Big experiments computing challenges

Claudio Grandi

INFN Bologna

Computing for the HEP experiments

HEP computing has different aspects

For instance the characteristics of an accelerator-based experiment are different from those of an astro-particle experiment

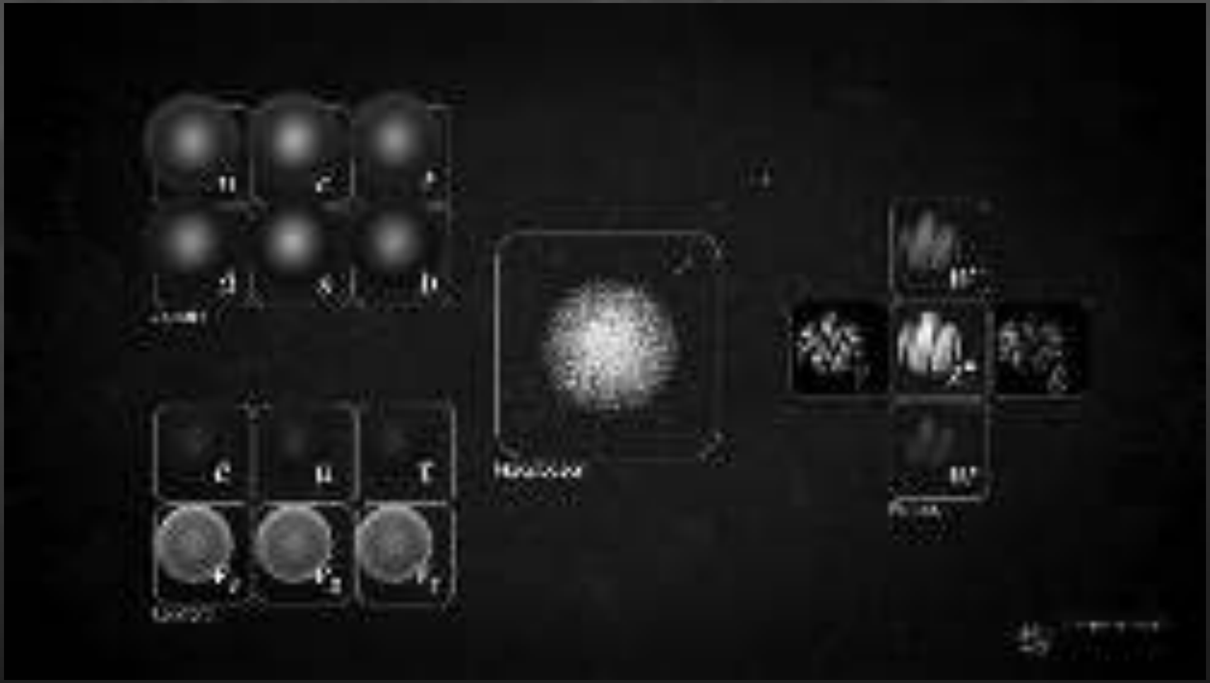
The infrastructure built by the community is tailored on the needs of LHC that is the most demanding user at the moment (but it serves all the HEP community and more)

What is HEP about?

High Energy Physics studies the fundamental constituents of matter and the forces that drive their interactions

One of the methods is to create very high energy densities

This reproduces the environmental conditions of the primordial universe



Particle accelerators

In order to create high energy densities we accelerate particles in opposite directions and make them collide one against the other

The CERN LHC accelerates protons. It has 27 km of circumference and is located in a tunnel about 100 m underground in the Geneva area

Particle detectors

Around collision points we have built particle detectors that can "see" the particle produced in the proton collision so that we can understand what happened.

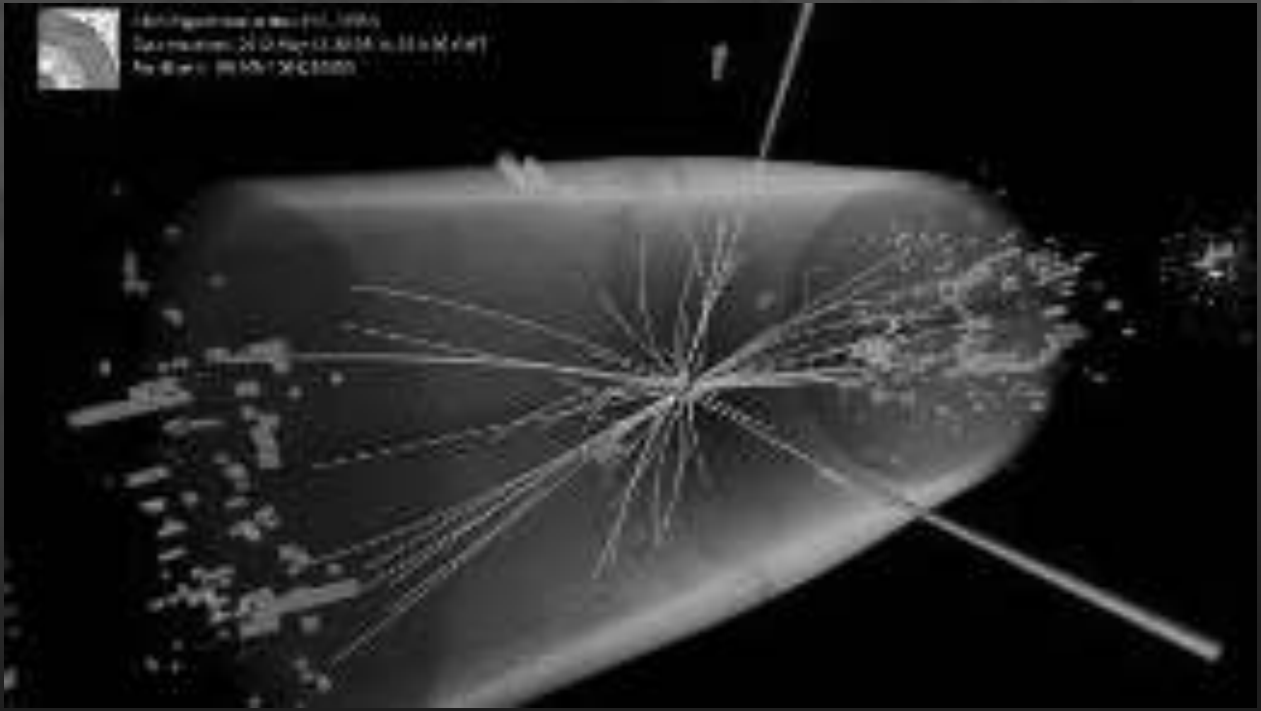
Detectors have about 100 million channels that are acquired at each collision

Collision events

We call "event" a single crossing of the proton bunches in the detector area.

For each event we reconstruct the particles produced in the collisions.

There are 40 millions crossings per second

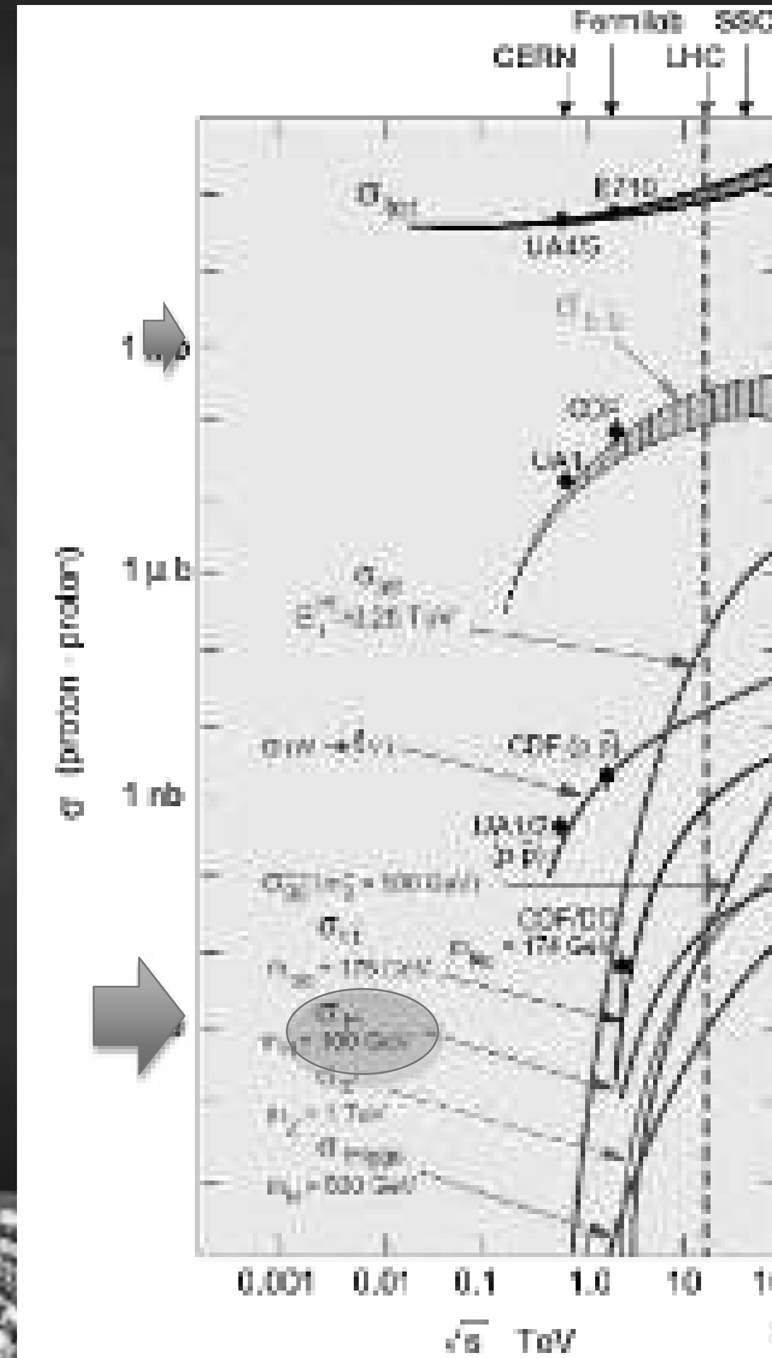


LHC Physics

The reason why in LHC we produce so many events is that experiments study rare events

- *For example the signal to noise ratio for Higgs events is ~ 10^{-13}*

Effective data reduction techniques are needed!



LHC data

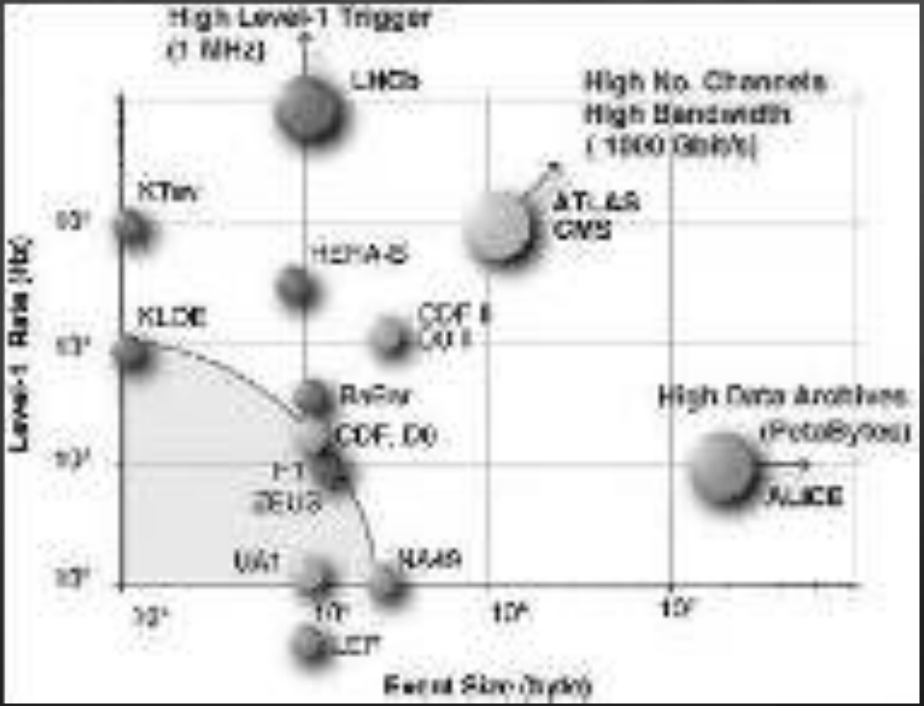
In each LHC experiment there are 40 million bunch crossings per second. Every time 100 million channels are acquired (100 MB)

➔ 40,000 EB/y (4x10^{22} Byte)

Obviously it is not affordable!

The data reduction process brings to 1000 events per second each ~ 1 MB

➔ ~10 PB/y (10^{16} Byte)



LHC Data processing

In general physicists do not like to work on RAW data coming from the detector

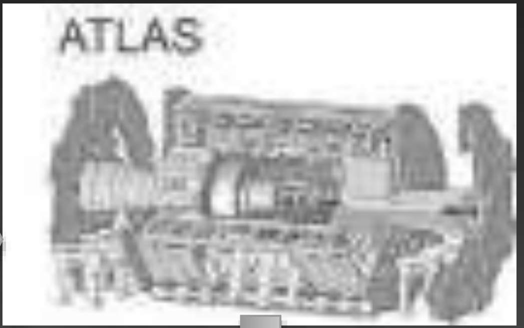
Typically they prefer to work with particles, jets, vertices, missing energy, etc…

The process that interprets RAW data in terms of physics objects is the reconstruction

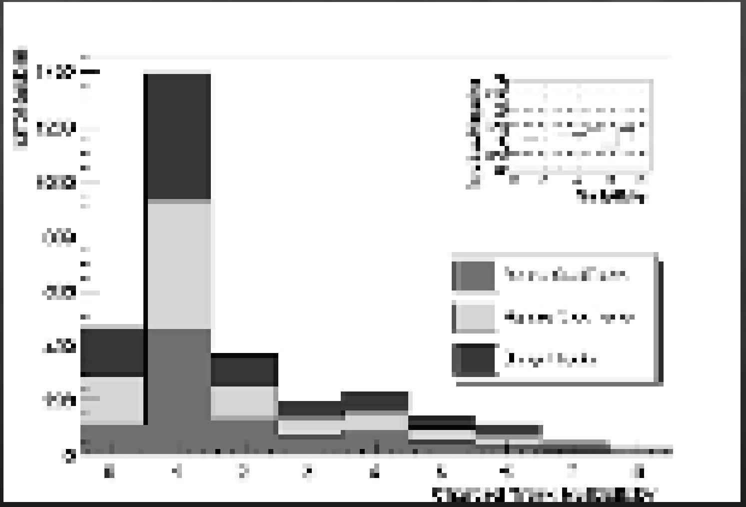
- Actually there are many reconstruction phases

Physicists do analysis on reconstructed data

LHC Real data



LHC collisions → Decay of unstable particles → Detector electronics → Trigger → Reconstruction → Analysis



LHC Simulation

Not just real data form detectors!

Since it is not possible to use analytical solutions of physic processes going from the proton interactions to the final state particles, we use simulations based on Monte Carlo techniques

Events are generated according to theoretical models and then simulated in order to reproduce the detector behaviour and then treated in the same way of the real data

The simulated data sample is 1 to 2 times the real data sample

LHC Simulated data

Theoretical model

Simulation of decays of unstable particles

Pythia, …

Simulation of interactions particle-detector

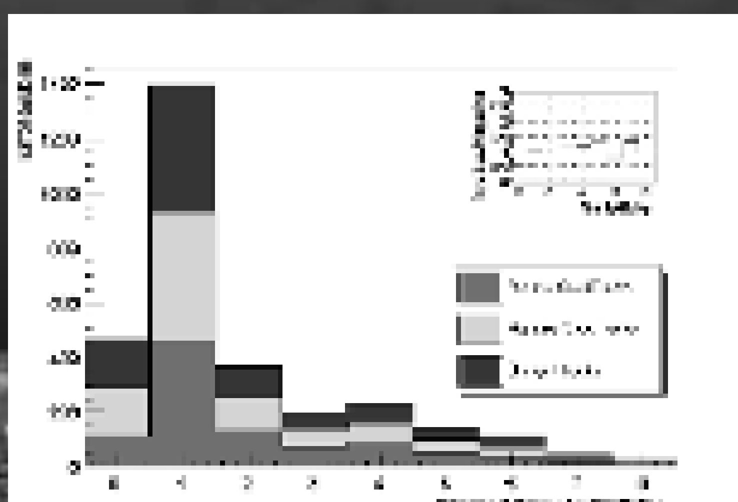
Geant4, …

Simulation of detector electronics

Trigger

Recontruction

Analysis



Computing infrastructure

Management of different kinds of data (raw, reconstructed, simulated, analysis products) and of processes (different phases of reconstruction, simulation, end-user analysis) is done on an infrastructure built by all countries participating to the LHC experiments

The project that coordinates the operations on the infrastructure is the

World-wide LHC Computing Grid (WLCG)

Units used

Storage

1 byte (B)= [0…255] = 8 bit

1 GB = 10^9 B

1 PB = 10^{15} B

1 EB = 10^{18} B

Today: Hard Disk ~ 7 TB

Network

Gb/s = 2^{30} bit/s ~ 100 MB/s

Today: sites are connected at n x 10 Gb/s to n x 100 Gb/s

CPU

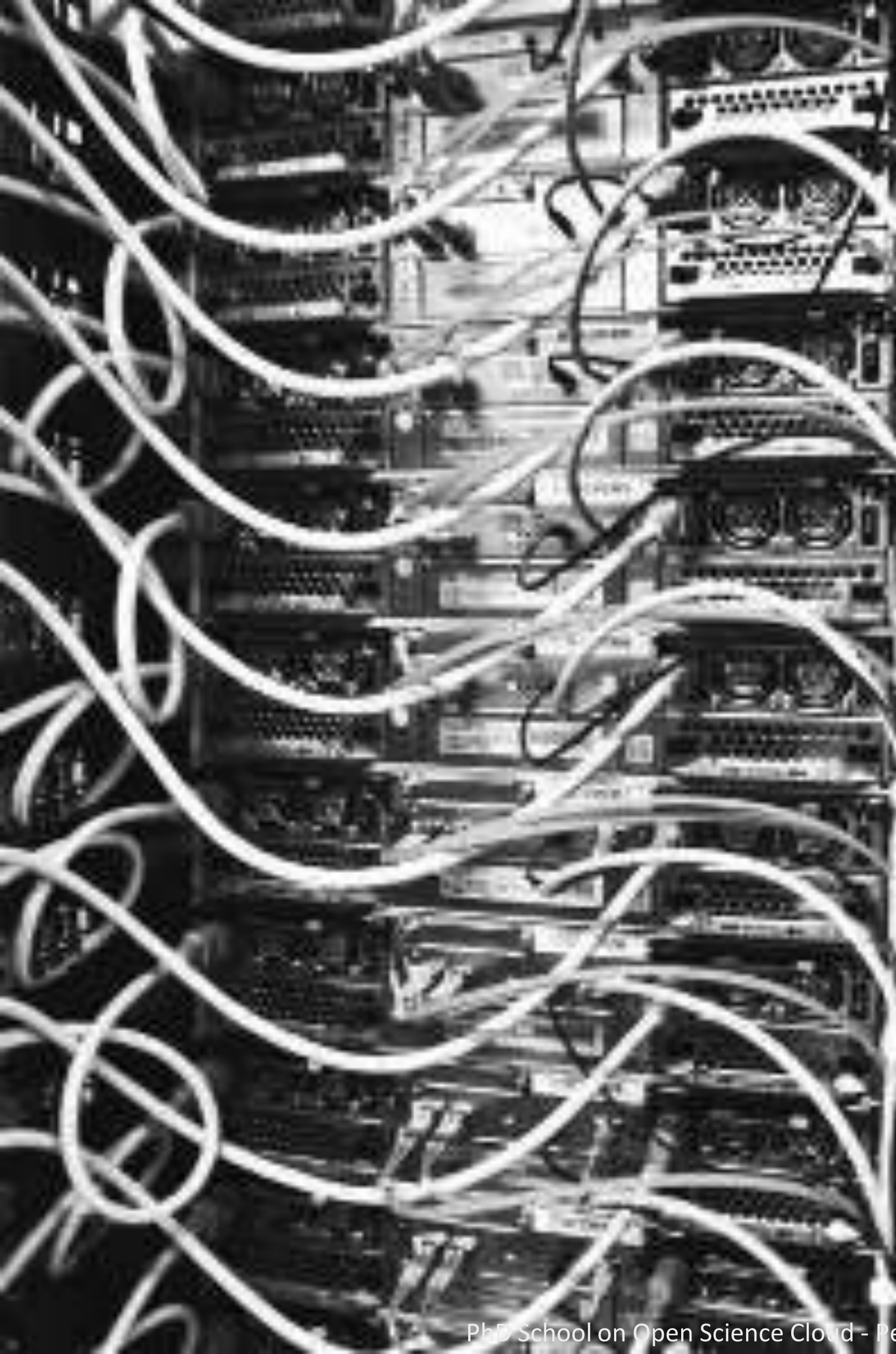
Using a unit specific for HEP: HepSpec06 (HS06)

Today:

1 computing core ~> 10 HS06

1 CPU (~12 cores) ~> 100 HS06

Data flow



Numbers from the movie (2013)

600 million collisions every second

Only 1 in a million collisions is of interest

Fast electronic preselection passes 1 out of 10 000 events and stores them on computer memory

100 GB/s transferred to the experiment computing farm

15 000 processor cores select 1 out of 100 of the remaining events

CERN Data Centre (Tier 0)

~ 100 000

- ~~*73.000*~~ *processor cores*
- *Data aggregation and initial data reconstruction*
- *copy to long-term tape storage and distribute to other data centres*

11 Tier 1 centres

- *Permanent storage, re-processing, analysis*

140 Tier 2 centres

- *Simulation, ent-useer analysis*

> 2 multicore

~~*1,5*~~ *million jobs running every day*

25

~~*10*~~ *GB/s global transfer rate*

…more numbers

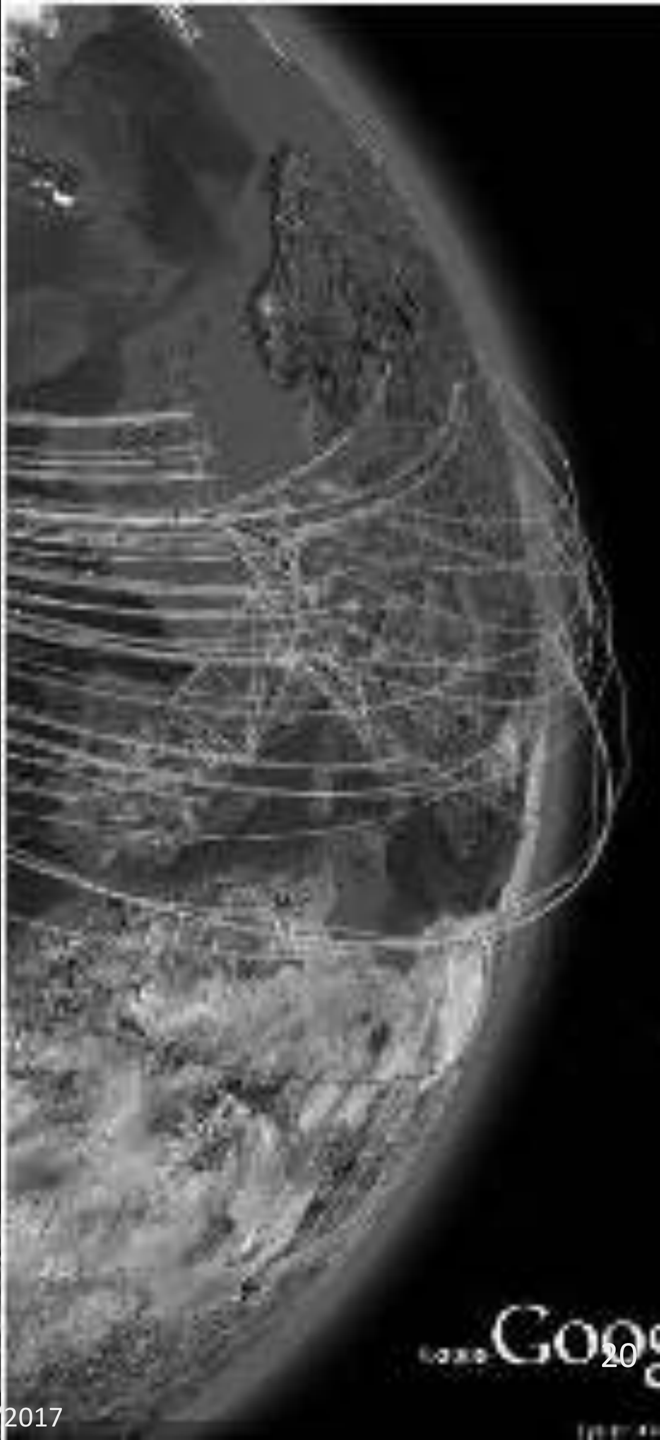
Global resources for 2017 are:

- 5,200,000 HS06 (~500.000 processor cores)
- 395.000 TB disk
- 590.000 TB tape
- Dedicated network connections (from multiples of 10 Gb/s to multiples of 100 Gb/s)

…and more available in collaborating institutes

More than 180 data centres in over 35 countries

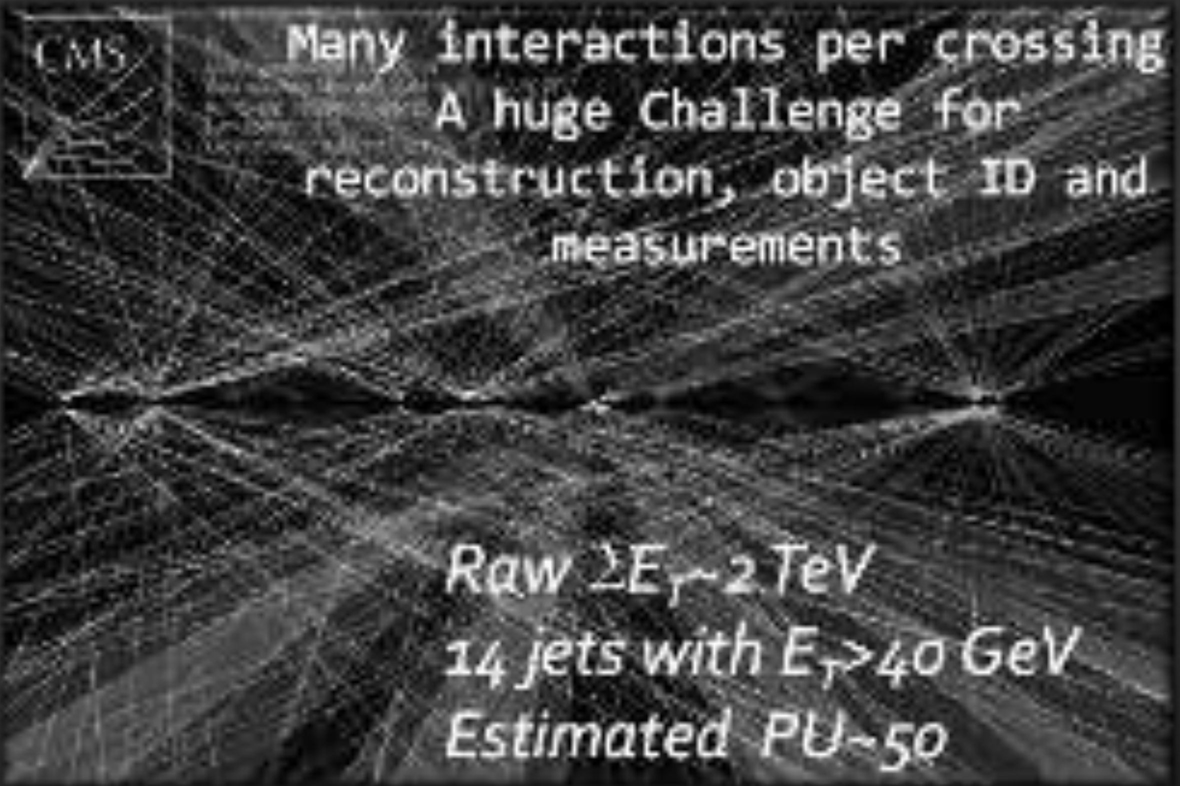
More than 8000 analysts all over the world



Pile-up

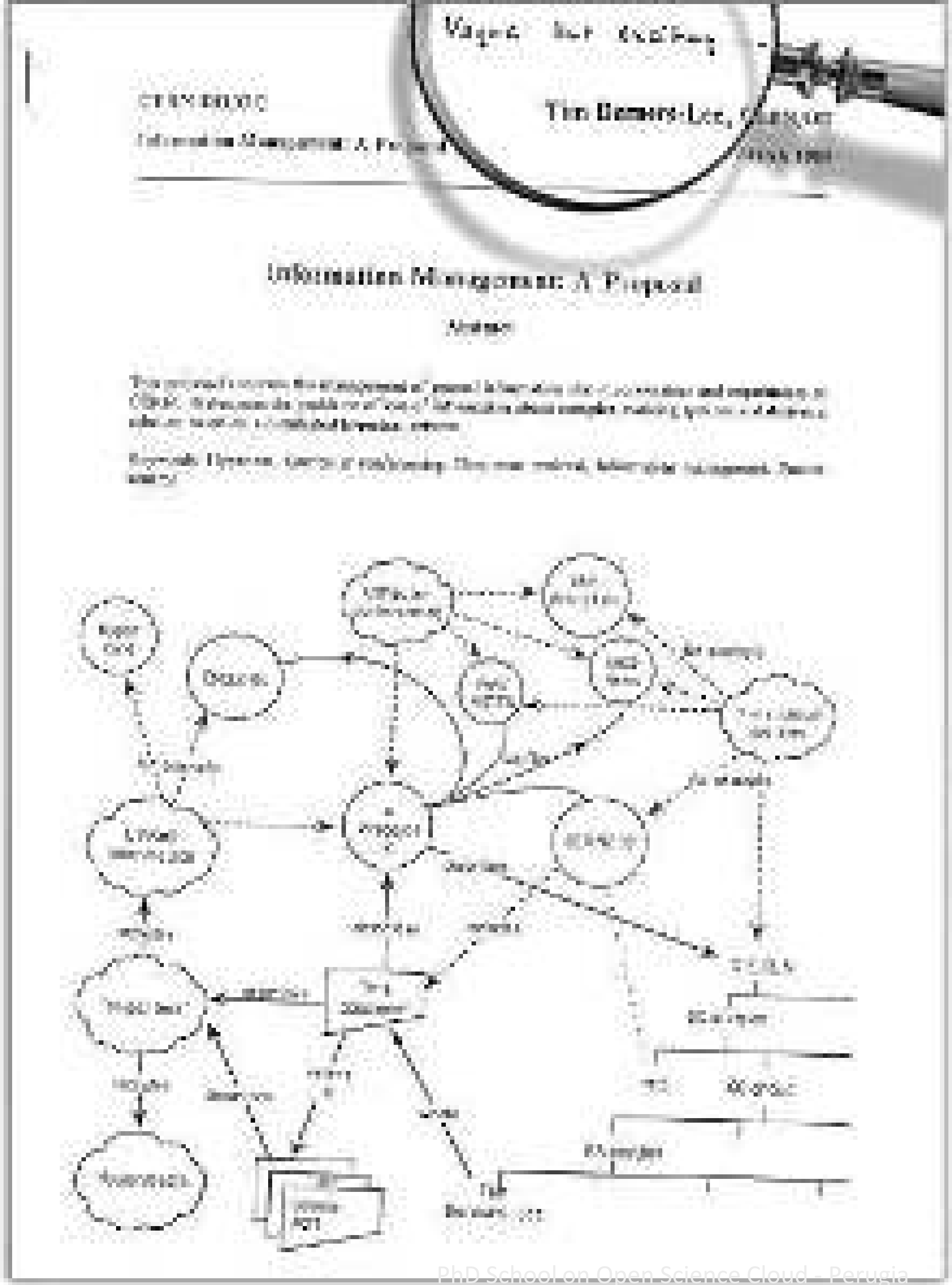
If you're wondering why a bunch crossing rate of 40 MHz produces 600 collisions per second:

- *Every bunch crossing (event) there are on average 15 p-p collisions (AKA pileup)*



Pileup is increasing to 50 and eventually to more than 150 in HL-LHC

How?



WWW

In 1989 CERN had needs that were not addressed by existing tools

Tim Berners-Lee proposed a mechanism for information sharing in the scientific community: the World Wide Web

Today WWW is available to the entire society for free!

Declaration

The following CERN software is hereby put into the public domain:

- W3 basic ("line-mode") client
- W3 basic server
- W3 library of common code.

CERN's intention in this is to further compatibility, common practices, and standards in networking and computer supported collaboration. This does not constitute a precedent to be applied to any other CERN copyright software.

CERN relinquishes all intellectual property rights to this code, both source and binary form and permission is granted for anyone to use, duplicate, modify and redistribute it.

CERN provides absolutely NO WARRANTY OF ANY KIND with respect to this software. The entire risk as to the quality and performance of this software is with the user. IN NO EVENT WILL CERN BE LIABLE TO ANYONE FOR ANY DAMAGES ARISING OUT THE USE OF THIS SOFTWARE, INCLUDING, WITHOUT LIMITATION, DAMAGES RESULTING FROM LOST DATA OR LOST PROFITS, OR FOR ANY SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES.

Geneva, 30 April 1993

W. Hoogland
Director of Research

H. Weber
Director of Administration

Copie certifiée conforme

faite à Genève le

The first picture on the web (1992)

Collider

I gave you a golden ring to show you my love
You went to stick it in a printed circuit
To fix a voltage leak in your collector
You plug my feelings into your detector
You never spend your nights with me
You don't go out with other girls either
You only love your collider
Your collider.

(CERN Hardronic Festival – 1990)

The first web-cam (1993)

Not verified…

Computer Laboratory,
University of Cambridge

From Web to Grid

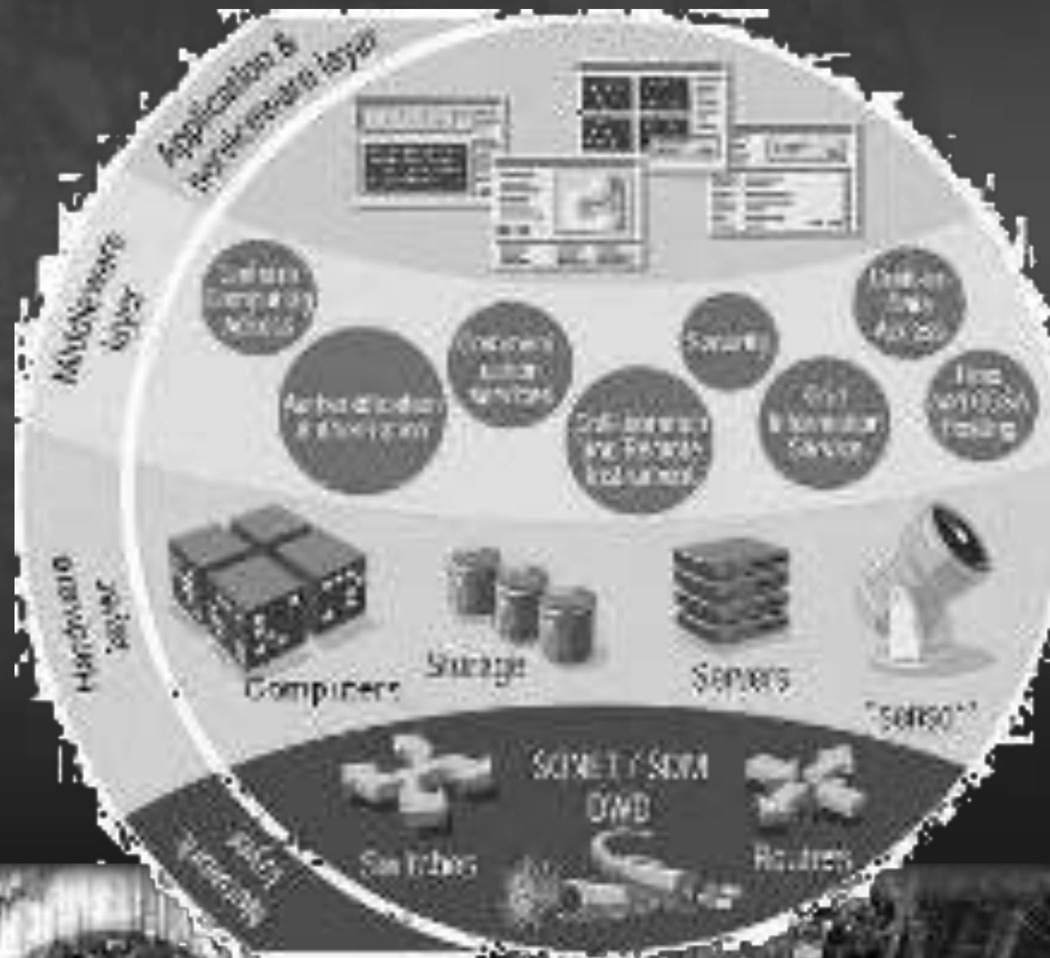
In the years 2000s the LHC community had to address the problem of how to manage the data that the experiments would produce

They started from an idea of a group of American computing scientists: the Computing Grid

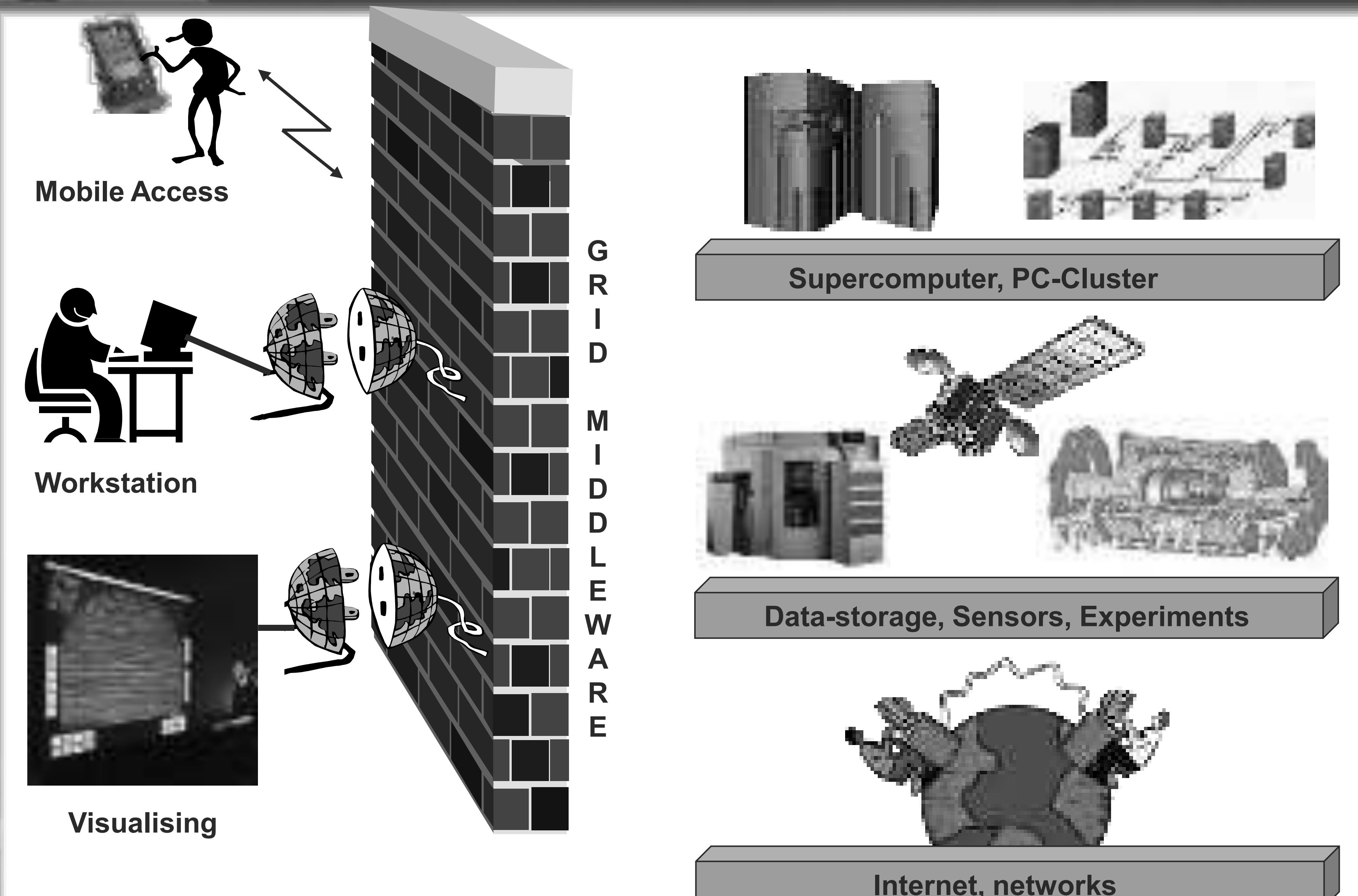
Computing resources are treated in the same way of the electrical power:

A computer is plugged to the network and gets what needed wothout knowing where it comes from

The middleware is a software layer between resources and users



The Grid metaphore



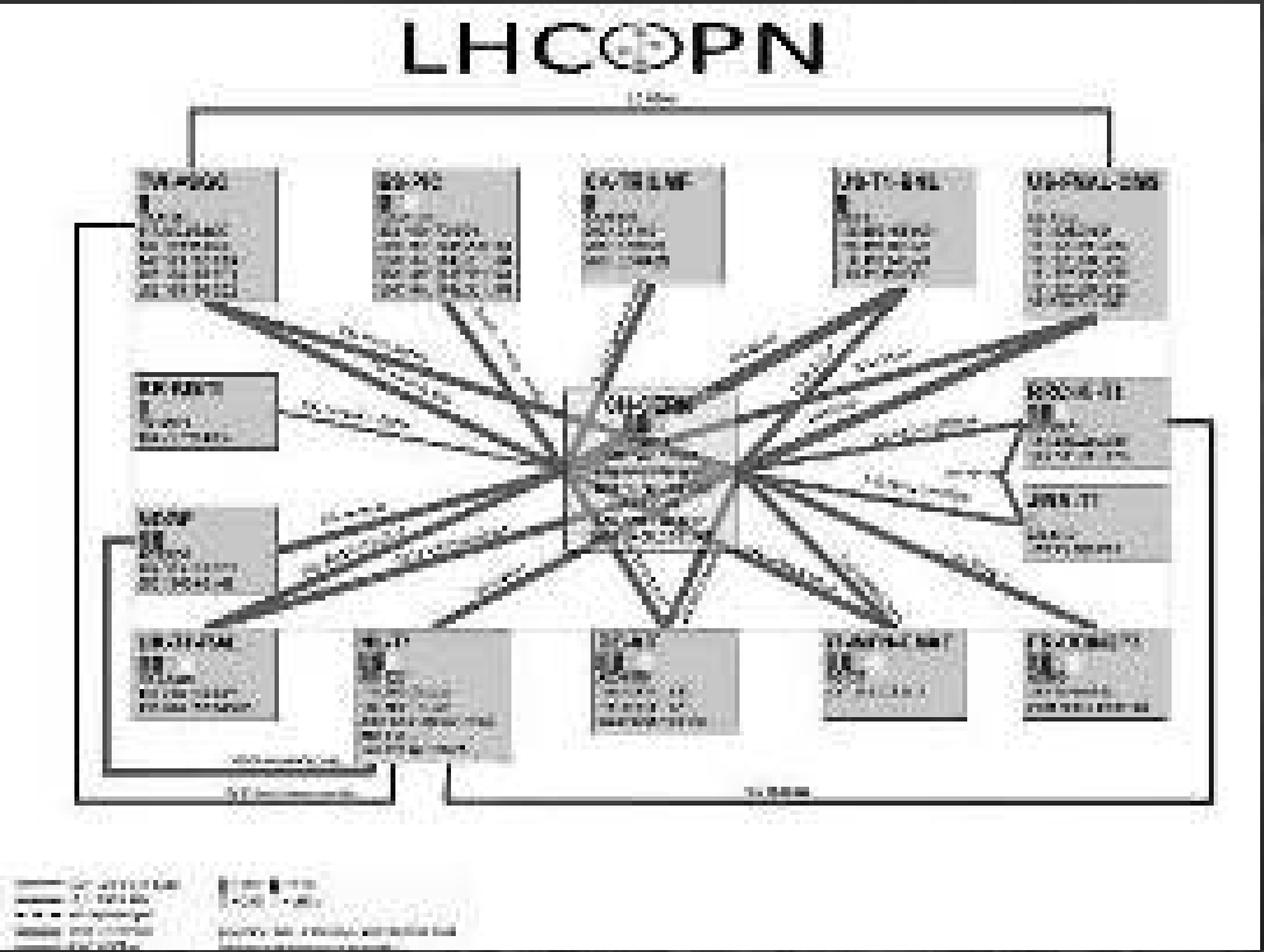
A distributed system

Advantages of a distributed system (w.r.t. a unique data centre)

- Avoid single point of failure
- Have access to local funding otherwise not provided by member states
- Investment on manpower available in different countries
- Build an adaptable system able to integrate external resources that are made available

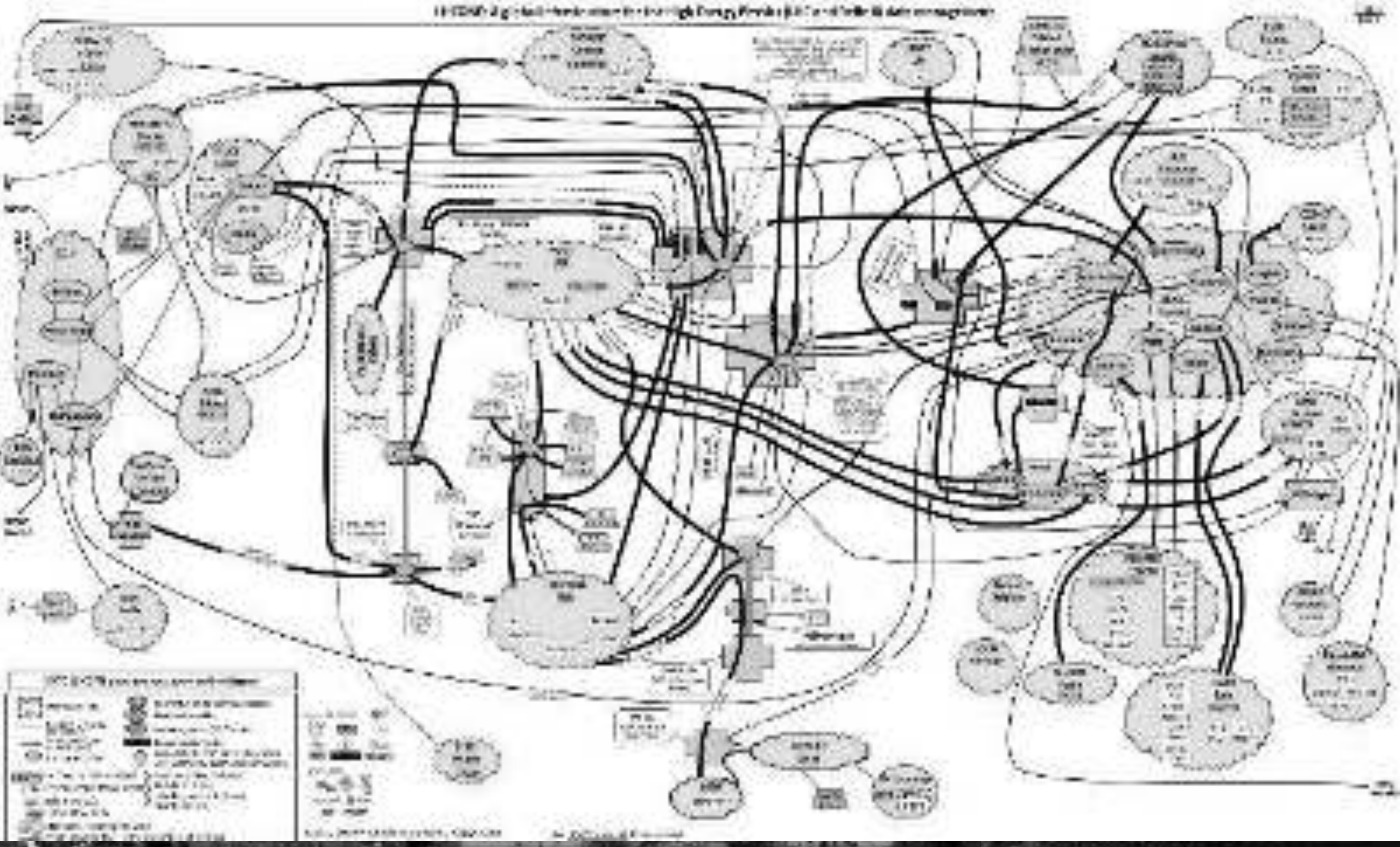
Only a few technical details…

The network - LHCOPN



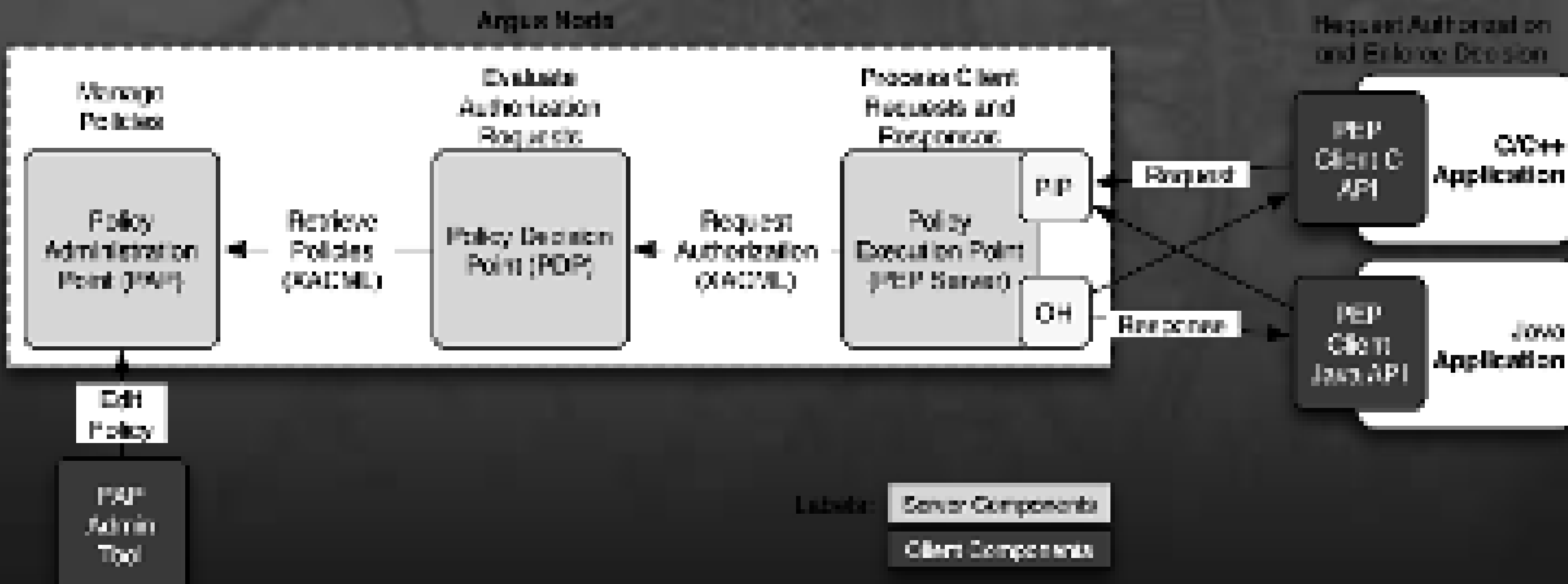
The network technology evolved significantly, offering adequate performance to support the distributed computing model

The network - LHCONE



Grid Security management

- Authentication based on x.509 certificates
- Authorization based on *attribute certificates* (VOMS)
- *Policy management* system (ARGUS)

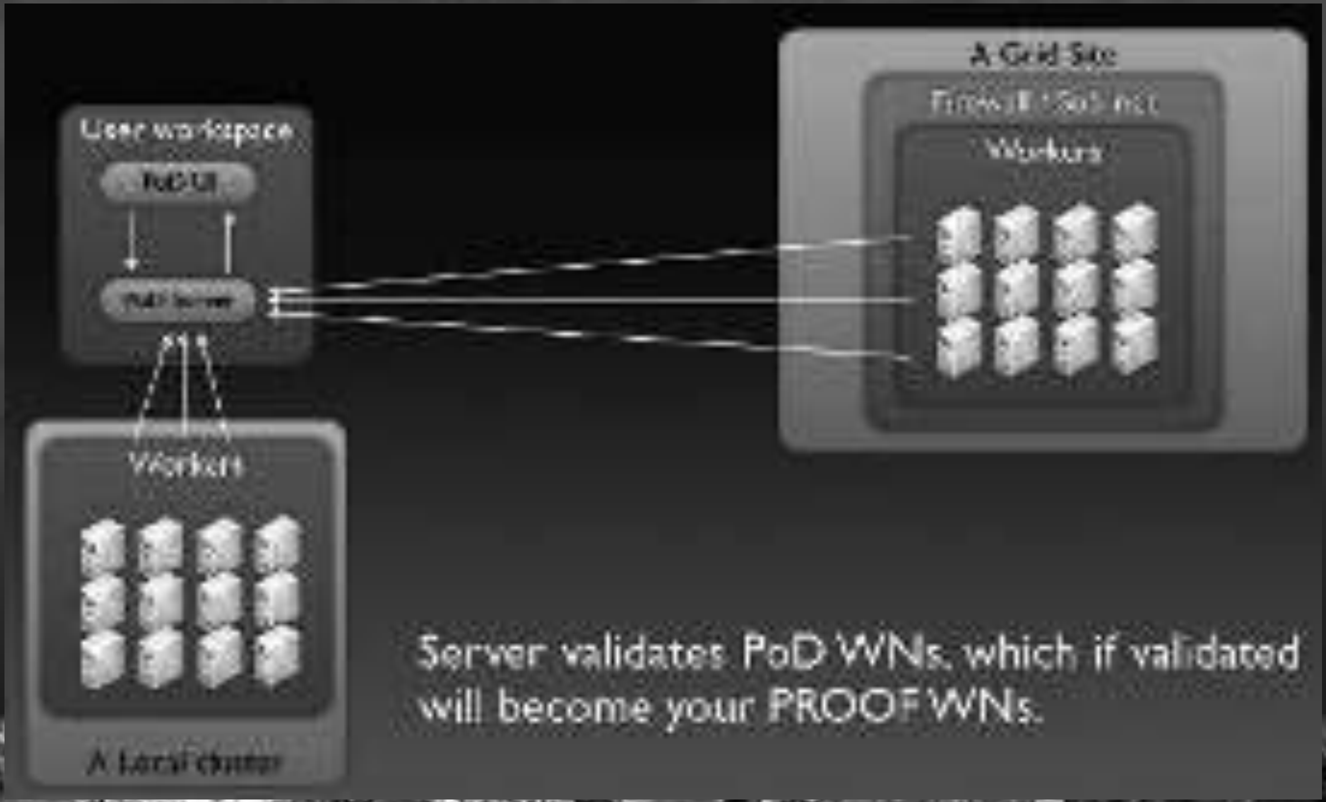


Grid Computing management

Access is based on batch jobs: asynchronous execution

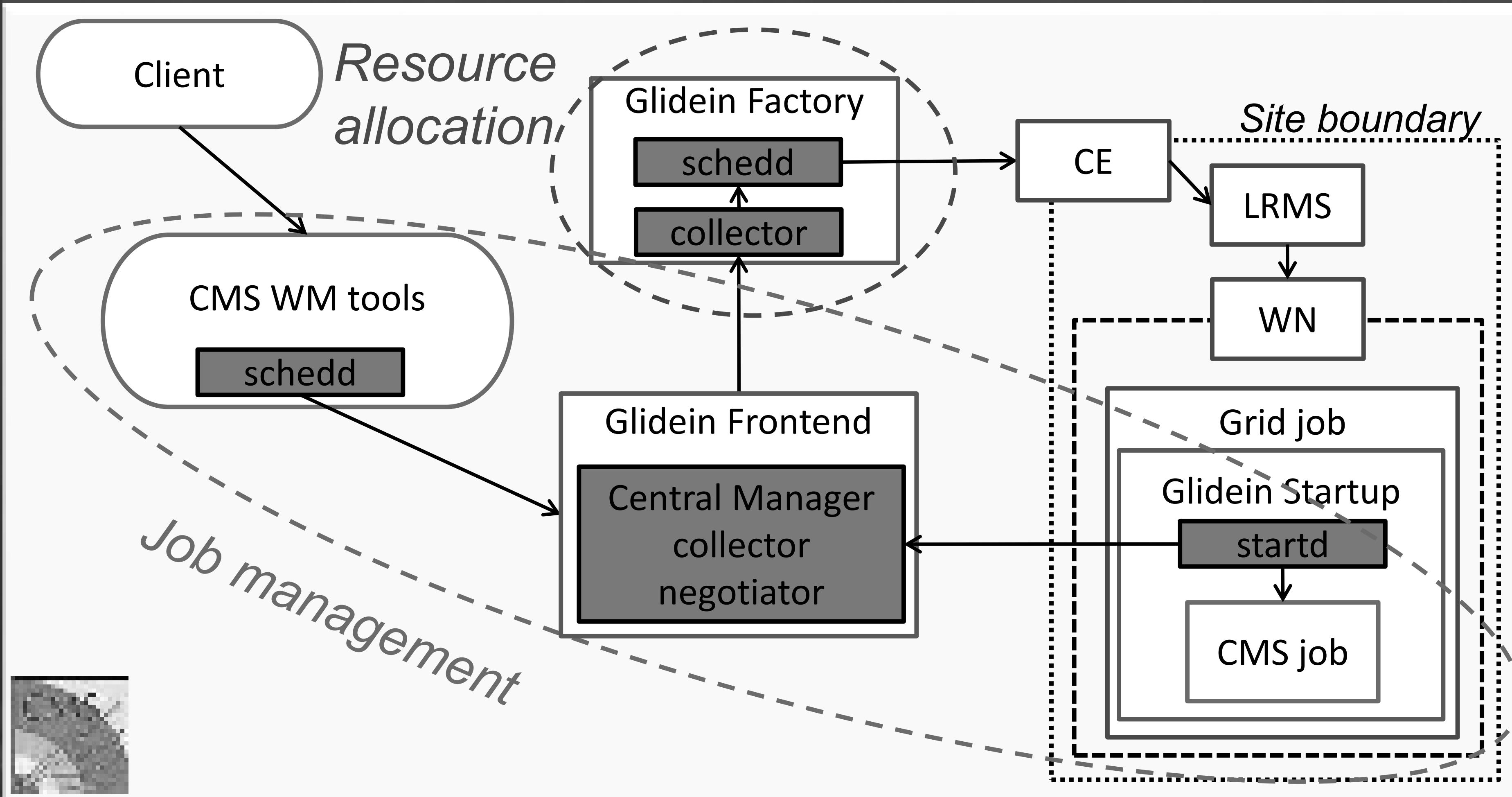
Dedicated interfaces allow to manage remote submissions as if local

Interactive processing is limited and based on local resources or on systems able to manage part of the load in batch mode (e.g. PoD)



The "pilot" model

Separation of resource allocation and job management



Grid Data management

Heavily relying on tape libraries for persistent data storage

- Accessible in a transparent way (nearline)
- Dedicated interfaces to uniformly manage data on disk and on tape

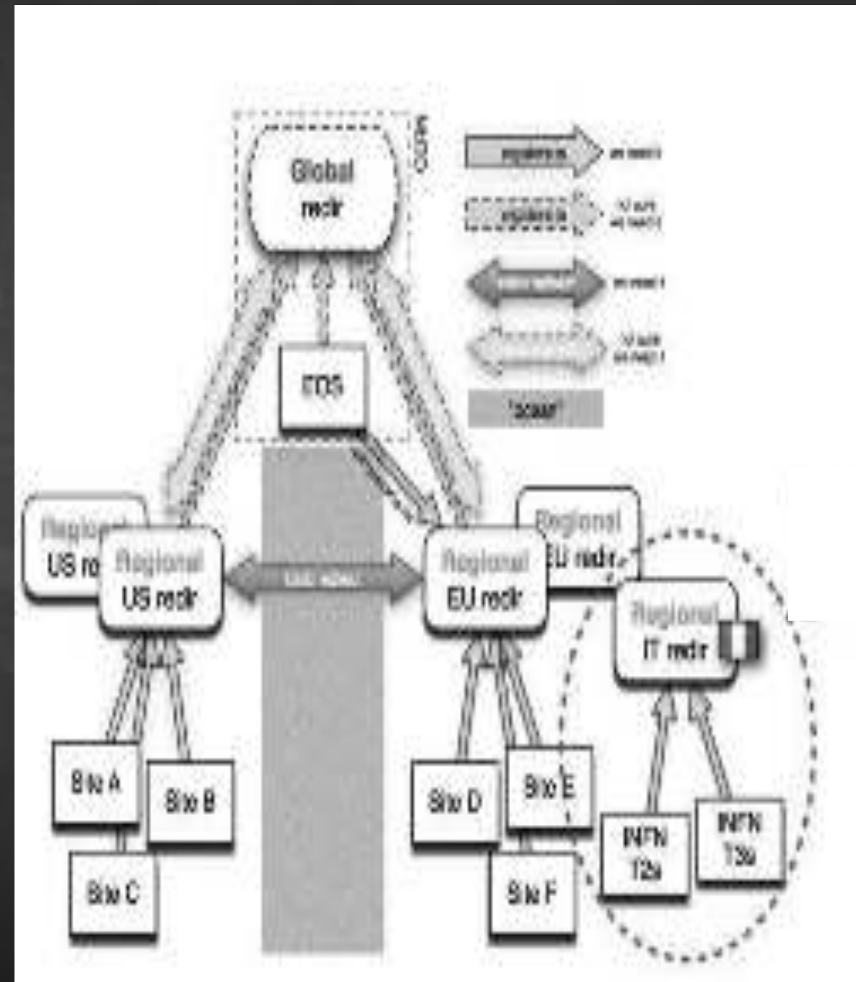
Tools to manage the transfer of large amounts of data

Local access to data by jobs but today network performances allow transparent remote access on the Wide Area Network

- Storage Federations

Storage Federations

- Starts from the possibility to have remote data access
- Clients always ask the closest location for files
- If the file is not available, the request is forwarded to a hierarchy of redirectors until it is satisfied (or fails globally)
- In production for xrootd and http



Let's see how it works…

Grid: an example of collaboration

Even though the HEP community has been dominant, the Grid has been thought and build for the whole scientific community

Projects as the European Grid Initiative (EGI), to which INFN participates, and the Open Science Grid (OSG) in the US provide computing resources to many scientific communities, and more.

Involvement also in the industrial world.

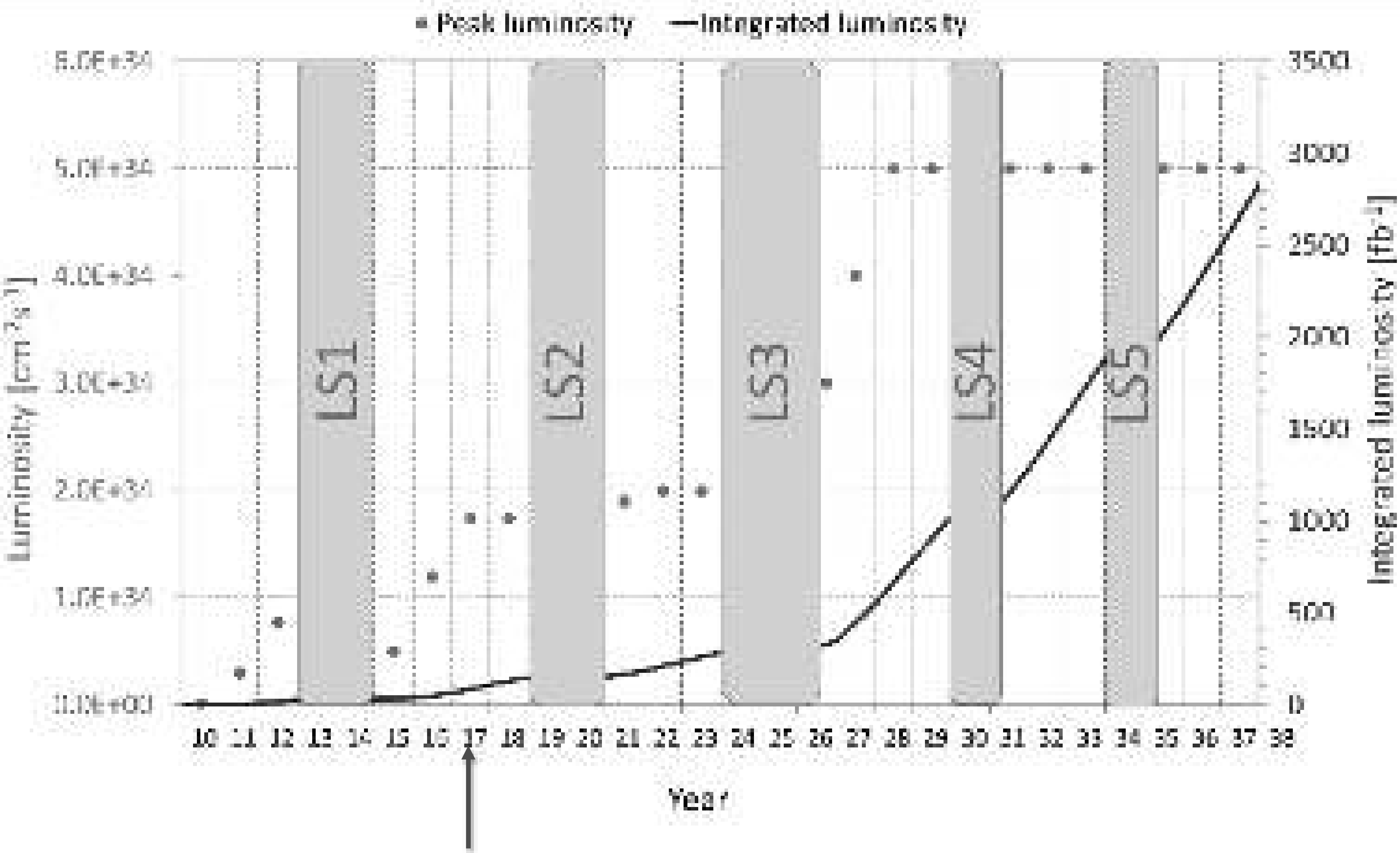
Was that enough?

July 4th 2012

[credits: D.Bonacorsi]

What about the years to come?

LHC roadmap



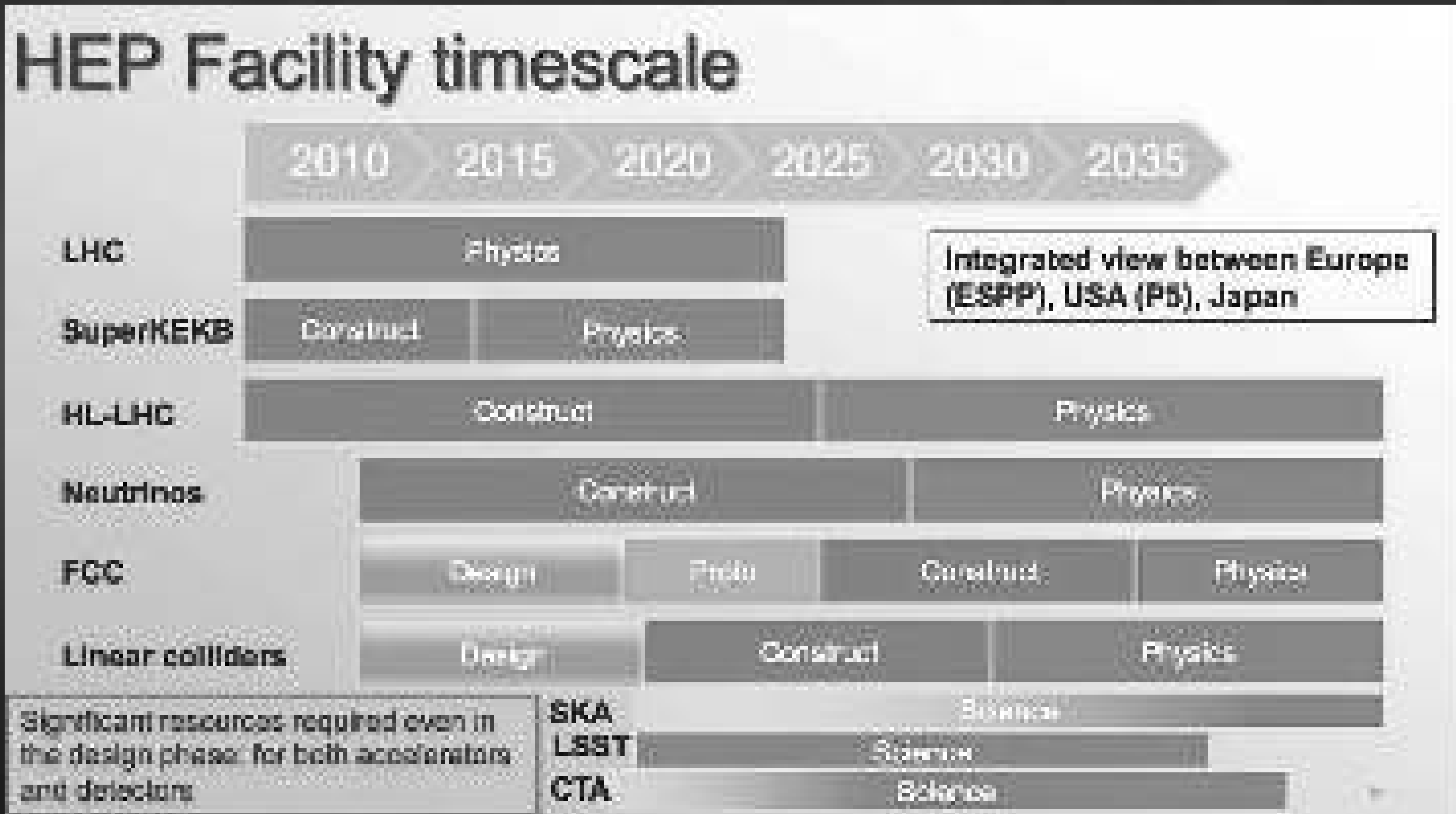
Peak luminosity
Integrated luminosity
6.0E+34
5.0E+34
4.0E+34
3.0E+34
2.0E+34
1.0E+34
0.0E+00
Luminosity [cm⁻²s⁻¹]
3500
3000
2500
2000
1500
1000
500
0
Integrated luminosity [fb⁻¹]
LS1
LS2
LS3
LS4
LS5
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Year

Resource requests for the future

Significant increase in experiments' requests in the coming years

…but the buzz-word is "flat-budget"!

Not only LHC...


HEP Facility timescale
2010
2015
2020
2025
2030
2035
LHC
Physics
SuperKEKB
Construct
Physics
HL-LHC
Construct
Physics
Neutrinos
Construct
Physics
FCC
Design
Proto
Construct
Physics
Linear colliders
Design
Construct
Physics
Integrated view between Europe (ESPP), USA (P5), Japan
Significant resources required even in the design phase for both accelerators and detectors
SKA
Science
LSST
Science
CTA
Science

Foreseen evolution – LHC Run 3

ATLAS and CMS

- Trigger rate is constant
- 50% increase in pile-up and luminosity → integrated luminosity doubles

ALICE

- DAQ rate in 50 kHz → 1 Tb/s…
- …but data reduction of a factor of 20 on the O^2 farm

LHCb

- Software trigger only (30 MHz) → 2-5 GB/s to offline

In addition the CTA (and SKA) experiments starts!

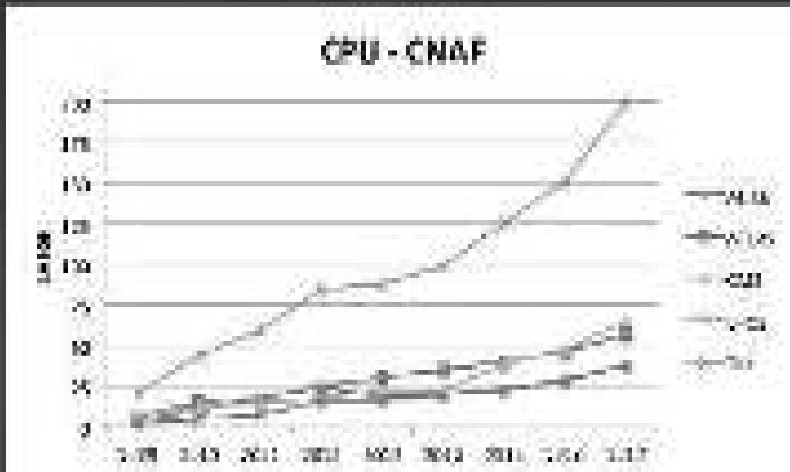
Italian resources in 2017

Let's take CNAF, the Italian Tier-1, as an example to understand what changes:

	CPU (kHS06)	Disk (PB)	Tape (PB)
All WLCG	5200	340	590
INFN Tier-1 & 2	520	38	57
% INFN	10	11	10

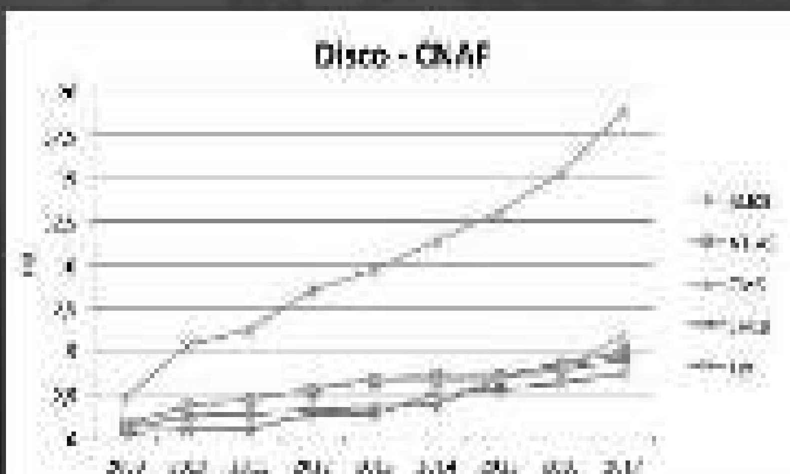
From: https://wlcg-rebus.cern.ch/

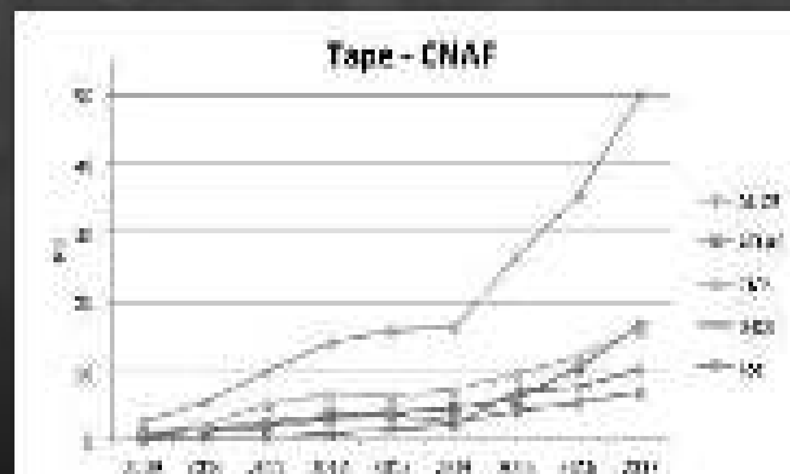
CNAF evolution - LHC Run 1 & 2



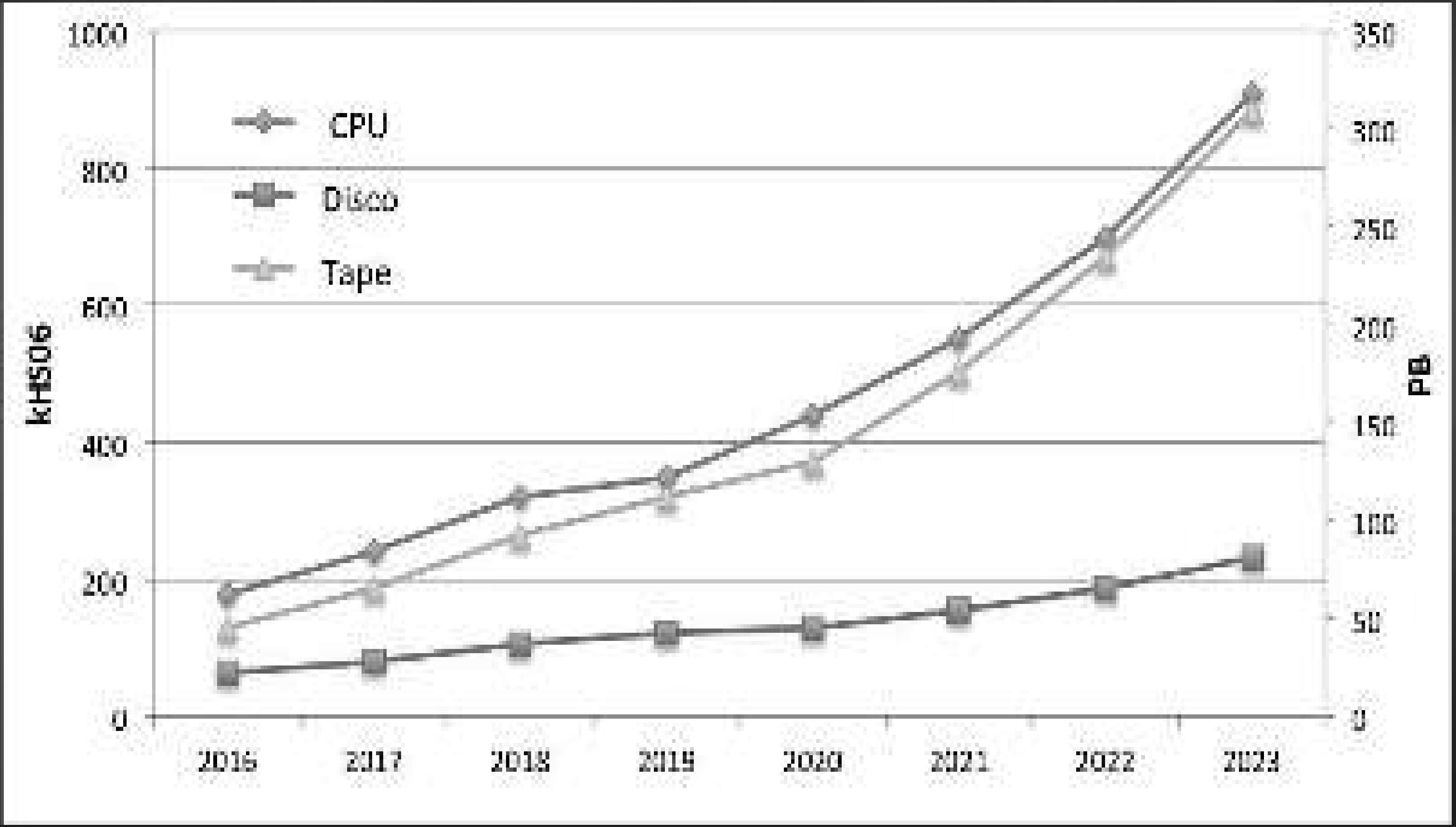
Run2 is ok with the flat budget hypothesis:

- *CPU + 20 - 30%*
- *Disk + 15 - 25%*
- *Tape + 30% - 60%*





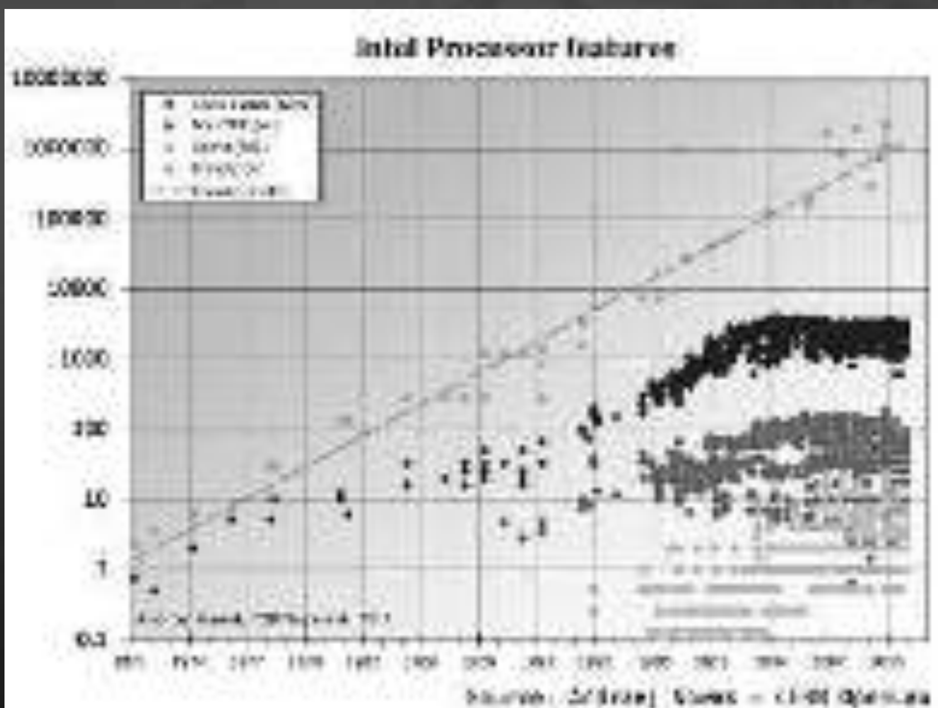
CNAF evolution up to LHC Run 3

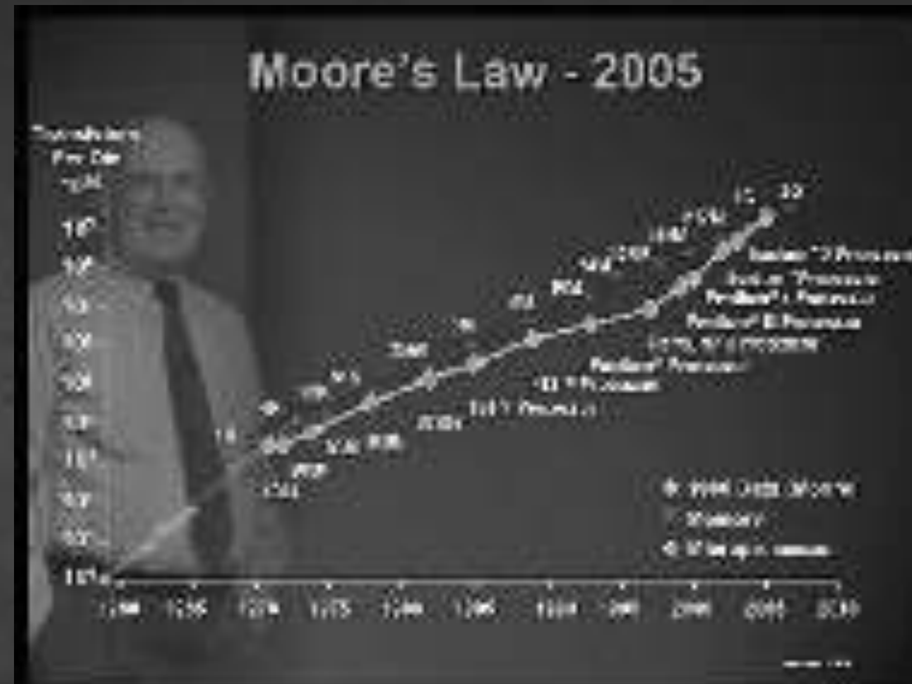


Does the technological evolution help?

CPU power

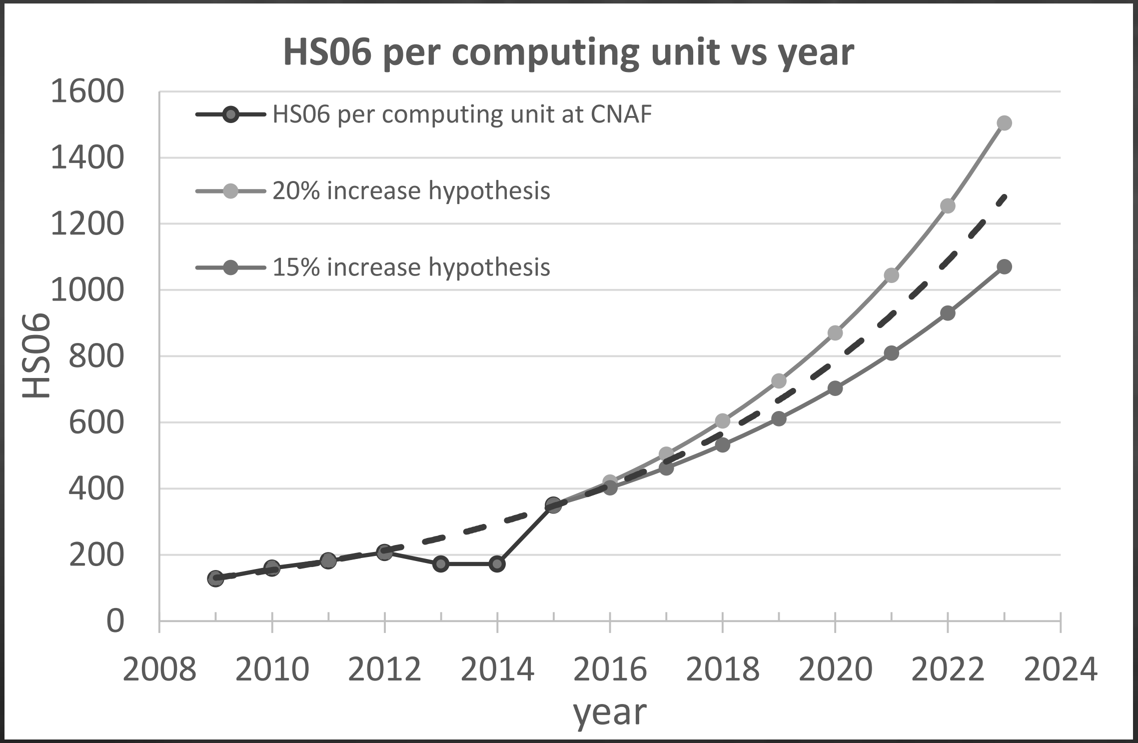
Moore's law (CPU performance doubles every 18 months at the same cost) does not hold any more





We may reasonably expect a 20% increase per year but we need to cope with multi-core systems

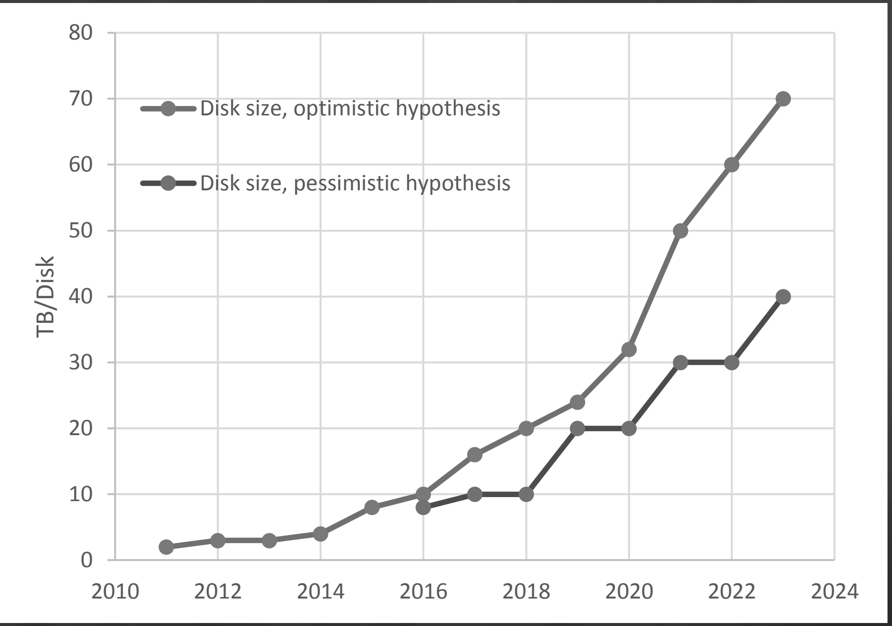
CPU power



Starting from the actual power of the nodes bought by CNAF in 2009-2015 we estimate an increase between 15 and 20%

Disk

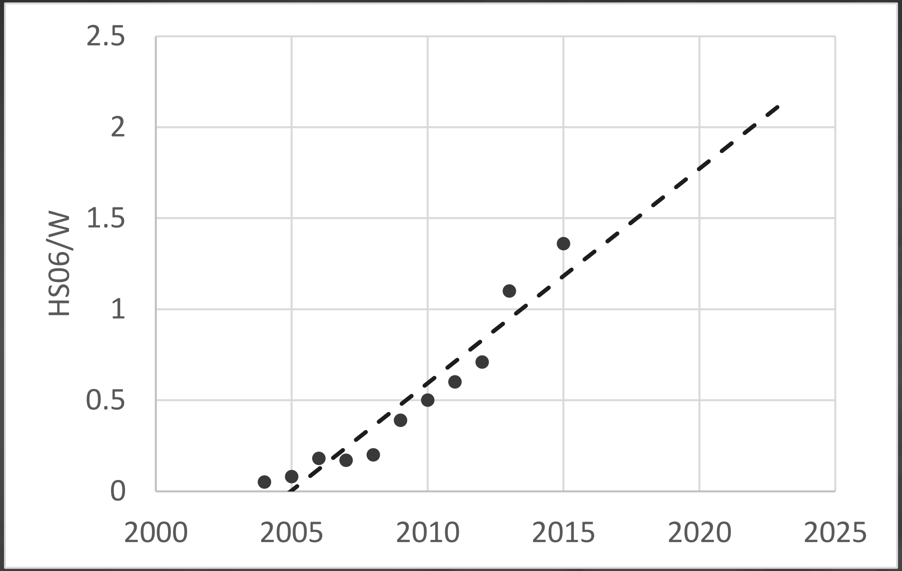
It is safe to assume that disk size in 2023 will be around 40 TB



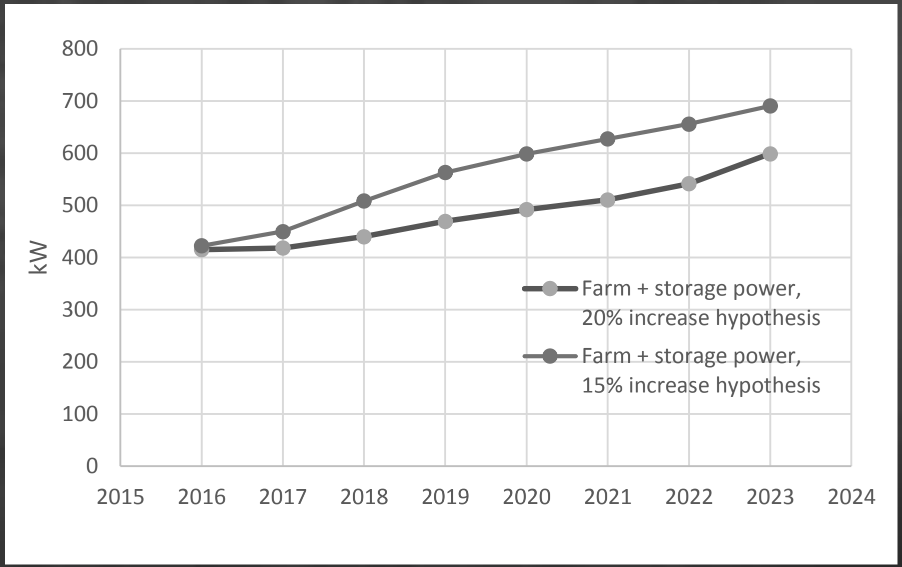
Extrapolation is more difficult for disk because there are technology changes foreseen

The number of disks may not need to increase

Electrical power



Disk power consumption does not depend on size in first approx.



CPU power to electrical power ratio increasing linearly. In 2023 foreseen 2 HS06/W
→ Low power architectures?

Total power (including services) in 2023 is foreseen to be ~ 1 MW

Costs

- *Provisioning of CPU, disk and tape*
- *Electrical power for IT*
- *Electrical power for cooling*

  ~60% of power for IT at CNAF (PUE 1.5 to 1.7 depending on the season)
- *Infrastructure maintenance*

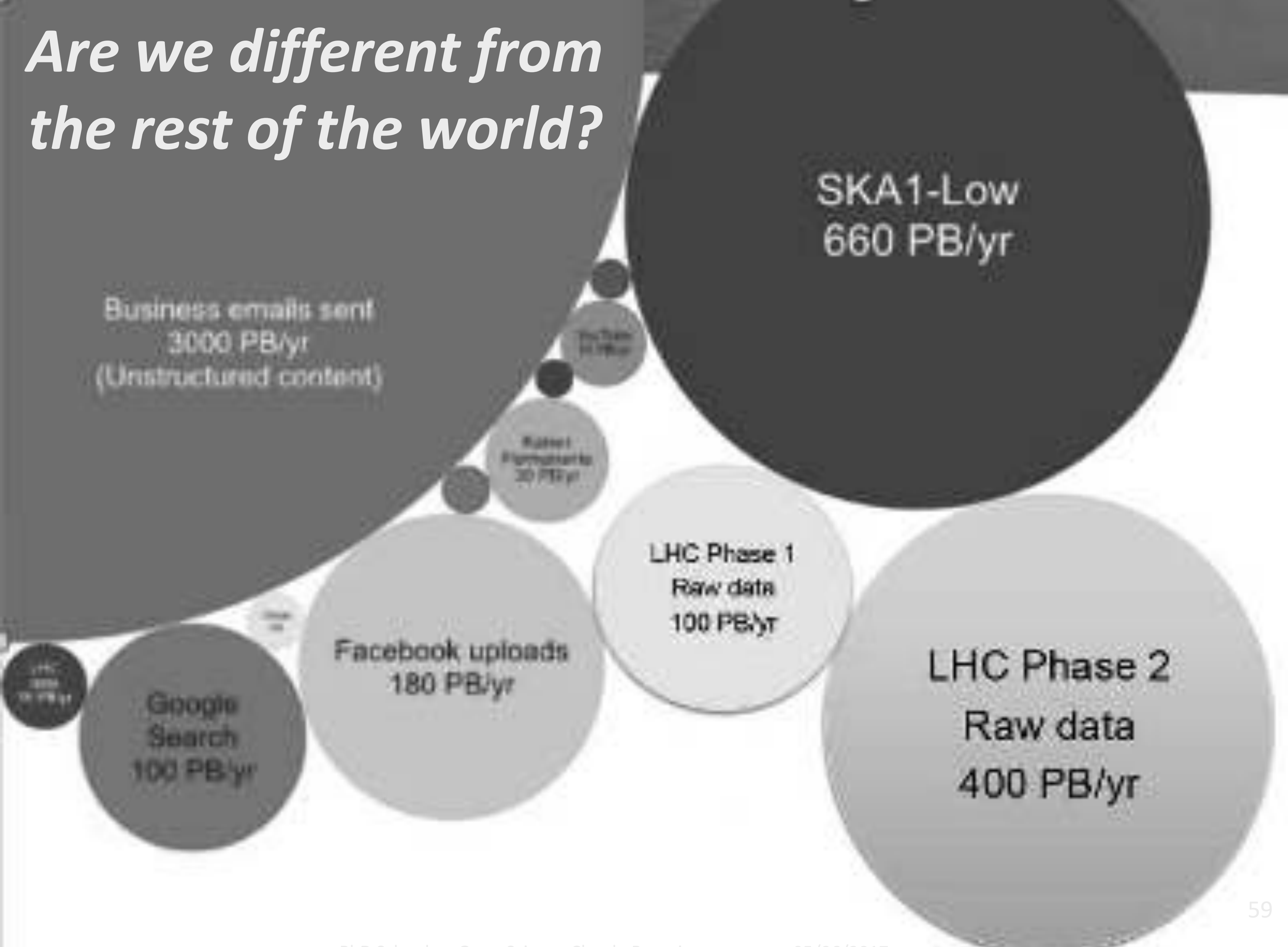
→ Far from a “flat budget” hypothesis for Run3

And Run4 is even worse!

Need to change models and exploit new technologies

Can’t just follow the evolution of currently used technologies!

Are we different from the rest of the world?



HEP is not different from the rest of the world

We can try to follow what others are doing

- *Even though Google, Facebook, & C. are making money out of investments while we have budget restrictions*

We can also try to exploit resources that others may make available to science in opportunistic mode

From Grid to Cloud

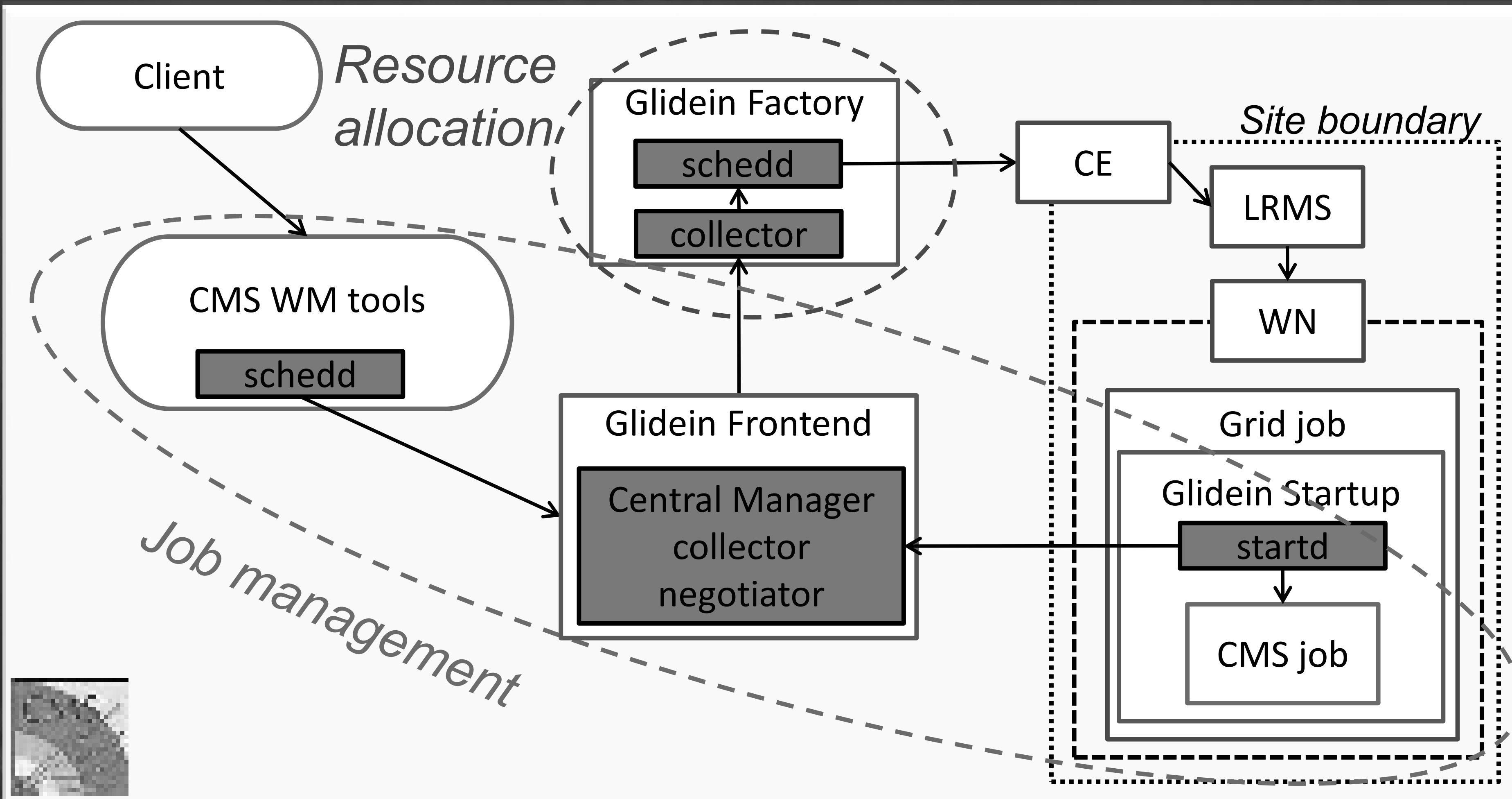
Cloud Computing offers most of the functionalities needed by HEP computing

Commercial and industrial world offers solutions that are being integrated

Actually there is a lot of Grid in the Cloud!

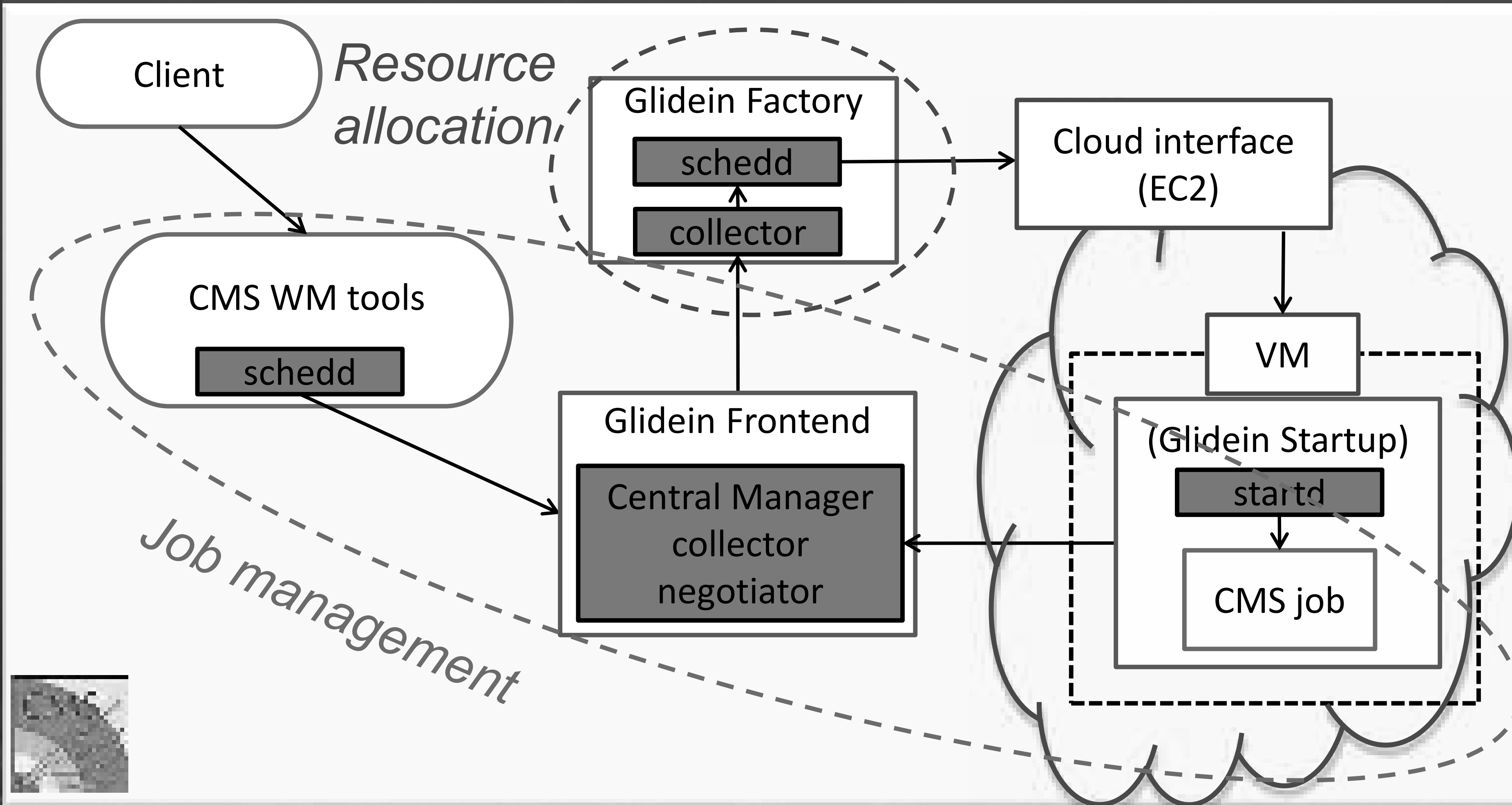
From the Grid…

The "factory" harvests job slots



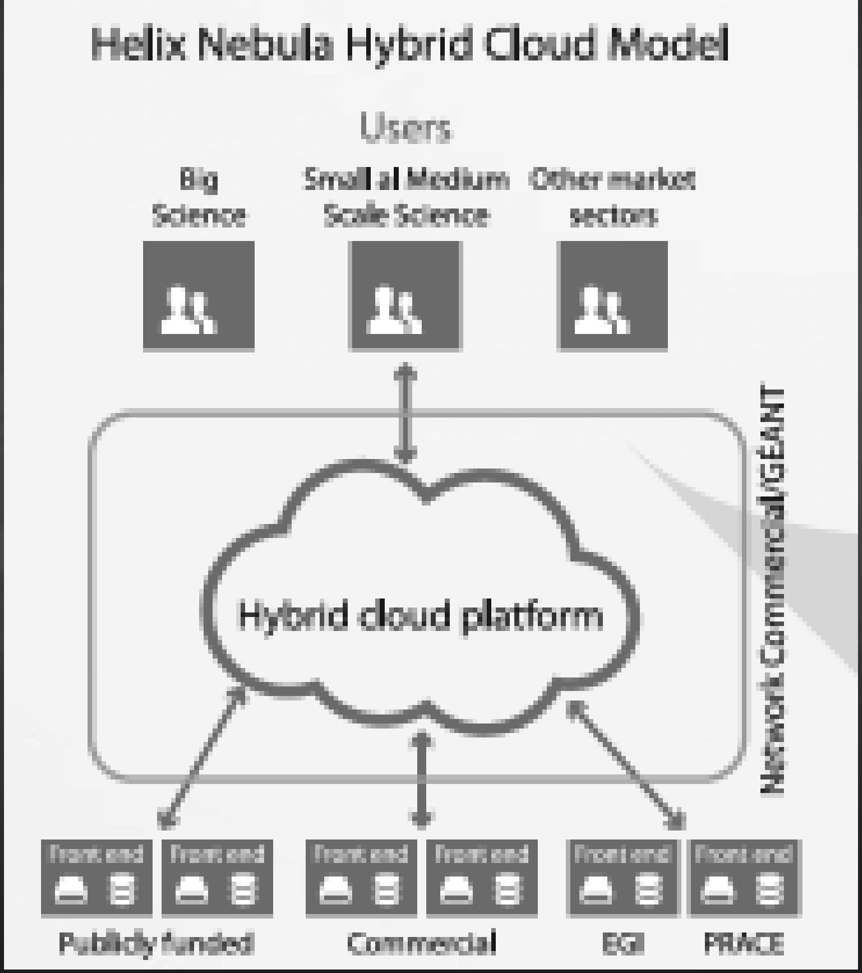
…to the Cloud

The "factory" harvests machines (or containers)



Hybrid Cloud model

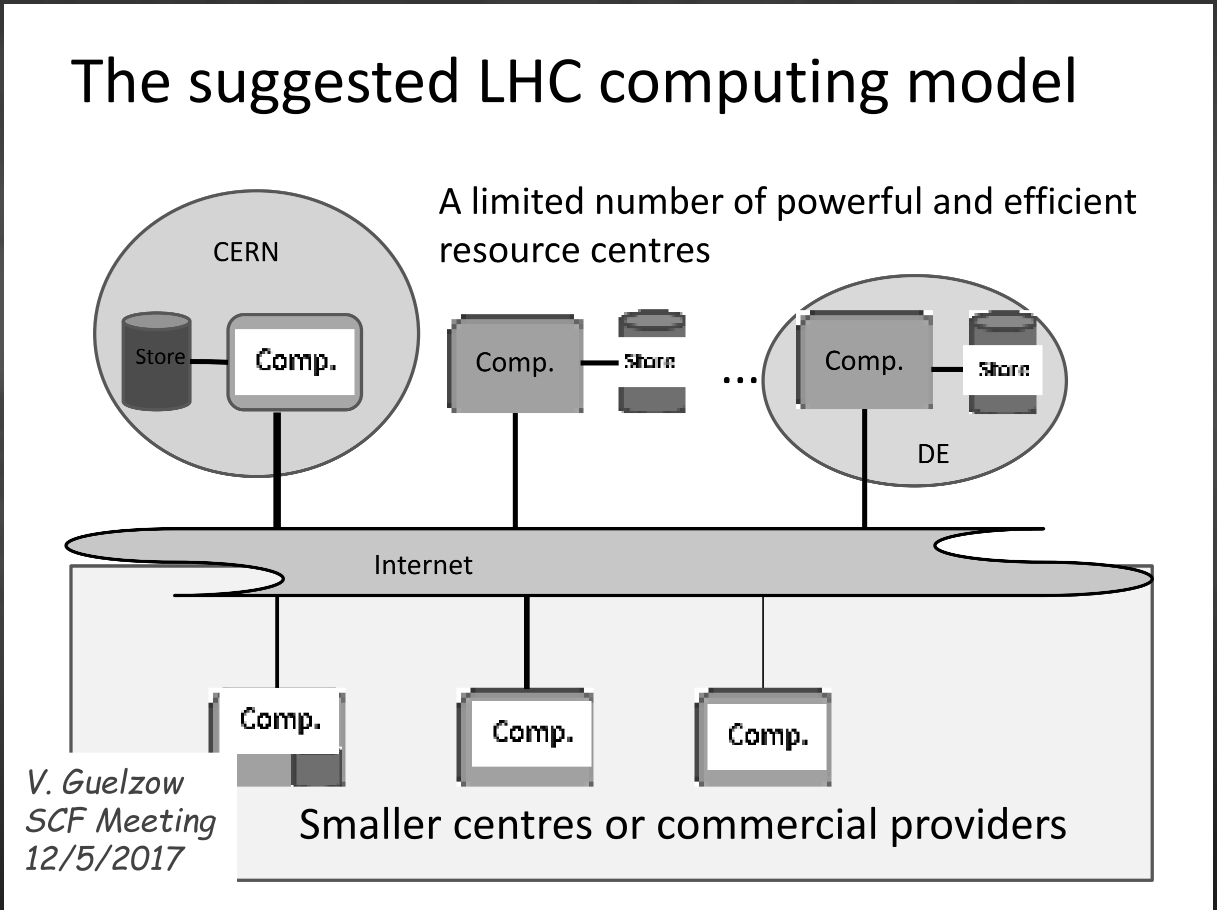
The use of standard cloud interface will allow to exploit private and commercial clouds at the same time



A new model for WLCG?

Decouple data and CPU management

- *Data is stored on a few, highly controlled, sites*
- *Most CPU is found elsewhere*



Not only distributed computing!

New architectures

Up to now HEP computing is based on a single architecture (x86-64)

→ Follow the market mainstream

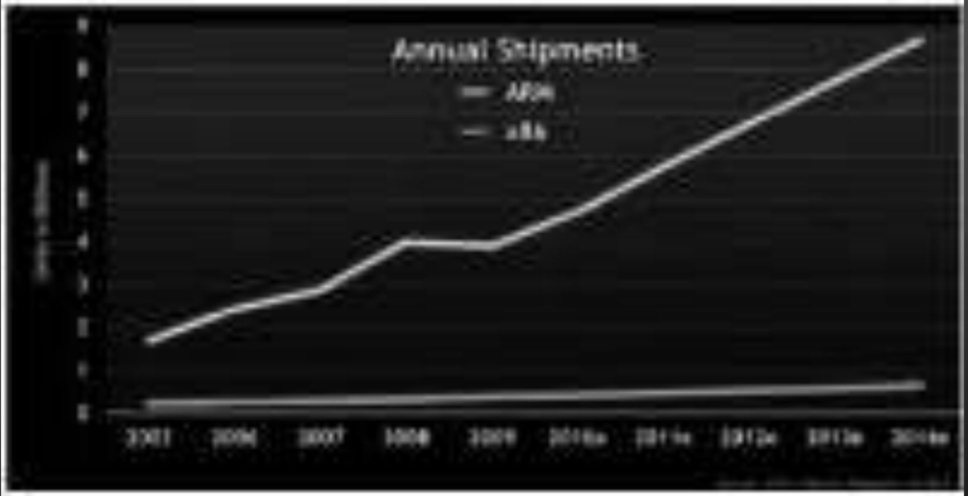
→ Use highly available architectures

ARM, …

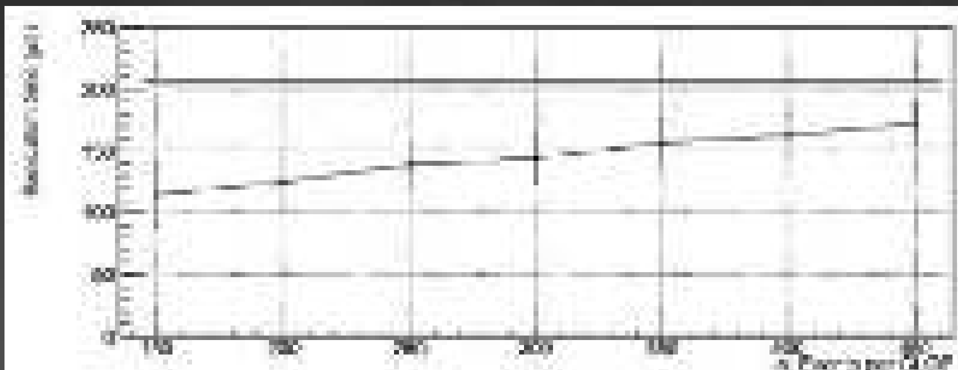
→ Exploit parallelization

Multi/many-core, GPGPU, …

→ Use low-power architectures



Exploit hardware capabilities

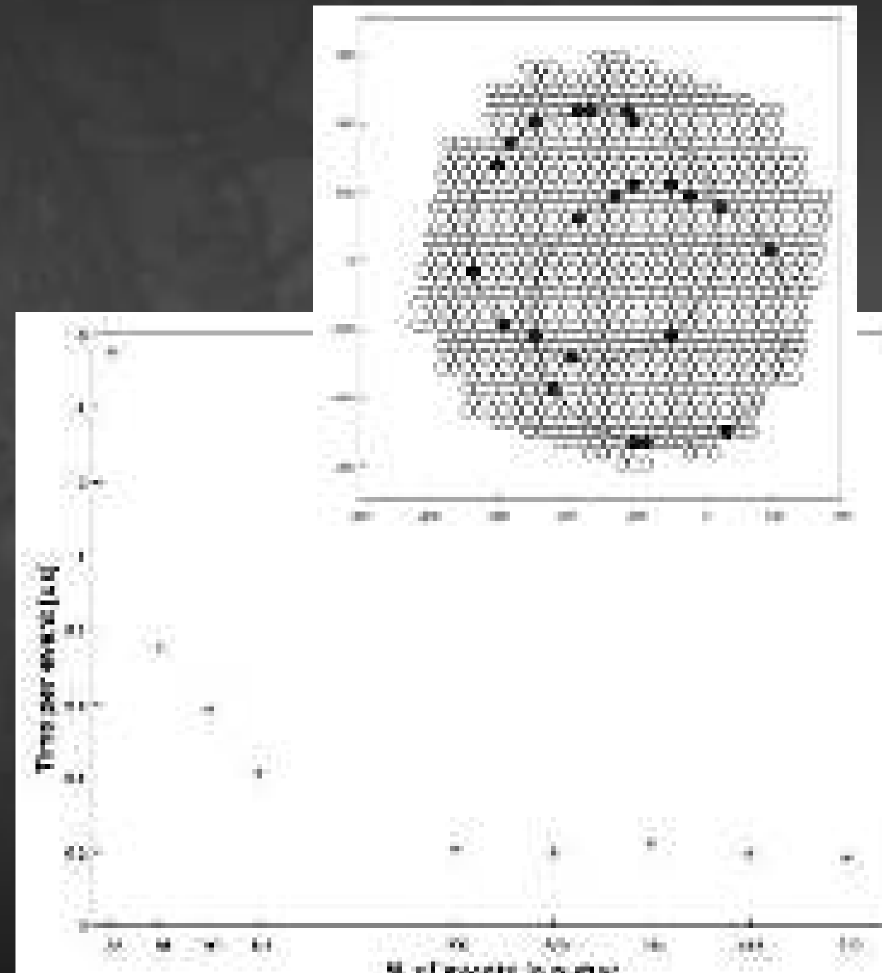


Algorithm parallelization in

NA64 RICH pattern recognition

Execution on NaNet-10

Based on GPU the Tesla K20c GPU



Effective programming!

Spread the knowledge

Software has considerably moved in the last years
- Consider that in few years we switched from C++98 to C++17

CPU power stopped moving towards higher frequencies, rather towards more tasks in parallel

The last generation of physicists was used to think in an OO way
- Huge chains of inheritance, big number of virtual functions

HEP programming seems to turn back to the old good functional paradigm
- → **Training is mandatory**

C++ offers a wide range of smart solutions to improve performances

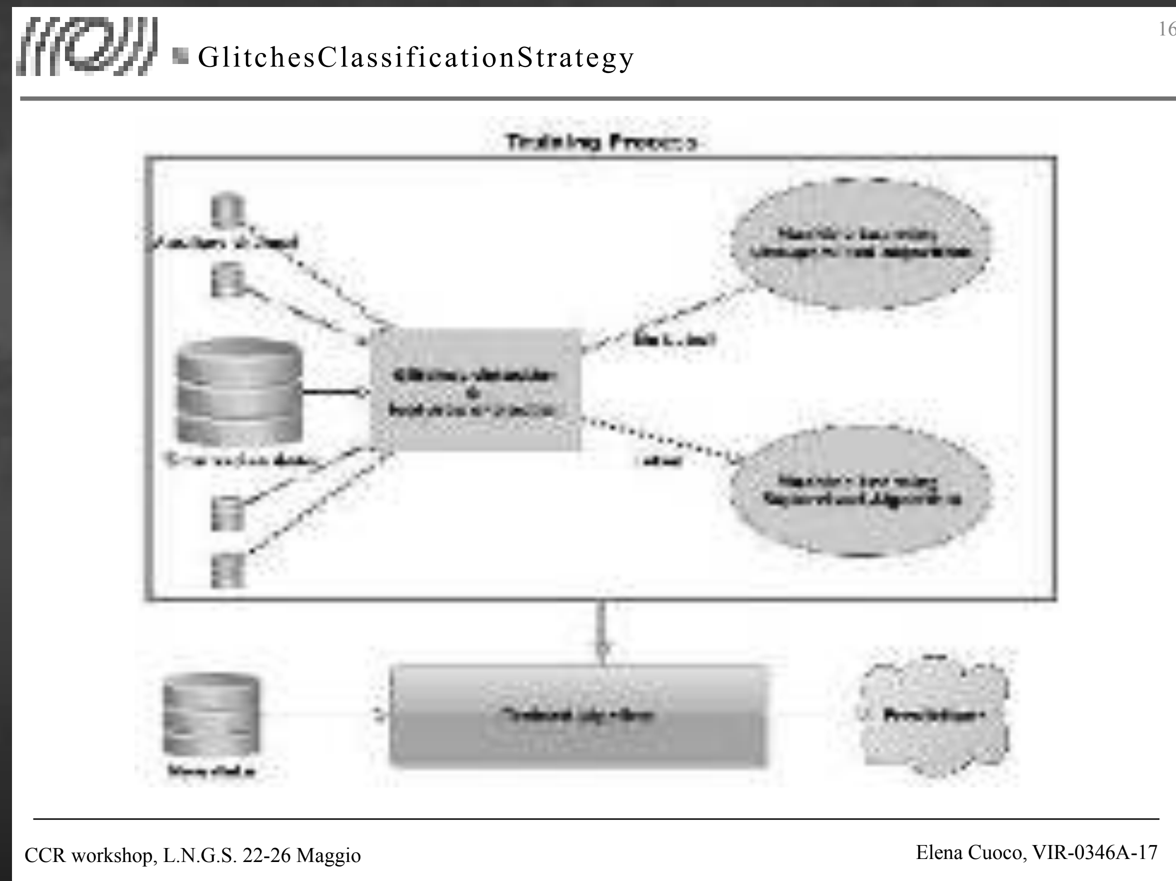
Compilers also are no more the *black magic boxes*
- Flags matter!

Machine Learning

Starting adopting Machine Learning & Deep Learning techniques for data processing

Example:

- *Glitches detection in Gravitational Waves searches*



Software: the key to the solution?

Why Software? Software is *the* Cyberinfrastructure

Computer hardware is a consumable.
Software is what we keep, and invest in, over time.

9/ 24

By P.Elmer
HSF workshop
23/1/2017

Concluding…

HEP computing is continuously evolving

Experiment requests impose an evolution of the model in order to comply with the (flat) budget

Need to understand and exploit new technologies

Software is the key to scalability and sustainability

There is room for new ideas and innovative projects!