Trends in computing for the INFN experiments

Claudio Grandi

INFN Bologna

Computing in INFN


4 National Laboratories
20 Divisions
11 Groups
1 Computer Science Center
1 Tier1 & 9 Tier2
3 Other Institutions
LHC-Tier2 Milano
LHC-Tier2 Torino
Milano-Bicocca
Milano
Trento
Brescia
LHC-Tier2 Legnaro (PD)
Udine
Trieste
Torino
Pavia
Padova
Alessandria
Ferrara
LNF-National Laboratory Legnaro (PD)
Parma
Genova
Bologna
INFN – CNAF Computer Science Center - Bologna
LHC-Tier1 CNAF-Bologna
LHC-Tier2 Pisa
Pisa
Firenze
Perugia
GGI Galileo Galilei Institute for Theoretical Physics - Arcetri (FI)
L'Aquila
LNGS-Gran Sasso National Laboratory - Assergi (AQ)
LHC-Tier2 Roma
LHC-Tier2 Frascati (RM)
Roma I
LNF-National Laboratory Frascati (RM)
Roma II
Roma III
Sanità
Bari
LHC-Tier2 Bari
Napoli
LHC-Tier2 Napoli
Salerno
Lecce
Cagliari
Cosenza
Ettore Majorana Foundation and Centre for Scientific Culture - Erice (TP)
Messina
Catania
LHC-Tier2 Catania
LNS- Sud National Laboratory - Catania
Portopalo di Capo Passero (SR) Control Centre
WWW.GARR.IT

Activities

Management of all INFN Operational Units (Sections, Laboratories, Centres)

- User support (network, e-mail, administration, …)

Management of resources for scientific computing

- Management of computing centres and of the distributed infrastructure

Development of technologies useful for the different activities

- Distributed computing, data preservation and open access
- Data acquisition and real-time, low-latency computing
- Data processing
  - Not just batch data analysis but also statistical analysis and visualization
- Theoretical computations and High Performance Computing

Computing for the experiments

HEP computing has different aspects

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

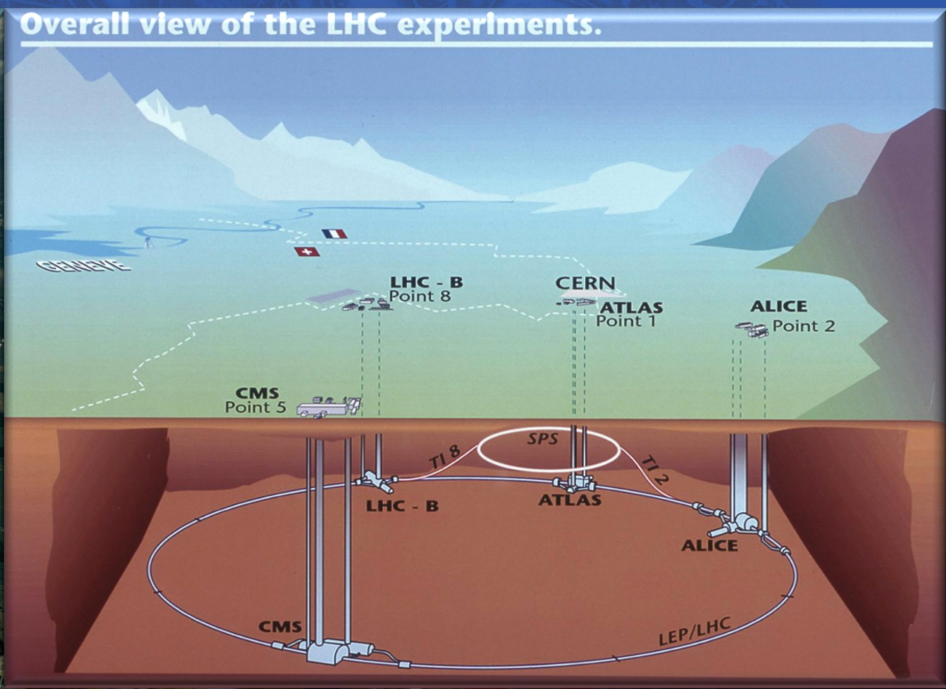
The infrastructure built by the community is tailored on the needs of LHC that is the most demanding user at the moment (but it serves all the HEP community and more)

LHC

27 km proton-proton collider, ~100 m underground

13 TeV c.m. energy, 10^{34} $cm^{-2}s^{-1}$ luminosity

40 MHz bunch crossing rate in each of the four experiments



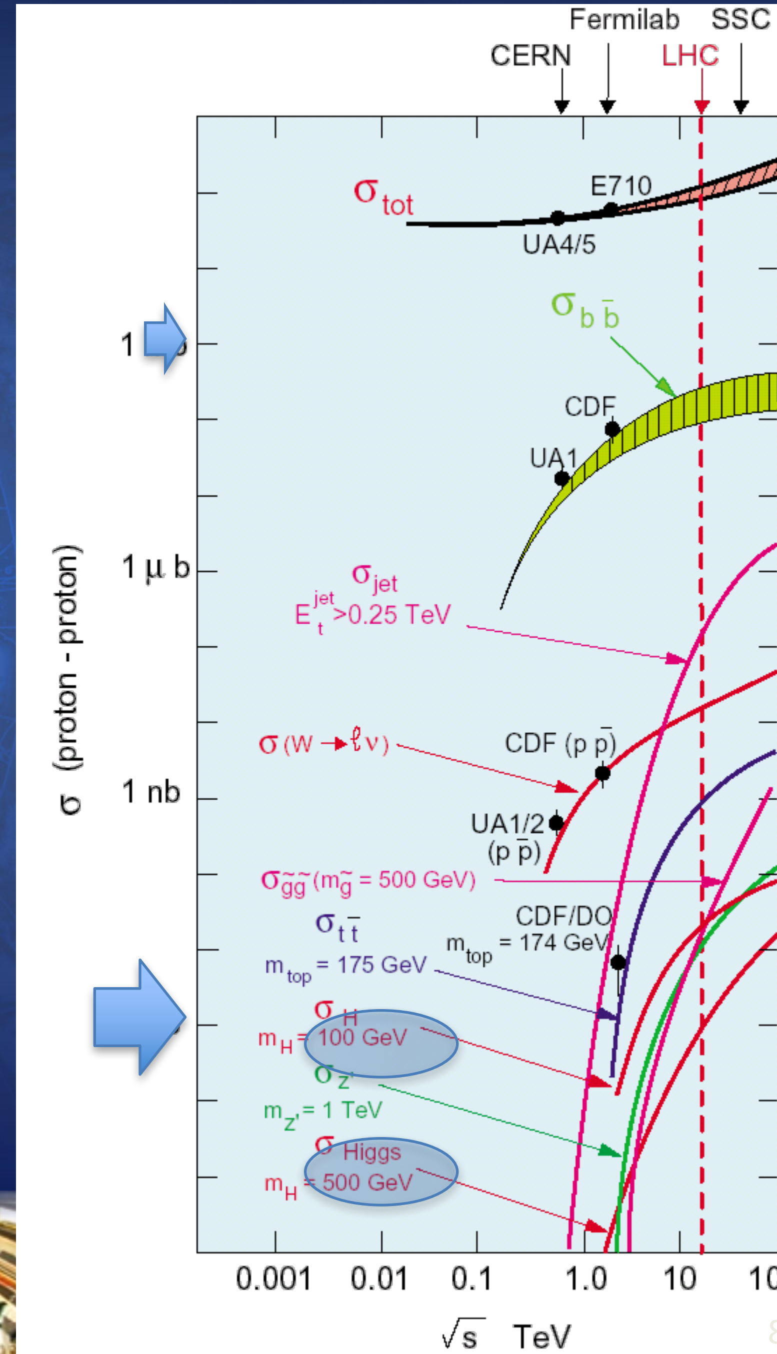
LHC

Not going to talk about physics and detectors. From the computing point of view:

LHC experiments study rare events!

Signal to noise ratio ~ 10^{-13}

Effective data reduction techniques are needed!



LHC

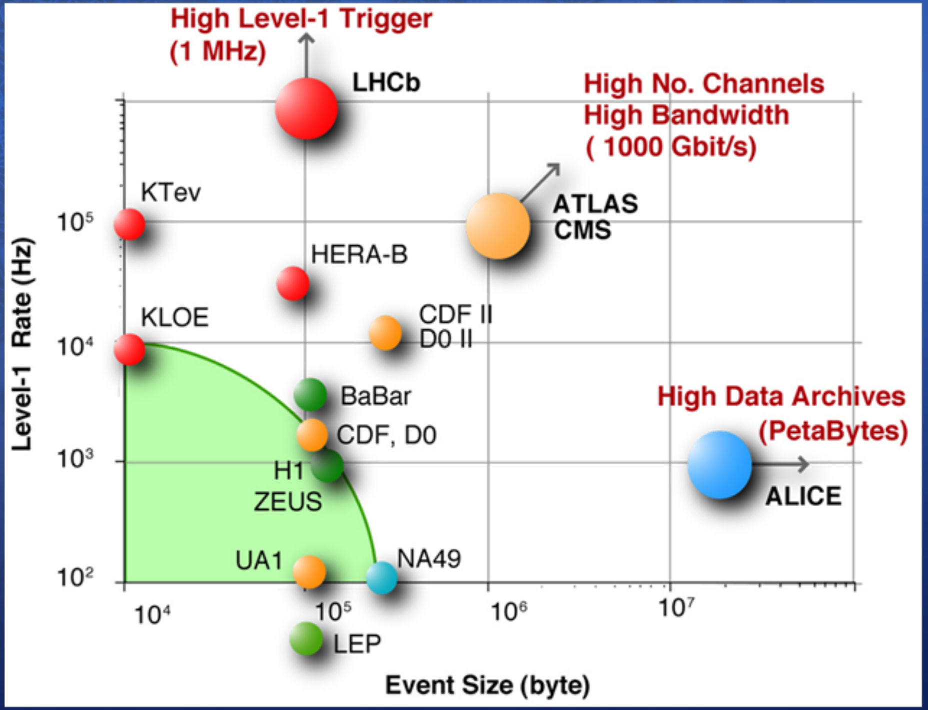
In each LHC experiment there are 40 million bunch crossings per second. Every time 100 million channels are acquired (100 MB)

➔ *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…

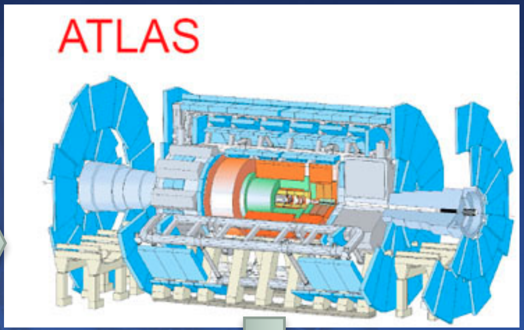
The process that interprets RAW data in terms of physics objects is the reconstruction

- *Actually there are many reconstruction phases*

Physicists do analysis on reconstructed data

LHC Real data

LHC collisions → Decay of unstable particles →



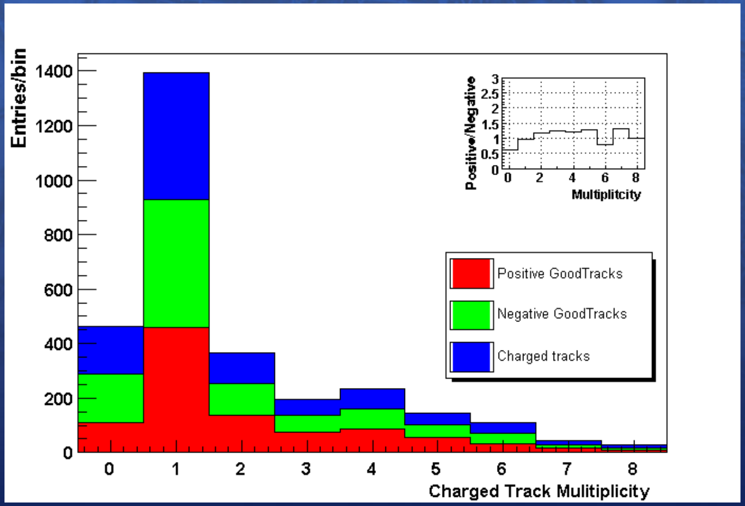
↓ Detector electronics

↓ 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

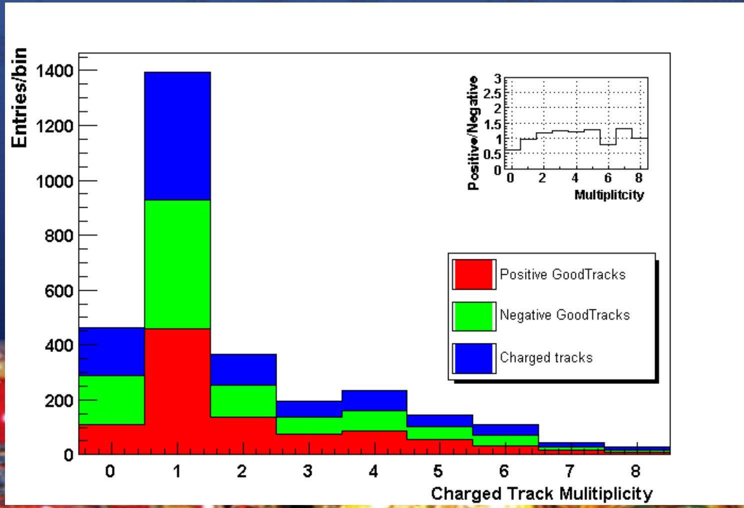
The simulated data sample is 1 to 2 times the real data sample

LHC Simulated data

Theoretical model → Simulation of decays of unstable particles → Simulation of interactions particle-detector → Simulation of detector electronics → Trigger → Recontruction → Analysis

Pythia, …

Geant4, …



Computing infrastructure

Management 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

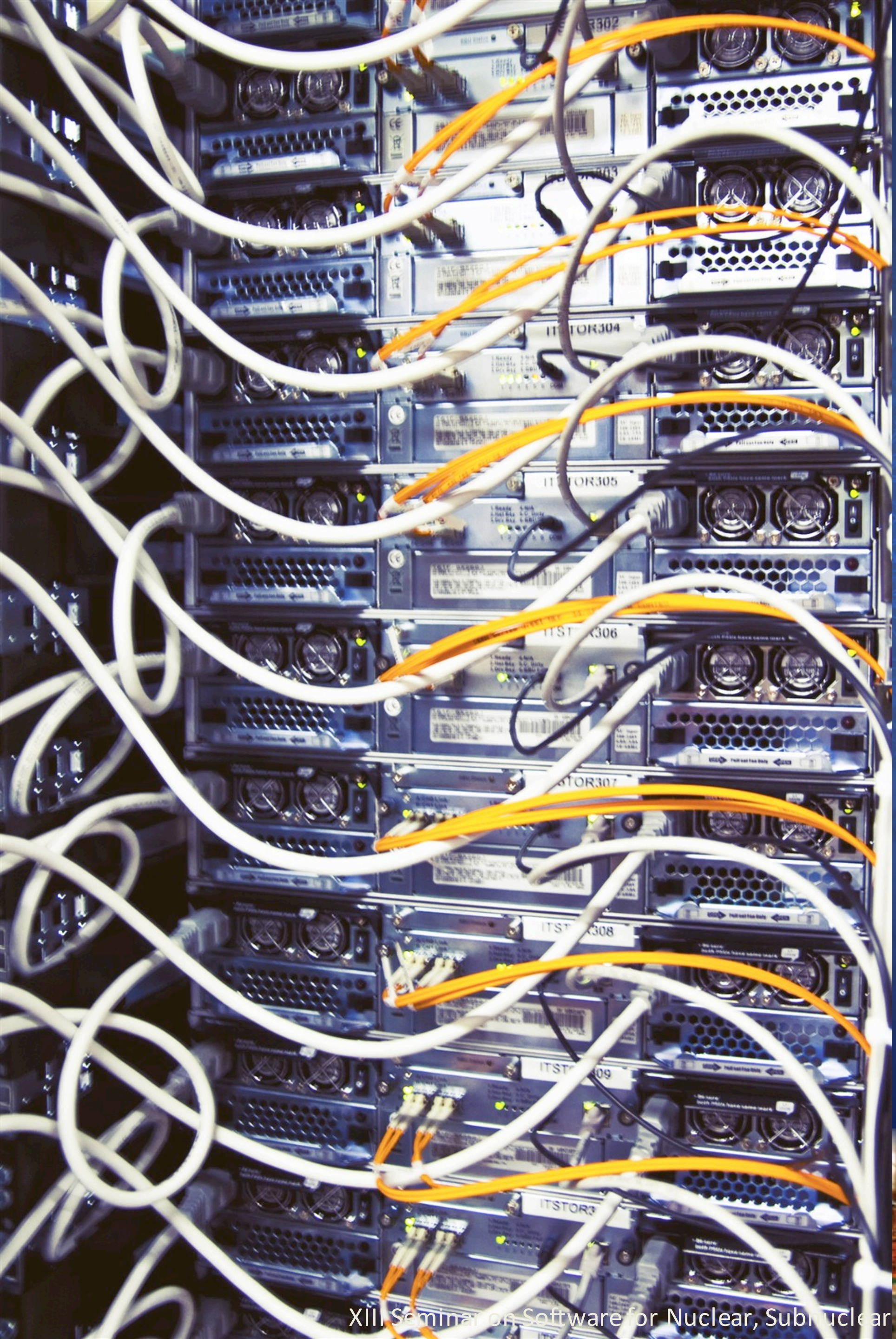
Using a unit specific for HEP: HepSpec06 (HS06)

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)

~~*73.000*~~ ~ 85 000 *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

~~*1,5*~~ > 2 *million jobs running every day*

~~*10*~~ 50 *GB/s global transfer rate*

…more numbers

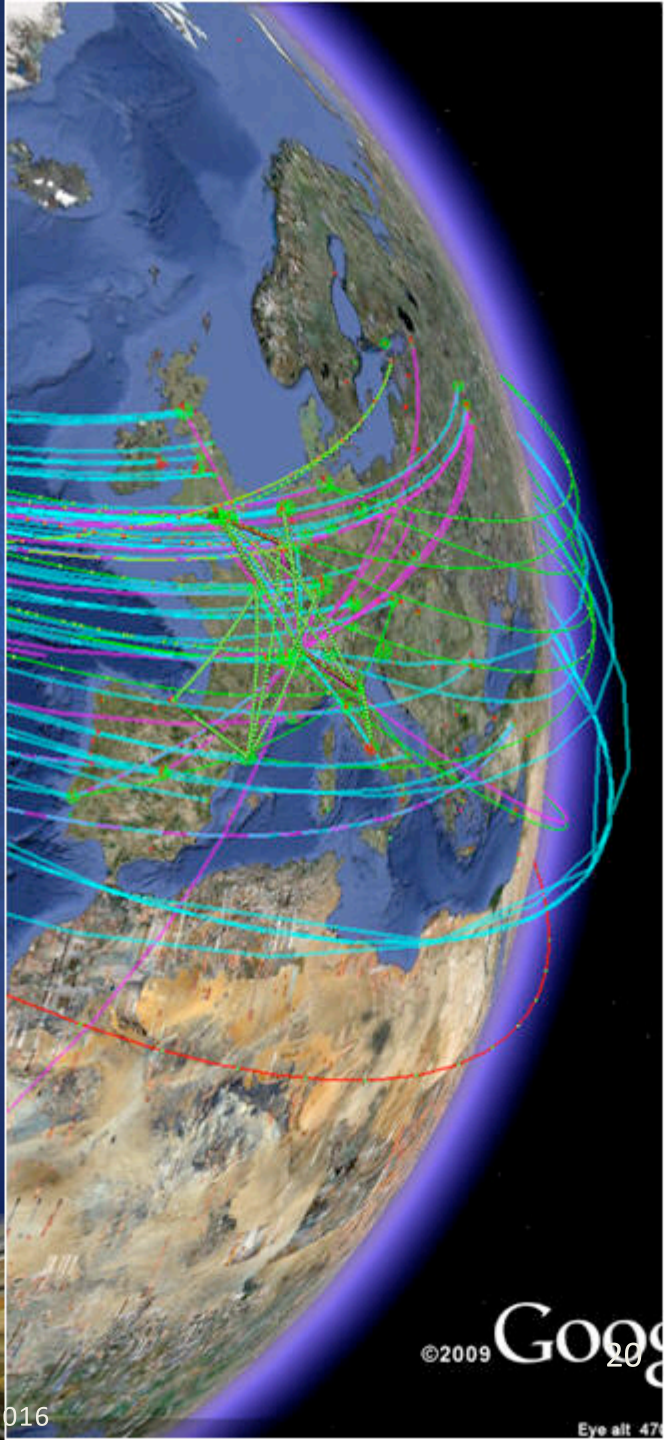
Global resources for 2016 are:

- *2,900,000 HS06 (~290.000 processor cores)*
- *240.000 TB disk*
- *250.000 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

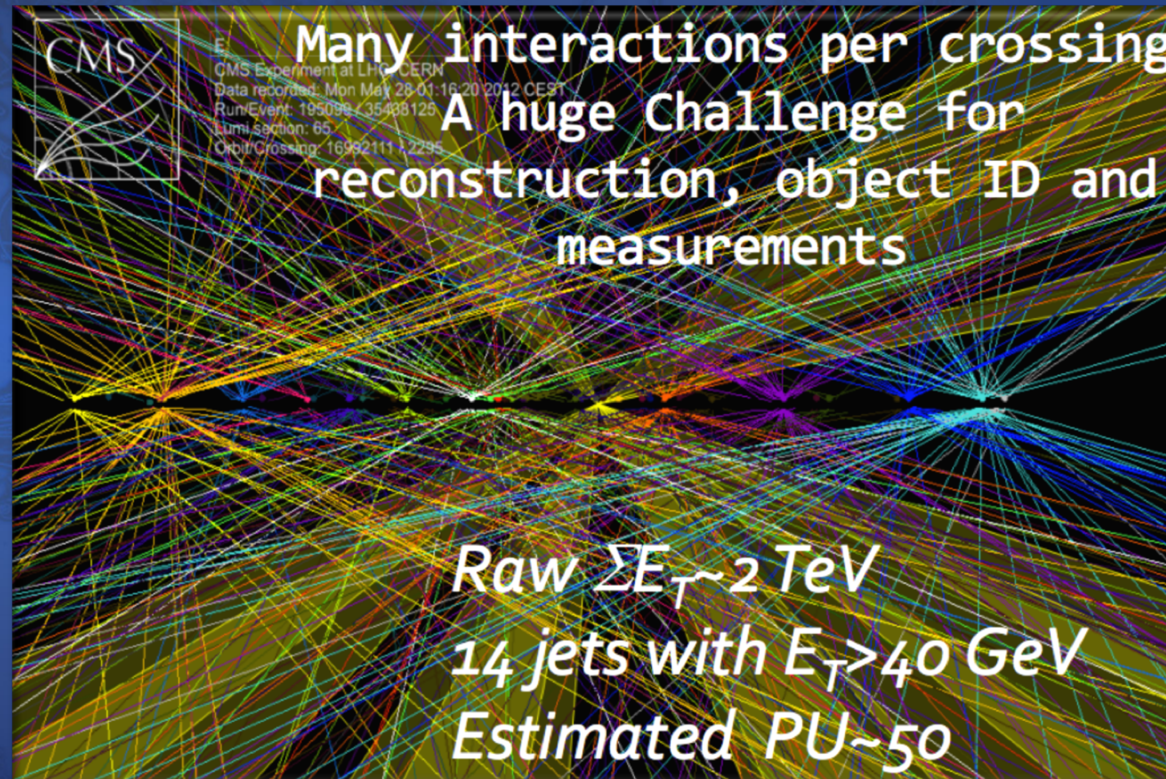
More than 8000 analysts all over the world



Pile-up

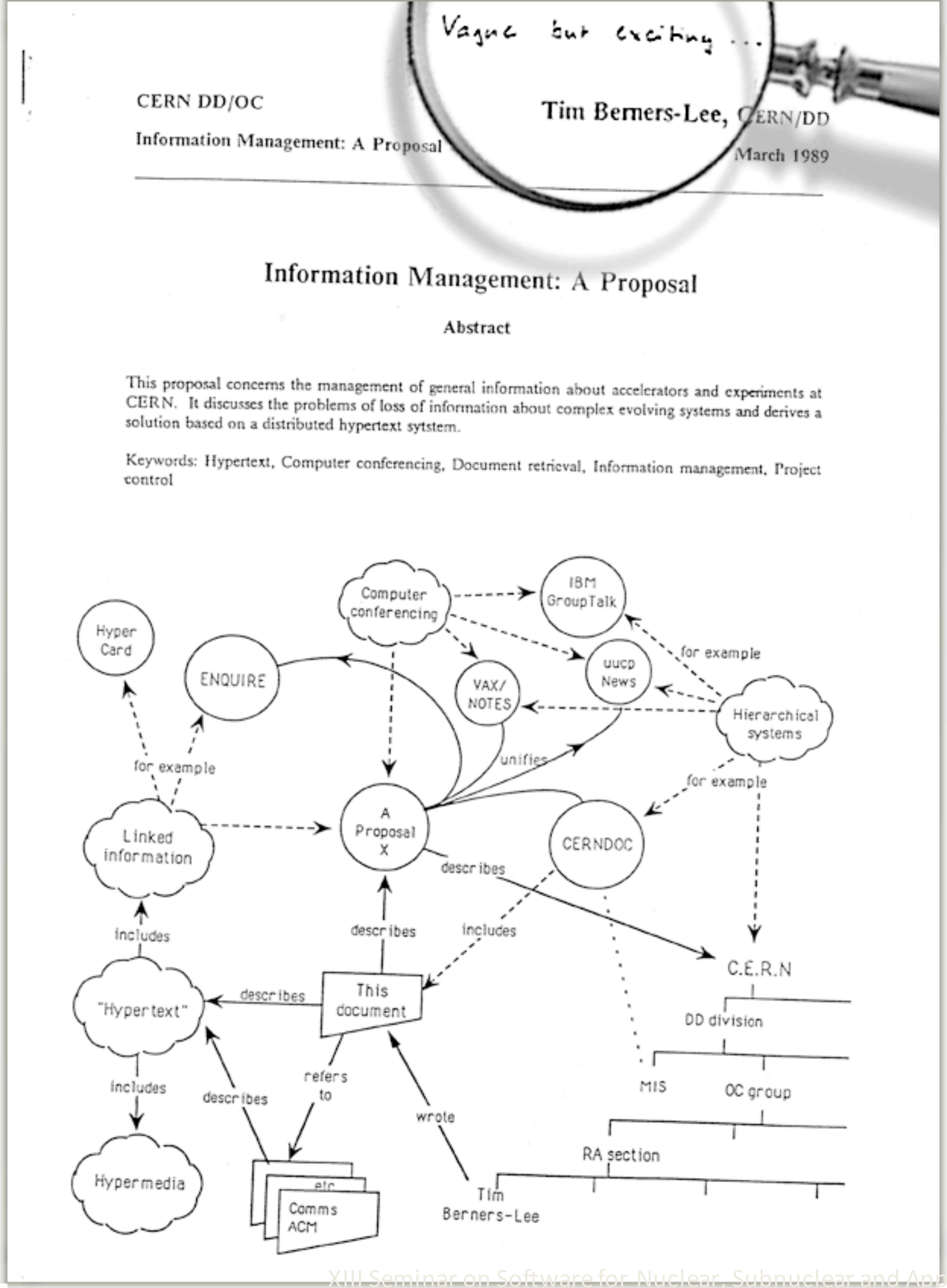
If you're wondering why a bunch crossing rate of 40 MHz produces 600 collisions per second:

> *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?



Vague but exciting ...

CERN DD/OC

Information Management: A Proposal

Tim Berners-Lee, CERN/DD

March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext sytstem.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

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!

2.

Declaration

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

- W 3 basic ("line-mode") client
- W 3 basic server
- W 3 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

opie certifiée conforme

ait à Genève le 03-05-93

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
DIVISION DES FINANCES
- CERN -

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)

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

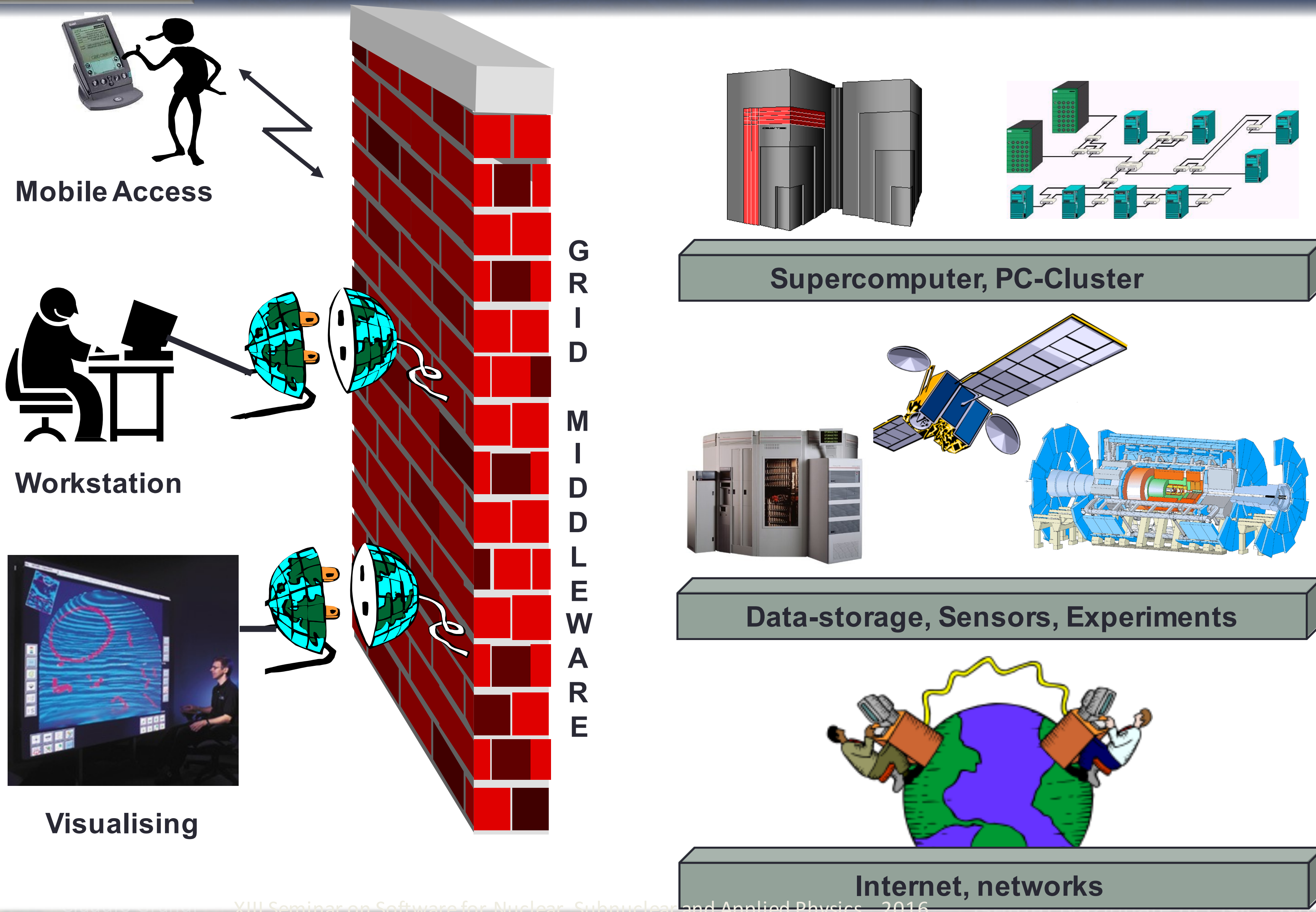
They started from an idea of a group of American computing scientists: the Computing Grid

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



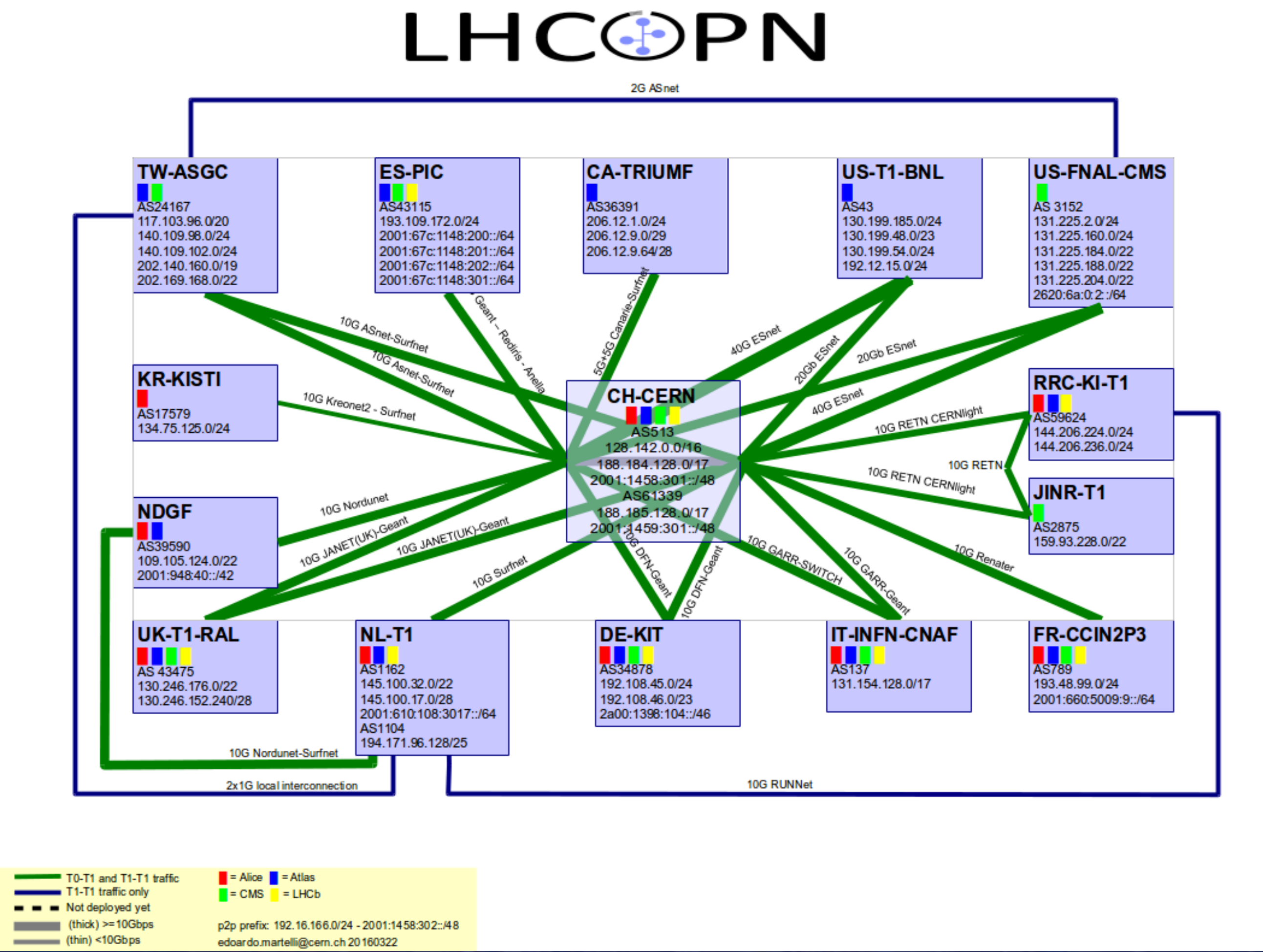
A distributed system

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*

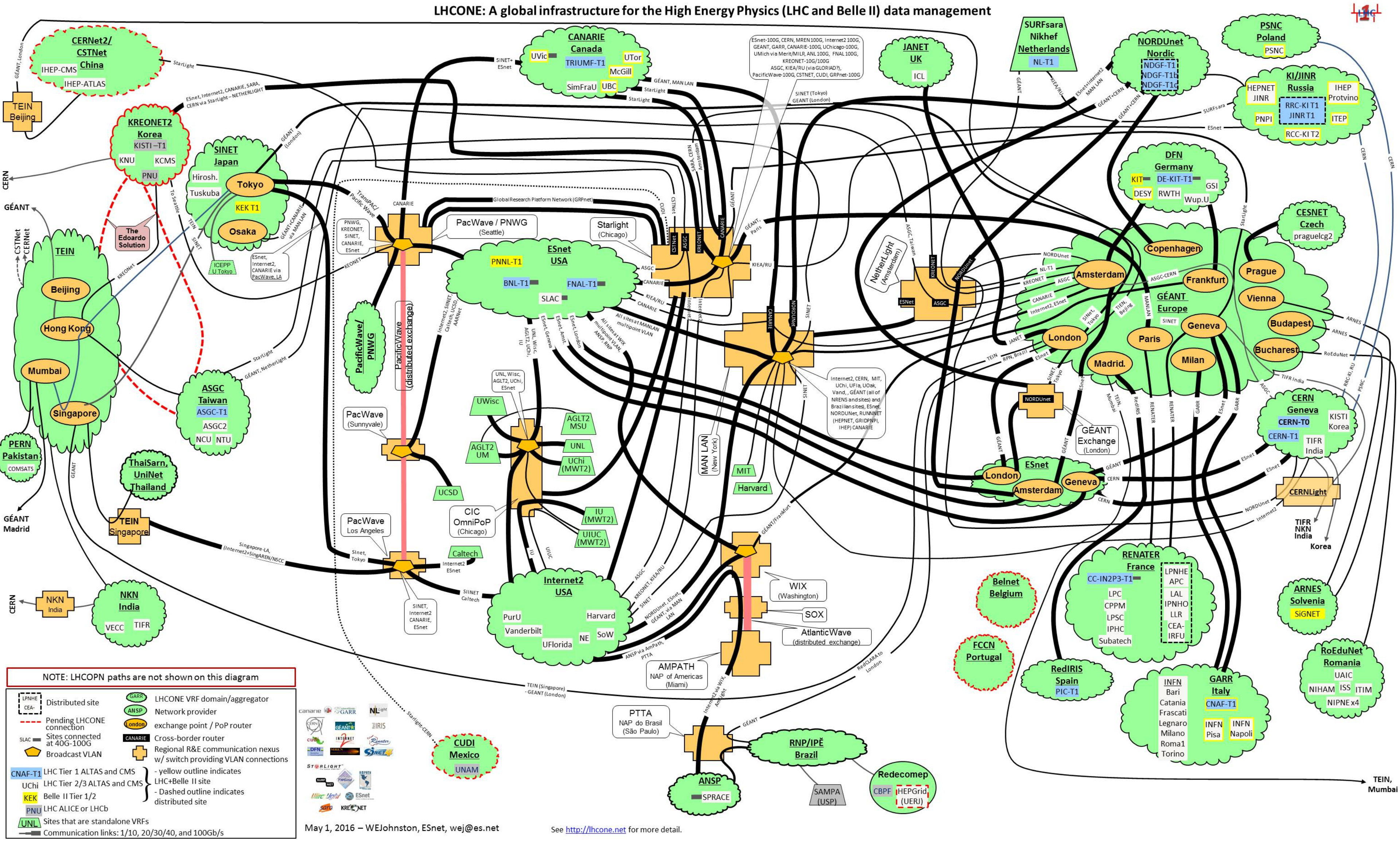
Only a few technical details…

The network - LHCOPN



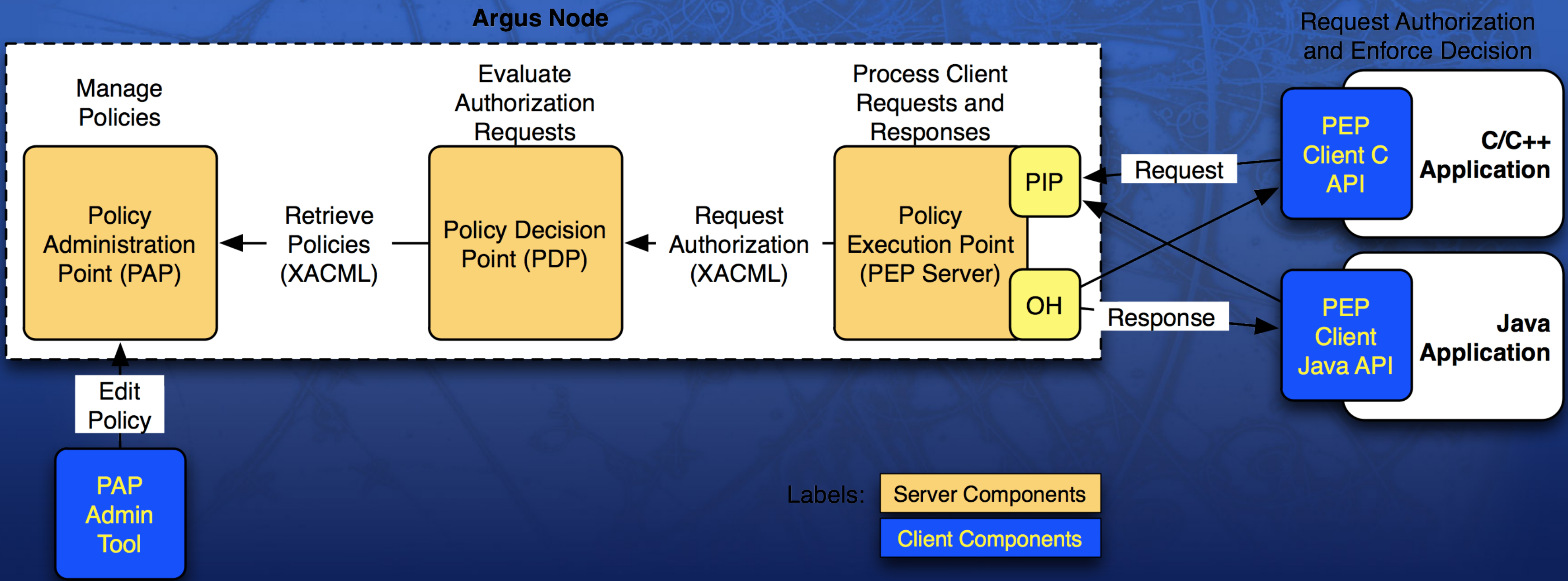
The network technology evolved significantly, offering adequate performance to support the distributed computing model

The network - LHCONE



Grid Security management

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

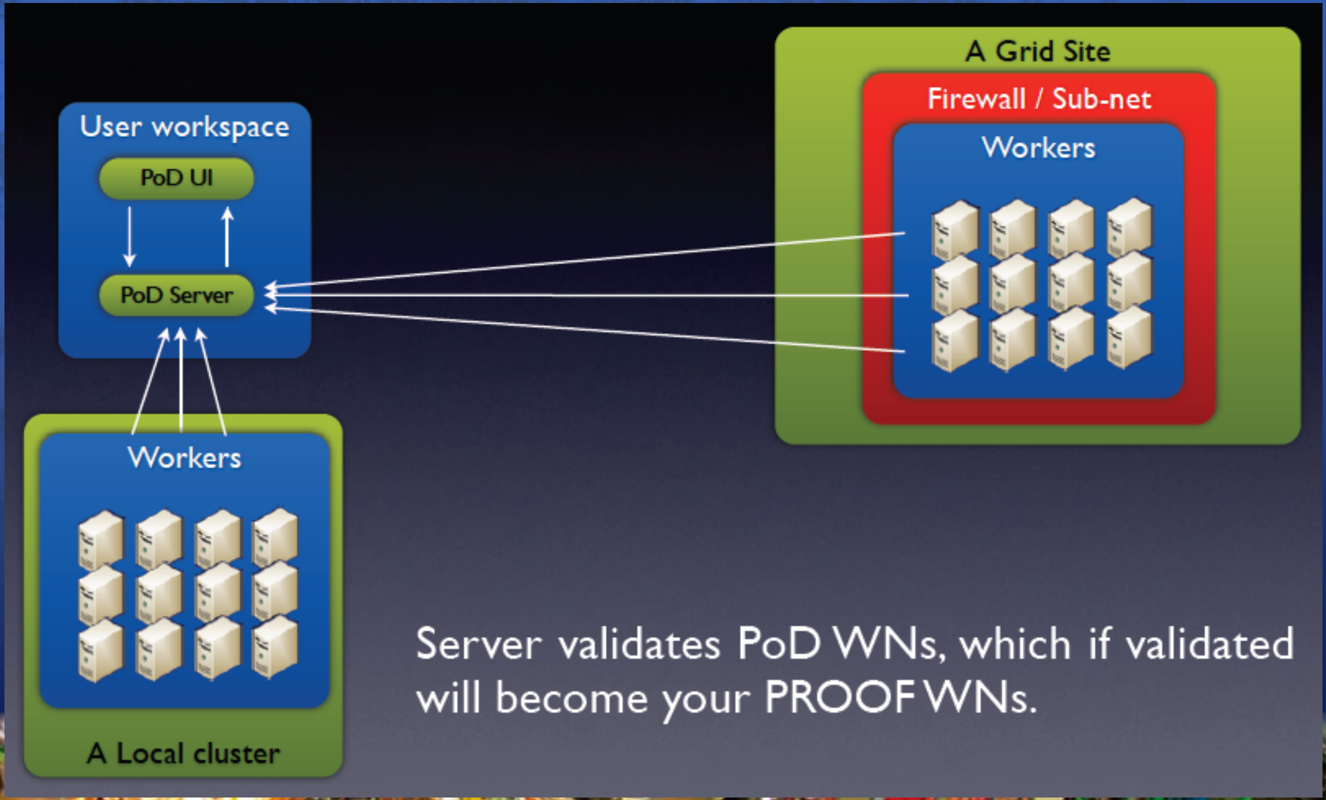


Grid Computing management

Access is based on batch jobs: asynchronous execution

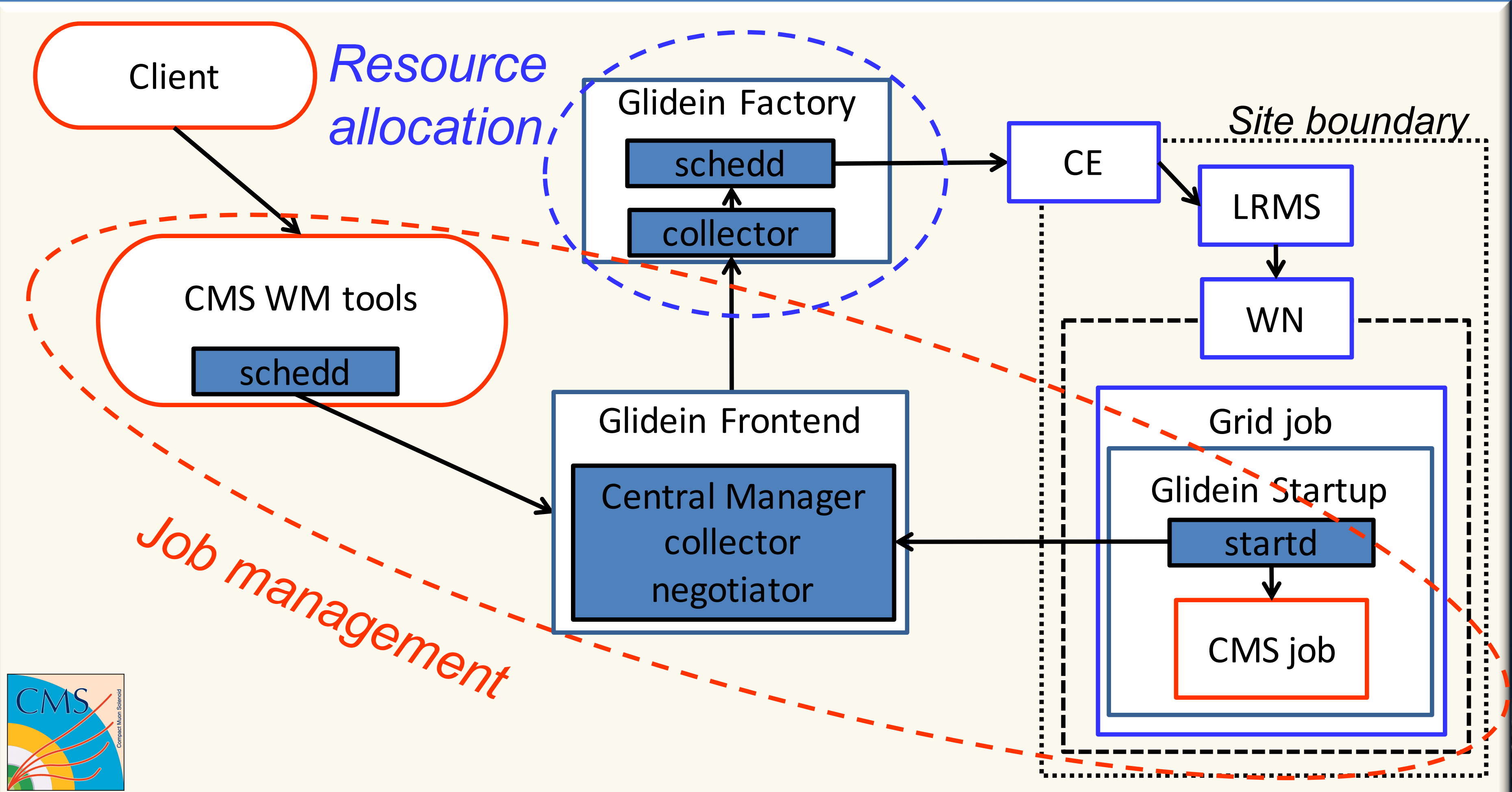
Dedicated interfaces allow to manage remote submissions as if local

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*

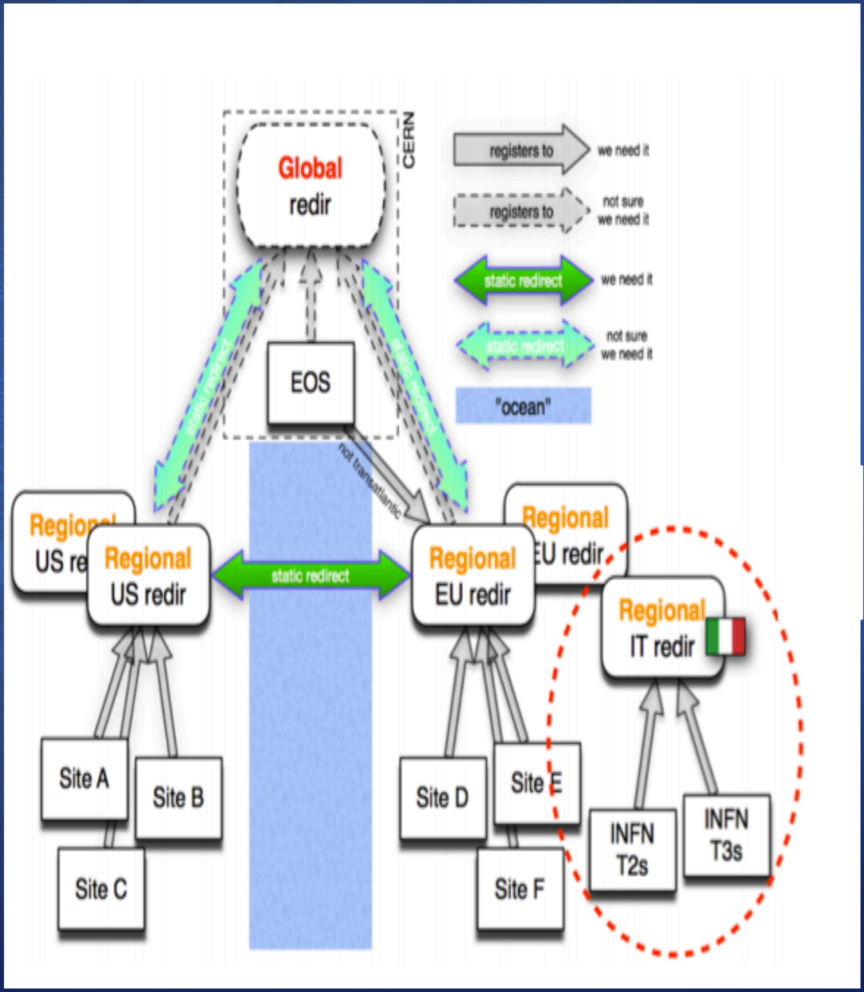
Tools to manage the transfer of large amounts of data

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

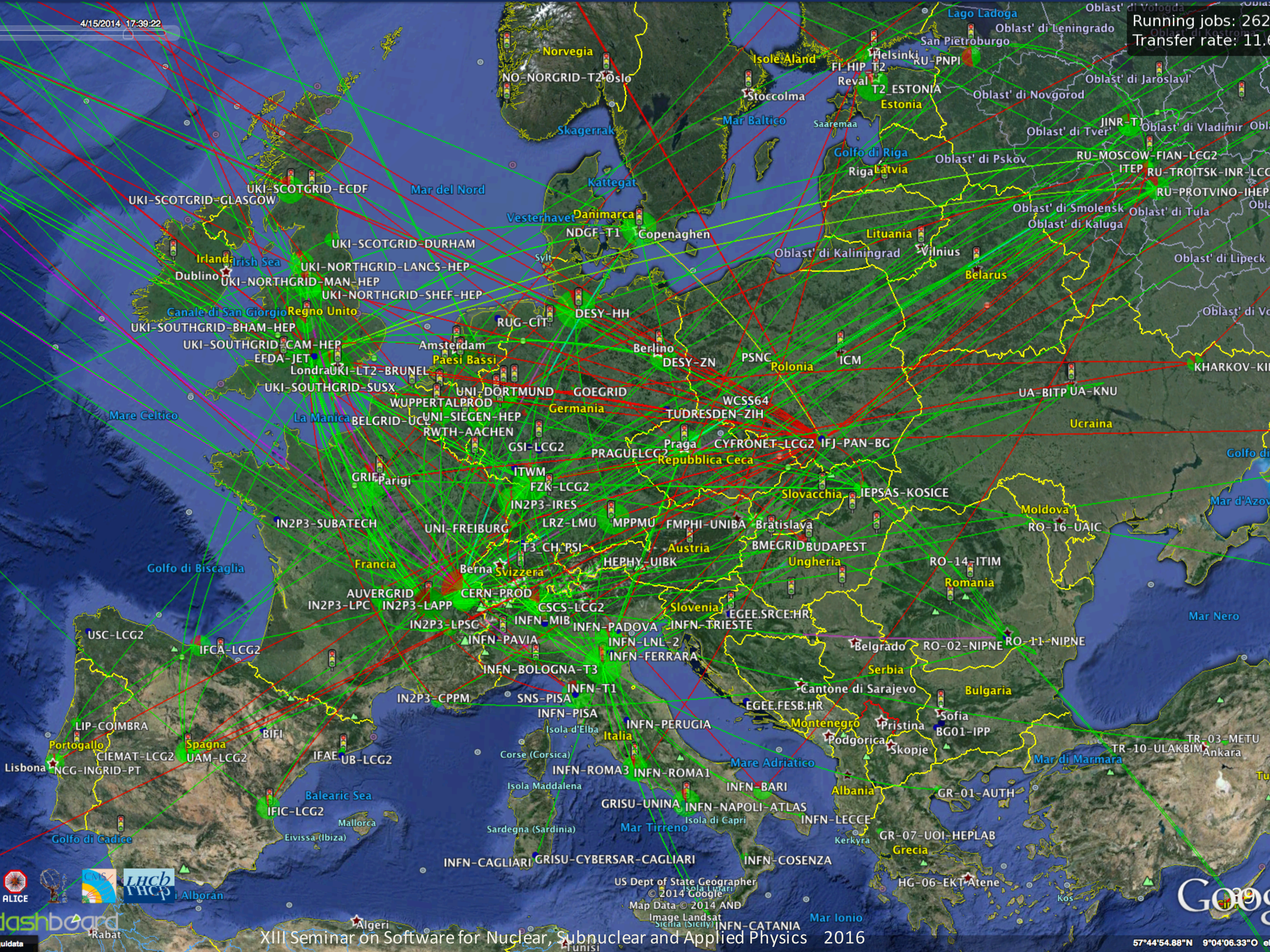
- *Storage Federations*

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)
- Currently in production for *xrootd*



Let’s see how it works…


4/15/2014 17:39:22
Running jobs: 262
Transfer rate: 11.
Lago Ladoga
Oblast' di Leningrado
San Pietroburgo
Norvegia
NO-NORGRID-T2 Oslo
Isole Aland
Helsinki
FI_HIP_T2
RU-PNPI
Reval
T2_ESTONIA
Estonia
Stoccolma
Oblast' di Jaroslavl'
Oblast' di Novgorod
Skagerrak
Mar Baltico
Saaremaa
Oblast' di Tver'
JINR-T1
Oblast' di Vladimir
Golfo di Riga
Riga
Latvia
Oblast' di Pskov
RU-MOSCOW-FIAN-LCG2
ITEP
RU-TROITSK-INR-LCG
RU-PROTVINO-IHEP
Oblast' di Smolensk
Oblast' di Tula
Oblast' di Kaluga
UKI-SCOTGRID-ECDF
UKI-SCOTGRID-GLASGOW
Mar del Nord
Kattegat
Vesterhavet
Danimarca
NDGF-T1
Copenaghen
Lituania
UKI-SCOTGRID-DURHAM
Irlanda
Irish Sea
Dublino
UKI-NORTHGRID-LANCS-HEP
UKI-NORTHGRID-MAN-HEP
UKI-NORTHGRID-SHEF-HEP
Oblast' di Kaliningrad
Vilnius
Oblast' di Lipeck
Belarus
Canale di San Giorgio
Regno Unito
DESY-HH
RUG-CIT
UKI-SOUTHGRID-BHAM-HEP
UKI-SOUTHGRID-CAM-HEP
Amsterdam
Berlino
EFDA-JET
Paesi Bassi
DESY-ZN
PSNC
Polonia
ICM
KHARKOV-KI
Londra
UKI-LT2-BRUNEL
UKI-SOUTHGRID-SUSX
UNI-DORTMUND
GOEGRID
WUPPERTALPROD
WCSS64
UA-BITP
UA-KNU
Germania
Mare Celtico
La Manica
BELGRID-UCL
UNI-SIEGEN-HEP
TUDRESDEN-ZIH
RWTH-AACHEN
Ucraina
GSI-LCG2
Praga
CYFRONET-LCG2
IFJ-PAN-BG
PRAGUELCG2
Repubblica Ceca
GRIF
Parigi
ITWM
FZK-LCG2
Slovacchia
IEPSAS-KOSICE
Moldova
IN2P3-IRES
IN2P3-SUBATECH
LRZ-LMU
MPPMU
FMPHI-UNIBA
Bratislava
RO-16-UAIC
UNI-FREIBURG
Austria
BMEGRID
BUDAPEST
T3_CH_PSI
HEPHY-UIBK
RO-14-ITIM
Golfo di Biscaglia
Francia
Ungheria
Berna
Svizzera
Romania
AUVERGRID
CERN-PROD
IN2P3-LPC
IN2P3-LAPP
CSCS-LCG2
Slovenia
EGEE.SRCE.HR
Mar Nero
INFN-MIB
IN2P3-LPSC
INFN-PADOVA
INFN-TRIESTE
USC-LCG2
INFN-PAVIA
INFN-LNL-2
RO-11-NIPNE
IFCA-LCG2
INFN-FERRARA
Belgrado
RO-02-NIPNE
INFN-BOLOGNA-T3
Serbia
INFN-T1
Cantone di Sarajevo
Bulgaria
IN2P3-CPPM
SNS-PISA
EGEE.FESB.HR
INFN-PISA
Sofia
LIP-COIMBRA
INFN-PERUGIA
Montenegro
Pristina
Isola d'Elba
BG01-IPP
Portogallo
BIFI
Spagna
Italia
Podgorica
TR-03-METU
CIEMAT-LCG2
Skopje
TR-10-ULAKBIM
Ankara
Lisbona
NCG-INGRID-PT
UAM-LCG2
IFAE
UB-LCG2
Corse (Corsica)
Mare Adriatico
Mar di Marmara
INFN-ROMA3
INFN-ROMA1
Isola Maddalena
Balearic Sea
INFN-BARI
Albania
GR-01-AUTH
IFIC-LCG2
GRISU-UNINA
INFN-NAPOLI-ATLAS
Mallorca
Isola di Capri
INFN-LECCE
Golfo di Cadice
Eivissa (Ibiza)
Sardegna (Sardinia)
Mar Tirreno
Kerkyra
GR-07-UOI-HEPLAB
Grecia
INFN-CAGLIARI
GRISU-CYBERSAR-CAGLIARI
INFN-COSENZA
US Dept of State Geographer
© 2014 Google
Map Data © 2014 AND
Image Landsat
HG-06-EKT
Atene
ALICE
LHCb
Alboran
dashboard
Algeri
Rabat
Mar Ionio
INFN-CATANIA
Sicilia (Sicily)
Tunisi
Google
57°44'54.88"N 9°04'06.33"O

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?

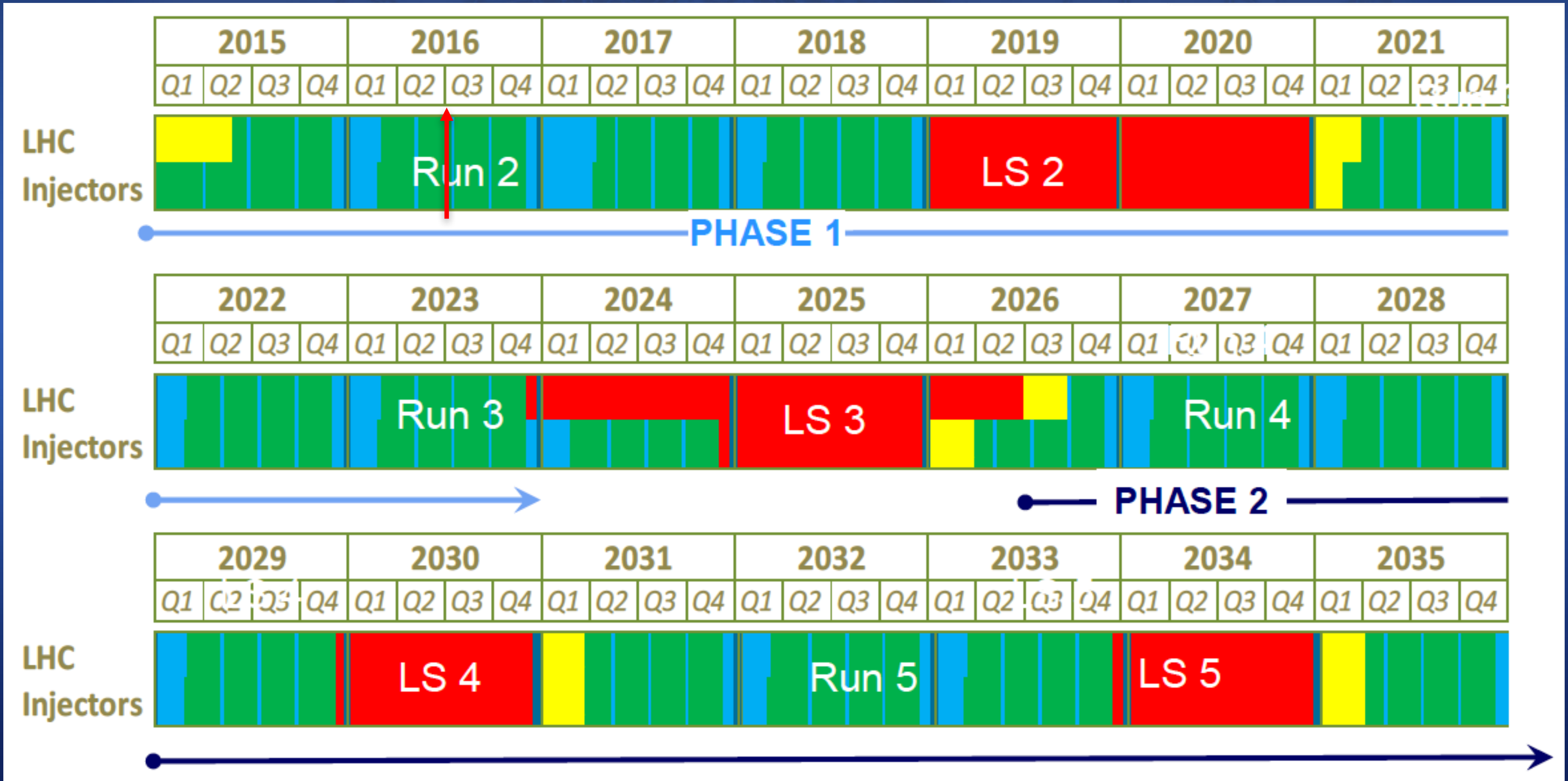
Commissione Calcolo Reti

July 4th 2012

[credits: D.Bonacorsi]

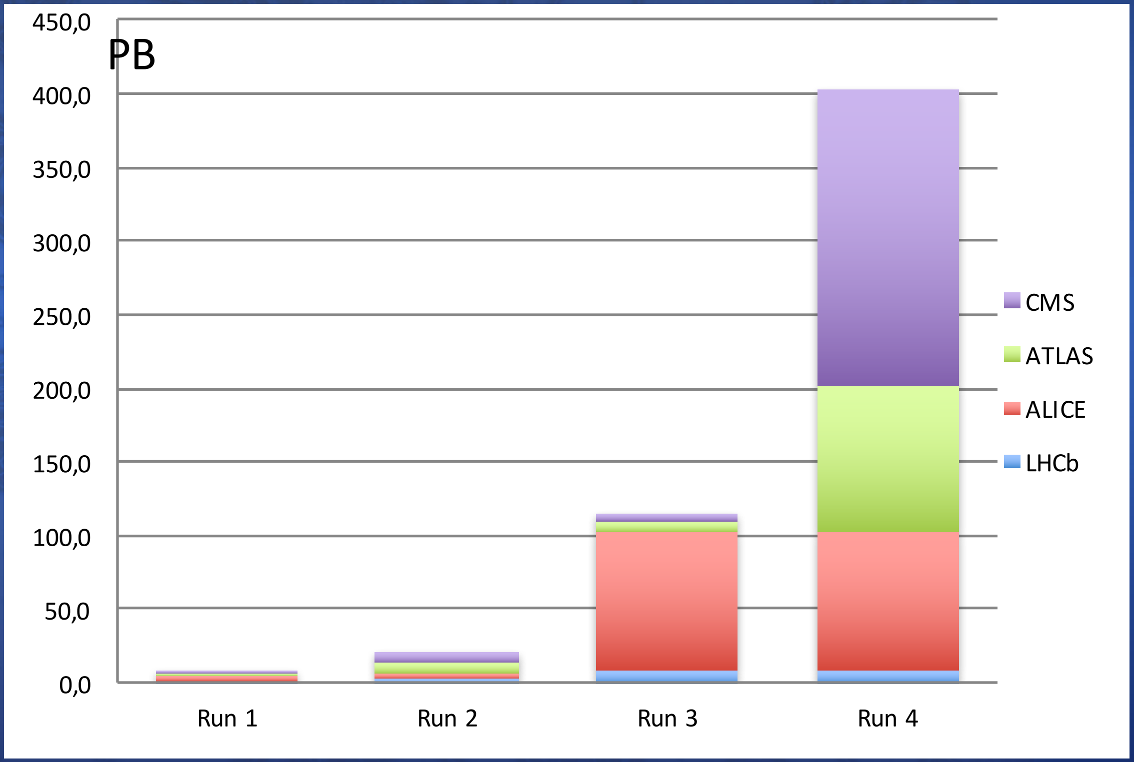
What about the years to come?

LHC roadmap


2015
2016
2017
2018
2019
2020
2021
Q1 Q2 Q3 Q4
LHC
Injectors
Run 2
LS 2
PHASE 1
2022
2023
2024
2025
2026
2027
2028
Run 3
LS 3
Run 4
PHASE 2
2029
2030
2031
2032
2033
2034
2035
LS 4
Run 5
LS 5

Resource requests for the future

Significant increase in experiments' requests in the coming years: the "scary plot"



…but the buzz-word is "flat-budget"!

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*

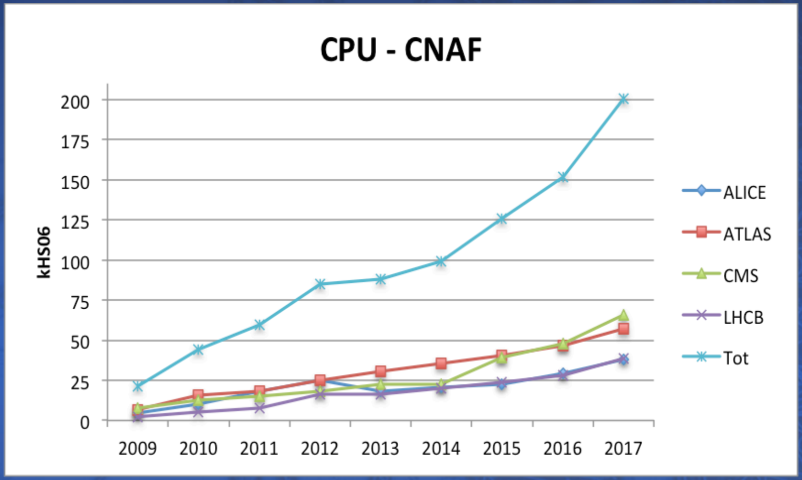
In addition the Cherenkov Telescope Aarray experiment starts!

Italian resources in 2016

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

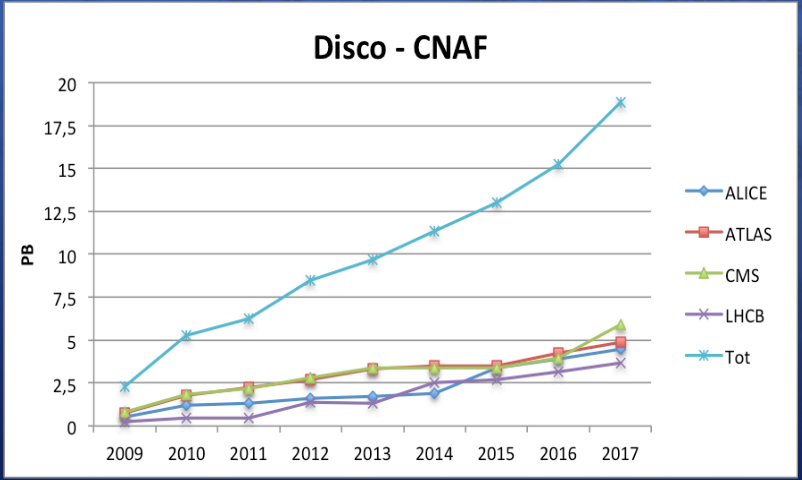
	CPU (kHS06)	Disk (PB)	Tape (PB)
WLCG	2900	240	250
INFN	306	30	35
% INFN	11	12	14

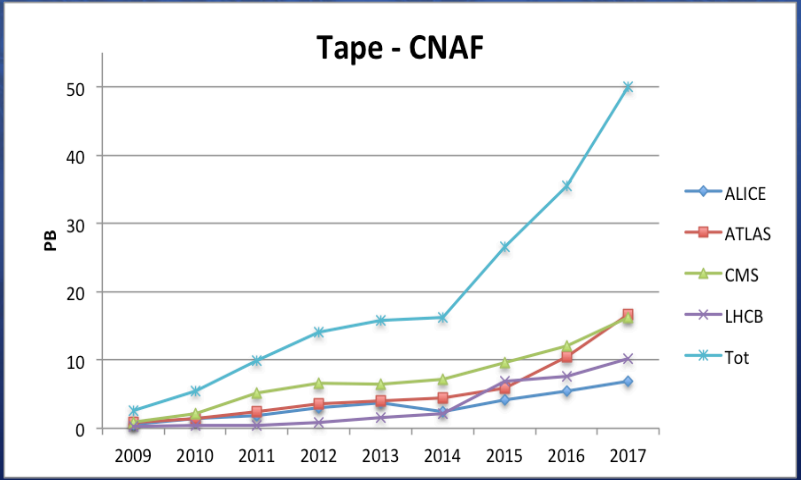
CNAF evolution - LHC Run 1 & 2



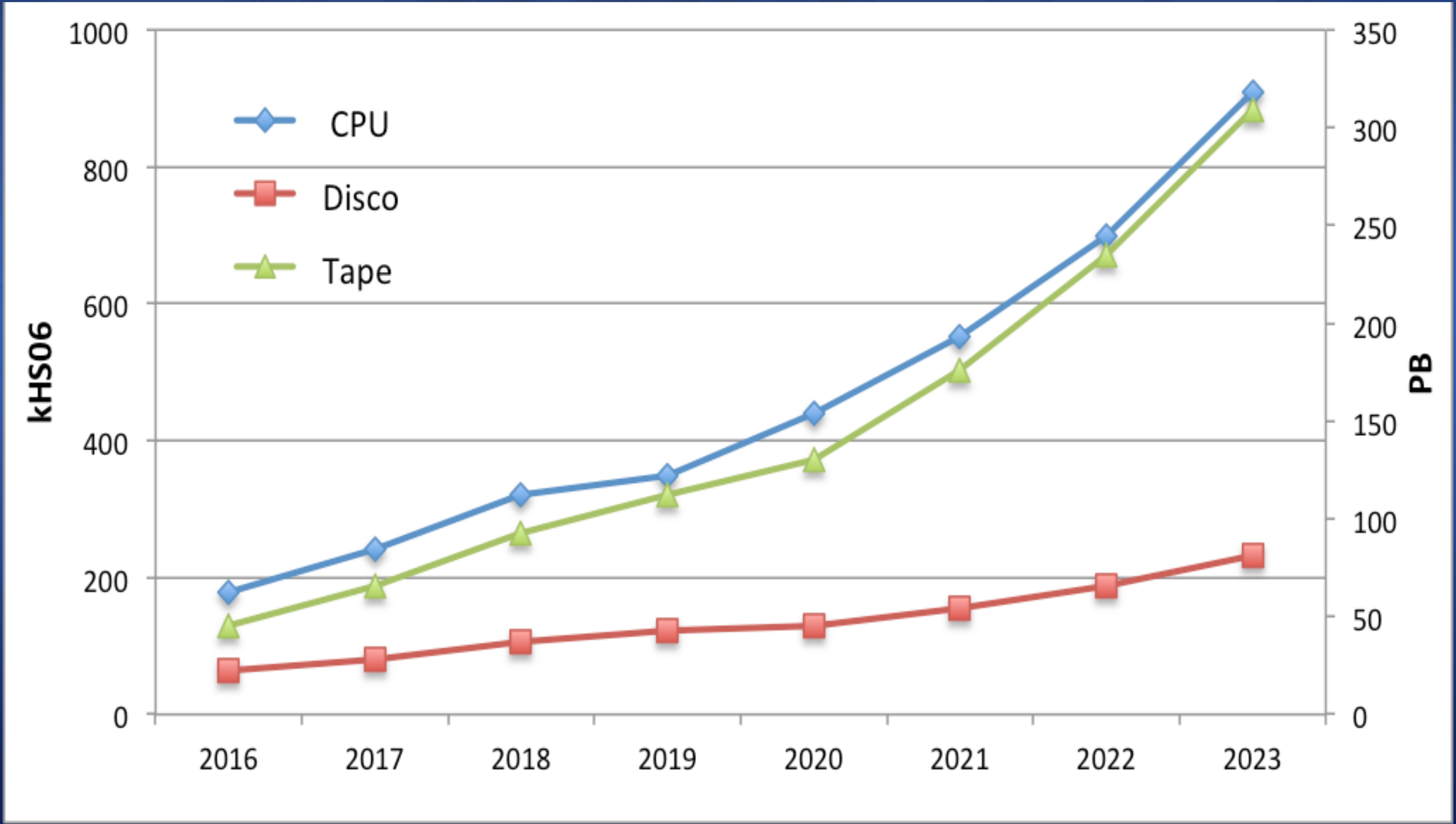
Run2 is ok with the flat budget hypothesis:

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





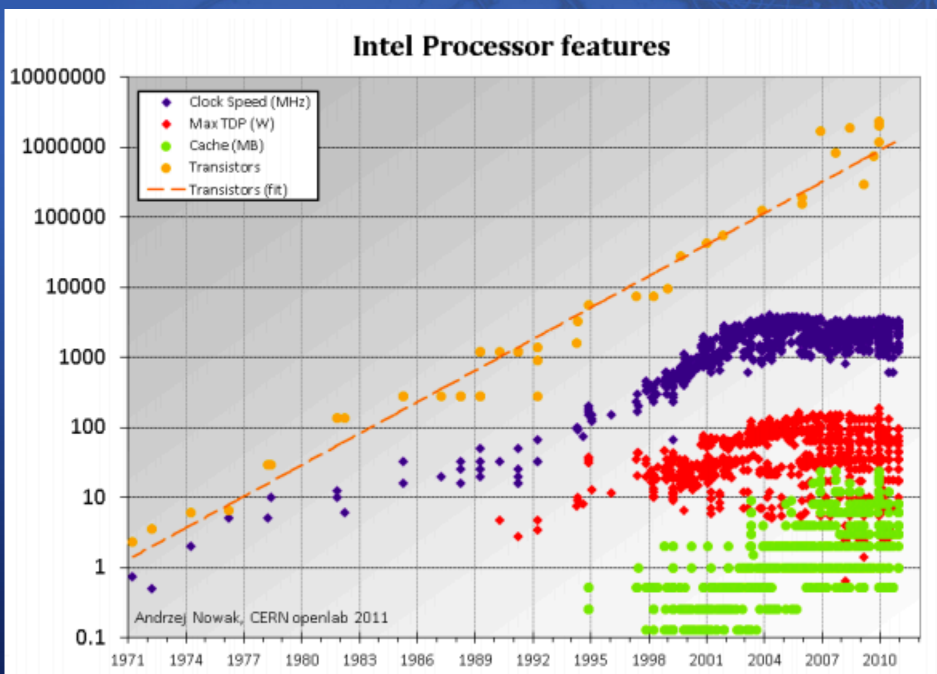
CNAF evolution up to LHC Run 3



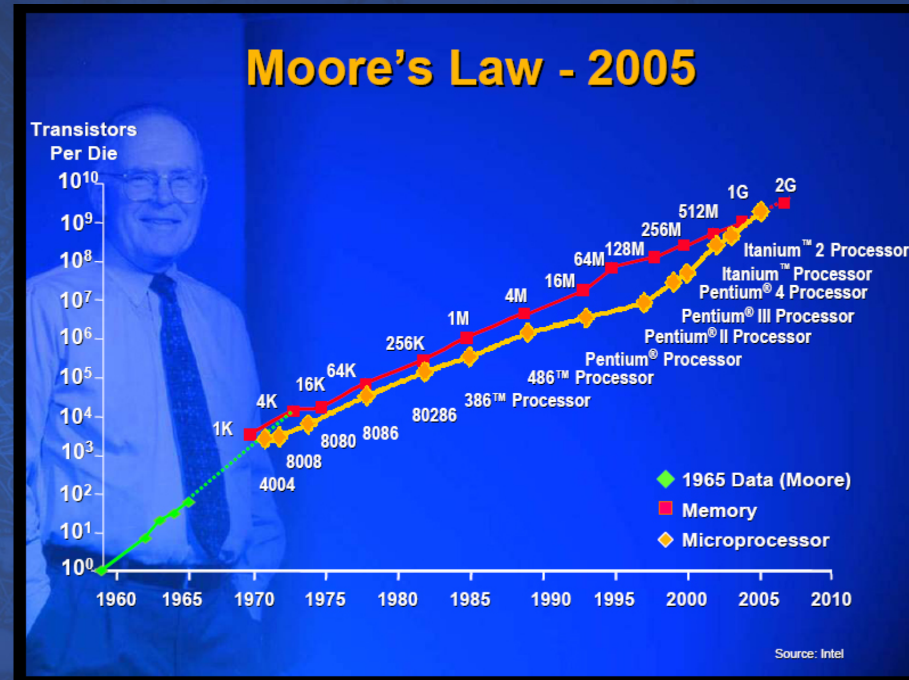
Does the technological evolution help?

CPU power

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

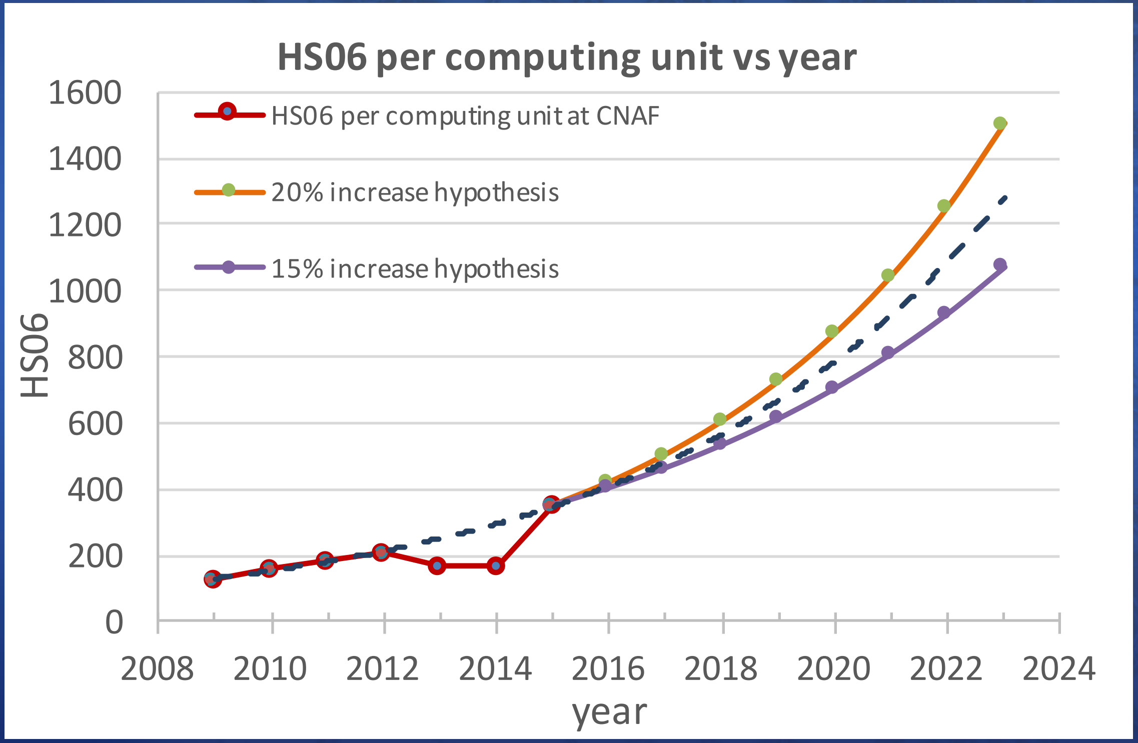


Source: Andrzej Nowak – CERN OpenLab



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

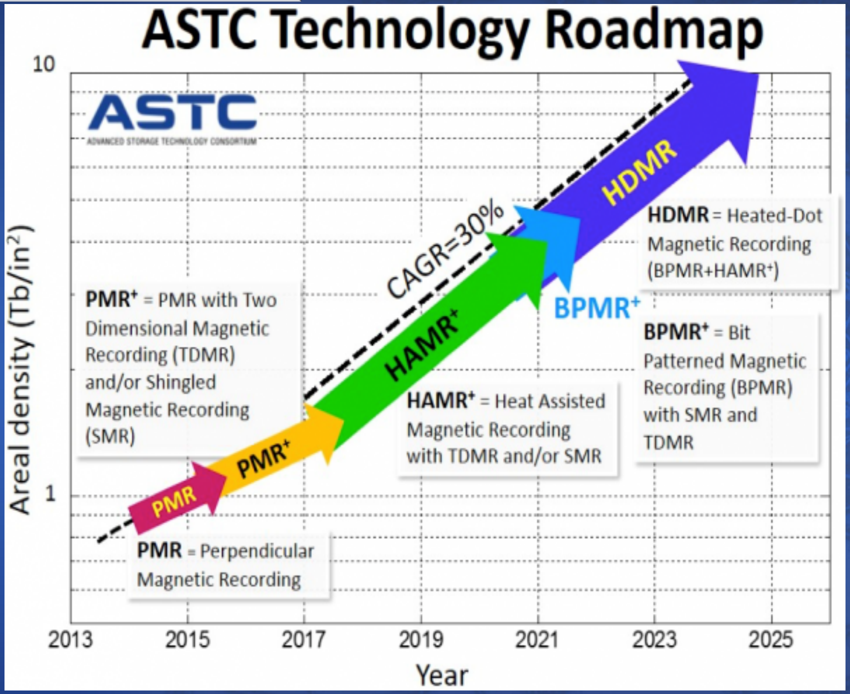
CPU power

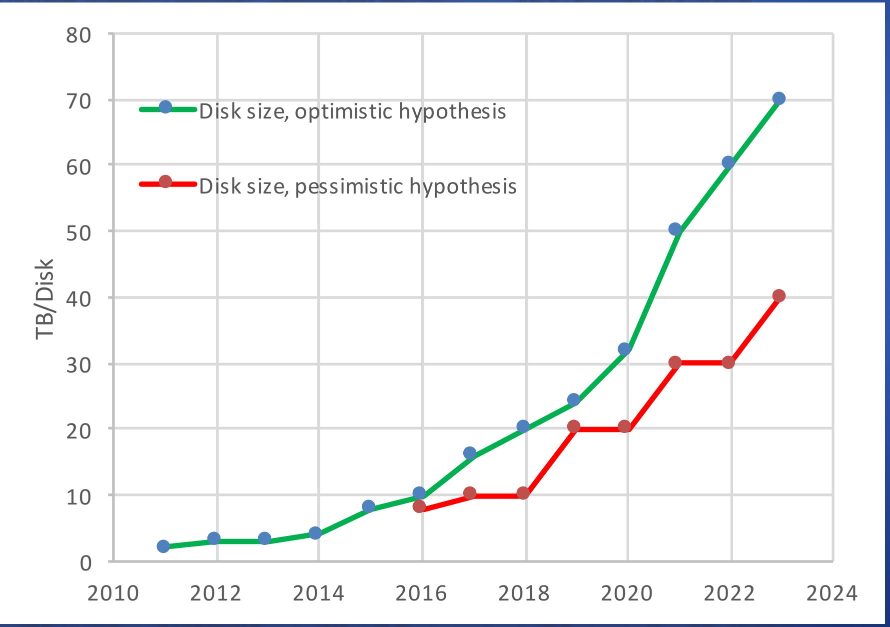


Starting from the actual power of the nodes bought by CNAF in 2009-2015 we estimate an increase between 15 and 20%

Disk

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

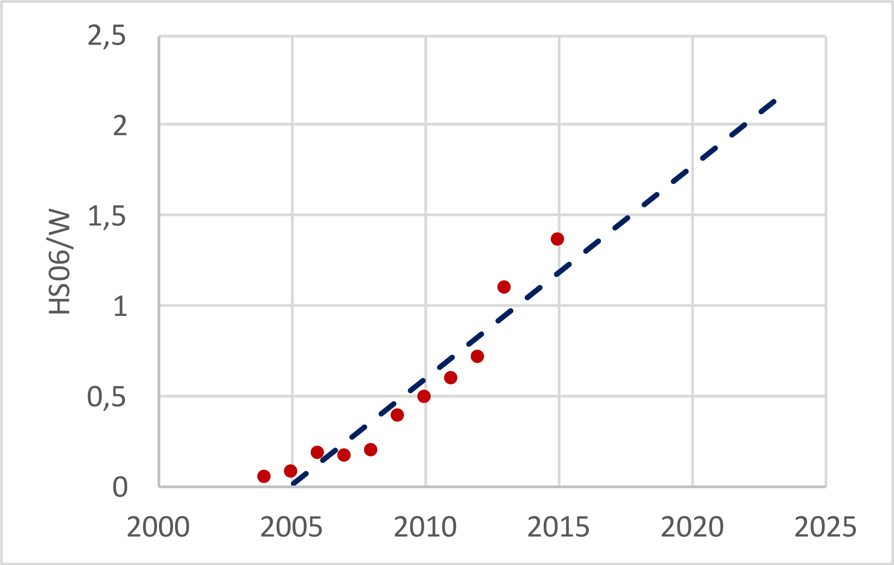




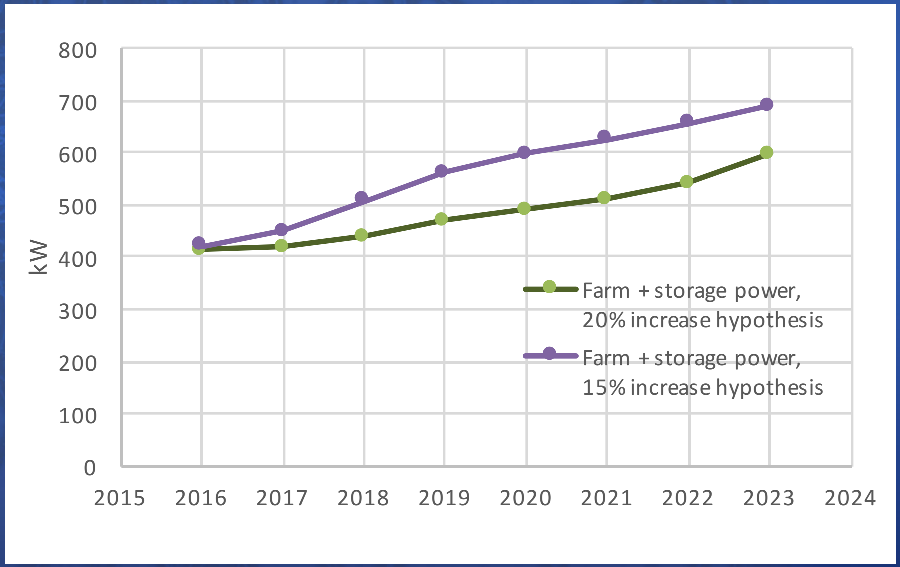
Extrapolation is more difficult for disk because there are technology changes foreseen

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*

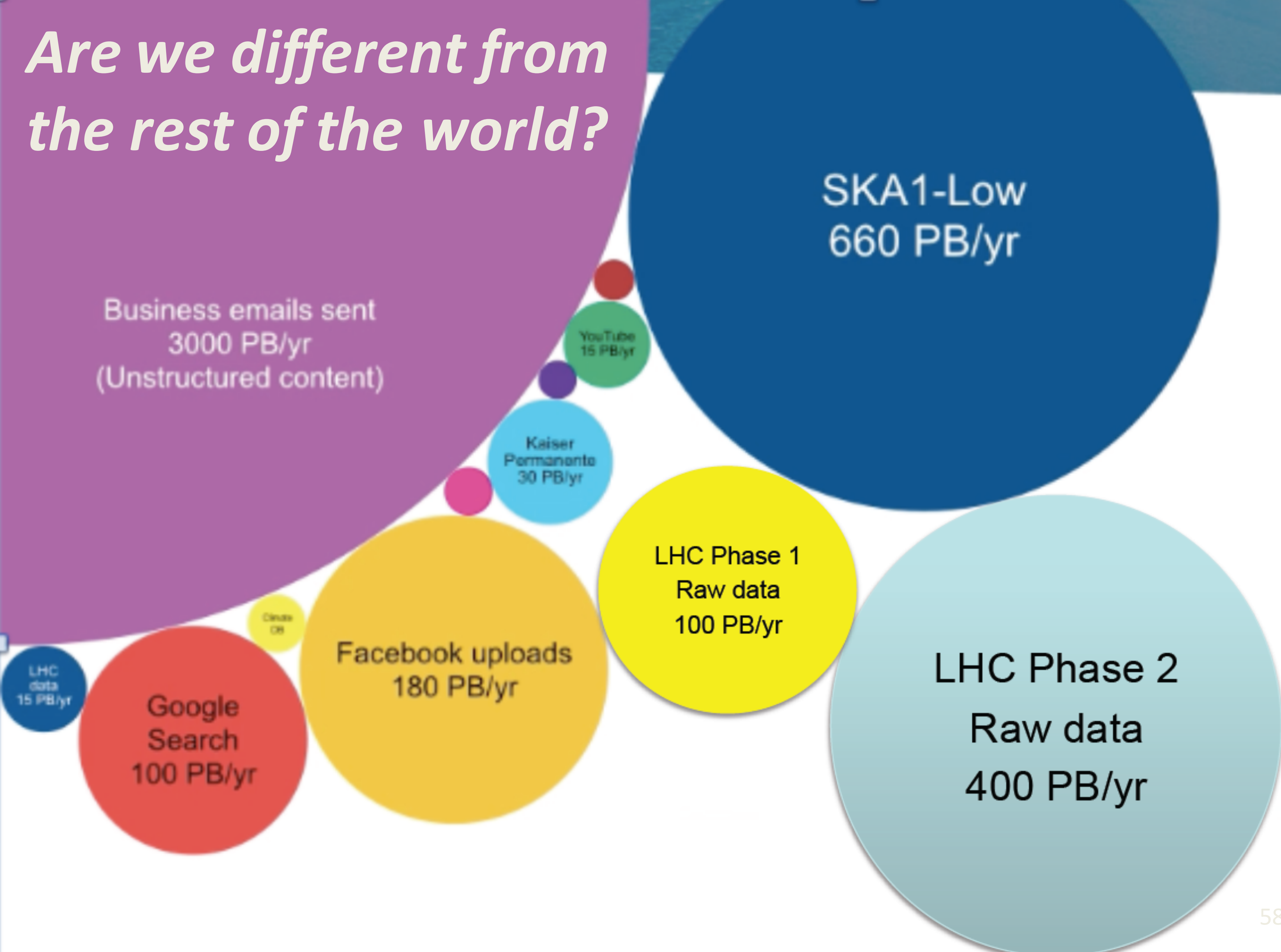
→ Far from a "flat budget" hypothesis for Run3

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

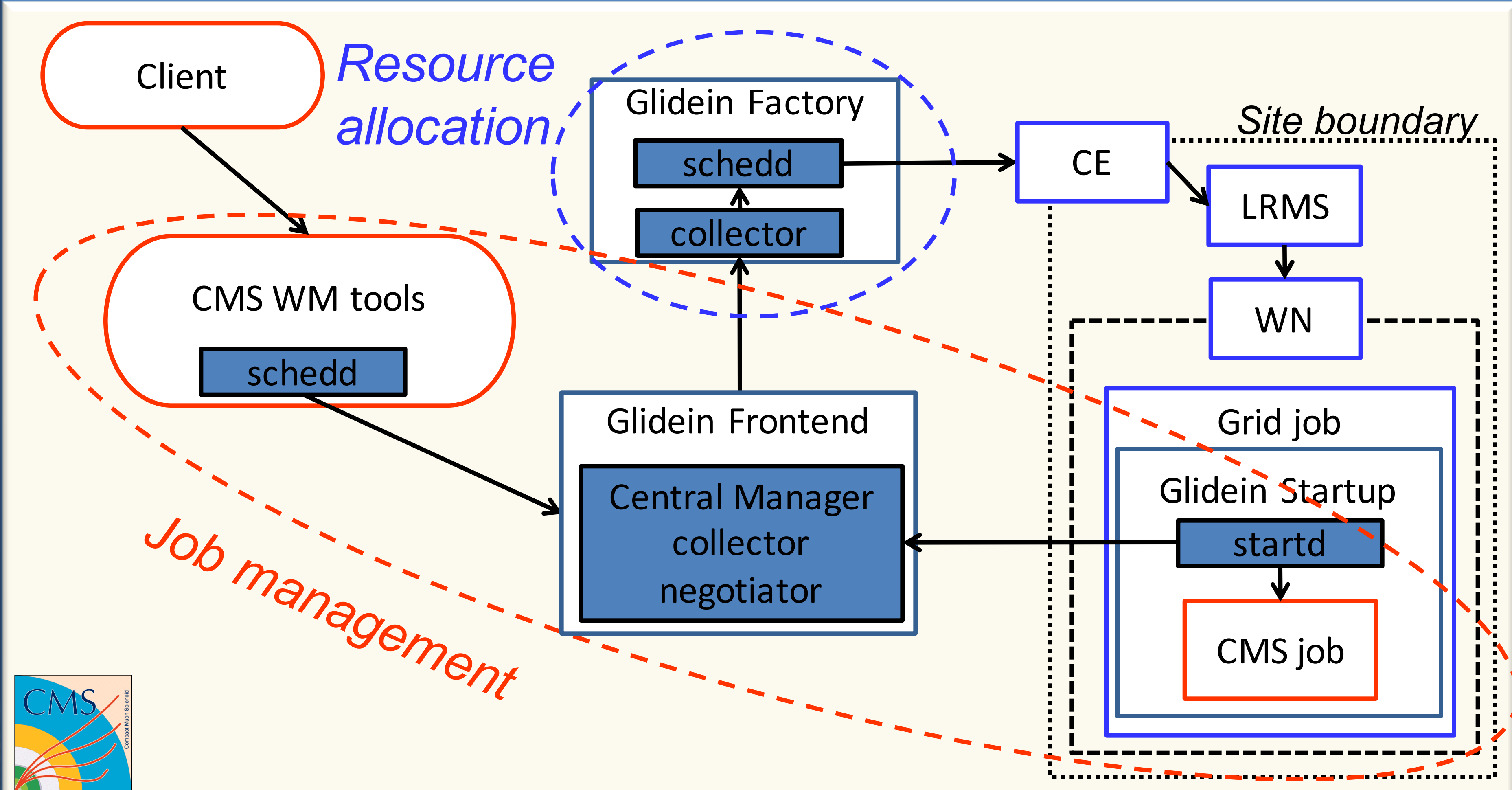
Cloud Computing offers most of the functionalities needed by HEP computing

Commercial and industrial world offers solutions that are being integrated

Actually there is a lot of Grid in the Cloud!

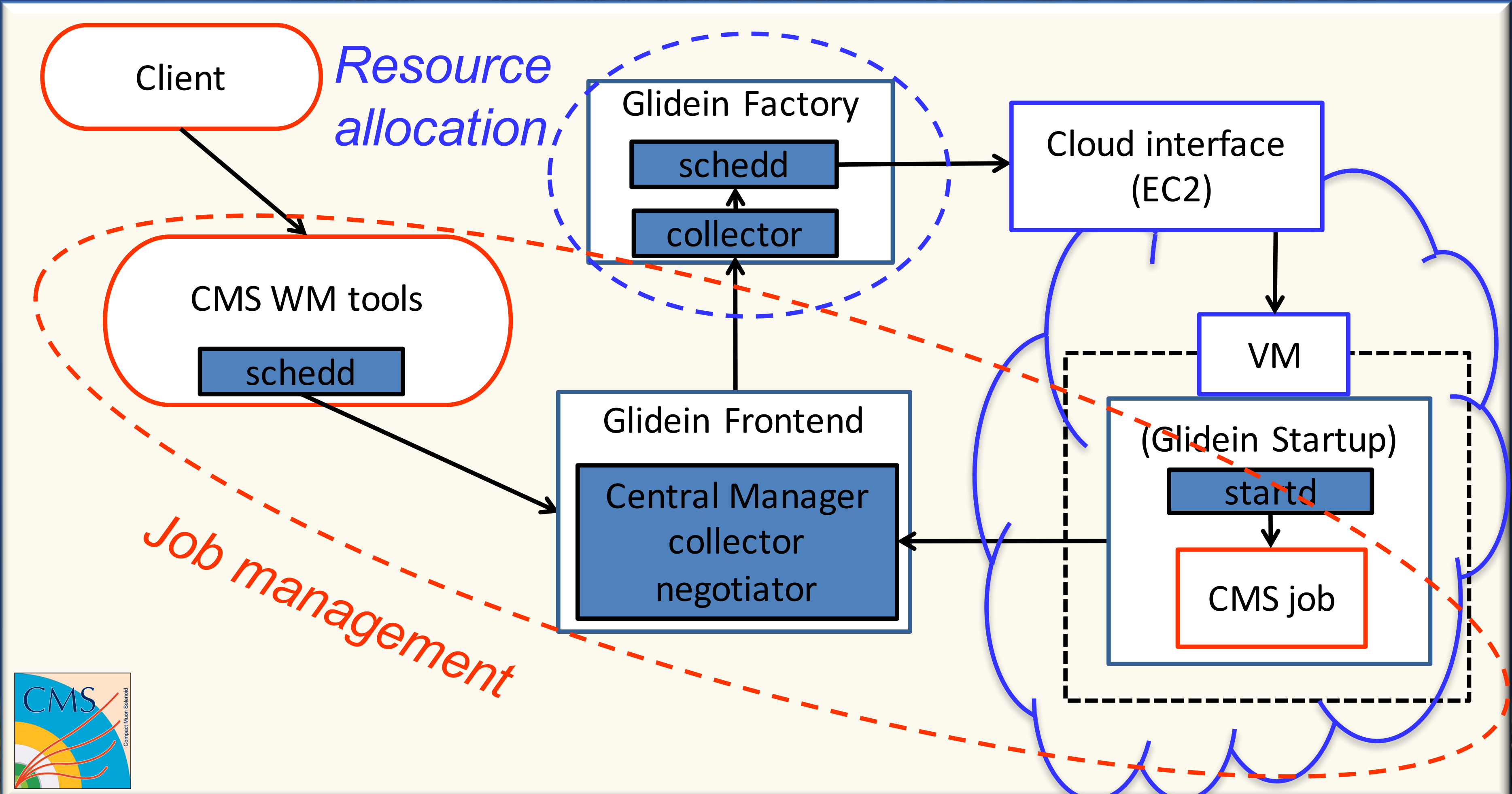
From the Grid…

The "factory" harvests job slots



…to the Cloud

The "factory" harvests machines (or containers)



Extension of a Computing Centre

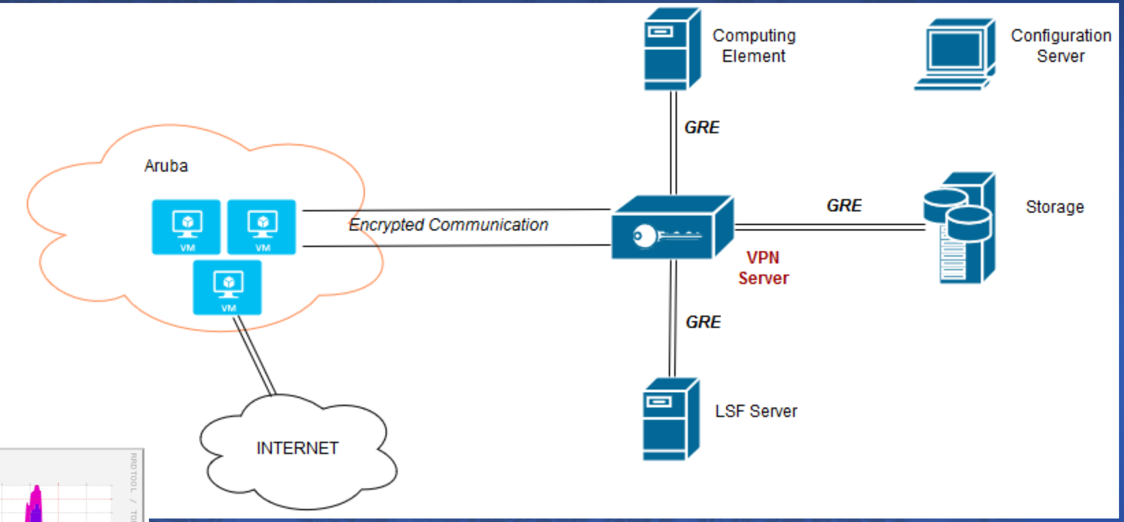
CNAF tested the ability to extend the computing center on external resources:

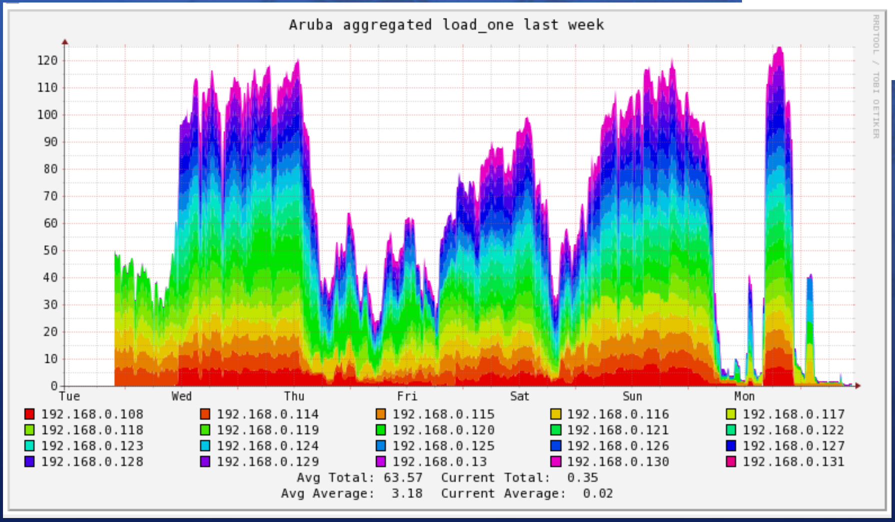
- *Opportunistic: with transient Aruba resources (Arezzo)*
- *Structural: ReCaS/Bari: extension and management of remote resources*
  - *These will become pledged resources for CNAF*

Extension on external resources

Use of gateways and caches over secure channels allowed to treat external resources as if internal

Remote access to data

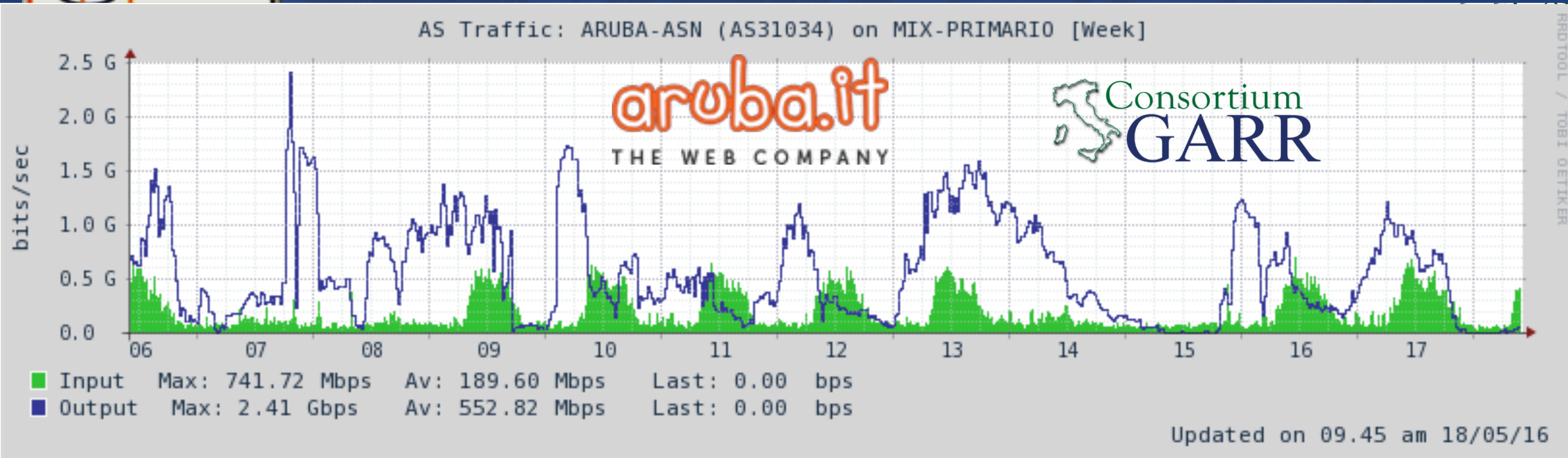


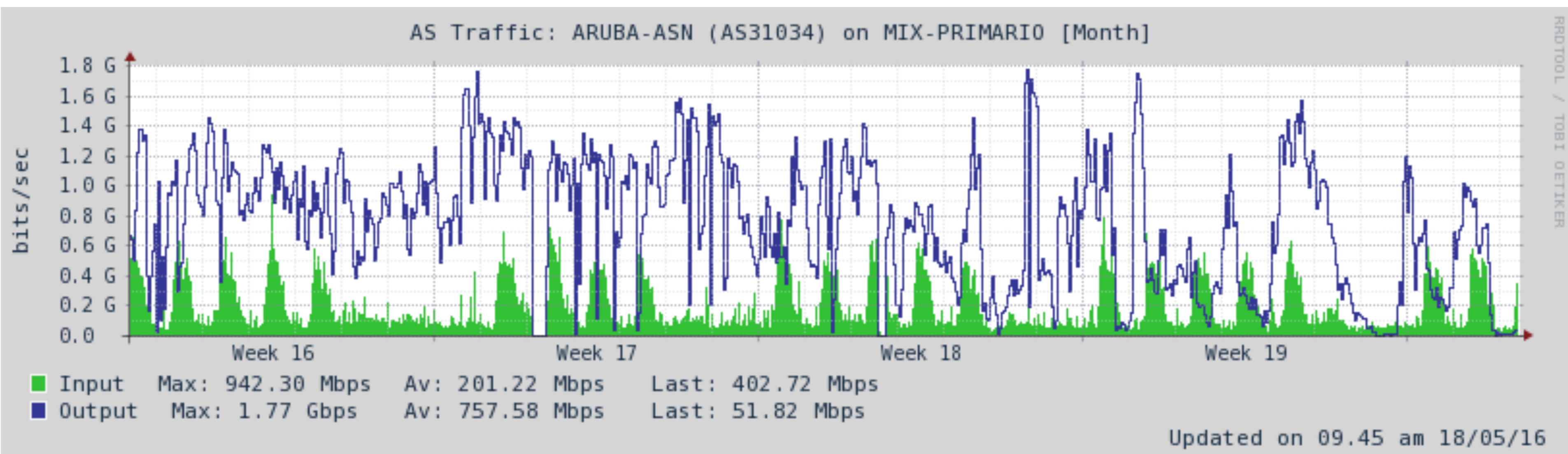


First tests on Aruba with CMS in 2015

Very good efficiency for specific jobs (Monte Carlo)

Aruba - network traffic





Remote management of resources

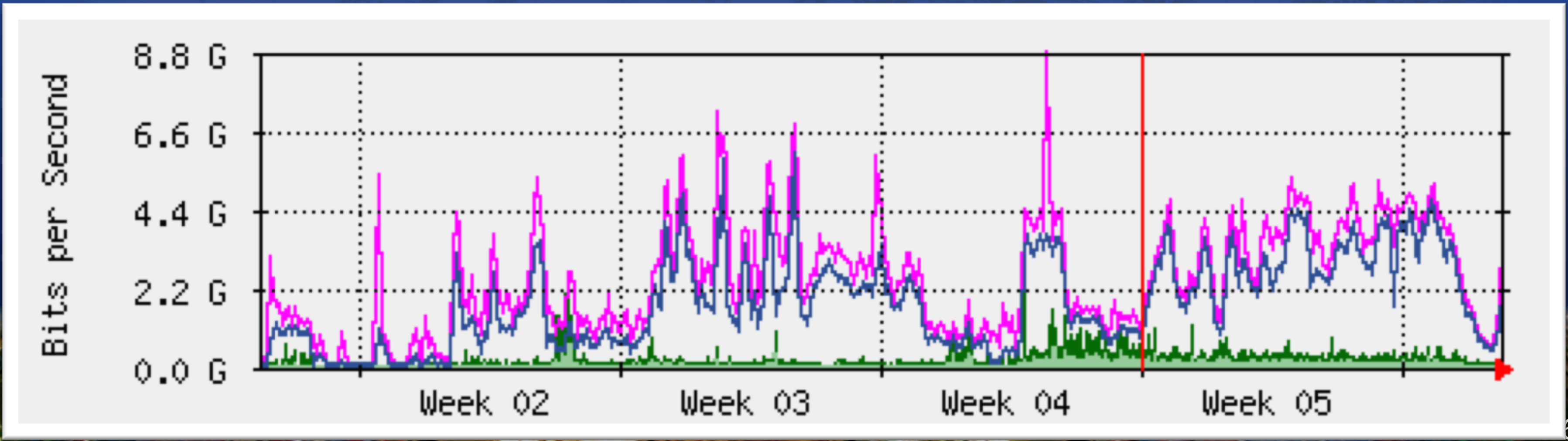
In principle require "transparent use" and performances comparable with a local execution

Special network configuration between CNAF and ReCaS Bari:

- *Level 3 Virtual Private Network and 20 Gb/s dedicated bandwidth*

Use of caches for data apparently not performing enough

- *Better results with remote access to data*



Commissione Calcolo Reti

Comparative Results

Queue	Nodetype	Njobs	Avg_eff	Max_eff	Avg_wct	Avg_cpt
Cms_mc	AR	2984	0.602	0.912	199.805	130.482
Alice	T1	98451	0.848	0.953	16.433	13.942
Atlas_sc	T1	1211890	0.922	0.972	1.247	1.153
Cms_mc	T1	41412	0.707	0.926	117.296	93.203
Lhcb	T1	102008	0.960	0.985	23.593	22.631
Atlas_mc	T1	38157	0.803	0.988	19.289	18.239
Alice	BA	25492	0.725	0.966	14.446	10.592
Atlas	BA	15263	0.738	0.979	1.439	1.077
Cms_mcore	BA	2261	0.444	0.805	146.952	69.735
Lhcb	BA	13873	0.916	0.967	12.998	11.013
Atlas_sc	BA	20268	0.685	0.878	24.378	15.658

Aruba (AR) — CNAF (local) (T1) — ReCaS Bari (BA)

Presented at ISGC 2016
By S. Dal Pra

New architectures

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

→ *Follow the market mainstream*

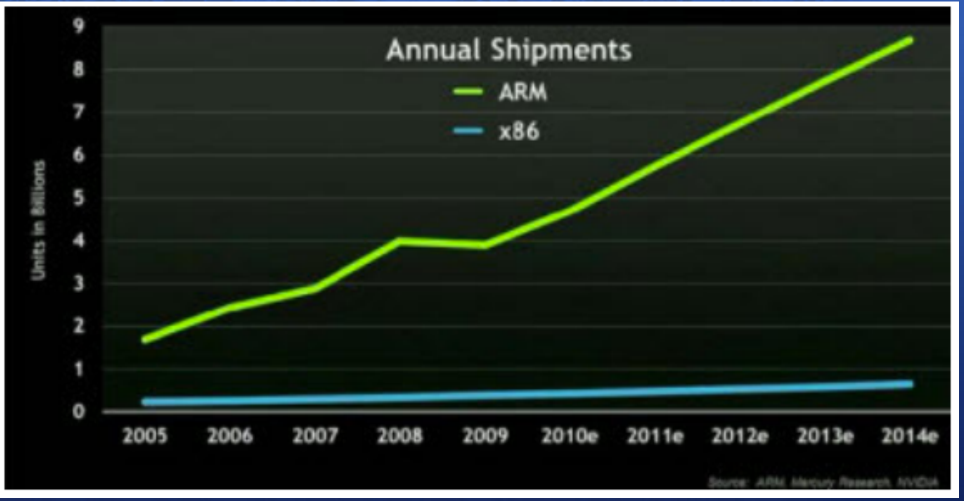
→ *Use highly available architectures*

ARM, …

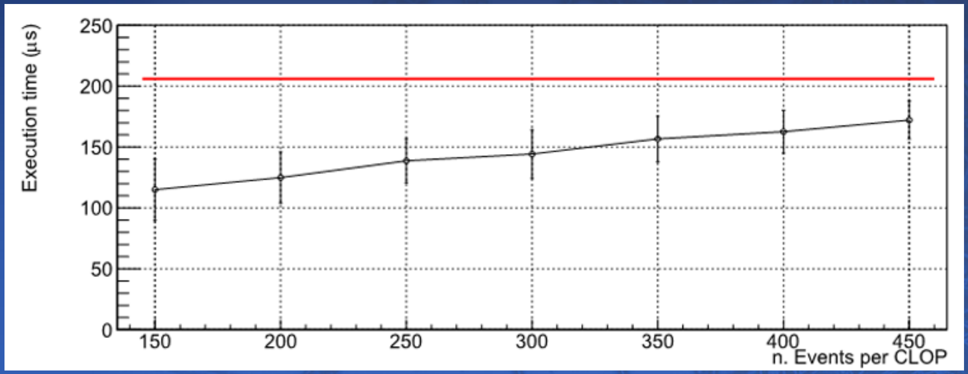
→ *Exploit parallelization*

Multi/many-core, GPGPU, …

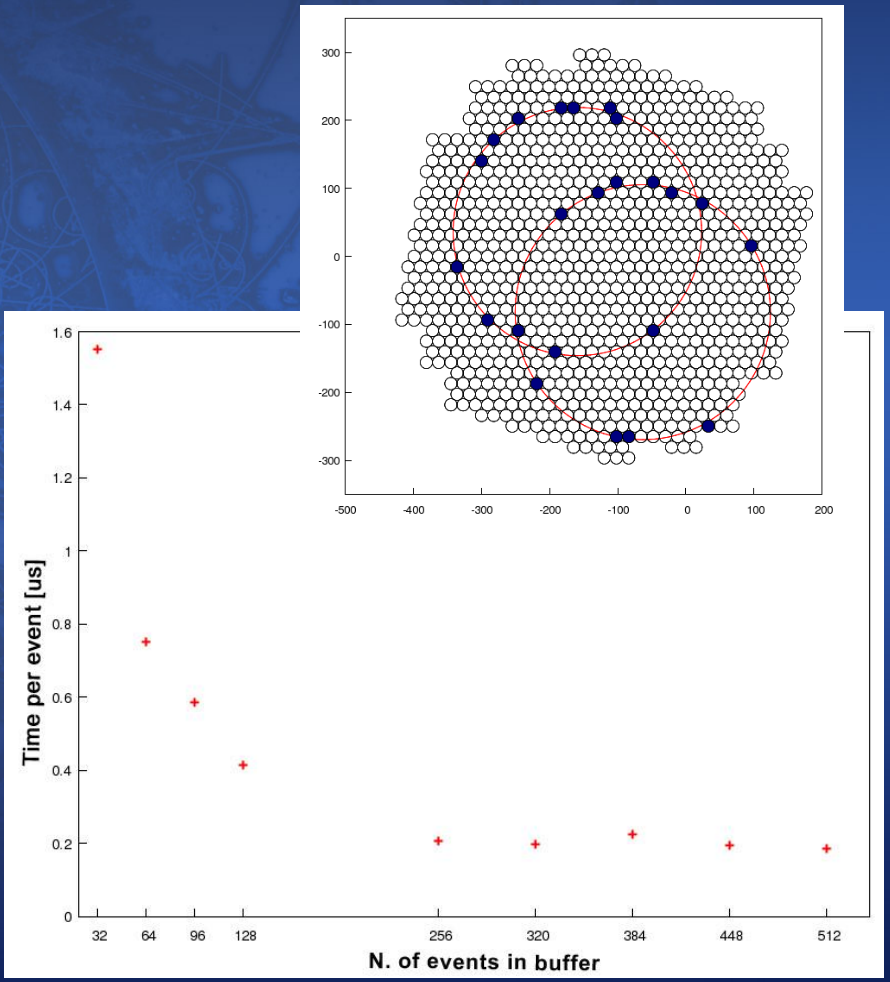
→ *Use low-power architectures*



NA64 RICH pattern recognition

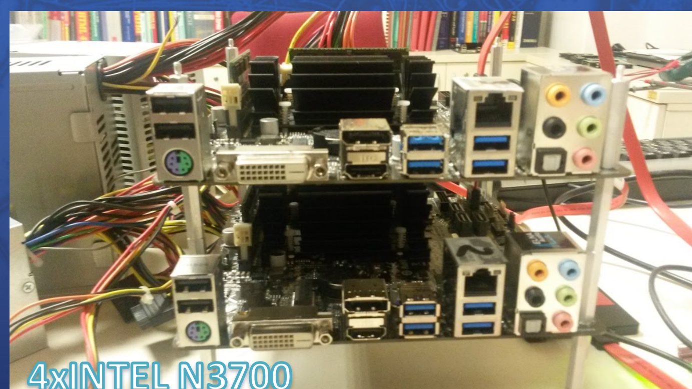


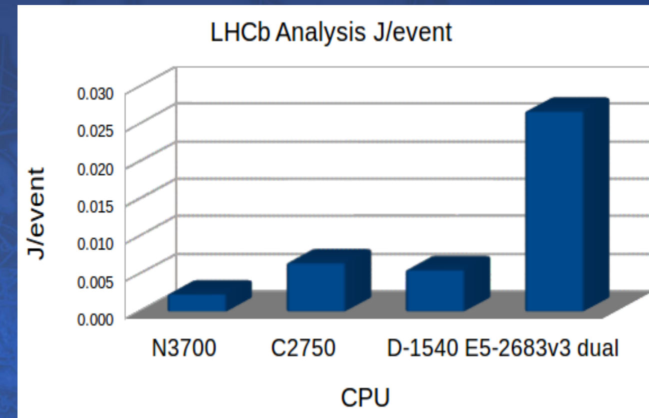
- *Algorithm parallelization*
  - *Almagest*
- *Execution on NaNet-10*
  - *Based on GPU the Tesla K20c GPU*

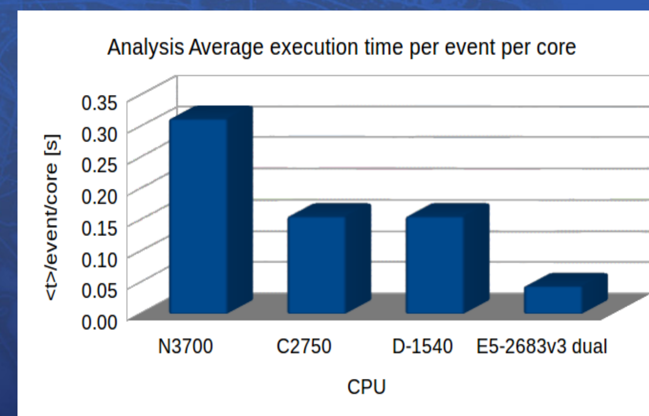


Test on LHCb analysis

Example of tests done by the INFN-COSA project on a small testbed at CNAF on Intel low power systems







CPU	Brand	Microarchitecture	Family	#	CORES	RAM (GB)	POWER (W)	HS06	HS/W
E5-2683v3	XEON	Haswell	(Reference)	2	56 (HT)	128	370	573	1.55
D-1540	XEON	Broadwell		1	16 (HT)	16	80	151	1.89
C2750	ATOM	Silvermont	Avoton	1	8	16	20	55	2.50
N3700	PENTIUM	Airmont	Braswell	1	4	16	7	28	4.00

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

The world-wide and the Italian communities are very active

There is room for new ideas and innovative projects!