LHC computing

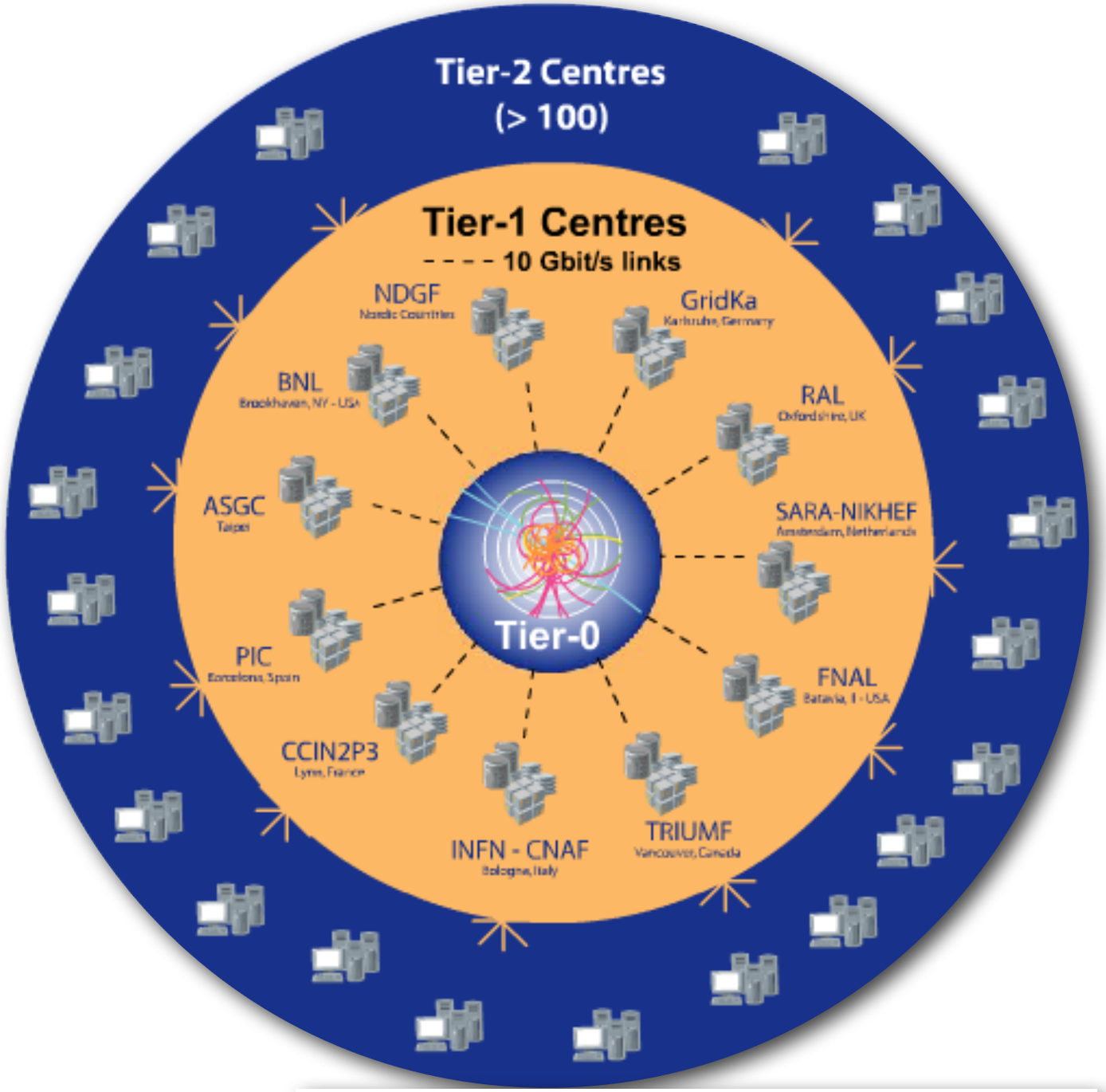
- History
- Current status
- Plans for LS1 and beyond
- The French contribution

Eric Lançon

Pre-History

Centric model

- Derived from MONARC (’99) model
- CERN-T0 the center
- 11 T1s, linked by dedicated 10Gb links (LHCOPN)
- >200 T2s each attached to a T1
- The data flows along the hierarchy
- Assumes poor networking
- Hierarchy of functionality and capability



Fear of the networks!

Pre-planned data distribution

Jobs-to-data brokerage

Central organization of systems and services

http://monarc.web.cern.ch/MONARC/

Models of Networked Analysis at Regional Centres for LHC Experiments

Evolutions

The LHC Distributed Computing & the Grid was doing very well

1,000's of users processing petabytes of data with > 1M jobs/day*

But at the same time, hitting some limits:

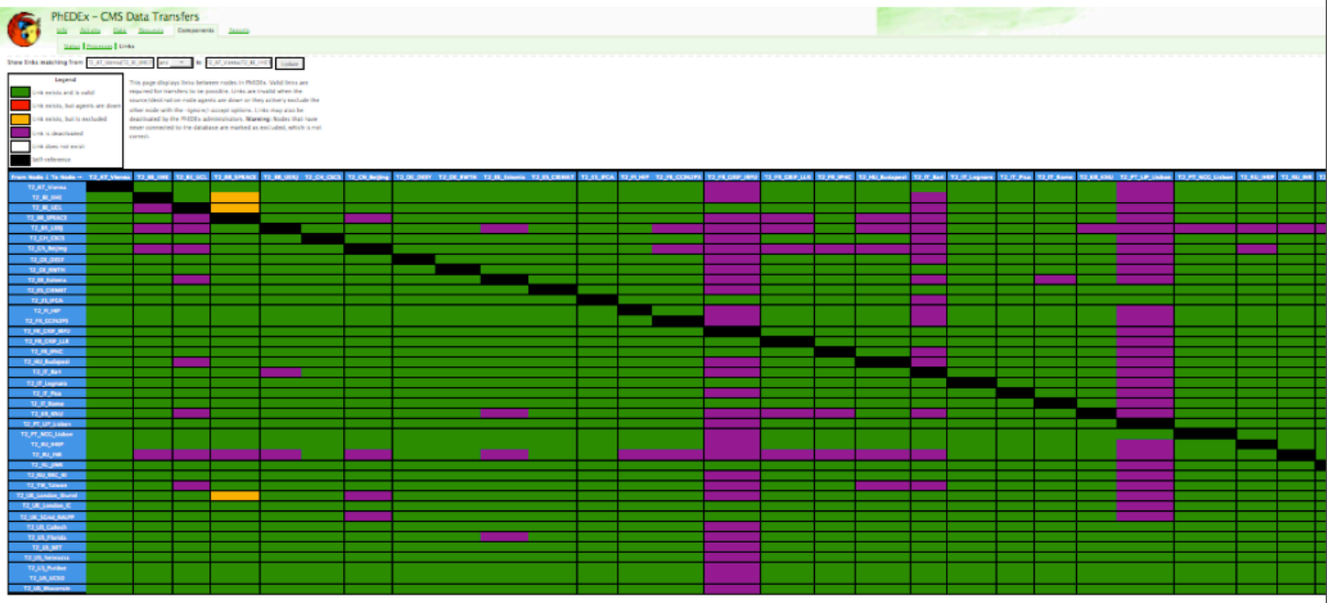
Scaling up, elastic resource usage, global access to data

(*) http://en.wikipedia.org/wiki/Petabyte
Internet: Google processed about 24 petabytes of data per day in 2009
At its 2012 closure of file storage services, Megaupload held ~28 petabyte of user uploaded data
Telecoms: AT&T transfers about 30 petabytes of data through its networks each day

Some of the many changes

- Hide grid complexity to users, **simplifications**, less middleware dependance, pilot jobs : pull model
- **Caching** opposed to centralized DB
  - Conditions data access from any site, not only at Tier-1s
  - No more need to pre-install software releases at sites
- **Dynamic** data placement and deletion based on **popularity**
  - Better usage of disk space
  - Reduced job waiting times
- **T2-T2 exchanges**

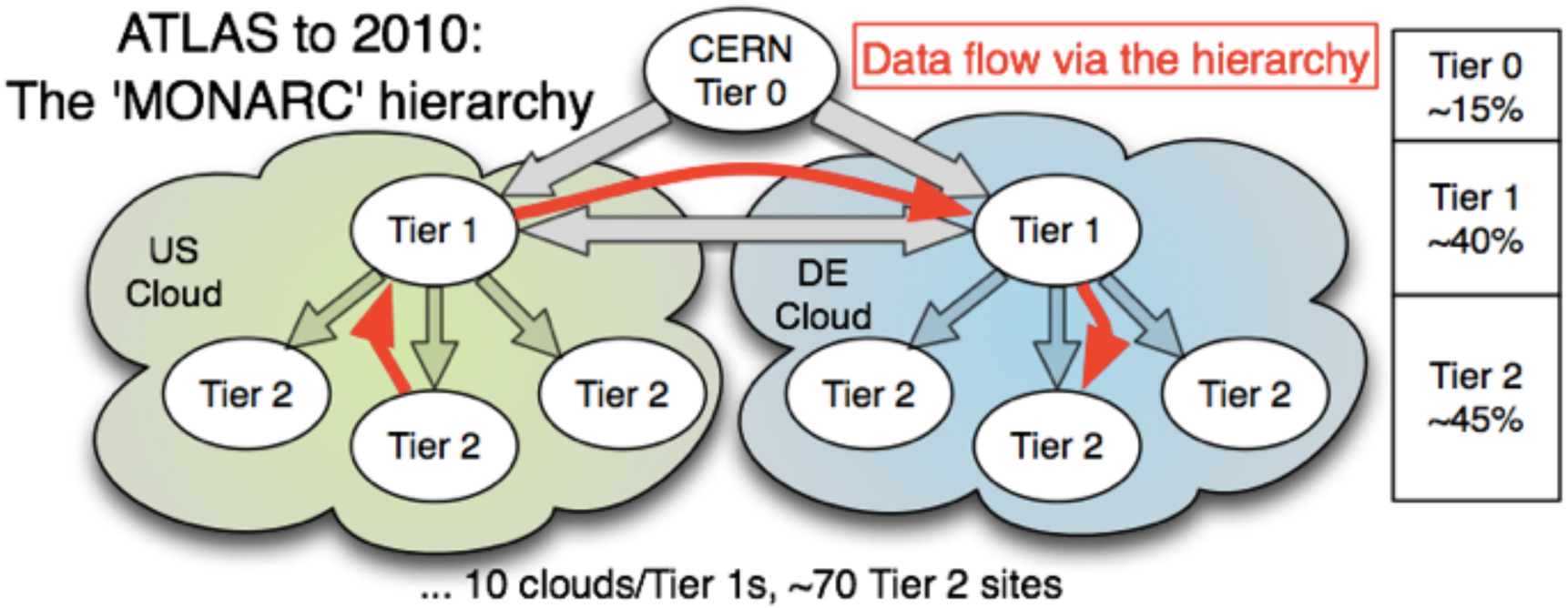
CMS T2-T2 mesh testing



Network performing over expectations

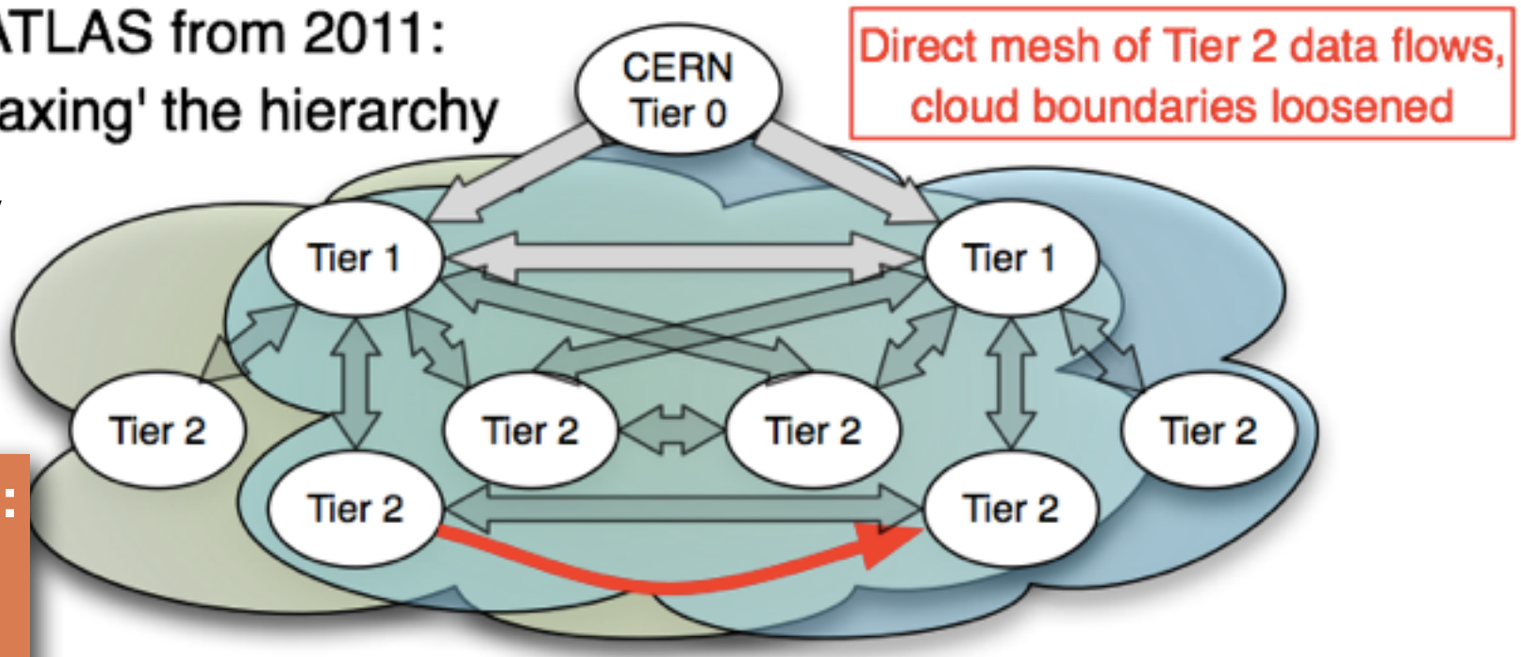
2011 LHCONE :
Dedicated network between (some) T2s

Computing Model Evolution in ATLAS



Originally:
Static, strict hierarchy
Multi-hop data flows
Lesser demands on
Tier 2 networking
Virtue of simplicity

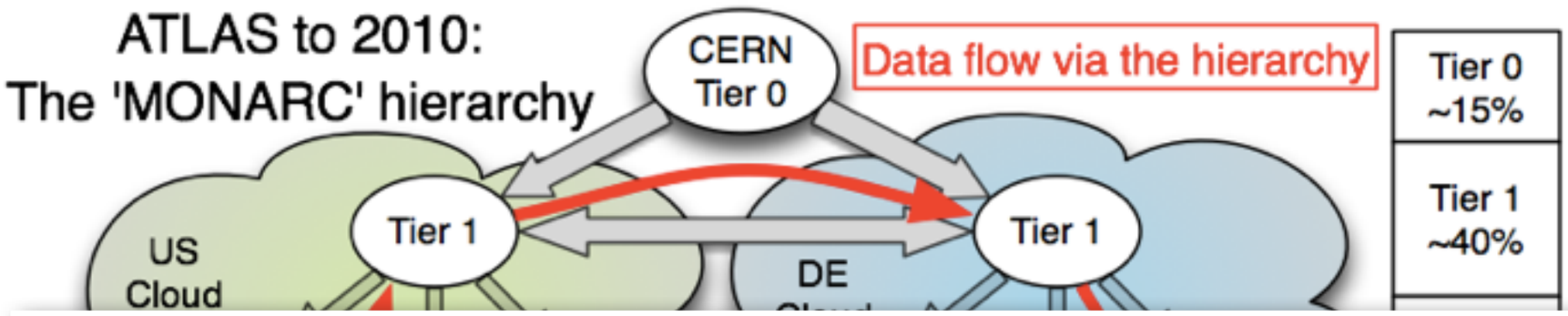
Today:
Flatter, more fluid, mesh-like
Sites contribute according to capability
Greater flexibility and efficiency
More fully utilize available resources



Principal enabler for the evolution: the network
Excellent bandwidth, robustness, reliability, affordability

©T. Wenaus

Computing Model Evolution in ATLAS



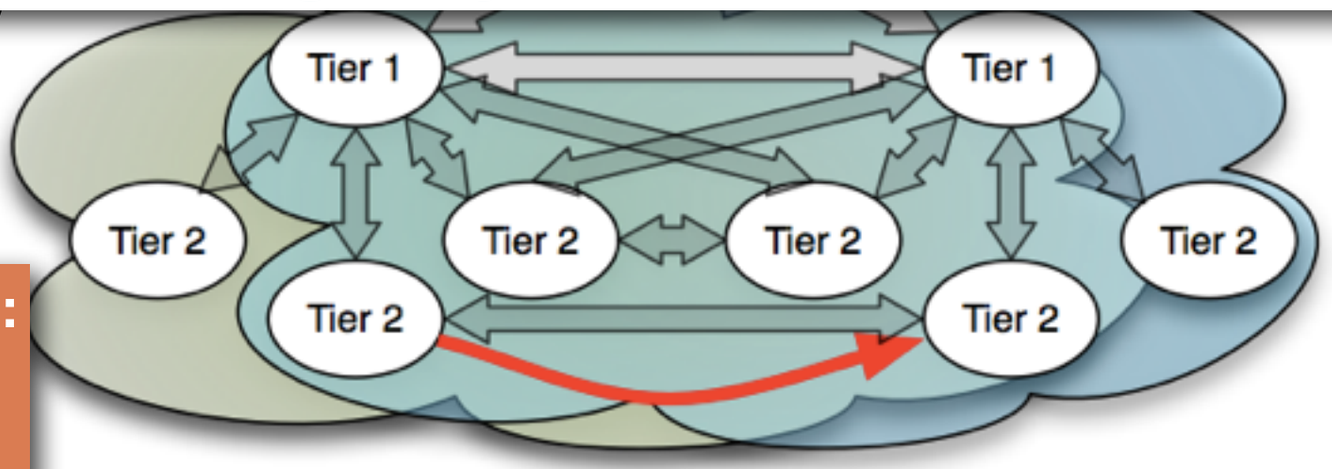
Originally:
Static, strict hierarchy
Multi-hop data flows
Lesser demands on
Tier 2 networking

Any site can use any other site as source of data

Analysis sites pull data from other sites “on demand”

User Analysis: Job-Splitting beyond cloud boundary

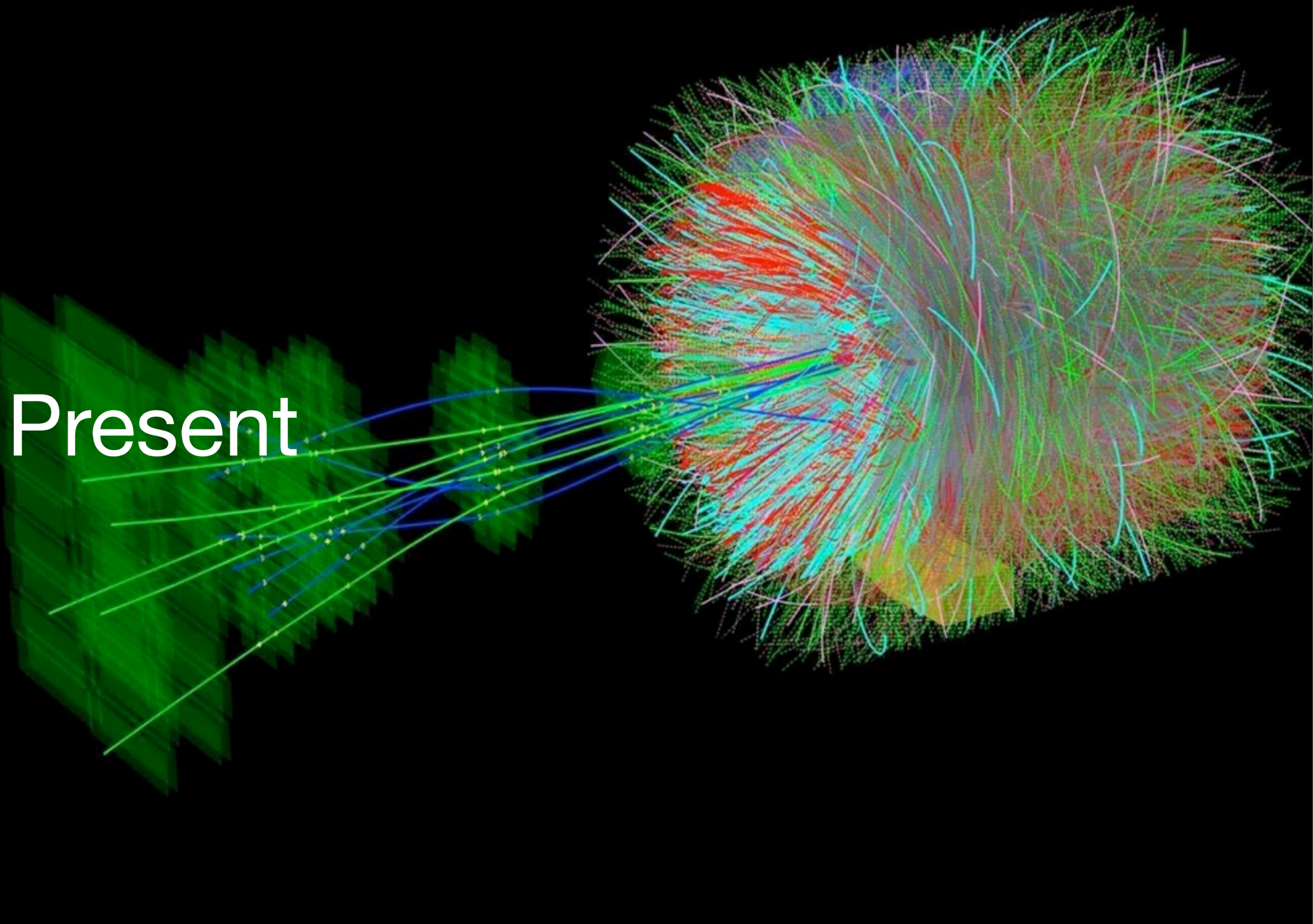
Sites contribute according to capability
Greater flexibility and efficiency
More fully utilize available resources



Principal enabler for the evolution: the network
Excellent bandwidth, robustness, reliability, affordability

©T. Wenaus

Present



July 4, 2012

Global Effort → Global Success

Results today only possible due to extraordinary performance of accelerators – experiments – Grid computing

Observation of a new particle consistent with a Higgs Boson (but which one…?)

Historic Milestone but only the beginning

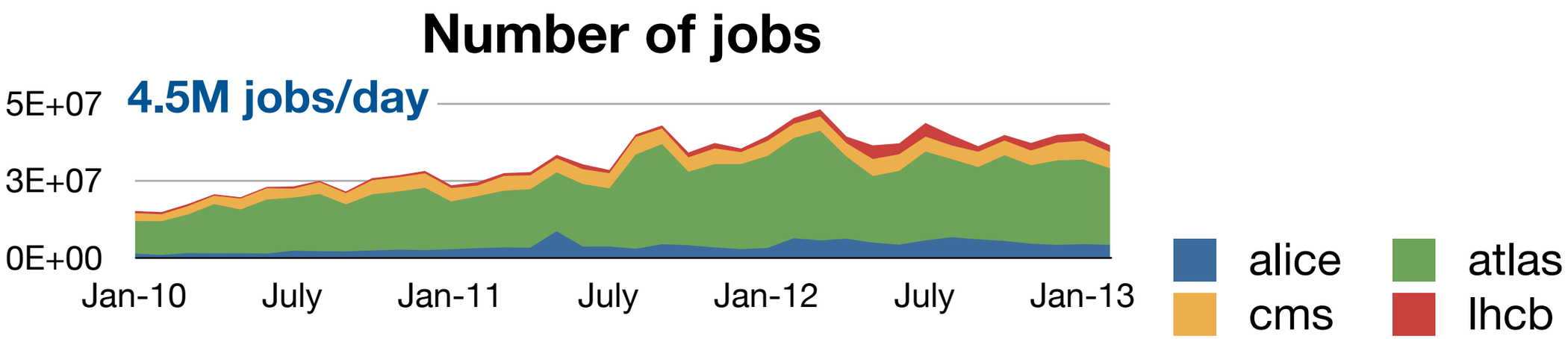
Global Implications for the future

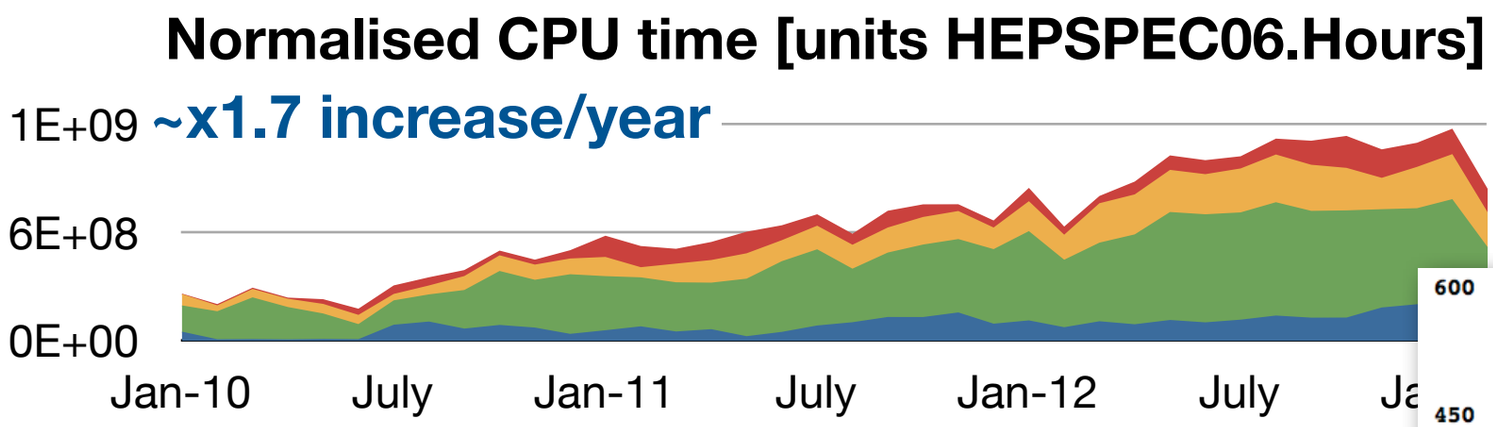
R-D Heuer

CERN

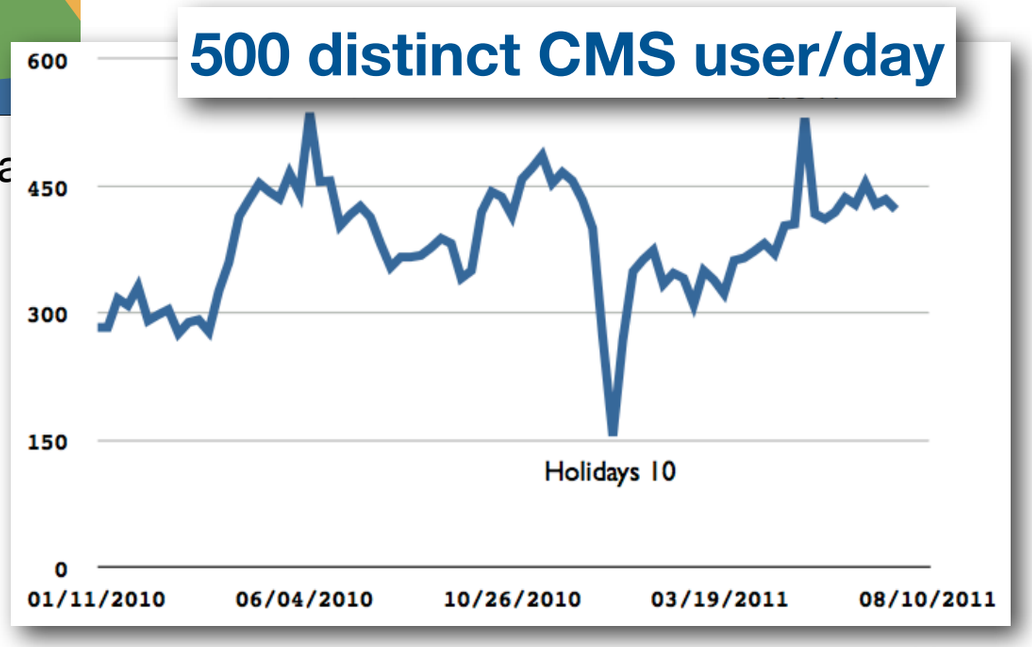
It works ! beyond expectations...

Jobs & CPU consumption



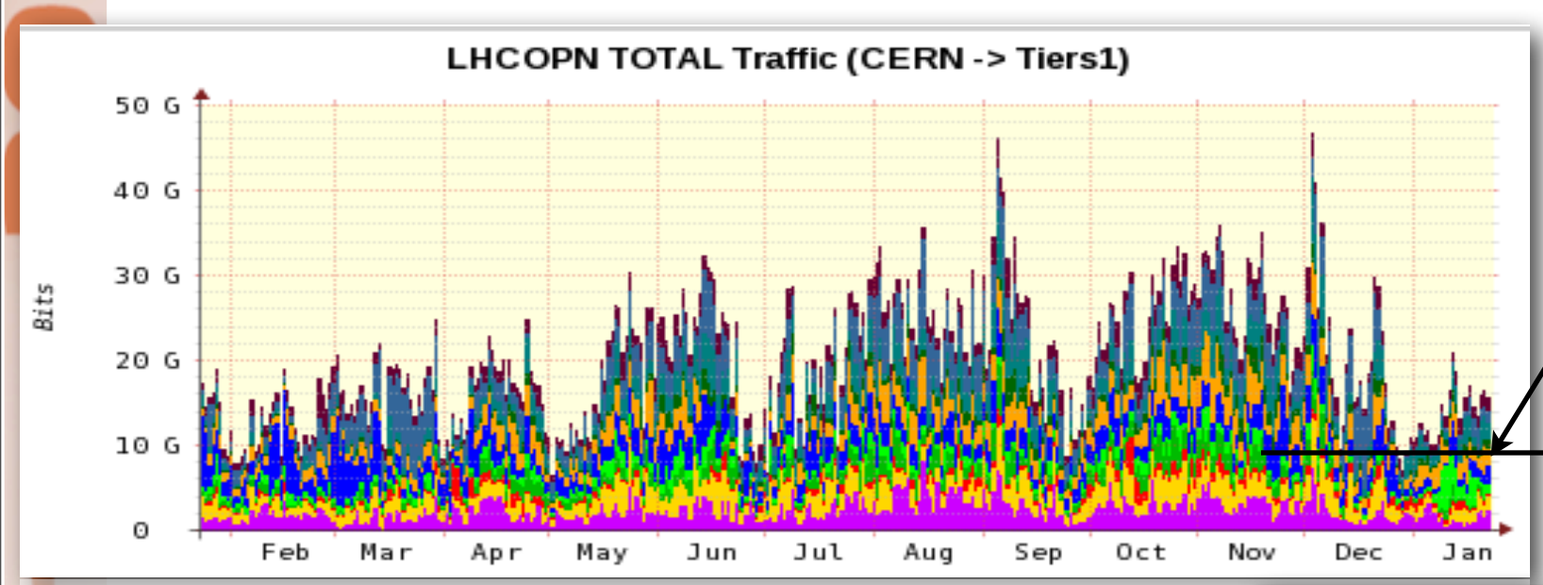


source : EGI accounting portal



User analysis: 15-20% of CPU

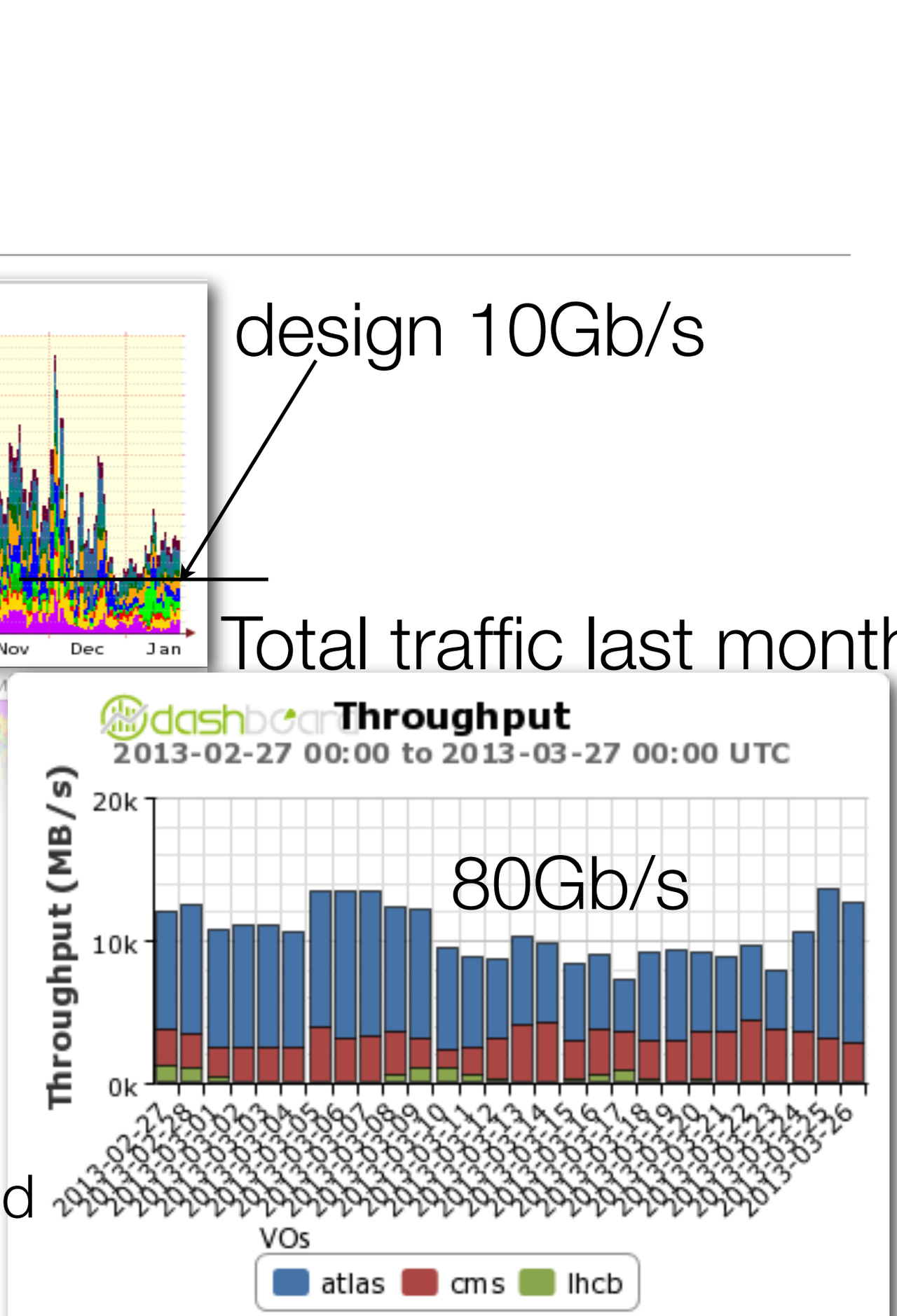
Network



design 10Gb/s

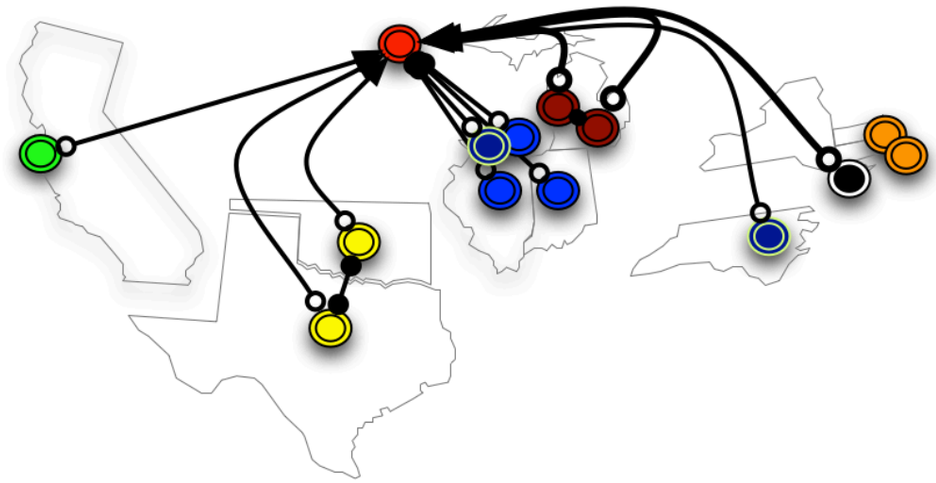
Total traffic last month

Performance well above initial expectations



Does not include Xrootd traffic (ALICE)

Near future



DISTRIBUTED STORAGE / REMOTE ACCESS

- Jobs access data on shared storage resources via WAN*
- Better usage of storage resources (disk prices!)
- Simplification of data management
- Possibly remote access (with caching at both ends); direct reading or file copy
- Bandwidth and stability needed

On going demonstrators for Xrootd & HTTP data access over WAN

(*)WAN : Wide Area Network

Virtualization / Cloud computing

- Cloud: extension of grid computing with delegation of QoS & ‘better’ reliability
- Prototypes going into production at pilot sites but also on academic clouds and Amazon or Google (in US & CA)
- HLT farms at CERN cloudified to run MC simulation during LS1
- Plan for use of ‘academic’ clouds and opportunistic use of ‘cheap’ commercial is possible
- However commercial cloud prices (even with special packages) very high
- CPU only cost on Amazon ~ 3x CPU cost at T1

openstack

amazon web services

Google Compute Engine

SuperMUC a PRACE Tier-0 centre :
155,000 Sandy Bridge cores, 2.8M HS06

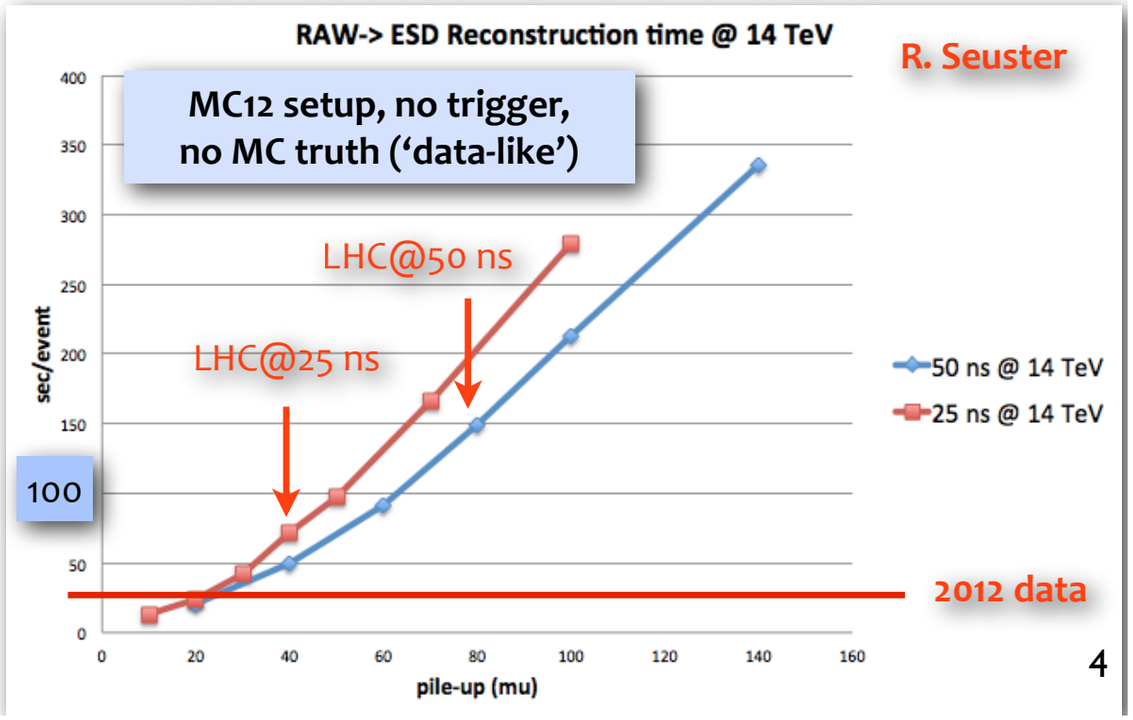
WLCG 2013 T0/1/2 pledges ~2.0M HS06

Opportunistic use of HPC (High-Performance Computing) resources

- Latest competitive supercomputers are x86 based (familiar linux cluster)
- ATLAS & CMS projects to use idle CPU cycles at HPC centers in US (Argonne, San Diego) & DE (Munich)
- Demonstrators working for simulation
- Difficult to use HPC centers for I/O intensive applications
- Outbound connectivity of HPC centers may be an issue

2015 :
- **New Energy**
- **New Pile-up**
- **New Trigger Rate**

reconstruction time/evt ATLAS



4

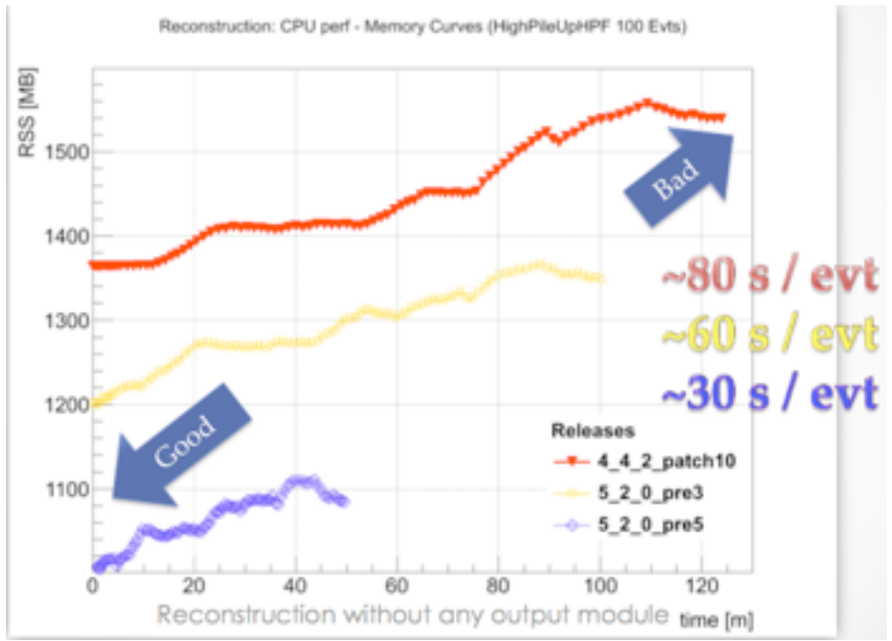
Future … after LS1

LHC computing resources increase will not follow the demands with current software

Software must gain several factors in speed

And new hardware coming...

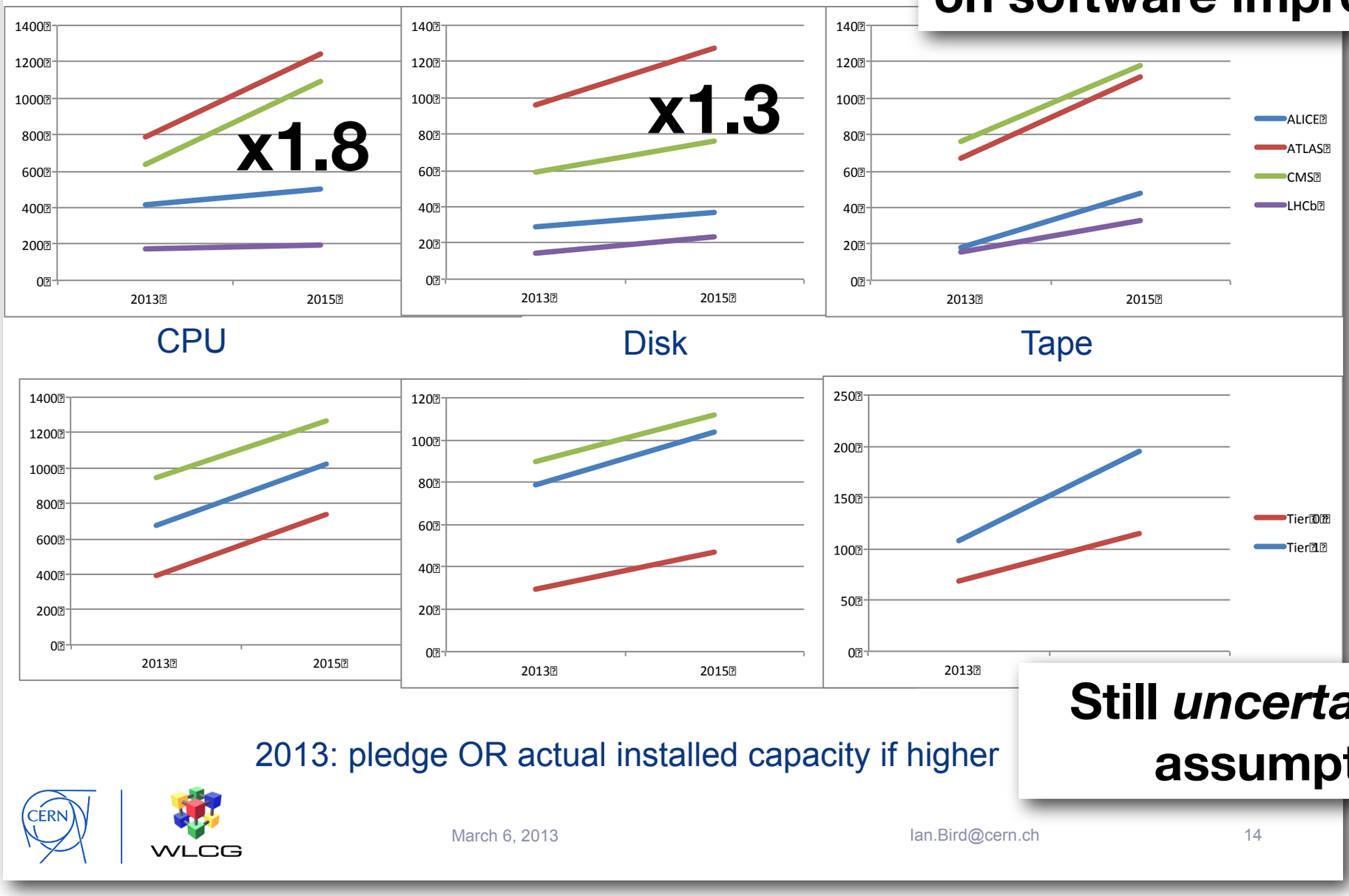
software speed improvement in CMS



Resources requests by experiments for 2015

Requests include some 'aggressive' *assumptions* on software improvements

2013 → 2015 resources



CPU

Disk

Tape

2013: pledge OR actual installed capacity if higher

Still *uncertainties* in assumptions

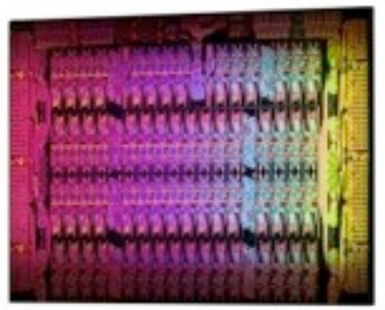
Requests not yet approved (RRB on April, 16)

New hardware are multi-core (<100)

- Speed of computer no more driven by clock speed
- Processor clock rates faster than memory clock rates
- No major change in throughput
- Data dependencies in software are very expensive
- No increase of memory/core

Complete revision of software mandatory to exploit new hardware

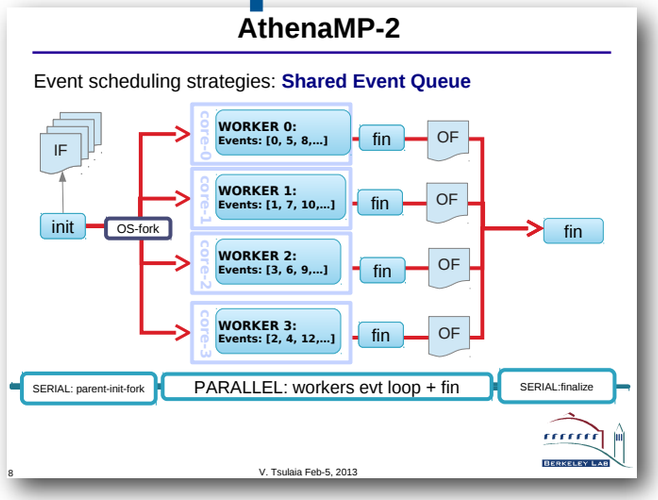
And many-core (>100) hardware coming...



Software changes

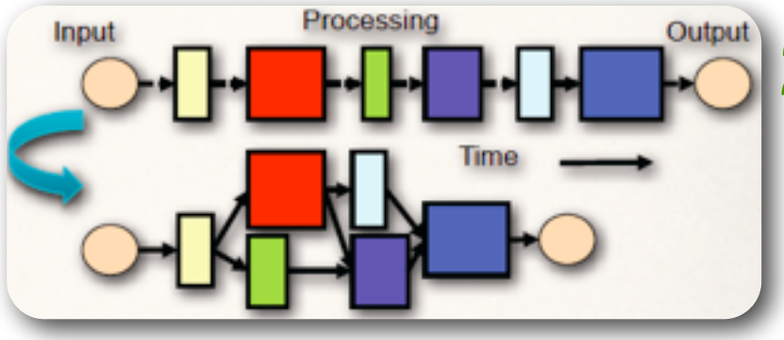
- All experiments embarked in profound software changes
- Geant team as well...
- Reduction of memory footprint
- Revision of data models
- Multithreating (memory sharing)
- **Vectorisation** (to exploit new architectures)
- I/O is a major concern

event parallelism



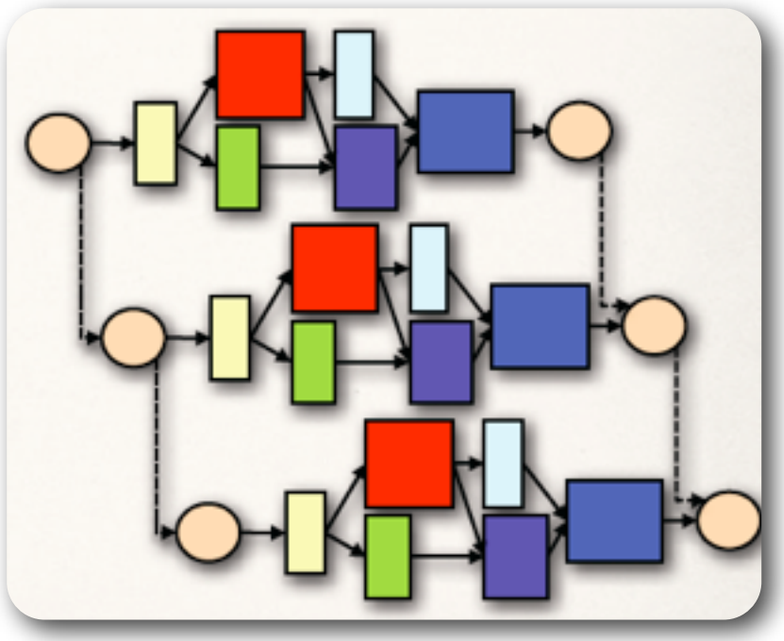
Today

algorithm parallelism



2015-2016 ?

event & algorithm parallelism



LS2 or before?

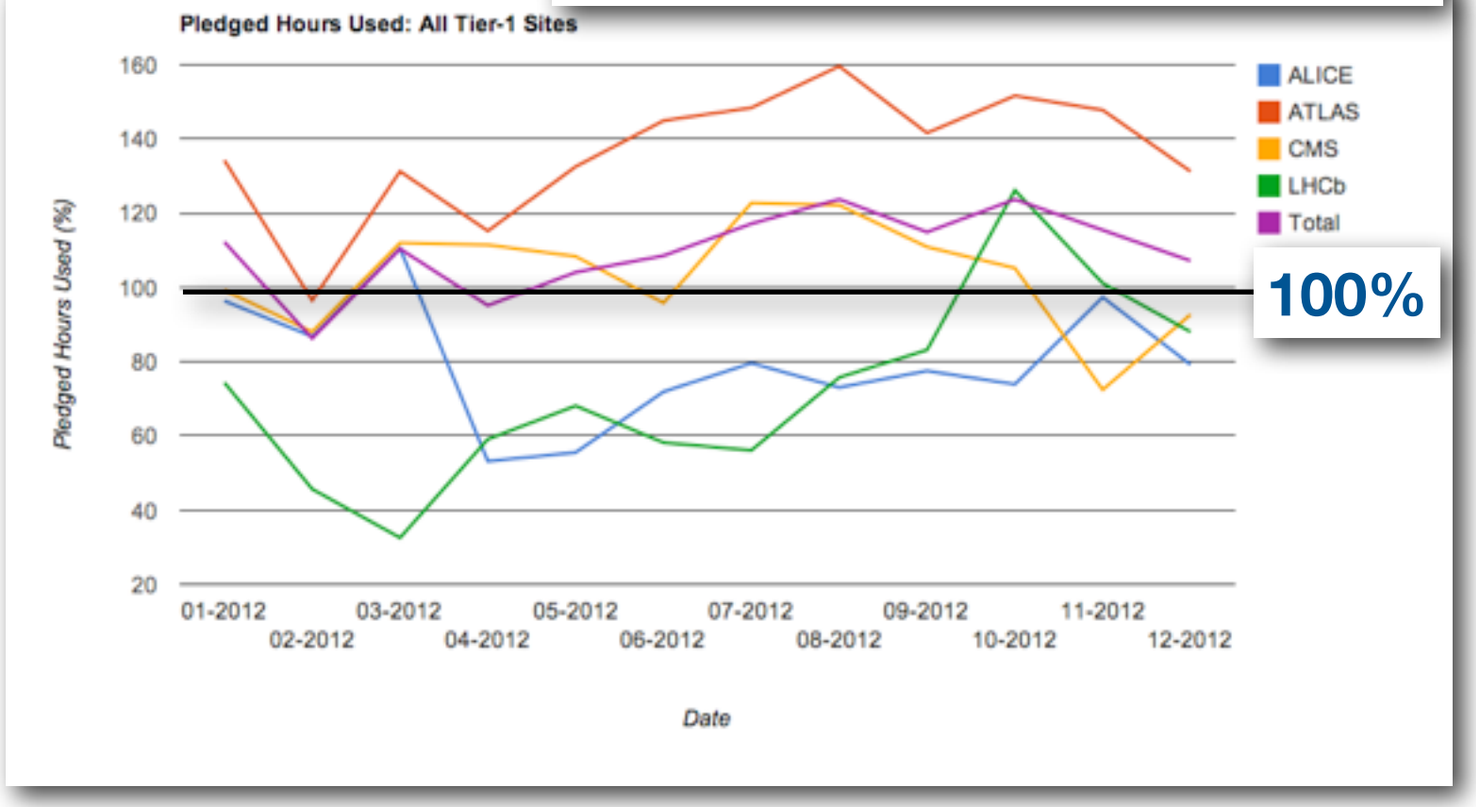
Some concern

CPU consumption above pledges both at T1s and T2s

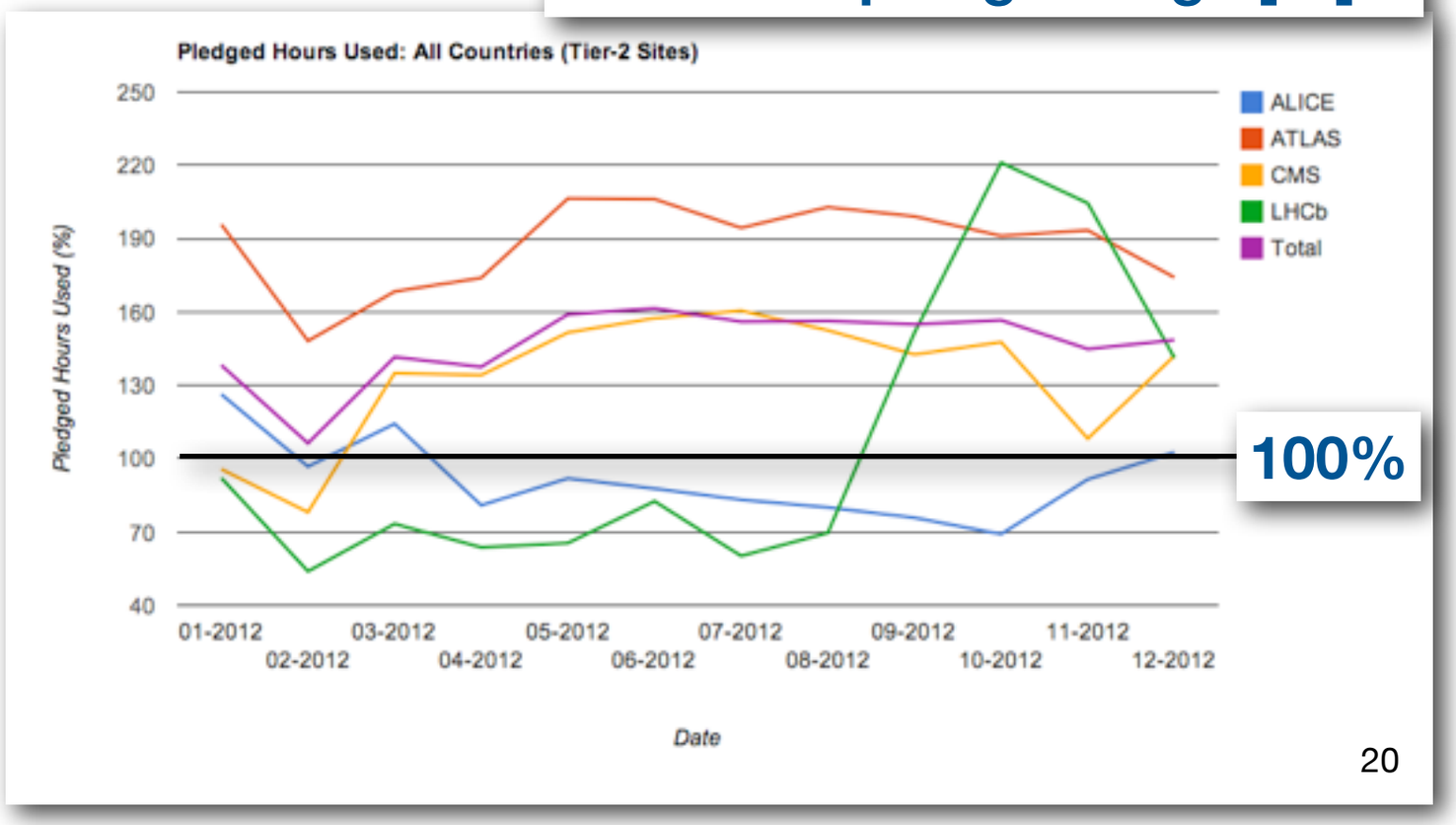
Sites provide unpledged resources (thank you!)

Experiments needs are larger than official requests

T1 CPU pledge usage [%]



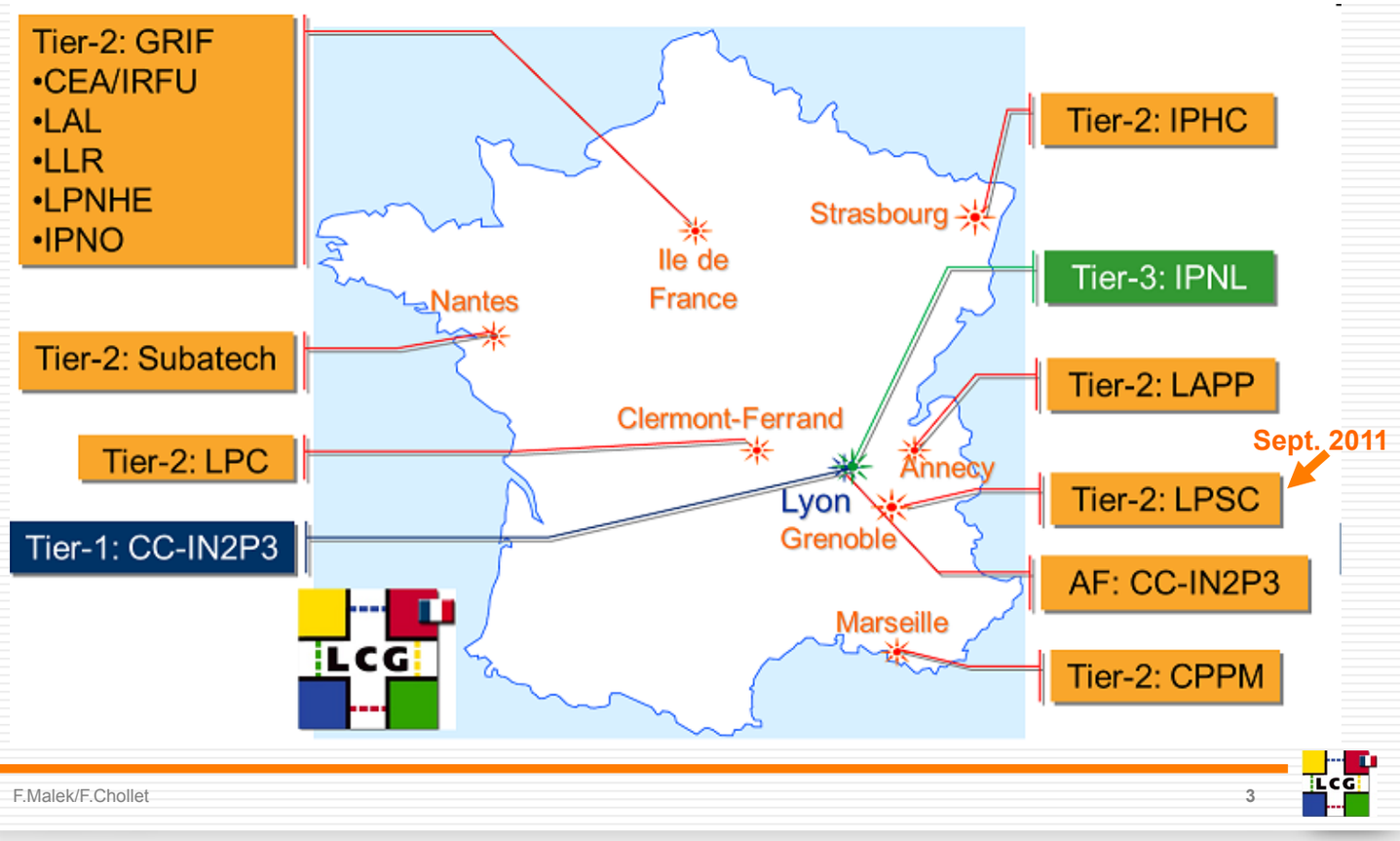
T2 CPU pledge usage [%]



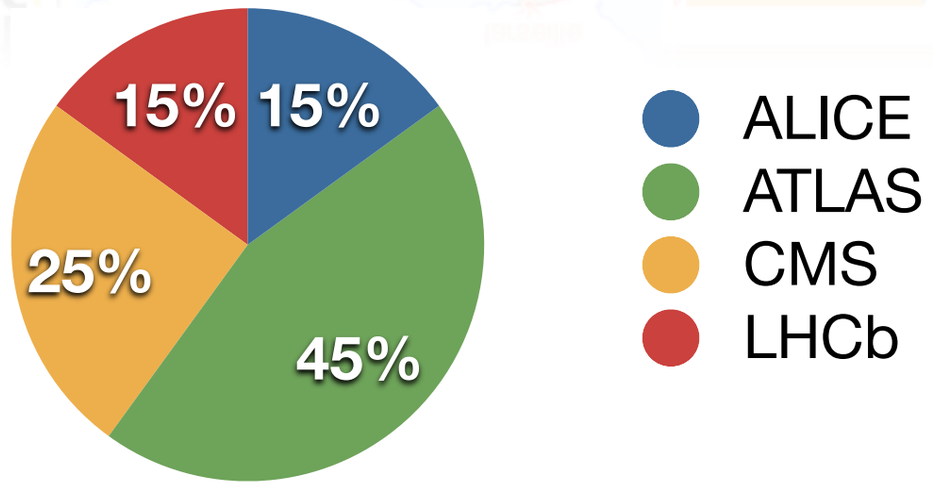
source : EGI accounting portal

The French contribution

Sites LCG-France 2012



LHC-France budget share



French sites contribution to LHC processing

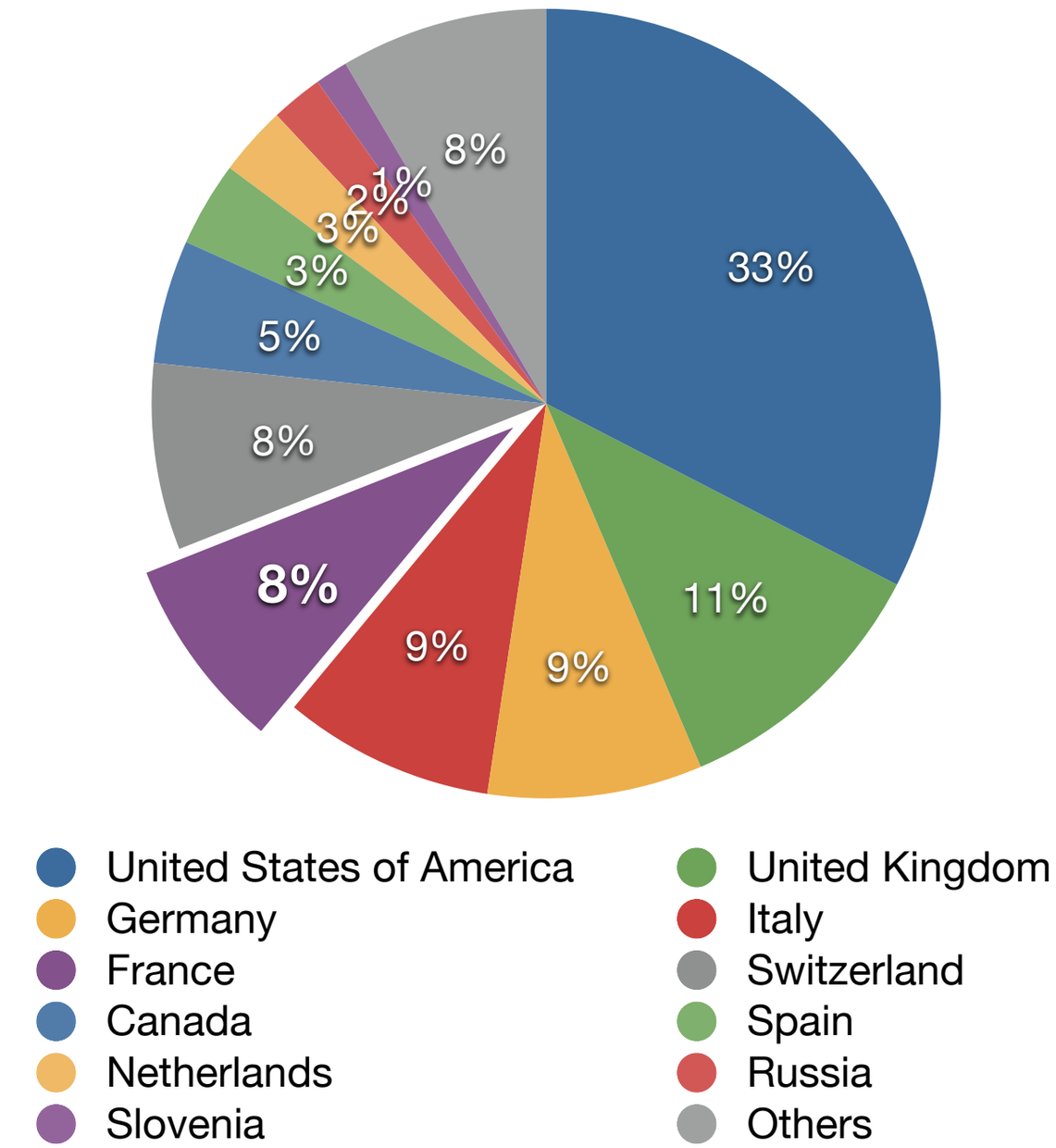
8% in 2012, T1 & T2s included

Was over 10% in 2010

Difficulties to follow the needs

Hardware is getting old

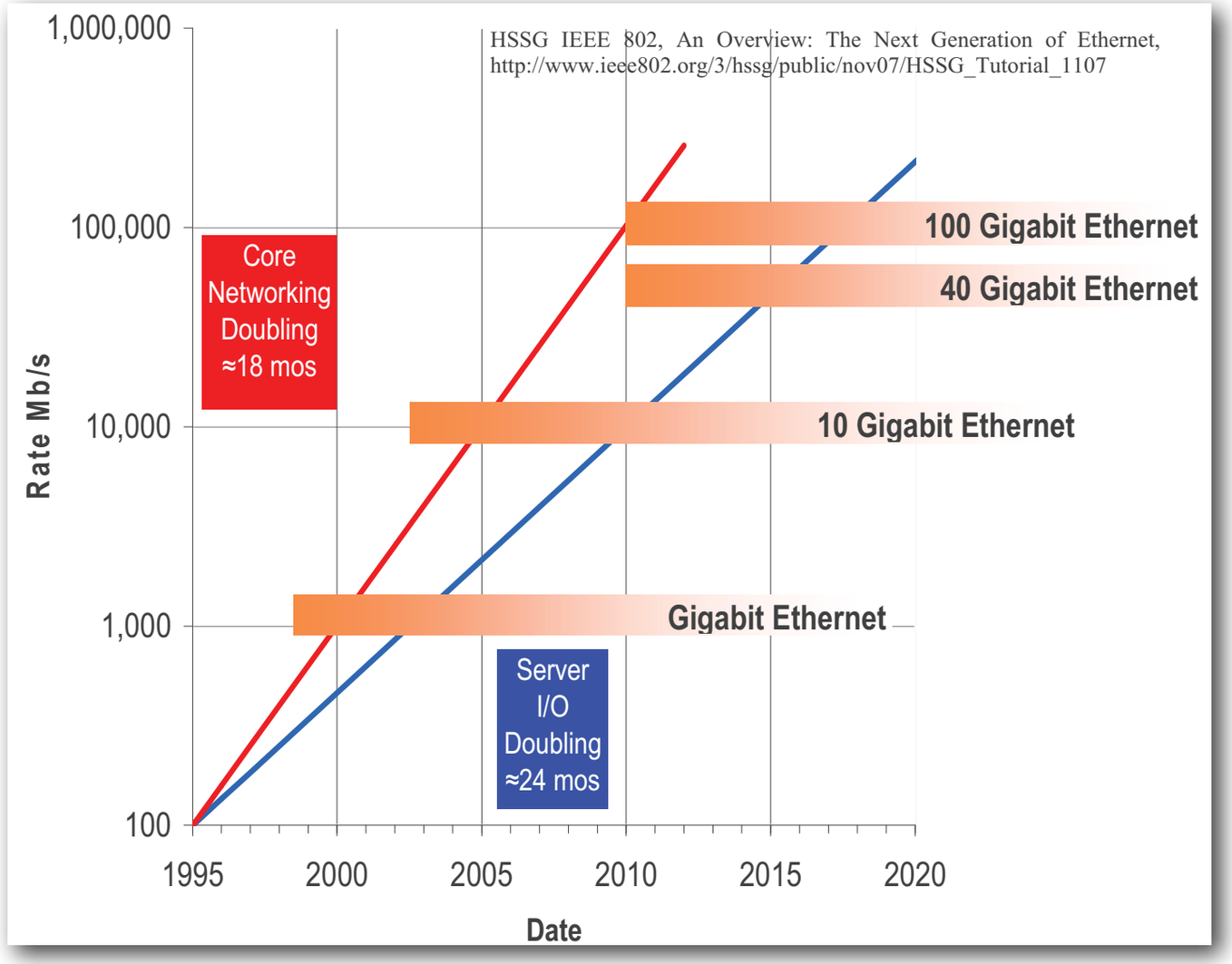
CPU delivered in 2012 WLCG year

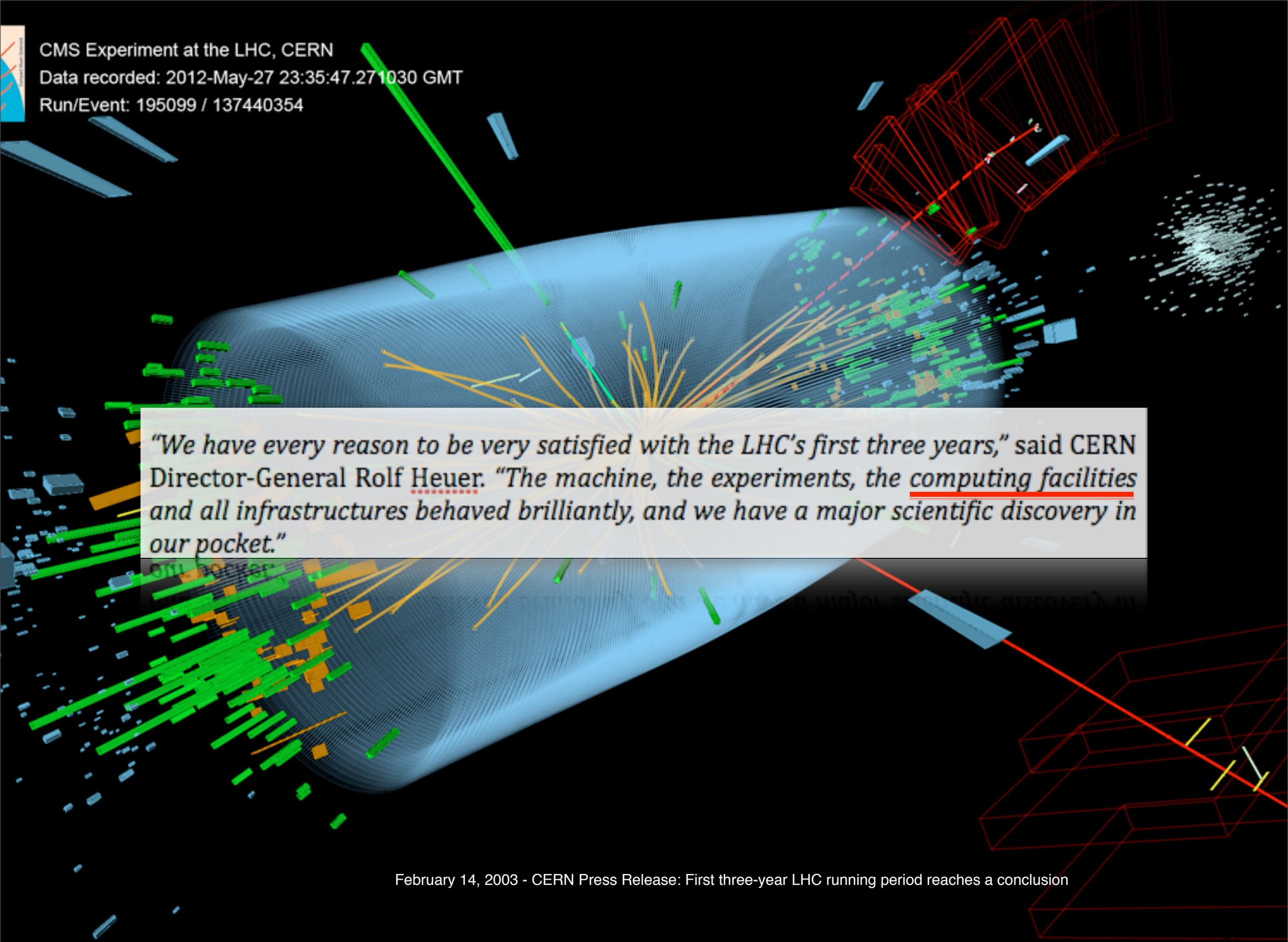


http://accounting.egi.eu/country.php

Outlook

- **Network** will continue to be the driving force
- **Virtualization** of services (cloud computing, distributed storage)
- On-going revolution in **software**
- Essential to maintain computing funding at decent **level**. LHC upgrades also include computing
- Computing should not be a **limiting factor** for physics output



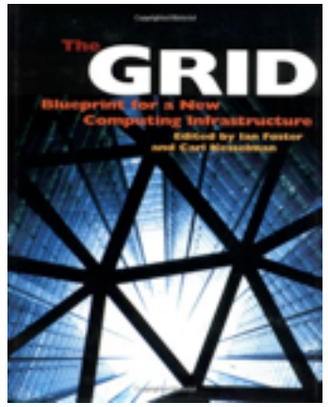


"We have every reason to be very satisfied with the LHC's first three years," said CERN Director-General Rolf Heuer. *"The machine, the experiments, the computing facilities and all infrastructures behaved brilliantly, and we have a major scientific discovery in our pocket."*

February 14, 2003 - CERN Press Release: First three-year LHC running period reaches a conclusion

Backup

Pre-history (<2000)



www.globus.org

1998 :

- The GRID by Ian Foster & Carl Kesselman (made the idea popular)
- **Globus**: first middleware widely available (proof of concept)

1999 : MONARC report

http://monarc.web.cern.ch/MONARC/

Models of Networked Analysis at Regional Centres for LHC Experiments

MONARC

GENERAL CONCLUSIONS on LHC COMPUTING

Following discussions of computing and network requirements, technology evolution and projected costs, support requirements etc.

- The scale of LHC "Computing" is such that it requires a worldwide effort to accumulate the necessary technical and financial resources
- The uncertainty in the affordable network BW implies that several scenarios of computing resource-distribution must be developed
- A distributed hierarchy of computing centres will lead to better use of the financial and manpower resources of CERN, the Collaborations, and the nations involved, than a highly centralised model focused at CERN
  - Hence: The distributed model also provides better use of physics opportunities at the LHC by physicists and students
- At the top of the hierarchy is the CERN Centre, with the ability to perform all analysis-related functions, but not the ability to do them completely
- At the next step in the hierarchy is a collection of large, multi-service "Tier1 Regional Centres", each with
  - 10-20% of the CERN capacity devoted to one experiment
- There may be Tier2 or smaller special purpose centres in some regions

June 22, 1999 MONARC Status Report Harvey Newman (CIT)

Fear of the networks!

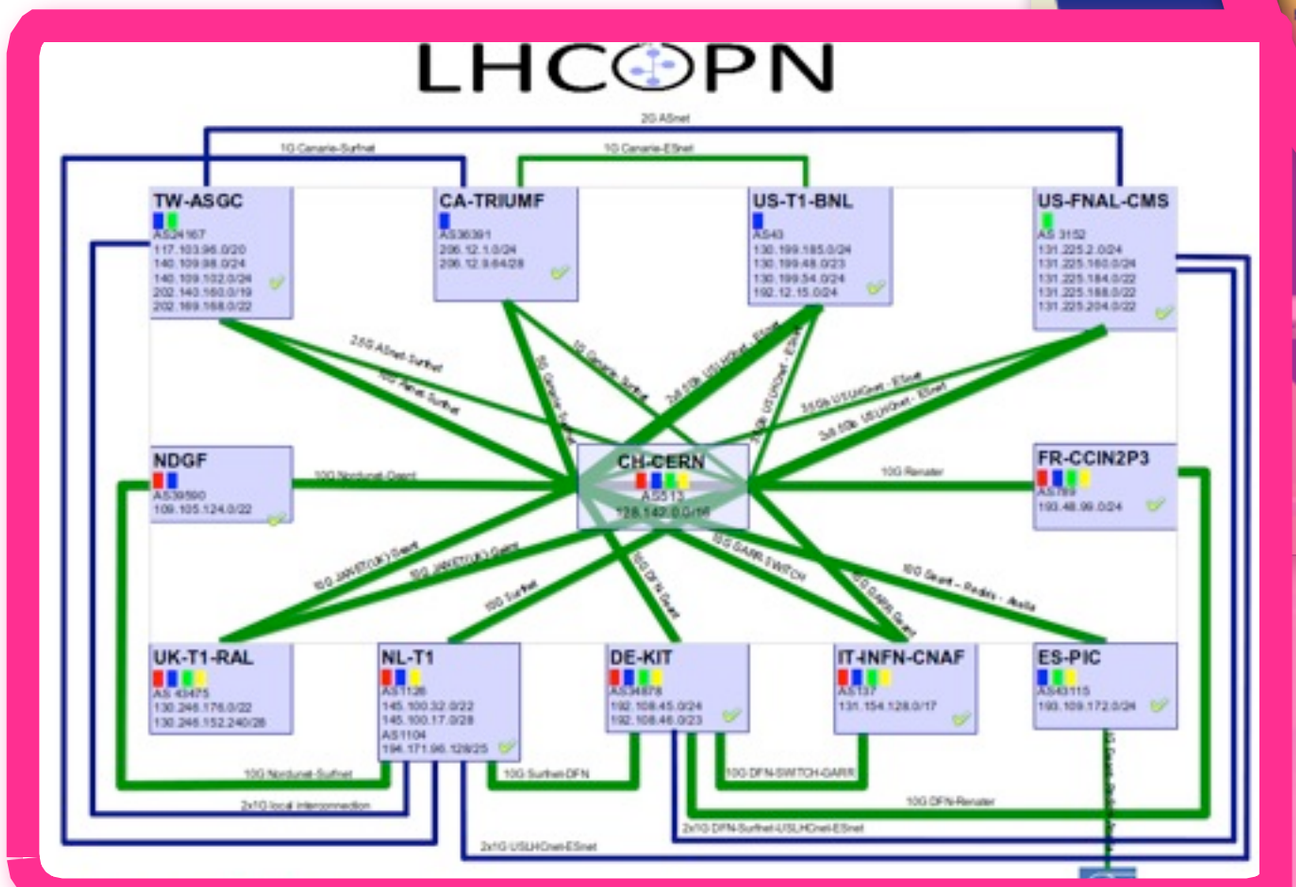
Tiered site architecture

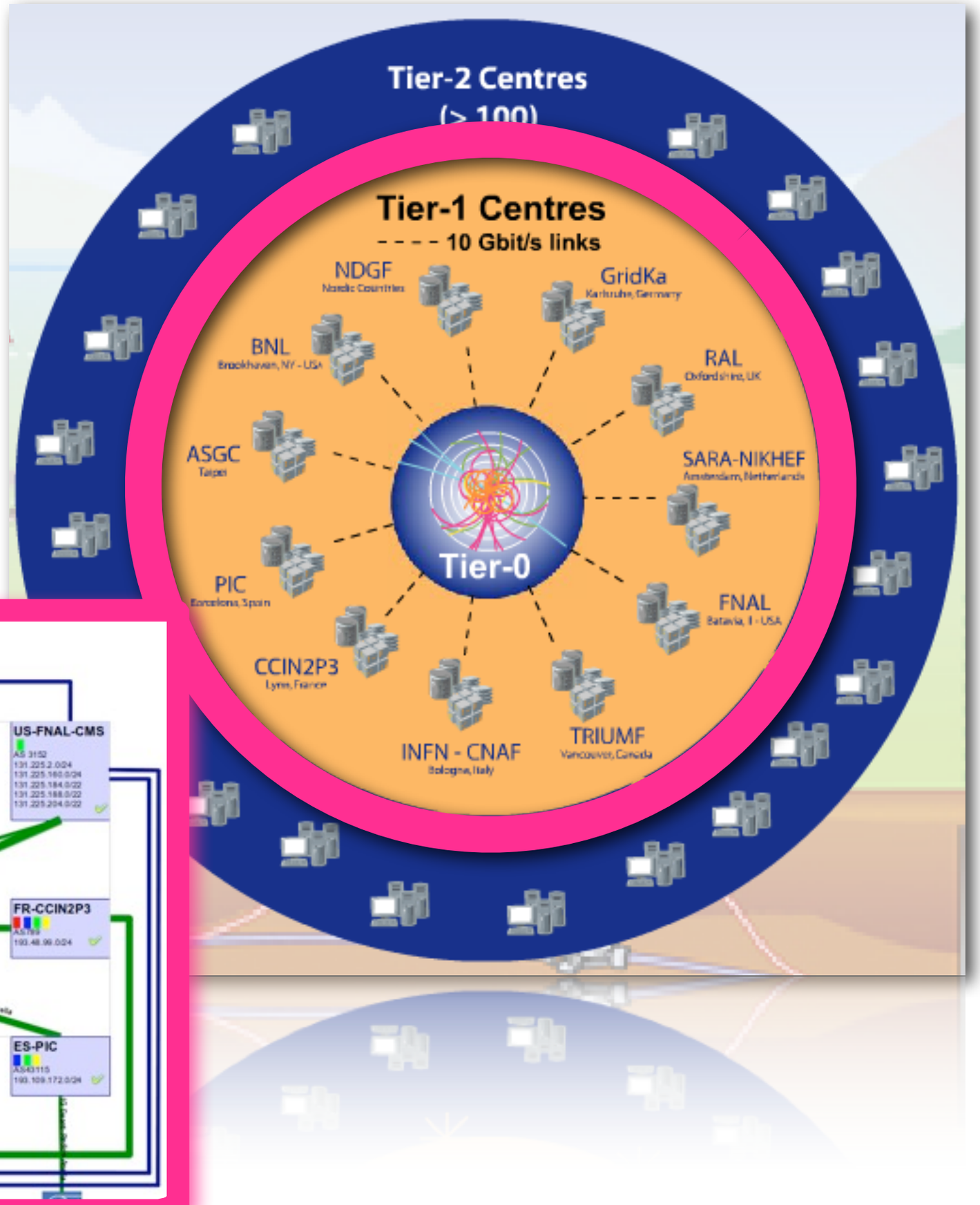
Pre-planned data distribution

Jobs-to-data brokerage

Centric model

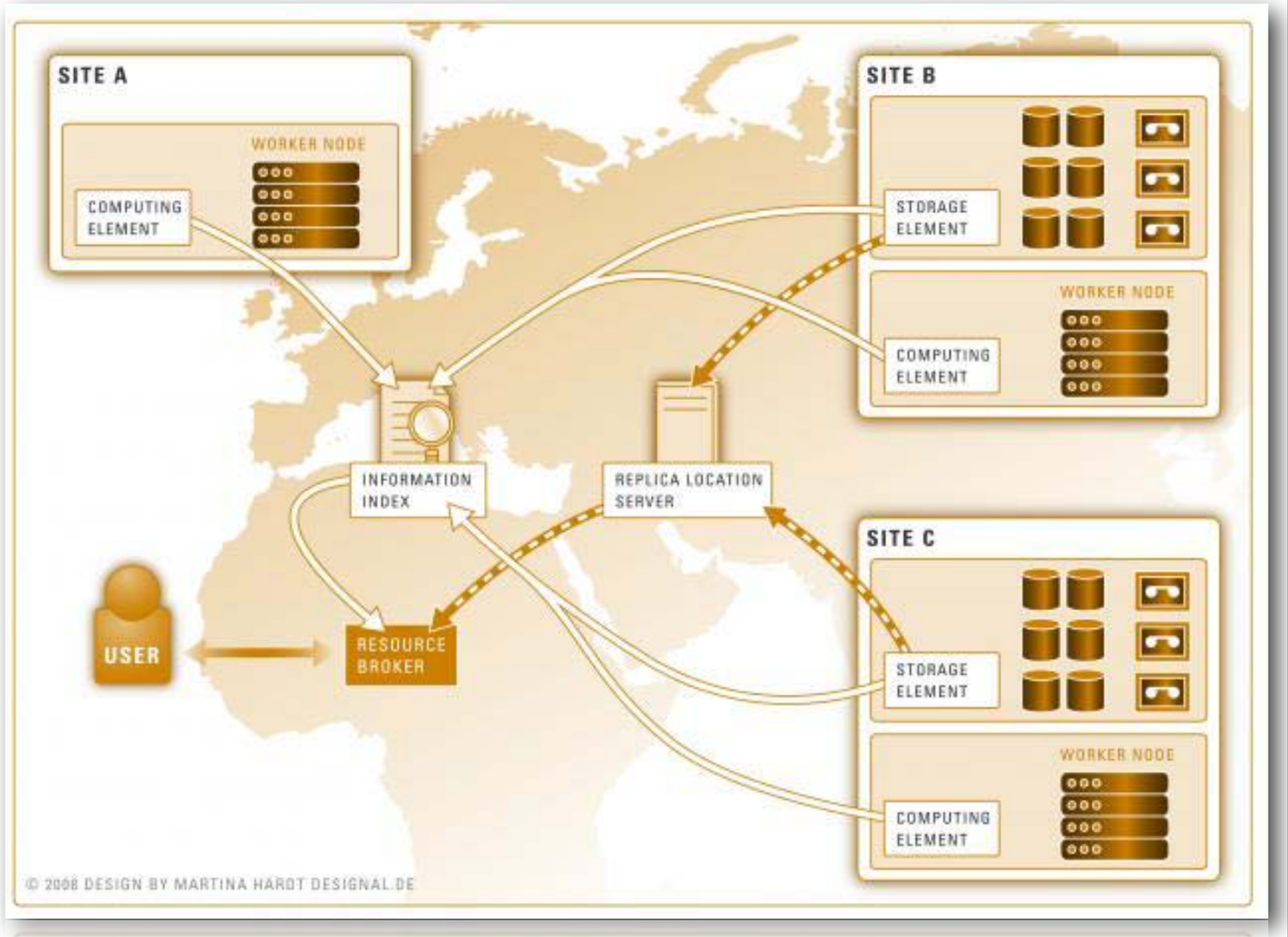
LHCOPN : T0 & T1s linked by dedicated 10 Gb networks



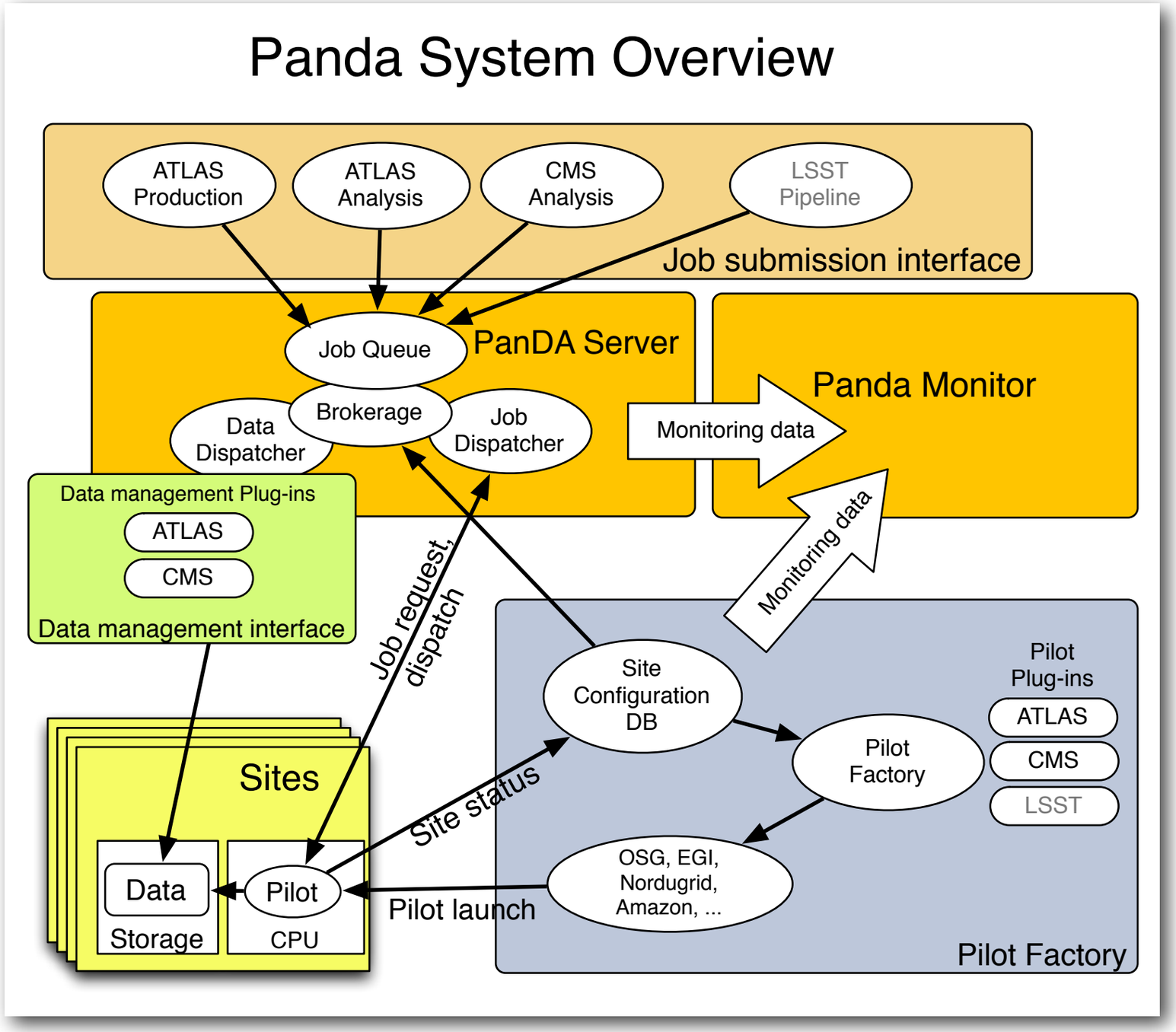


Grid software

- Layers of (complex) software developed in EU and US (derived from **Globus**: 1998)
- Information system, authentication & authorization system
- File catalogs
- File transfers
- Job brokering
- Interfaces to Storage & batch systems
- etc...

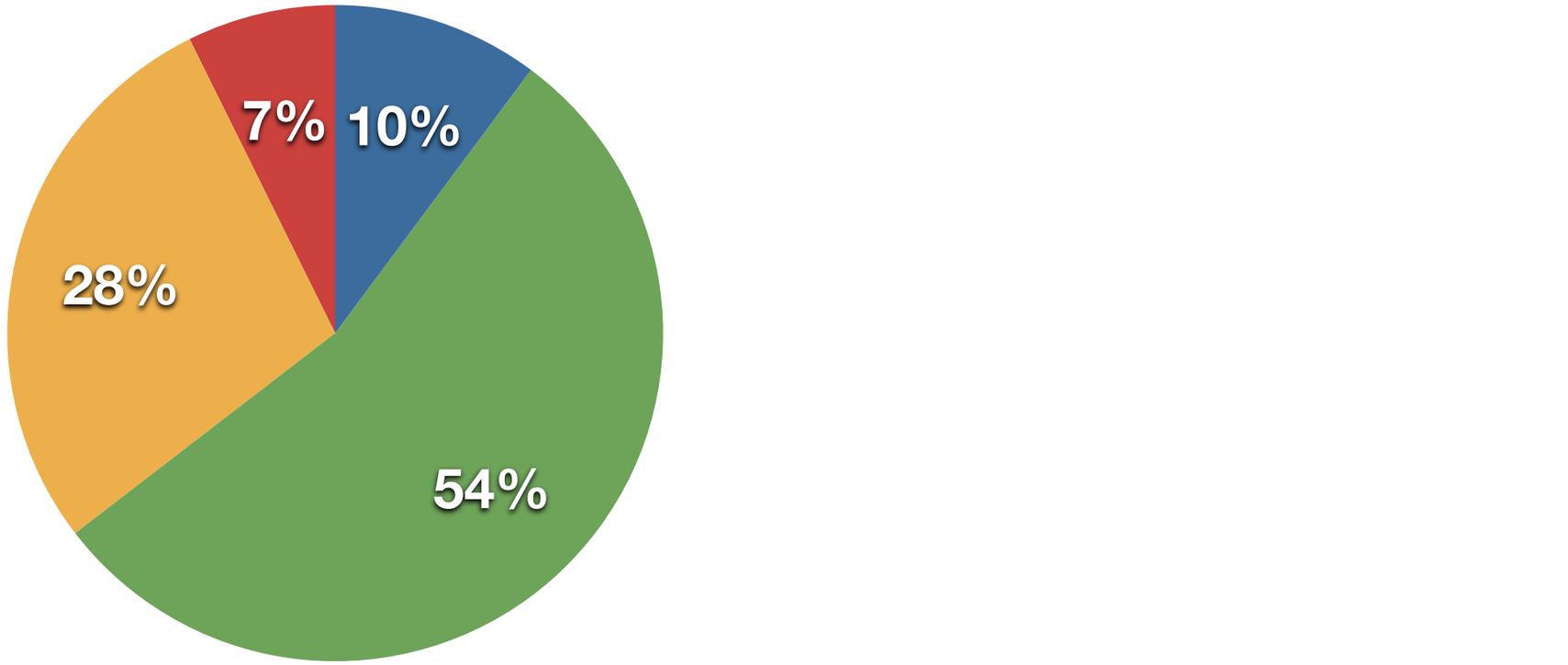


Panda system in a nutshell

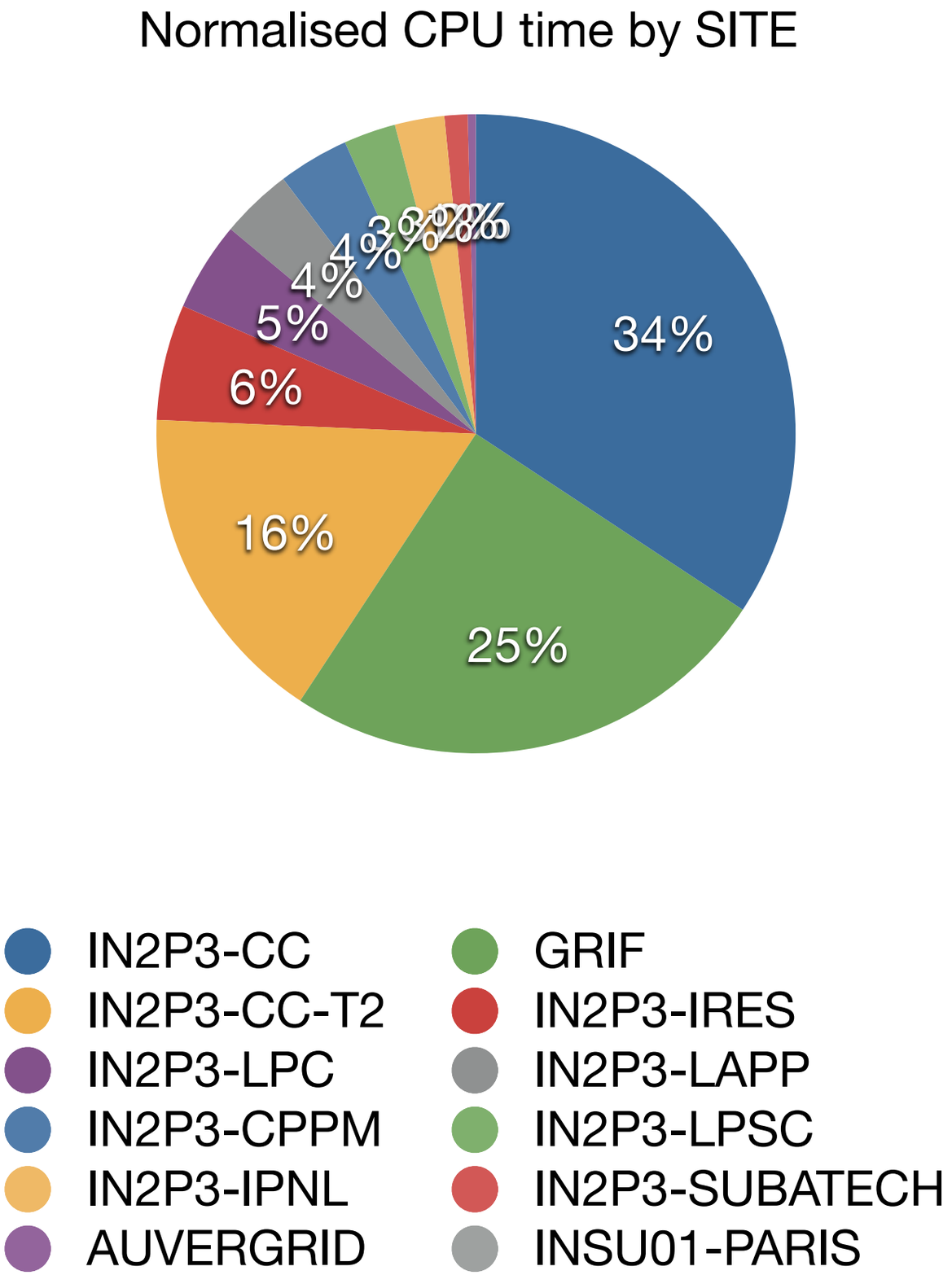


CPU consumption by experiment

Normalised CPU time by LHC experiment



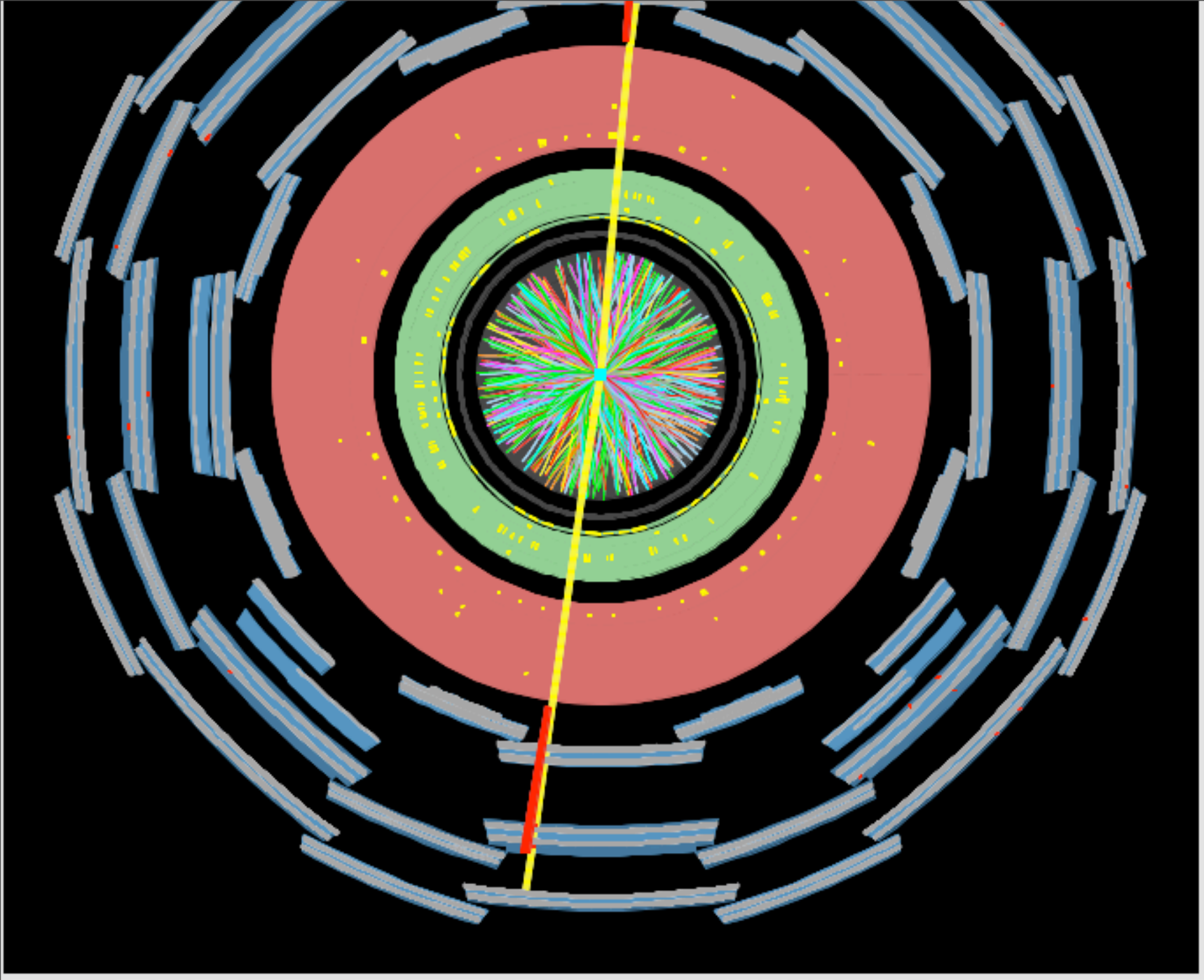
CPU delivered by French sites

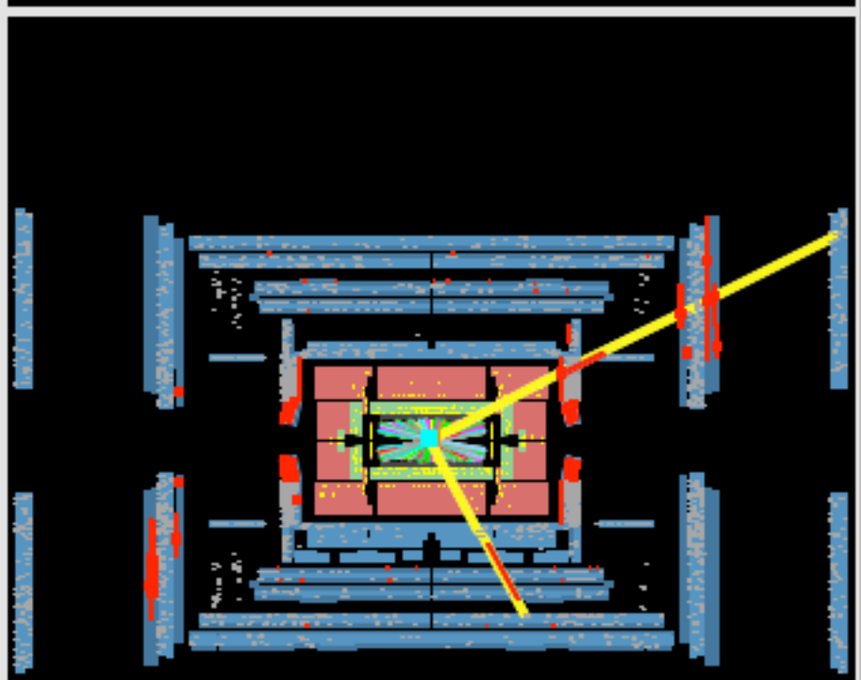


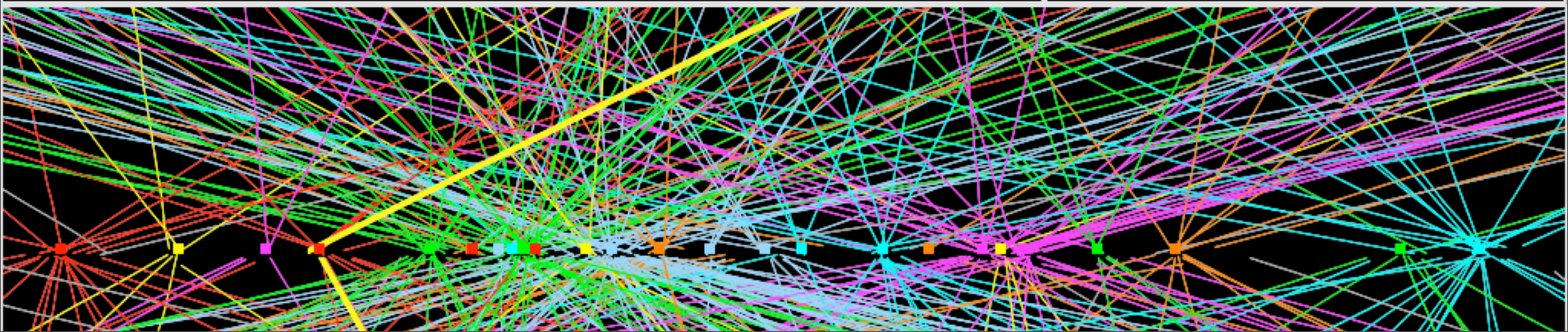
ATLAS EXPERIMENT

Run Number: 201289, Event Number: 24151616

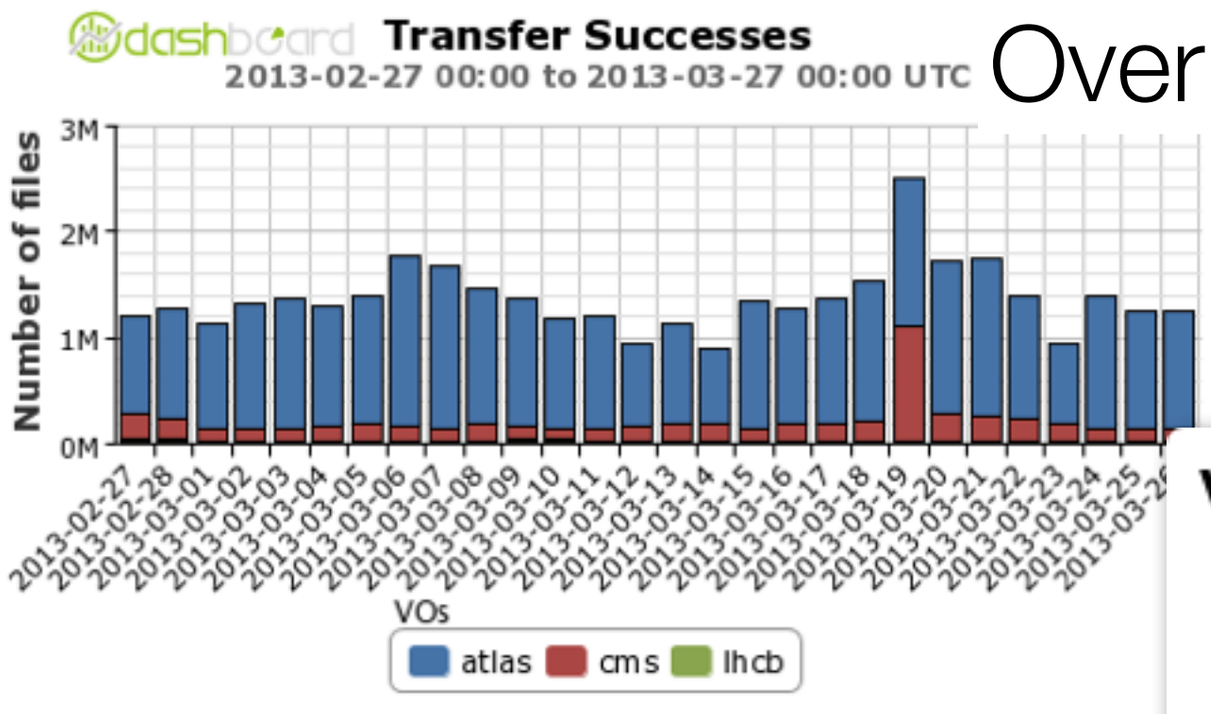
Date: 2012-04-15 16:52:58 CEST





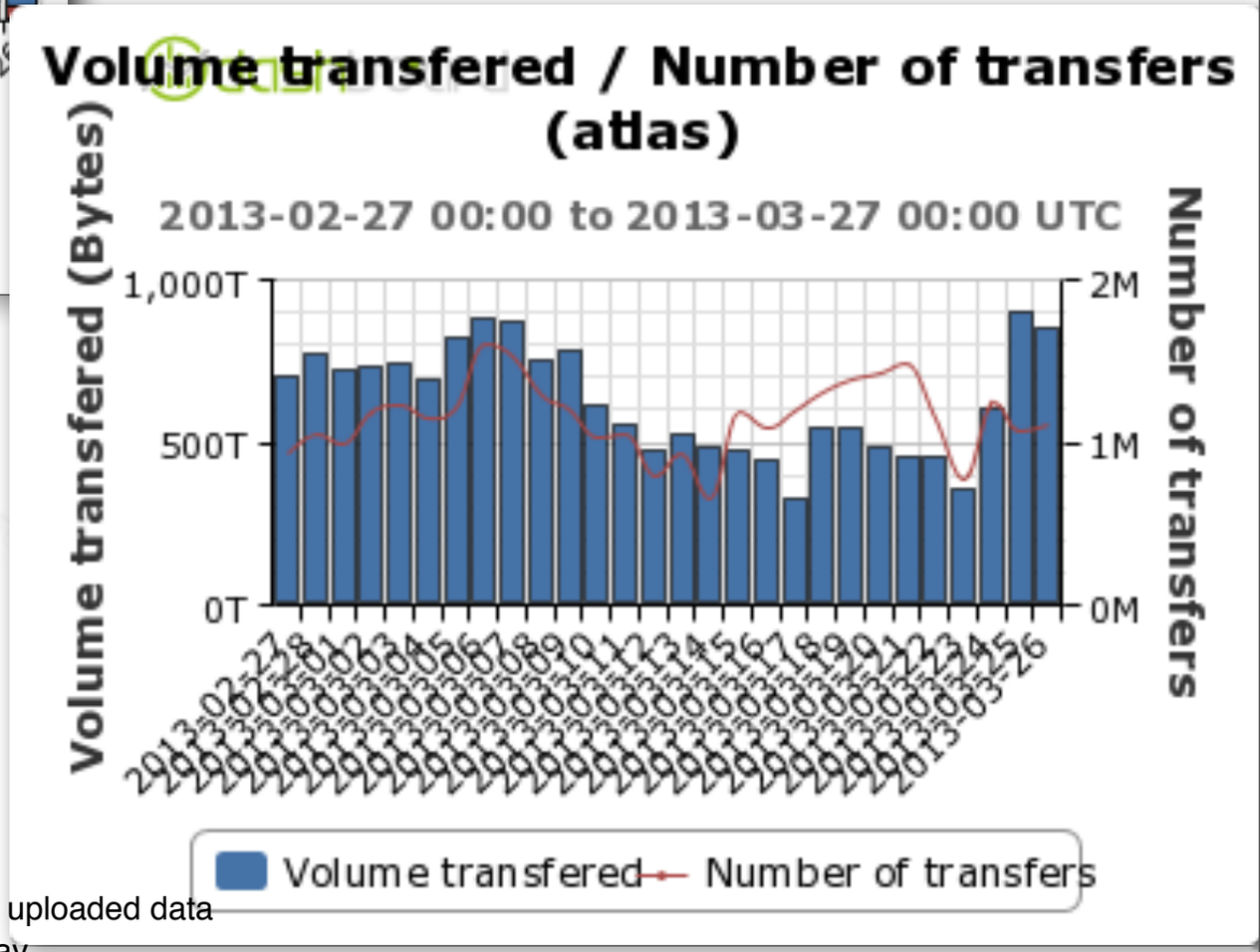


Data volume transferred



Over 1M files/day

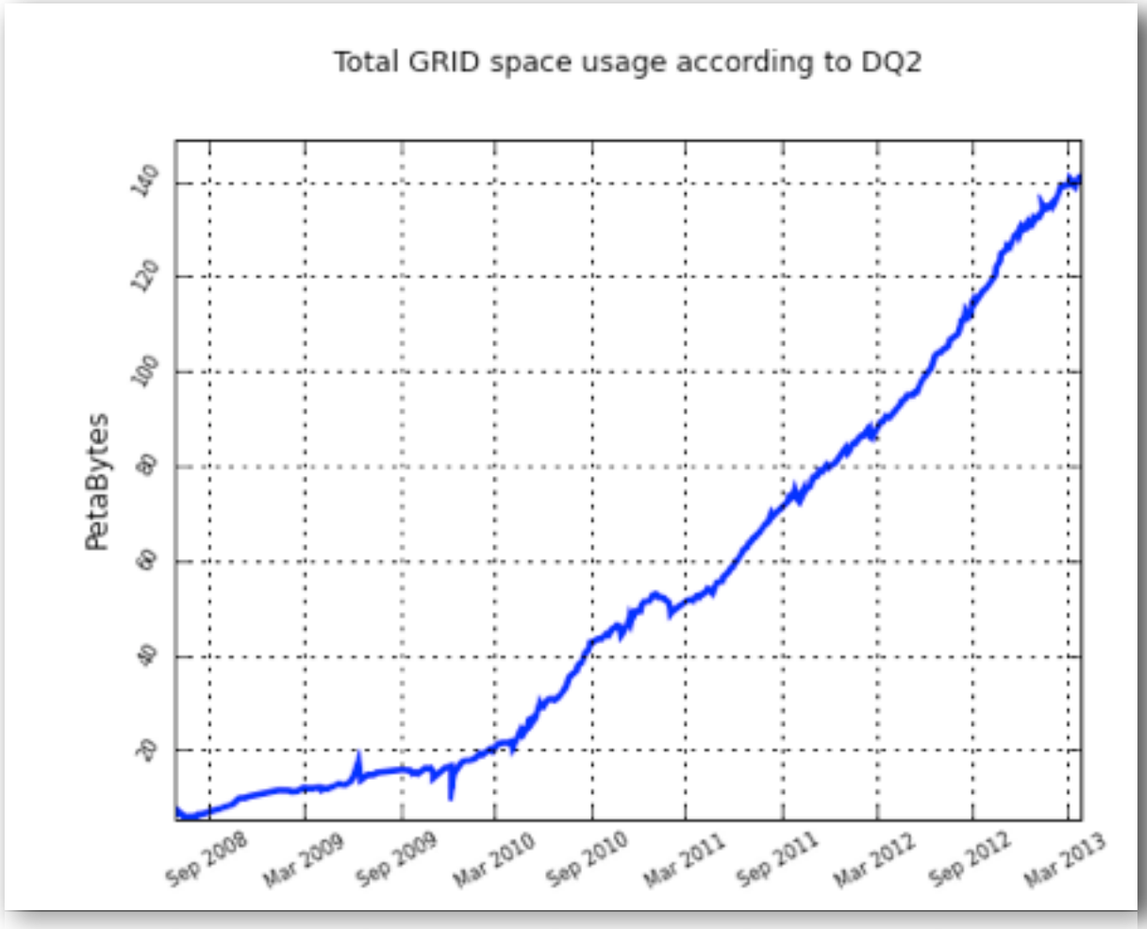
Over 1PB/day



Internet: Google processed about 24 petabytes of data per day in 2009
At its 2012 closure of file storage services, Megaupload held ~28 petabyte of user uploaded data
Telecoms: AT&T transfers about 30 petabytes of data through its networks each day
http://en.wikipedia.org/wiki/Petabyte

ATLAS file catalog

>140 PB



> 4B files