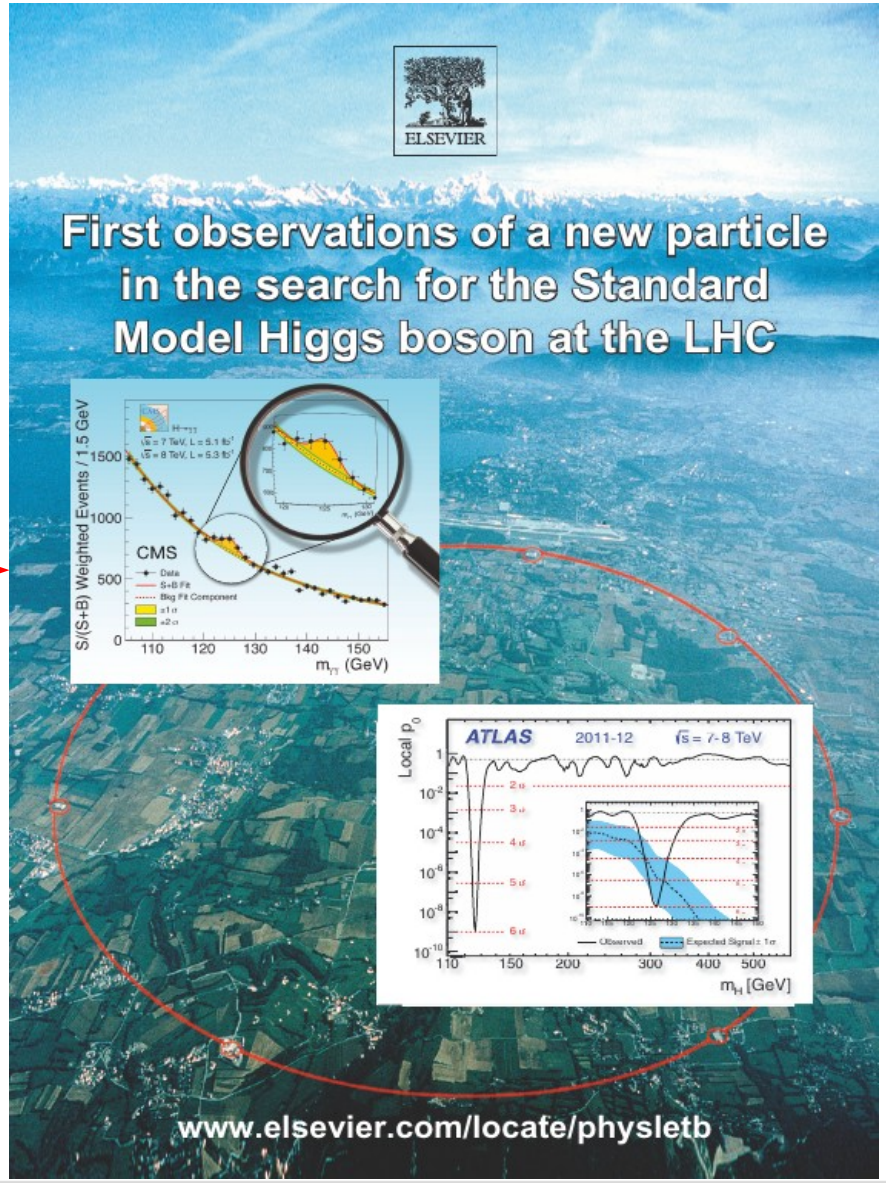
Did it work ?

Obviously it did!

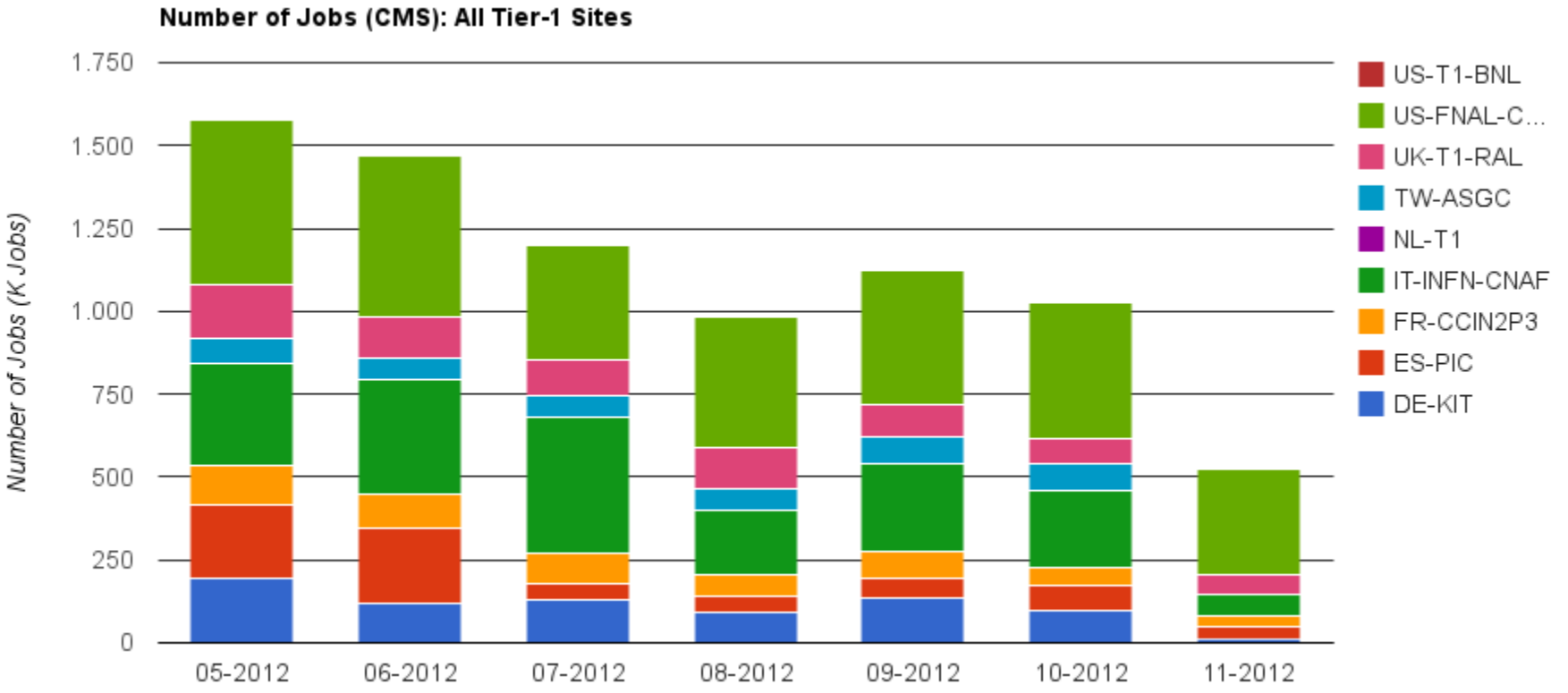
- Grid infrastructure for the LHC performed extremely well
- physics results from freshly recorded data

- but: effort for running computing infrastructure is high!

Let`s have a look in detail ...

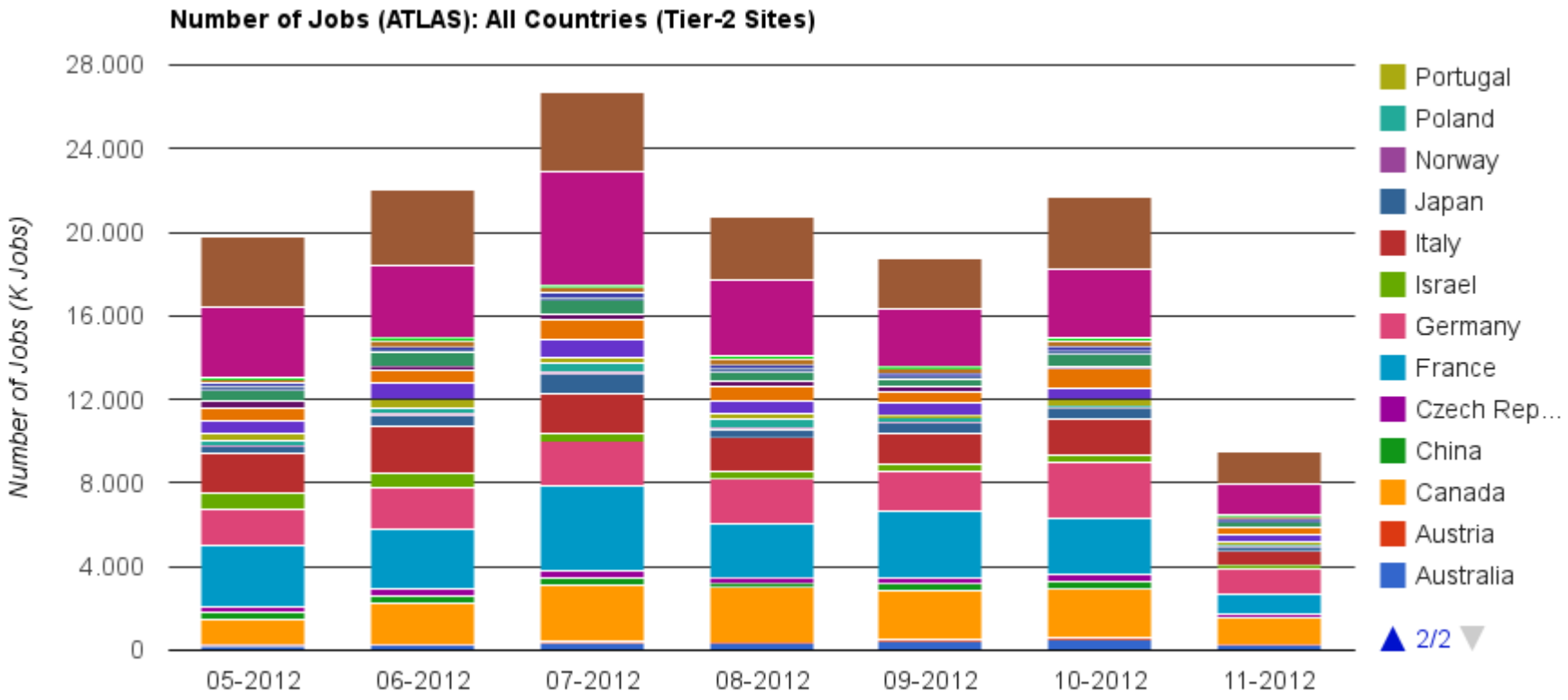


Tier1 Jobs CMS



CMS: ~1 million jobs/month
ATLAS: ~5 million (shorter) jobs/month on Tier1s

Tier2 Jobs ATLAS



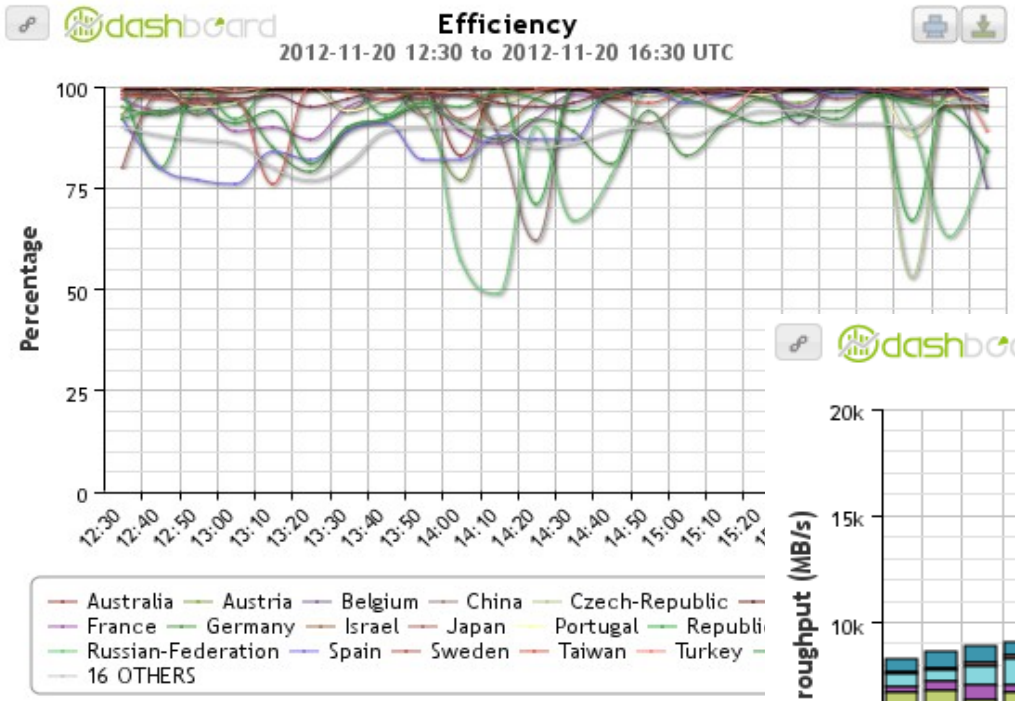
ATLAS: ~20 million jobs/month
CMS: ~ 7 million jobs/month on Tier2s

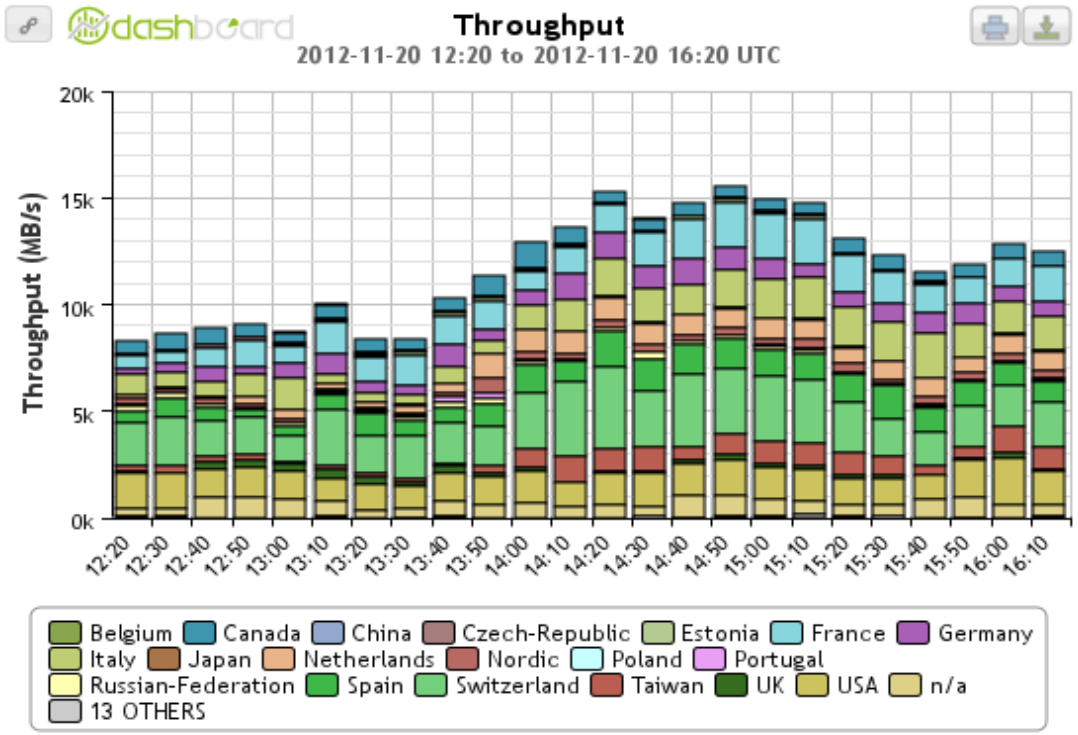
Monitoring – the key to success

	TOTAL-	Australia+	Austria+	Belgium+	Brazil+	Canada+	Chile+	China+	Czech-Republic+	Estonia+	Finland+	France+	Germany+	Hungary+	Israel+	Italy+	Japan+	Netherlands+	Nordic+	Poland+	Portugal+	Republic-of-Korea+	Romania+	Russian-Federation+	Spain+	Sweden+	Switzerland+	Taipei+	Taiwan+	Turkey+	UK+	USA+	Ukraine+	n/a+
TOTAL-	93 %	22 %	76 %	13 %	28 %	99 %	38 %	94 %	100 %	98 %	74 %	96 %	98 %	93 %	100 %	98 %	100 %	94 %	97 %	89 %	100 %	63 %	94 %	97 %	99 %	100 %	97 %	100 %	99 %	91 %	49 %	96 %	79 %	93 %
Armenia+	0 %																	0 %																
Australia+	100 %					99 %						100 %	100 %			100 %		100 %						100 %	100 %		100 %				100 %	100 %		100 %
Austria+	98 %	100 %		100 %	50 %	100 %		71 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	50 %	100 %	100 %	100 %	100 %	100 %		100 %	100 %	88 %	100 %		
Belgium+	99 %		100 %		100 %			100 %		100 %	100 %	100 %	100 %	100 %		98 %				100 %	100 %	100 %		100 %	100 %		100 %		100 %		100 %	100 %		
Brazil+	6 %		0 %	0 %	1 %			25 %		25 %	10 %	20 %	6 %	25 %		25 %				25 %	0 %	8 %		19 %	30 %		26 %		12 %	33 %	24 %	6 %	33 %	
Canada+	89 %	30 %				100 %						100 %	100 %			98 %		100 %	84 %						100 %		98 %		100 %		3 %	99 %		100 %
Chile+	89 %																								95 %							33 %		
China+	96 %		0 %	100 %	100 %					0 %	100 %	100 %	88 %	100 %		92 %					100 %	100 %	57 %	67 %	21 %		80 %		88 %	100 %	100 %	33 %		
Czech-Republic+	96 %		100 %			100 %		100 %				100 %	86 %		100 %	100 %	100 %	100 %	100 %		100 %		100 %	100 %	100 %	100 %	100 %		100 %	100 %	100 %	100 %		100 %
Estonia+	100 %											100 %	100 %			100 %									100 %		100 %							
Finland+	100 %											100 %	100 %			100 %									100 %		100 %					100 %		
France+	96 %		0 %	100 %	100 %	98 %		97 %		100 %	50 %	98 %	99 %	100 %		98 %	100 %	99 %	98 %	50 %		50 %	96 %	100 %	98 %		100 %		100 %	100 %	34 %	98 %		100 %
Germany+	93 %	33 %	98 %	100 %	100 %	100 %		73 %	100 %	100 %		100 %	98 %	100 %	100 %	99 %	100 %	94 %	100 %	92 %	100 %	100 %	100 %	98 %	99 %	100 %	99 %		100 %	100 %	52 %	92 %	100 %	85 %
Greece+	0 %															0 %																		
Hungary+	91 %		100 %	100 %	100 %			0 %		100 %	100 %	100 %	100 %			63 %				100 %	100 %	100 %		75 %	100 %		100 %		100 %		100 %	100 %		
Israel+	Folie 36					98 %							97 %		0 %	100 %		100 %	97 %						100 %		98 %				98 %	100 %		
Italy+	90 %		0 %	100 %	67 %	93 %		50 %		63 %	100 %	100 %	98 %	100 %		100 %		97 %	33 %	100 %	100 %	100 %	87 %	99 %	100 %		100 %		100 %	100 %	39 %	97 %	100 %	37 %
Japan+	98 %					100 %						99 %				97 %		100 %	88 %				81 %		100 %				100 %			100 %		
Netherlands+	92 %	100 %				100 %		100 %				68 %	98 %		100 %	100 %		89 %	100 %					100 %	100 %		100 %		100 %		98 %	98 %		
Nordic+	91 %	100 %	75 %			86 %						50 %	93 %		100 %	95 %	100 %	97 %	96 %		87 %		100 %	80 %	100 %		95 %		100 %		35 %	93 %		90 %
Pakistan+	75 %											0 %				100 %									100 %				100 %					
Poland+	88 %											100 %	99 %			100 %				46 %					100 %		100 %		100 %			100 %		
Portugal+	98 %	0 %	100 %	100 %	60 %	100 %		86 %	100 %	100 %	100 %	92 %	100 %	100 %	100 %	93 %	100 %	96 %	71 %	100 %	100 %	100 %	100 %	98 %	100 %	100 %	100 %		100 %	71 %	100 %	100 %	0 %	
Republic-of-Korea+	95 %			100 %	100 %			0 %		100 %	100 %	100 %	100 %	100 %		100 %				100 %				75 %	100 %		100 %		100 %		50 %	100 %	100 %	
Romania+	83 %							96 %				82 %	86 %				100 %	80 %					67 %									100 %		
Russian-Federation+	93 %		100 %	100 %	100 %	100 %		0 %		100 %	100 %	99 %	100 %	100 %	100 %	100 %		100 %	99 %	100 %	100 %	100 %		96 %	100 %		98 %		100 %	100 %	1 %	99 %	100 %	
Spain+	92 %	14 %	75 %	100 %	100 %	100 %	38 %	69 %	100 %	100 %	100 %	97 %	99 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %		100 %	83 %	18 %	98 %	100 %	0 %
Sweden+	100 %					100 %						100 %	100 %		100 %			100 %	100 %				100 %	100 %	100 %		100 %			100 %	100 %	100 %		
Switzerland+	88 %	0 %	25 %	100 %	100 %	72 %		28 %		100 %	100 %	98 %	99 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	94 %	99 %		100 %		100 %	100 %	13 %	100 %		98 %
Taiwan+	99 %		0 %	100 %	40 %	98 %		99 %		100 %	100 %	98 %	76 %	100 %		96 %		100 %	99 %	0 %	100 %	100 %		93 %	100 %		100 %		100 %	100 %	9 %	98 %		100 %
Thailand+	2 %												100 %			1 %											0 %							
Turkey+	100 %		100 %	100 %	100 %			0 %		100 %	100 %	100 %	100 %	100 %		100 %		100 %			100 %	100 %		100 %	100 %		100 %				100 %	100 %		
UK+	51 %											100 %	86 %			100 %									100 %		2 %		100 %			95 %		
USA+	94 %	0 %	0 %	67 %	81 %	100 %		45 %		90 %	100 %	98 %	98 %	100 %		90 %		100 %	96 %	67 %	100 %	100 %	100 %	100 %	97 %		100 %	100 %	98 %	100 %	15 %	94 %		99 %
Ukraine+	85 %		0 %	100 %	0 %			0 %		100 %	100 %	57 %	100 %	100 %		100 %				100 %	100 %			60 %	100 %		100 %		100 %		100 %	100 %		
n/a+	98 %	90 %				100 %		100 %	100 %			100 %	100 %			100 %		100 %	100 %				100 %	100 %	100 %		100 %	100 %	100 %		74 %	98 %		100 %

WLCG Quality Monitoring of Data Transfer Matrix (by country)

Data rates





Routinely run data transfers across the world at a transfer volume of ~50 TBytes/h

Again: Does it work ?

YES !

- Routinely runing ~1 million jobs per day on the Grid
- Shipping over 1 PB/day of data around
- data distribution to T2 works well
- very little is known about T3 usage and success rates -
  responsibility of the institutes
- plenty of resources were available at LHC start-up,
  now reached “resource limited operation”
- Users have adapted to the “GridWorld” -
  Grid is routinely used as a huge batch system, output is transferred home

but ...

Does it work?

Message:

it worked better than expected by many,

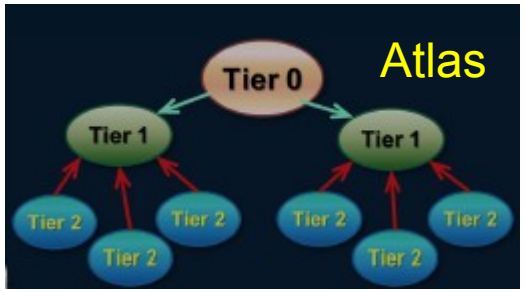
but running such a complex computing infrastructure as the WLCG is tedious (and expensive!)

Reliability and cost of operation can be improved by

- simplified and more robust middleware
- redundancy of services and sites,
  requires dynamic placement of data and investment in network bandwidh
- automated monitoring and triggering of actions
- use of commercially supported approaches to distributed computing:
  - private clouds are particularly important for shared resources at universities
  - eventually off-load simple tasks (simulation, statistics calculations) to commercial clouds

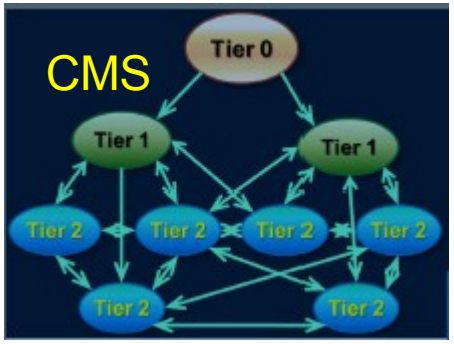
Let`s have a look at some future developments ...

LHCONE



ATLAS and CMS
computing models differ slightly

CMS already more "distributed"



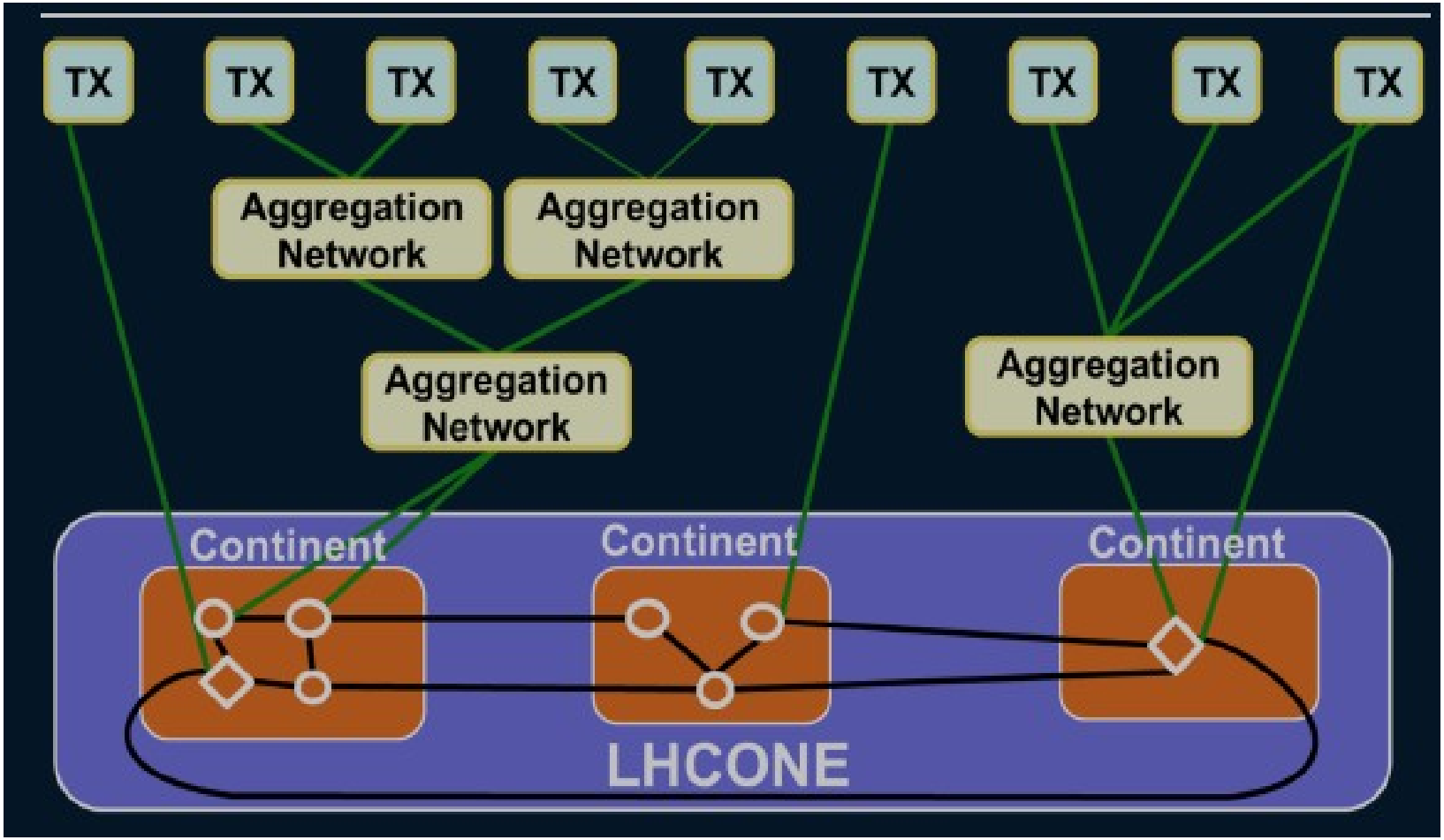
Aim of LHC ONE project is

better trans-regional networking for data analysis,
complementary to **LHCOPN** network connecting LHC T1s

- **flat(er) hierarchy:** any site has access to any other site's data
- **dynamic data caching:** pull data "on demand"
- **remote data access:** jobs may use data remotely

by interconnecting open exchange points between regional networks

LHCONE (2)



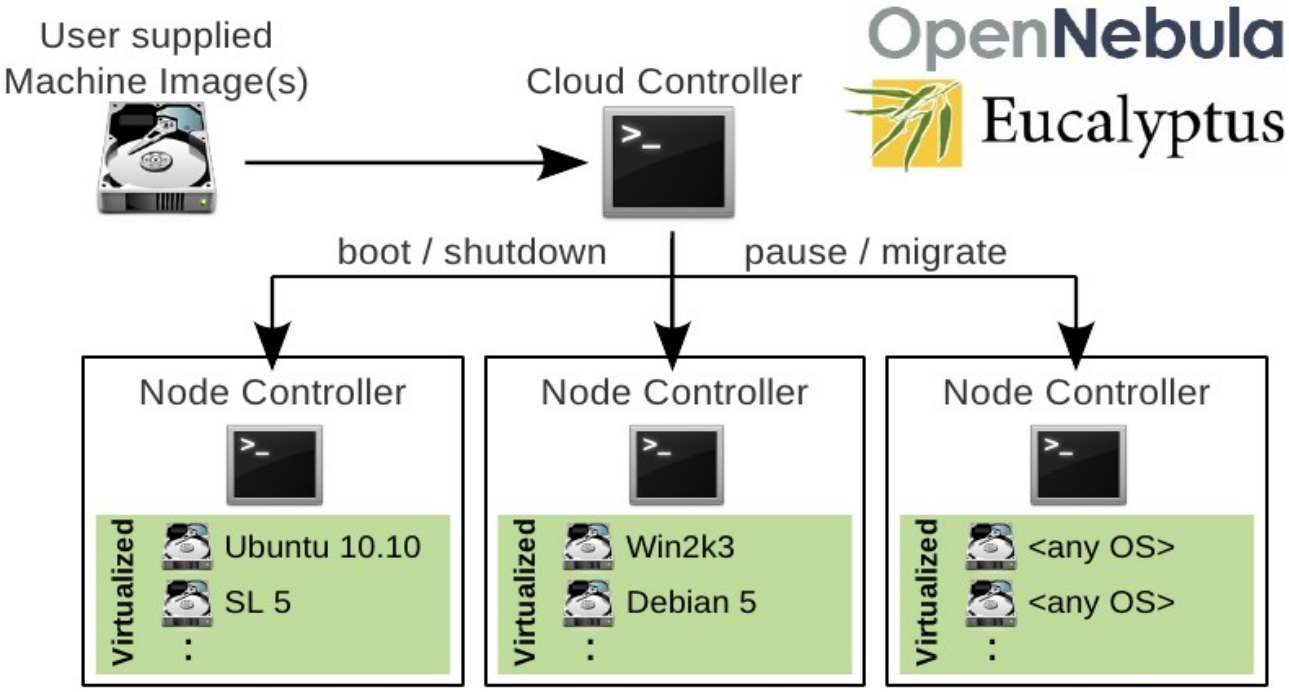
Schamatic layout of LHCONE network infrastrucure

A dedicated HEP network infrastrucure – what is the cost ?

Virtualisation

in a nutshell:

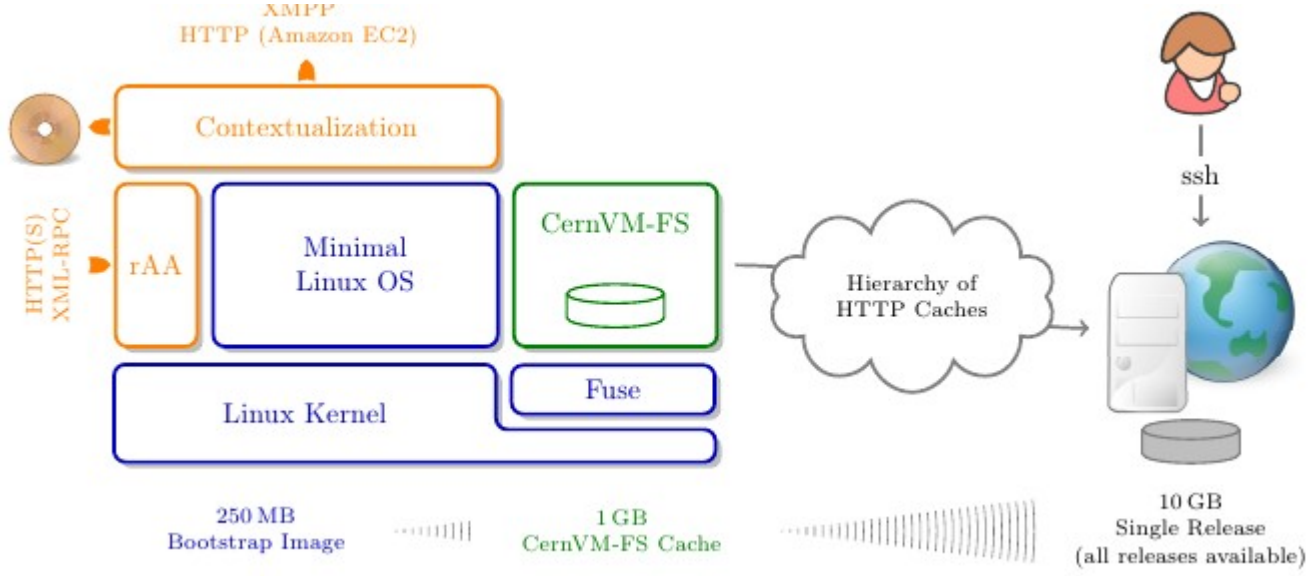
- **Clouds** offer "Infrastructure as a Service"
  - easy provision of resources "on demand"
    even by including (private) clould resources as a classical batch queue
    (e.g. ROCED project developed at EKP, KIT)
- independent of local hardware and operating system
  (Scientific Linux 5 for Grid middleware and experiment software)



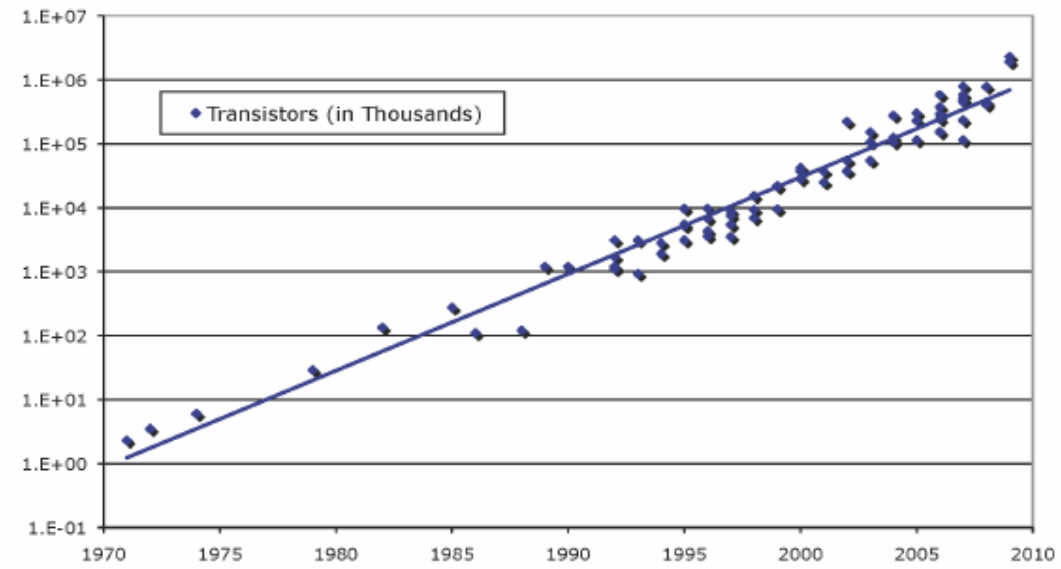
CernVM & CernVM-FS

CernVM is a virtual machine ("Virtual Software Appliance") based on Scientific Linux with CERN software environment, runs on

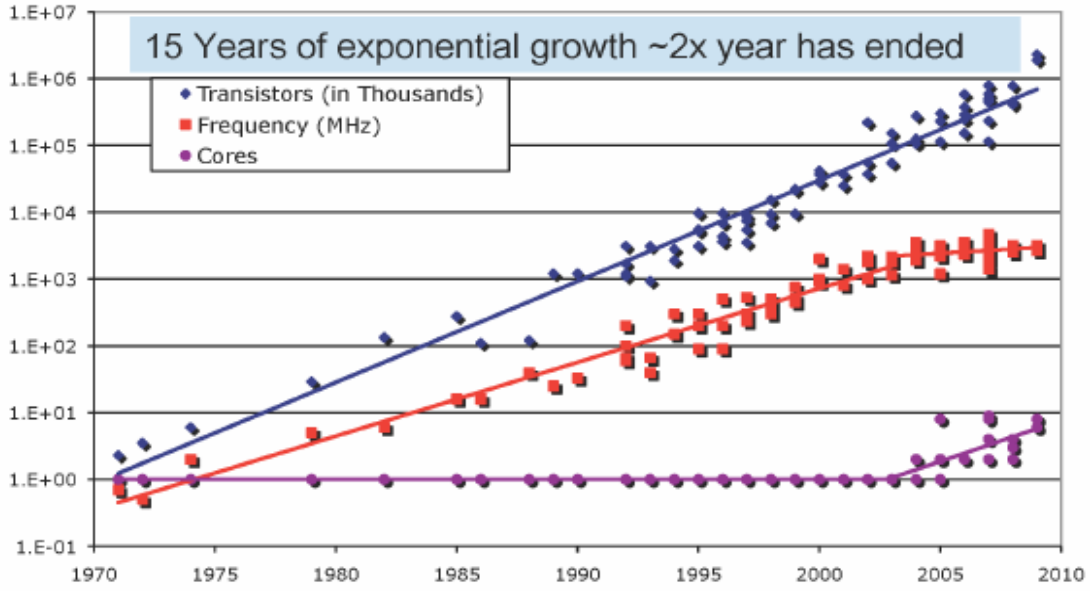
CernVM-FS is a client-server file system based on http and implemented as as user-space file system optimized for read-only access to software repositories with a performant caching mechanism. Allows a CernVM instance to efficiently access software installed remotely.



Most important challenge: Hardware evolves



# of transistors per chip keeps growing exponentially



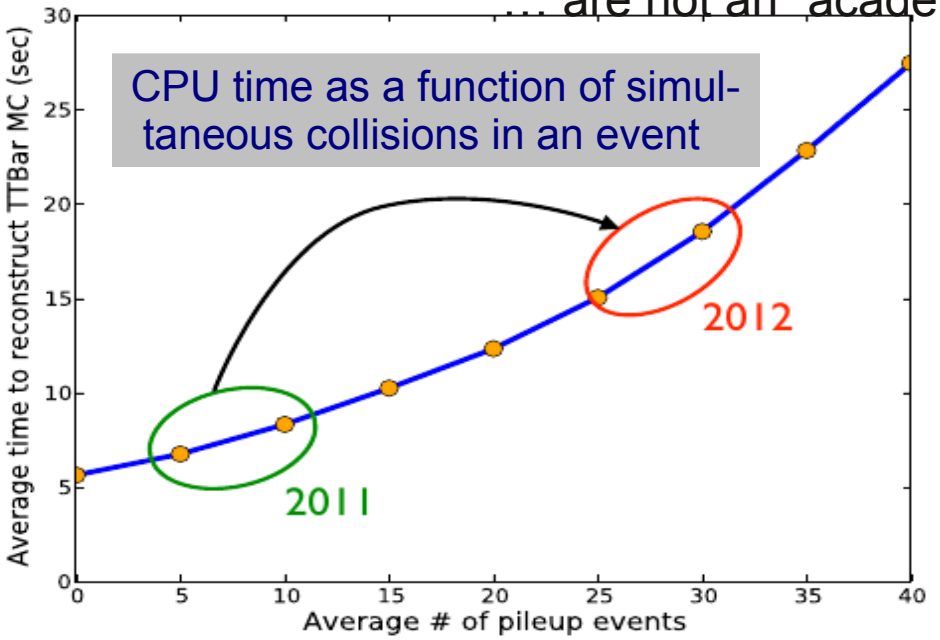
but clock rates saturate

instead:
more CPU cores on a chip !

→ **affects the way we (have to) write software**

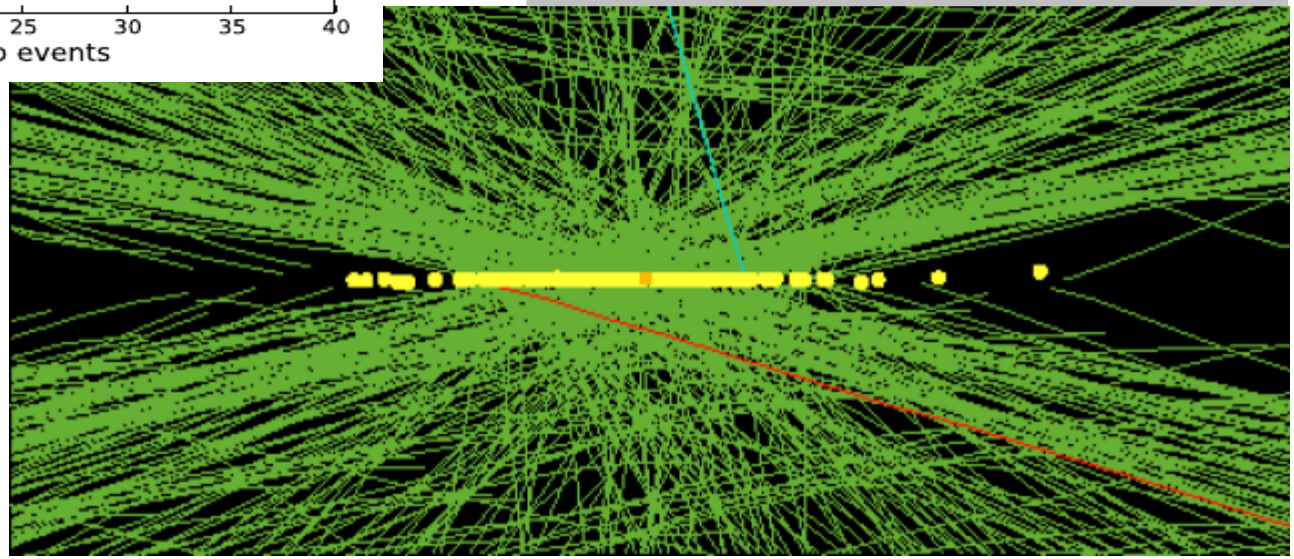
Software challenges …

… are not an “academic” problem, already hit us:



Such a **“high pile-up” environment** will become “normal” with rising LHC luminosity

Event recorded by CMS with 78 simul-taneously colliding protons in a bunch



Software Challenges

Need (to learn) new ways of writing code

– exploit vector units on CPU cores: “data oriented programming”
(to replace “object oriented programming in time-critical algorithms)
exploit auto-vectorisation in new gcc to use
Single Instrucion Multiple Data (MMX, SSE(2), AVX,...) features of modern CPUs

– memory / core remains roughly constant

– multi-threading and multi-cores
“one job per core”, i.e. “trivial parallelism” does not work any more
requires clever workflows and thread control

– use of Graphical Processing Units (GPUs) for time-critical parts
development environments get more user-friendly, will have to learn this!

Hint: *there will be a course on Parallel Software Development at the “1st KSETA topical days”* in February 2013 by Benedikt Hegner (CERN) et al.
(from whom I've stolen some of the material just shown)

Outlook

- **existing LHC Grid provides exellent performance** to particle physics
- situation was rather comfortable in the past:
  "sit and wait" for better hardware

  this will change: have to adapt to new hardware environments

  ??? Will PhD students be able to master **parallel programming** & **physics** & **detecotor service tasks** ???

  ⟶ need new, parallel-capable frame works
- **resource demands of expriments will rise:**
  - detector upgrades with more readout channels
  - higher trigger rates
  - higher luninosity and more crowded events

  will count on an evolutionary ansatz**, as we did so far**
- **new, open and standard ways for distributing computing emerged:**
  various variants of cloud computing to be integrated in computing models
- **network bandwith will continue rising,**
  decouples storage from CPU,
  enables clever, transregional data management