



User communities on production grids

Johan Montagnat, CNRS / I3S





User communities

➤ User communities well represented in France

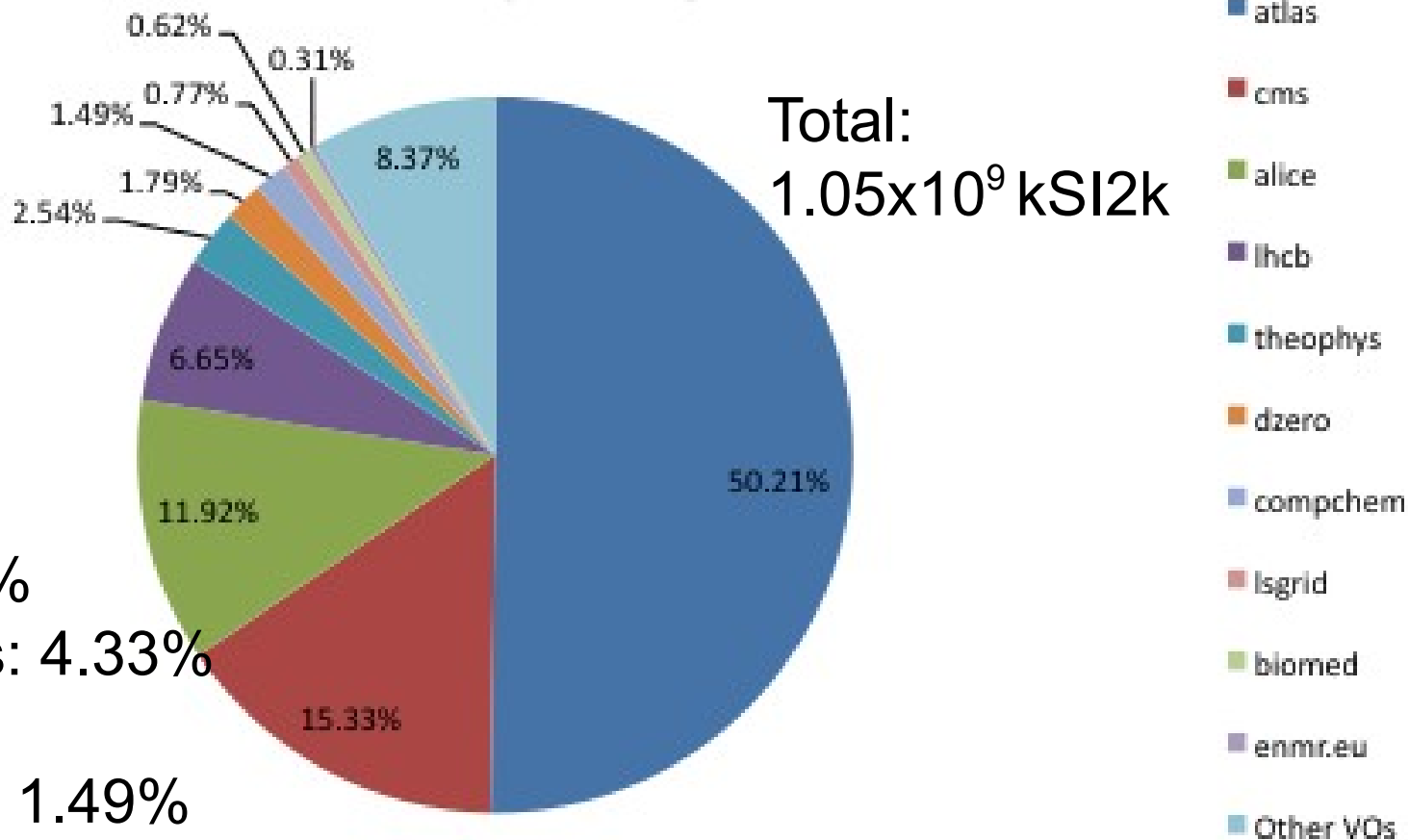
- Particle Physics
- Life Sciences (medical imaging, bioinformatics, drug discovery)
- Earth Sciences (satellite data, seismology)
- Astronomy & astrophysics
- Grid Observatory

➤ All have an international activity

- Little tooling to monitor national activity...
- ...but does it make sense anyway?
 - A federation of clusters is a grid. What is a grid sliced in regional sectors?

➤ EGI CPU resources consumed per-VO

Top 10, all region



LHC: 84.11%

Part. physics: 4.33%

LS: 1.70%

CompChem: 1.49%

Others: 8.37%

➤ Without LHC experiments

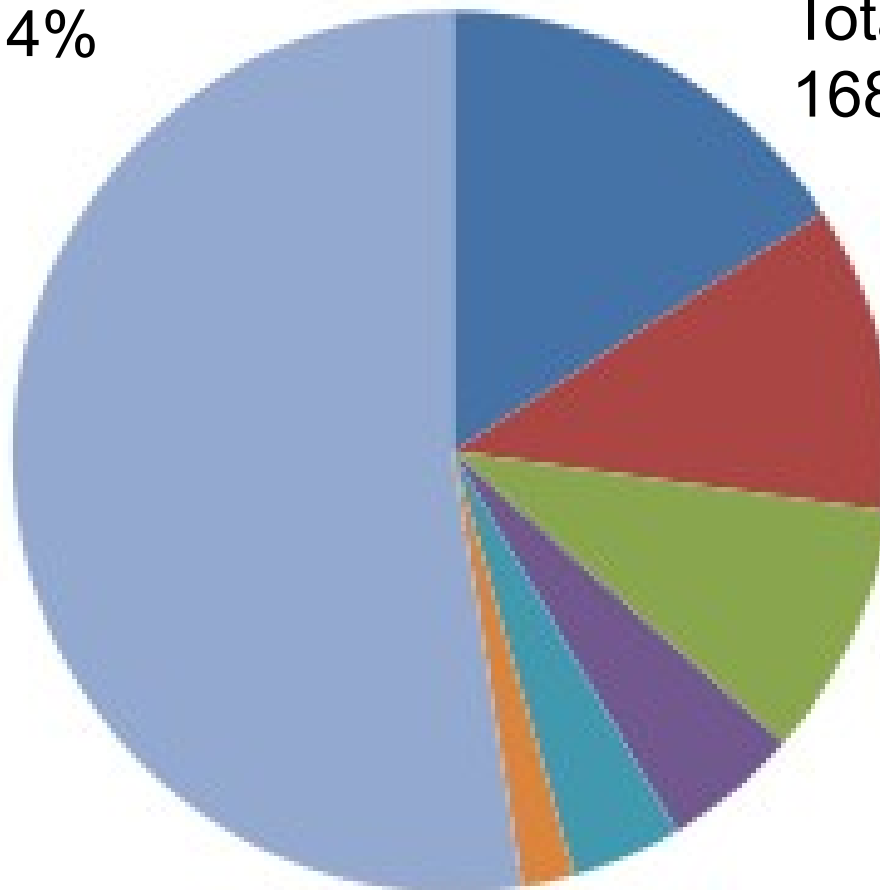
Particle physics: 27.14%

LS: 10.99%

CompChem: 9.34%

Others: 52.53%

Total:
 168×10^6 kSI2k



- theophys
- dzero
- compchem
- lsgrid
- biomed
- enmr.eu
- Other VOs



French resources, year 2010

➤ French CPU resources consumed per-VO

LHC: 86.53%

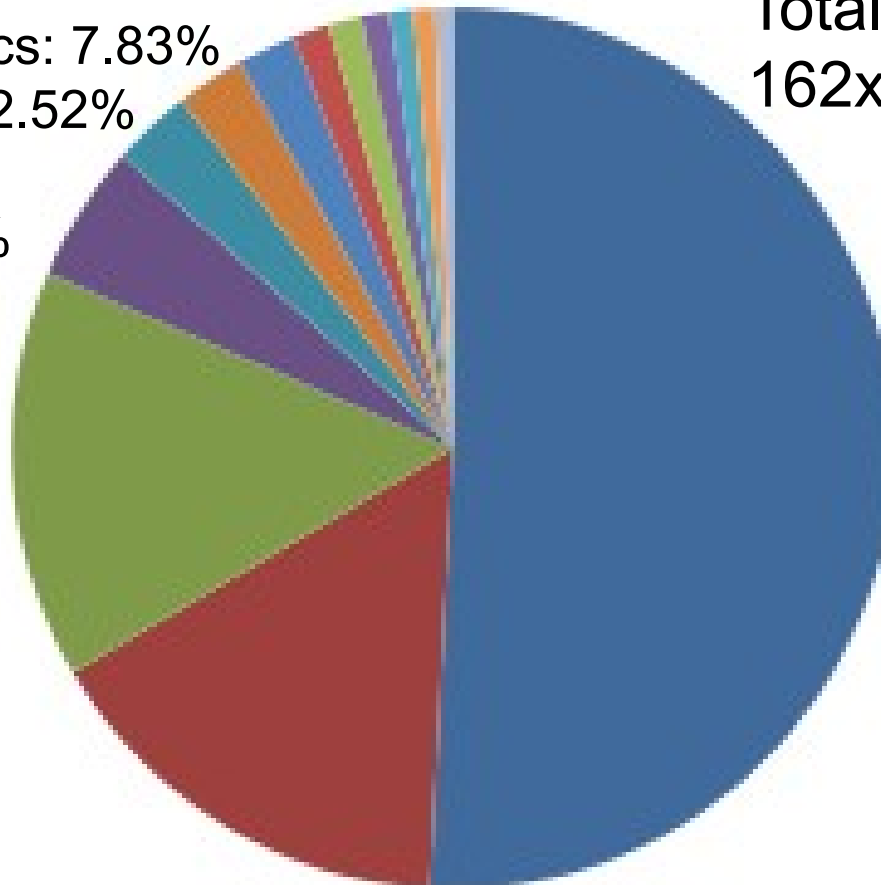
Particle physics: 7.83%

CompChem: 2.52%

LS: 1.34%

Others: 1.77%

Total:
 162×10^6 kSI2k



- atlas
- alice
- cms
- lhcb
- dzero
- compchem
- vo.lal.in2p3.fr
- biomed
- vo.cta.in2p3.fr
- vo.iscplif.fr
- vo.ipno.in2p3.fr
- fkppl.kisti.re.kr
- hone

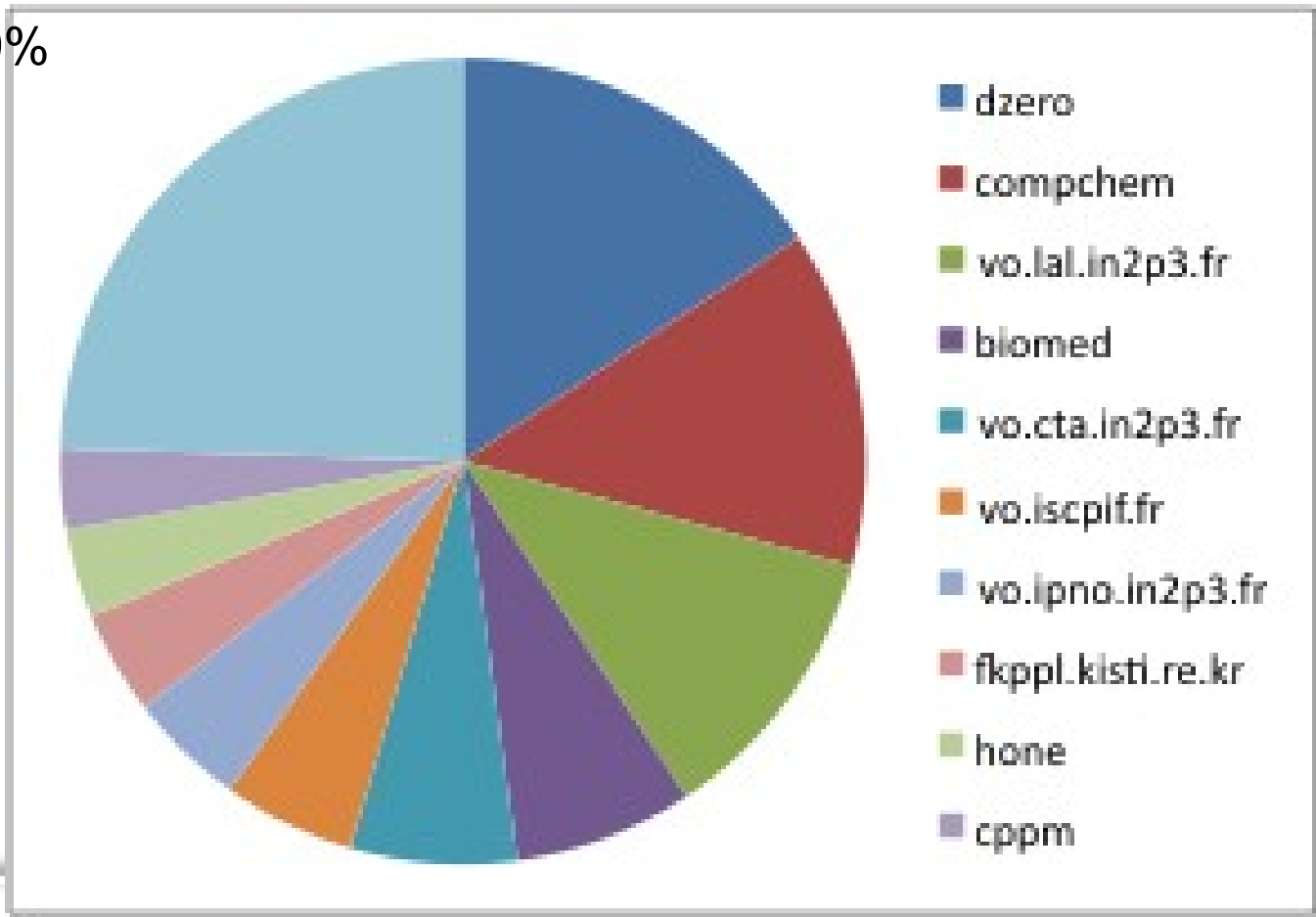


French resources, year 2010

➤ Without LHC experiments

Total:
 28.6×10^6 kSI2k

Particle physics: 45.29%
CompChem: 13.54%
LS: 7.21%
Others: 33.96%

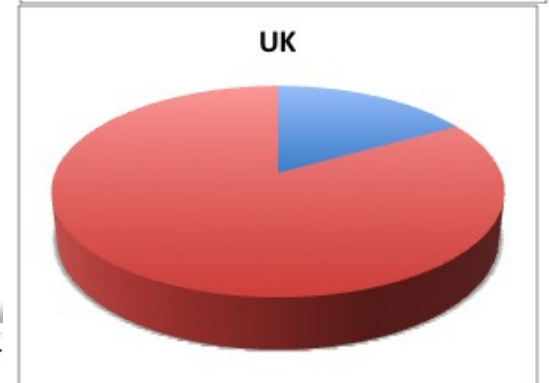
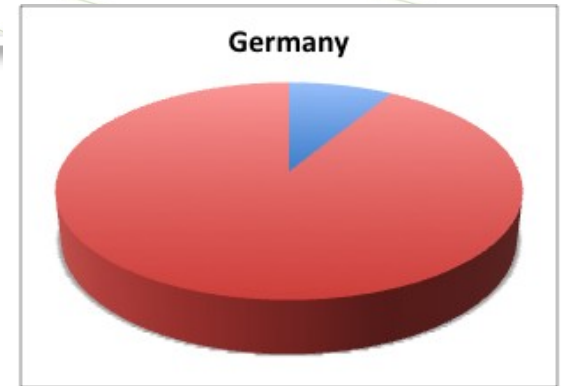
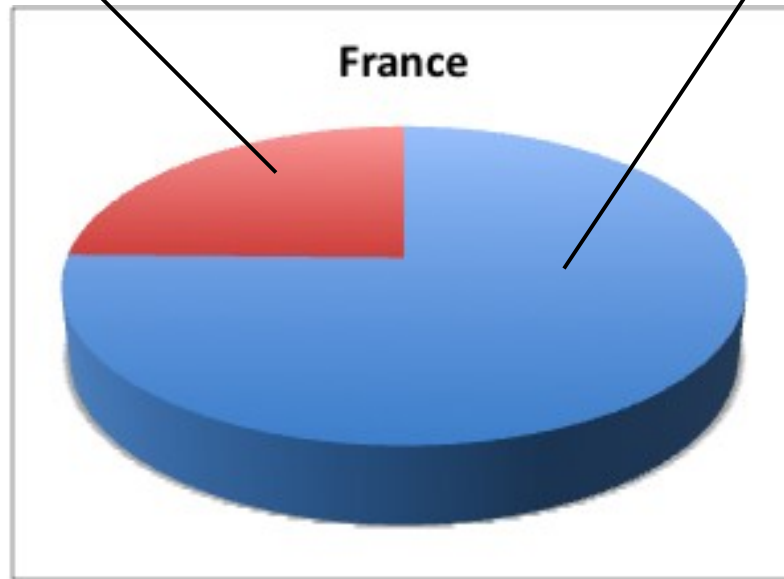


Resource consumption

➤ Resources used by

Foreigners

National users





French users

➤ Ratio of national users in VOs

French VOs

HEP

Others

%	VO name
6	atlas
96	vo.formation.idgrilles.fr
29	biomed
95	vo.lal.in2p3.fr
12	dteam
96	egeode
3	cms
9	alice
51	esr
96	vo.ipno.in2p3.fr
98	auvergrid
100	vo.renabi.fr
87	astro.vo.eu-egee.org
3	gilda
97	vo.sbg.in2p3.fr
100	c ppm
90	vo.llr.in2p3.fr
93	vo.u-psud.fr
5	lhcb
48	vo.cta.in2p3.fr
100	vo.irfu.cea.fr
100	vo.lpsc.in2p3.fr
89	vo.ipnl.in2p3.fr
100	vo.rhone-alpes.idgrilles.fr
100	vo.grif.fr
22	calice
100	vo.apc.univ-paris7.fr
11	ilc
100	vo.ucad.sn
100	vo.lpnhe.in2p3.fr
100	vo.mcia.fr
100	vo.msfg.fr
17	auger

Tier-2: GRIF

- CEA/IRFU
- LAL
- LLR
- LPNHE
- IPNO

Tier-2: IPHC

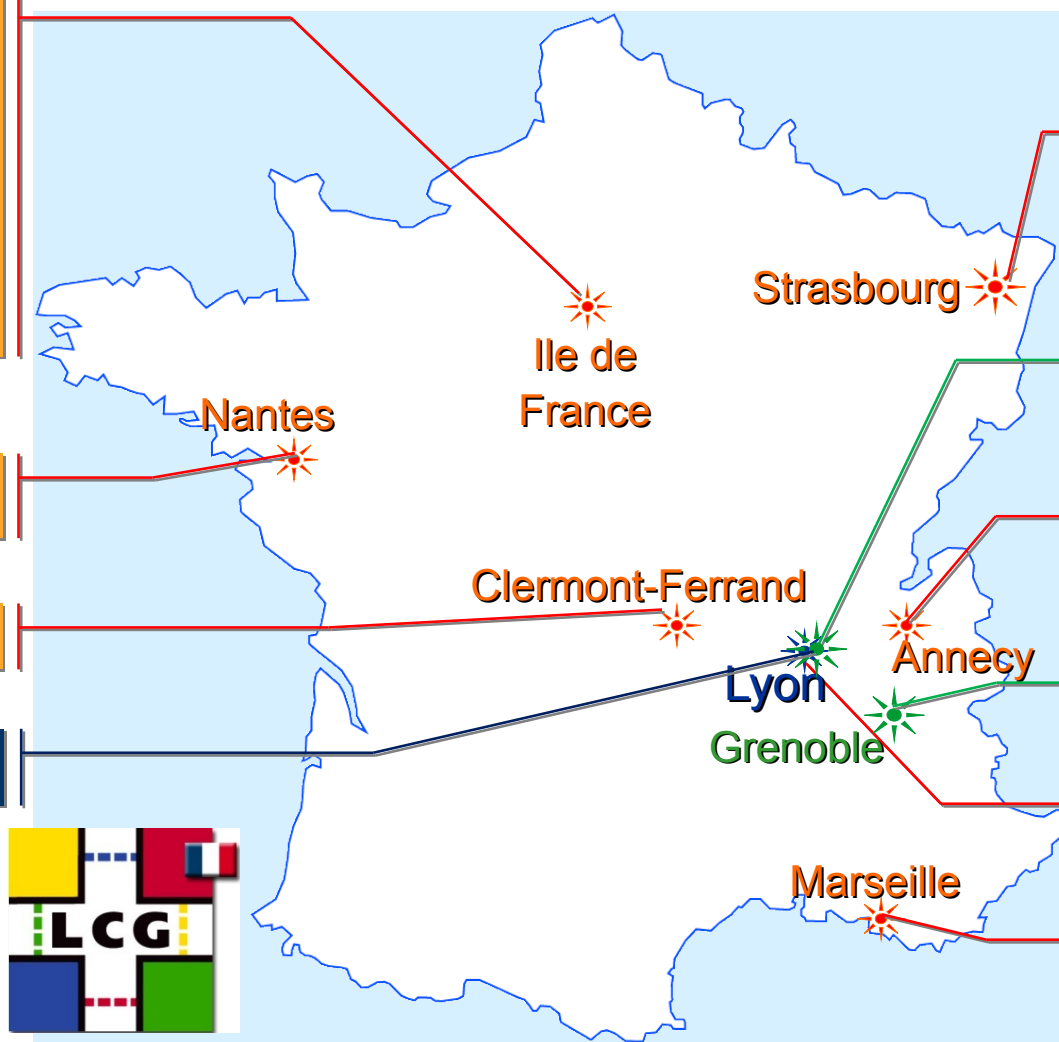
Tier-3: IPNL

Tier-2: LAPP

Tier-3: LPSC

AF: CC-IN2P3

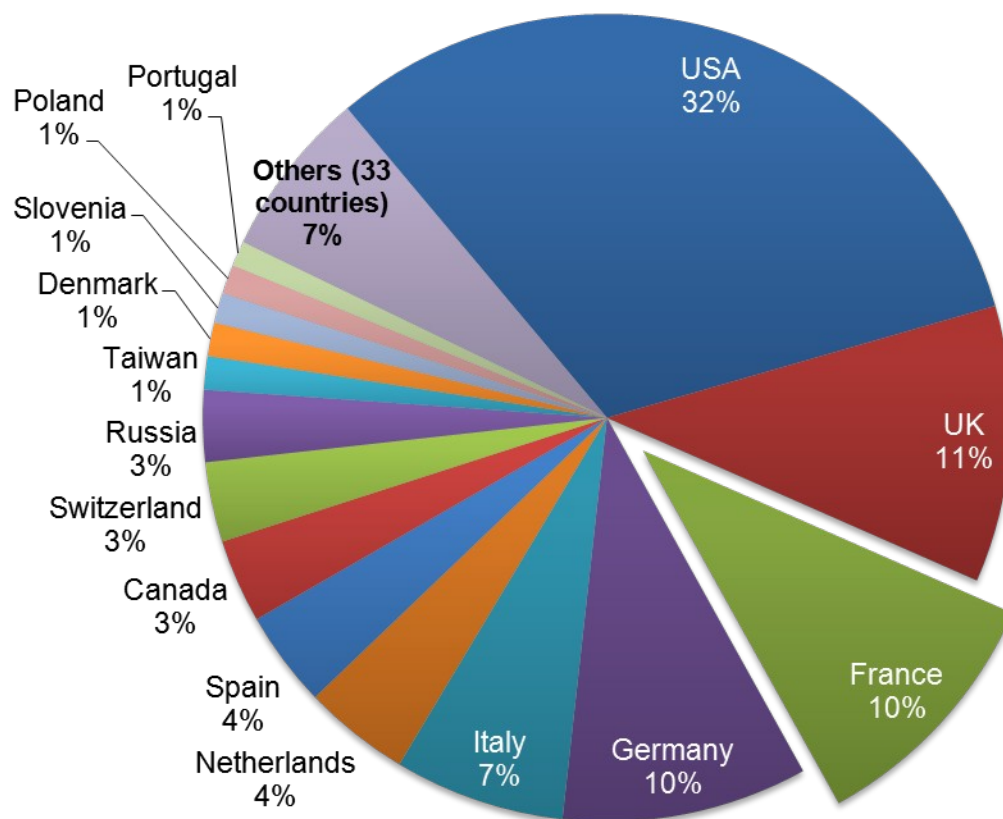
Tier-2: CPPM



French contribution

CPU contribution per country

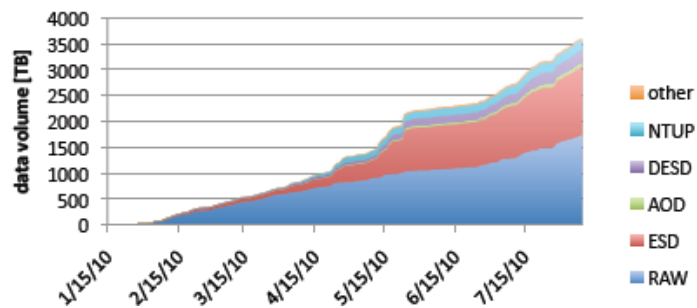
*Normalised CPU time (HEP-SPEC06)
All LHC experiments - Jan-Dec. 2010*



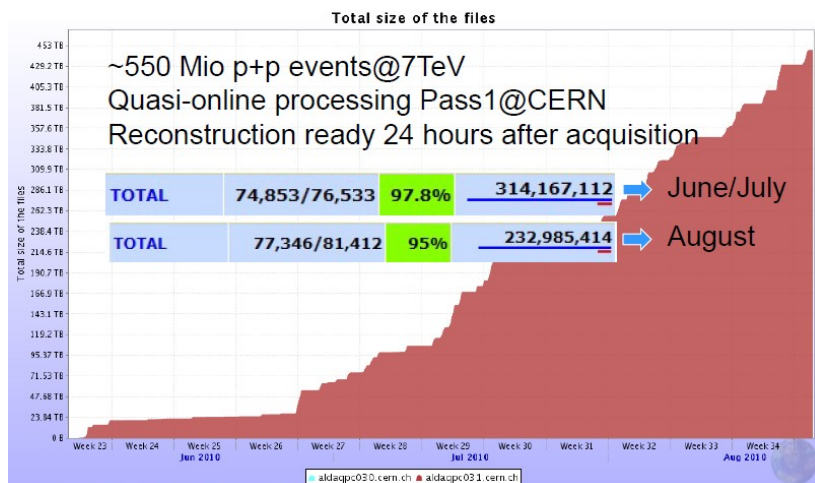
➤ Manpower

- 35 FTE
 - 30 FTE IN2P3
 - 4 FTE CEA
- 2.5 M€ (salary)

logical data volume



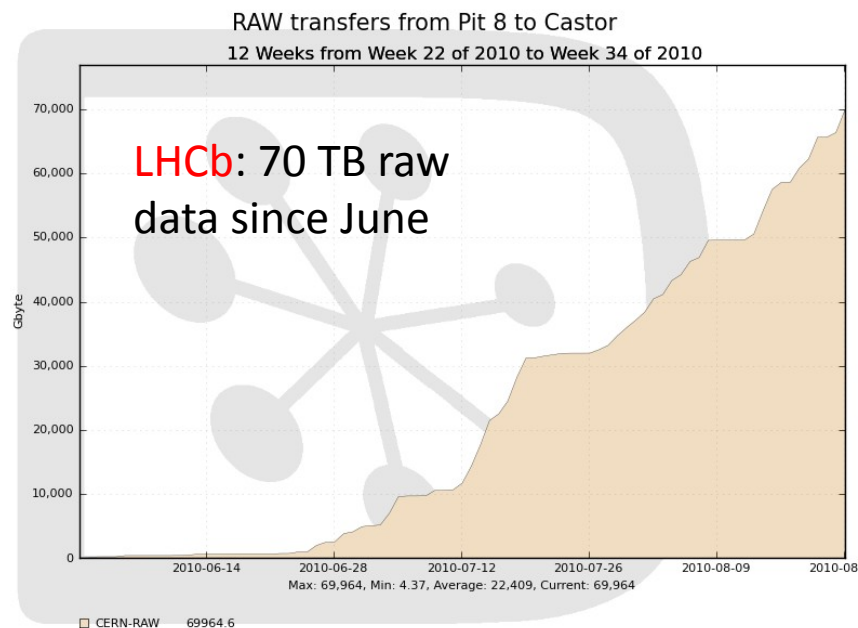
ATLAS: 1.7 PB raw



ALICE: 550 TB

CMS:

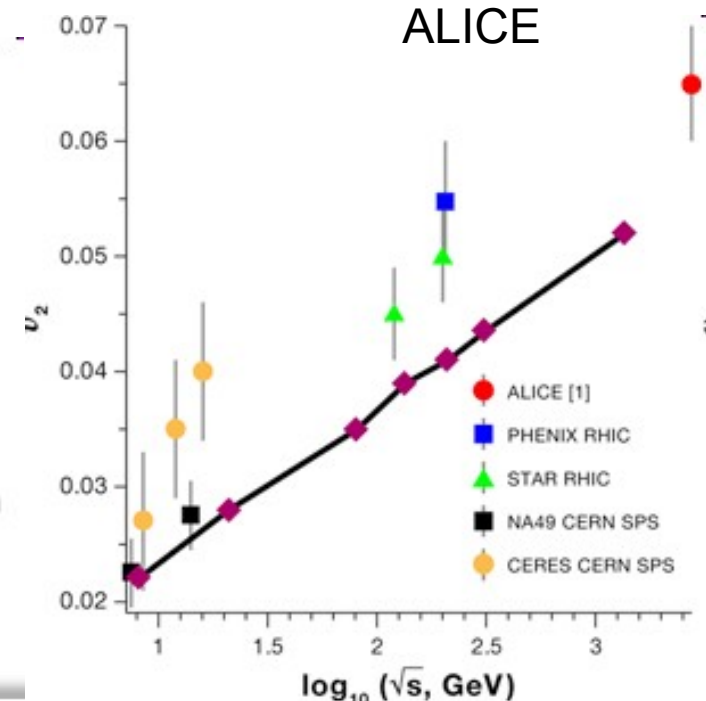
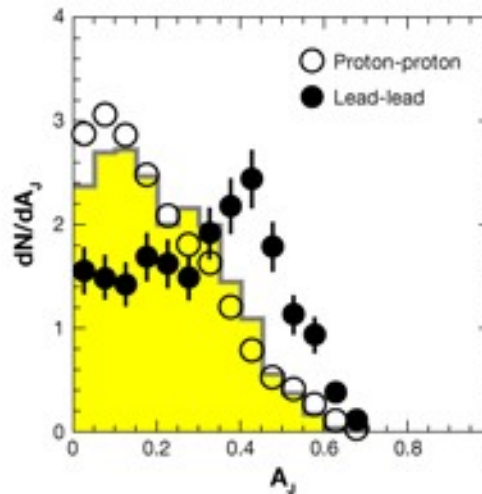
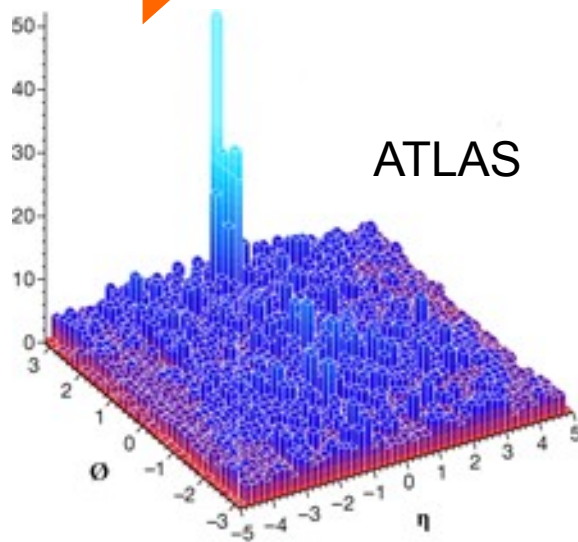
- 220 TB of RAW data at 7 TeV
- 70 TB Cosmics during this period
- 110 TB Tests and exercises with trigger.



LHCb: 70 TB raw data since June

Approx. **70** international publications after **12 months** of LHC data acquisition:
<http://lpsc.web.cern.ch/LPCC/index.php?page=lhc-articles>

Exploring ATLAS and ALICE first results, a “Little Bang” arrives at the LHC: unexpected event in quarks and gluons plasma



Observation of a Centrality-Dependent Dijet Asymmetry in Lead-Lead Collisions
 at $\sqrt{s_{NN}} = 2.76$ TeV with the ATLAS Detector at the LHC
 Phys. Rev. Lett. **105**, 252303 (2010) – Published December 13, 2010



Astronomy & Astrophysics

➤ Large data processing capability needed

- Current and future telescopes data (PB / year)
- Hybrid supercomputing – grid computing needs

➤ Current applications

- Modeling of chemical structure in star nesting areas
- Solar system orbits stability
- Tracking of spatial debris
- X-rays emission sources modeling



A&A community diff culties

➤ **Grid remains diff cult to exploit**

- Grid application development is a complex process
- Hybrid computing needs
- Need to interface to community data standards

➤ **Manpower needed to**

- Integrate legacy application
- Interface to community data sources
- Pre-deploy grid-enabled user-friendly domain applications



Earth Sciences

➤ **ES deals with wealth of data**

- Earth monitoring networks and satellite observations continuously transmitted
- Community-standard DMS and formats
- Paradigm shift: data-driven → data intensive

➤ **Manpower needed to create a SOA**

- Community data sources and services integration
- Hybrid grid and HPC computing needs
- Hide infrastructure complexity to users and deliver community services



ES Community

➤ Applications to

- Parallel data mining; Large scale synthetic data analysis; Non-linear and parametric methods.

➤ International Virtual Organizations

- **ESR**: Earthquake and Seismology; Atmospheric sciences; Hydrology & Hydrodynamics; Climate; Biodiversity
- **EGEODE**: seismic and geophysics exploration

➤ Community building

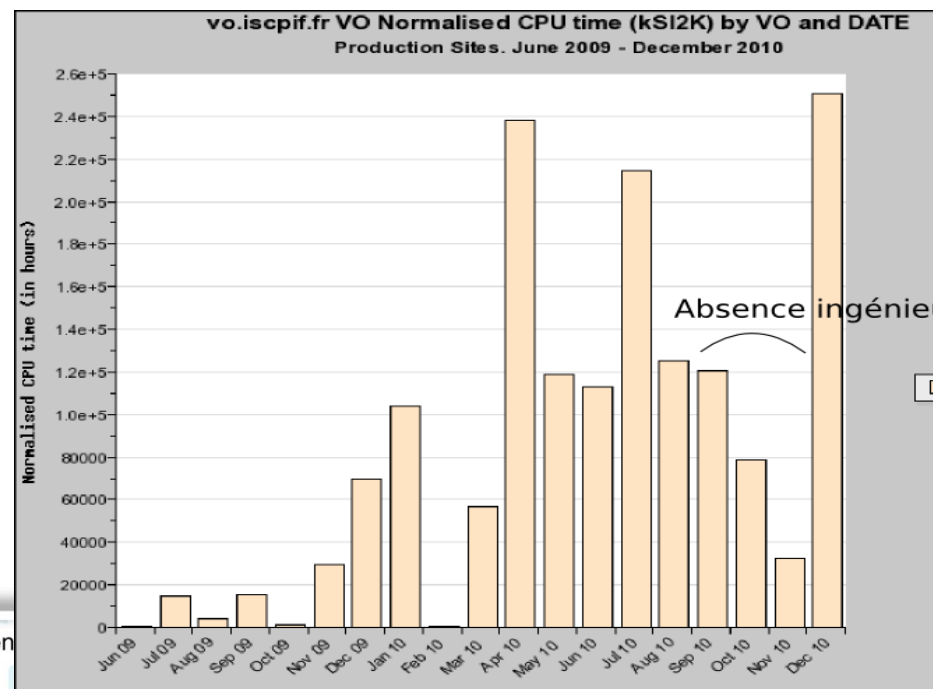
- EGI VRC
- ESFRI-PP projects: EPOS, EMSO...
- Other in Europe (SeeGrid, EnviroGrid), Asia (Taiwan, Thailand, Vietnam, Japan), Latin America (GISELA)

➤ Broad community

- Biology, Humanities, Physics, Geography...
- Methods: stochastic optimization, system variability, multi-scale modeling, image analysis...
- VO vo.iscpif.fr

➤ Computing infrastructure needs

- Large-scale computing application
- Generic, accessible infrastructure





Complex Systems

➤ **OpenMOLE grid-enabled generic numerical models experimental platform**

- No pre-deployment
- Robust submission algorithms
- Data flow control
- Extensible (plugins)

➤ **Perspectives**

- GP-GPU resources integration
- Stratuslab integration for deploying application VM



Grid Observatory

➤ Grid traces collection portal

- Started in EGEE-III, online since October 2008
- Jobs traffic, files access, middleware activity

➤ Tool for studying grid behavior

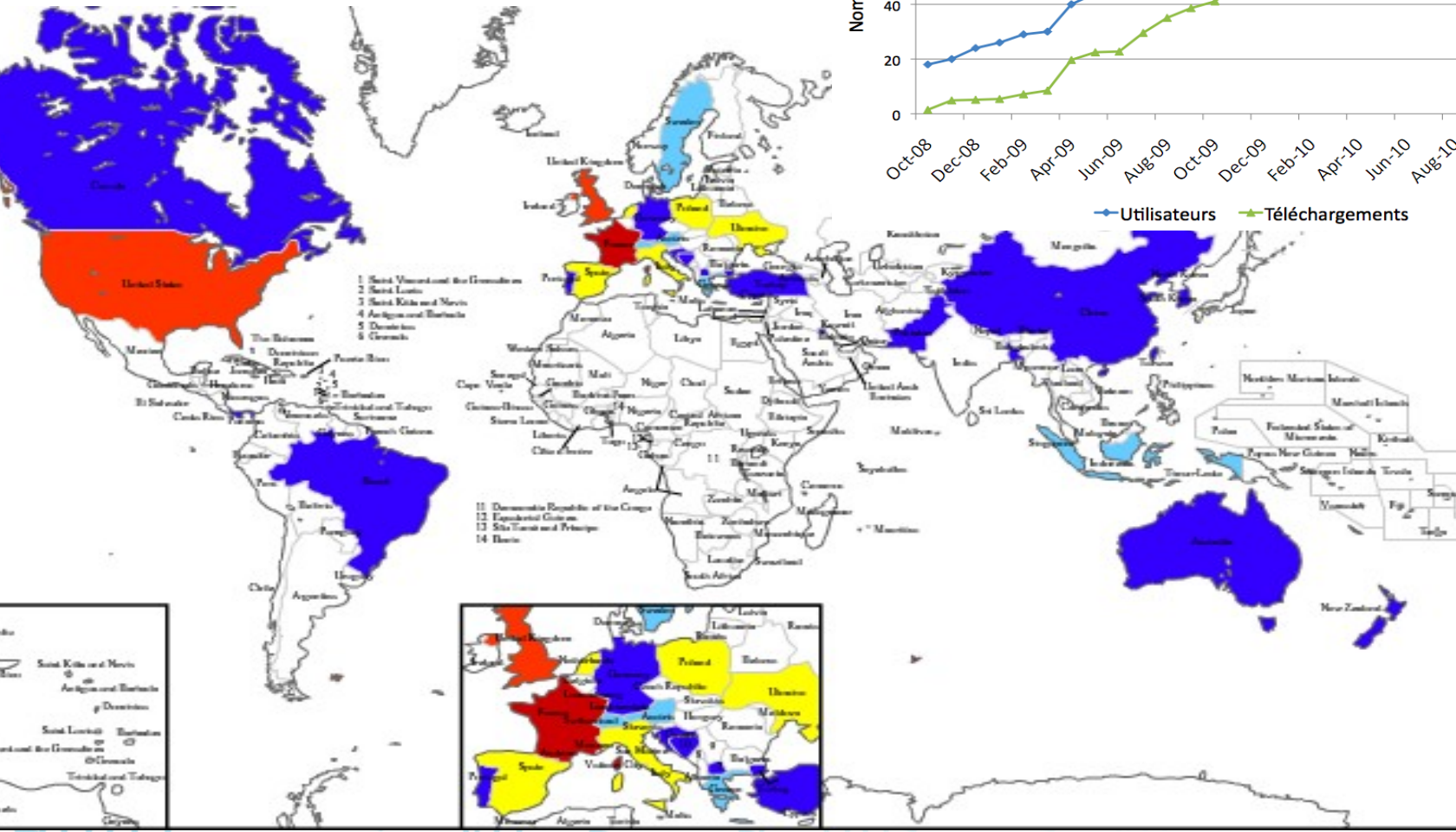
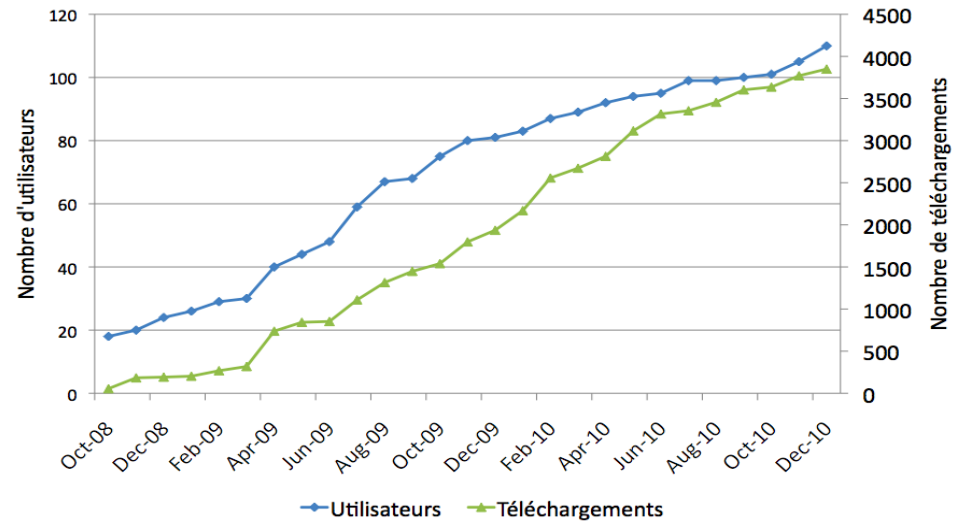
- Research-production link



www.grid-observatory.org



Usage statistics





Grid Observatory work

- **Establish a digital curation process for the behavioral data of the EGEE/EGI grid**
 - Collection, preservation, indexation, querying
 - Continuous and exhaustive datasets
 - For scientific and engineering usage
- **Contribute to analysis and modeling**
 - Complex systems description
 - Statistical and Machine Learning models and optimization

[C. Germain-Renaud, A. Cady, P. Gauron, M. Jouvin, C. Loomis, J. Martyniak, J. Nauroy, G. Philippon, M. Sebag. The Grid Observatory. 11th IEEE Int. Symp. on Cluster, Cloud and Grid Computing, 2011]

➤ **Grids Meet Autonomic Computing Workshop with IEEE/ACM Int. Conference on Autonomic Computing**

- Panel published at
<http://www.computer.org/portal/web/computingnow/panel>

➤ **21 peer-reviewed publications**

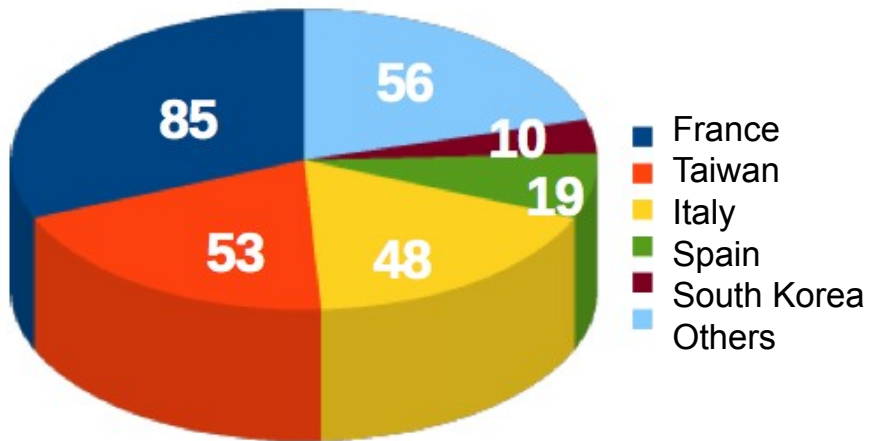
- <http://grid-observatory.org/index.php?id=72>

➤ **PhD theses, postdocs, collaborations**

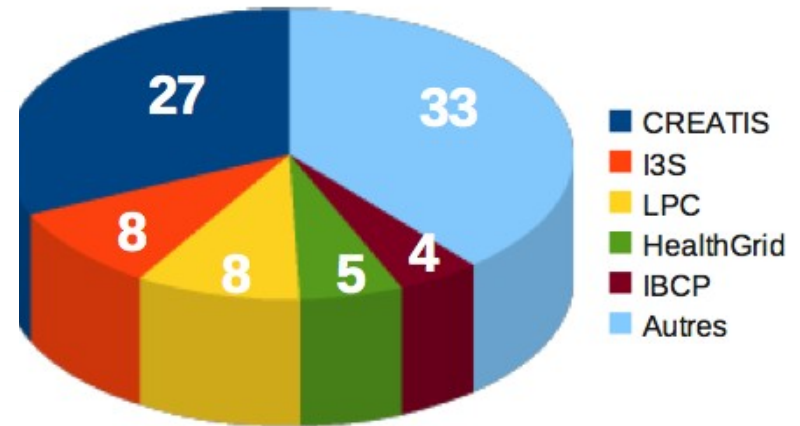
On-going work

- **Grid Observatory v2.0: "services make the repository"**
 - Semantic data organization – national project FROGS (approval pending)
 - On-line visualization
- **Keep-on with monitoring standardization effort at EMI**
- **Data collection on energy efficiency**
 - Unique facility reporting detailed data at the motherboard level

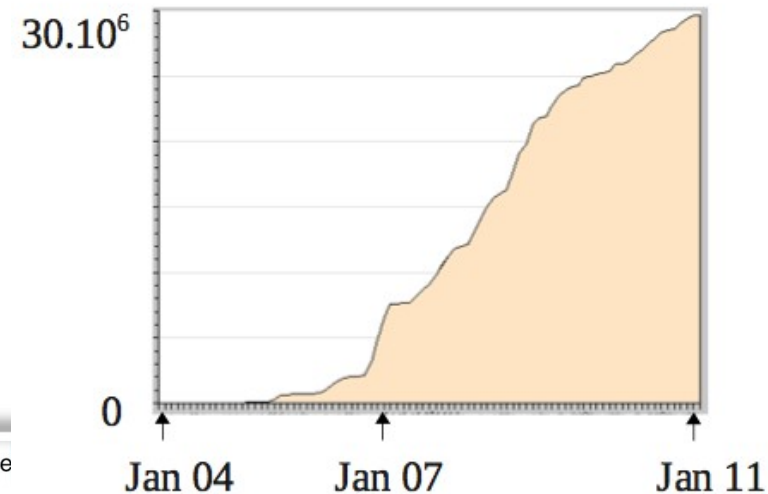
➤ Biomed VO users / country



French users / laboratory



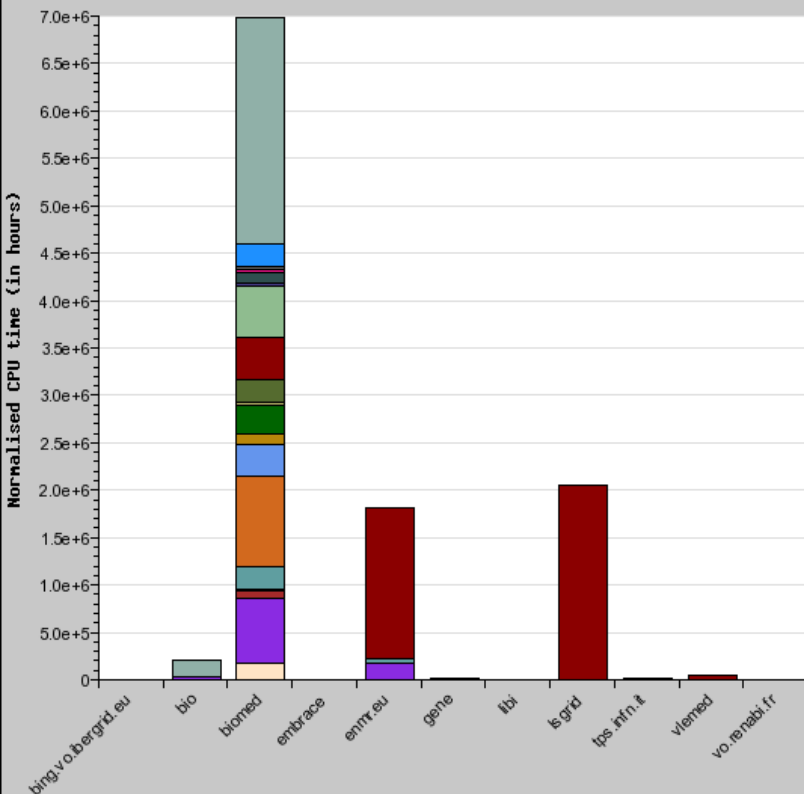
➤ Biomed VO CPU consumption





Life sciences resources consumption

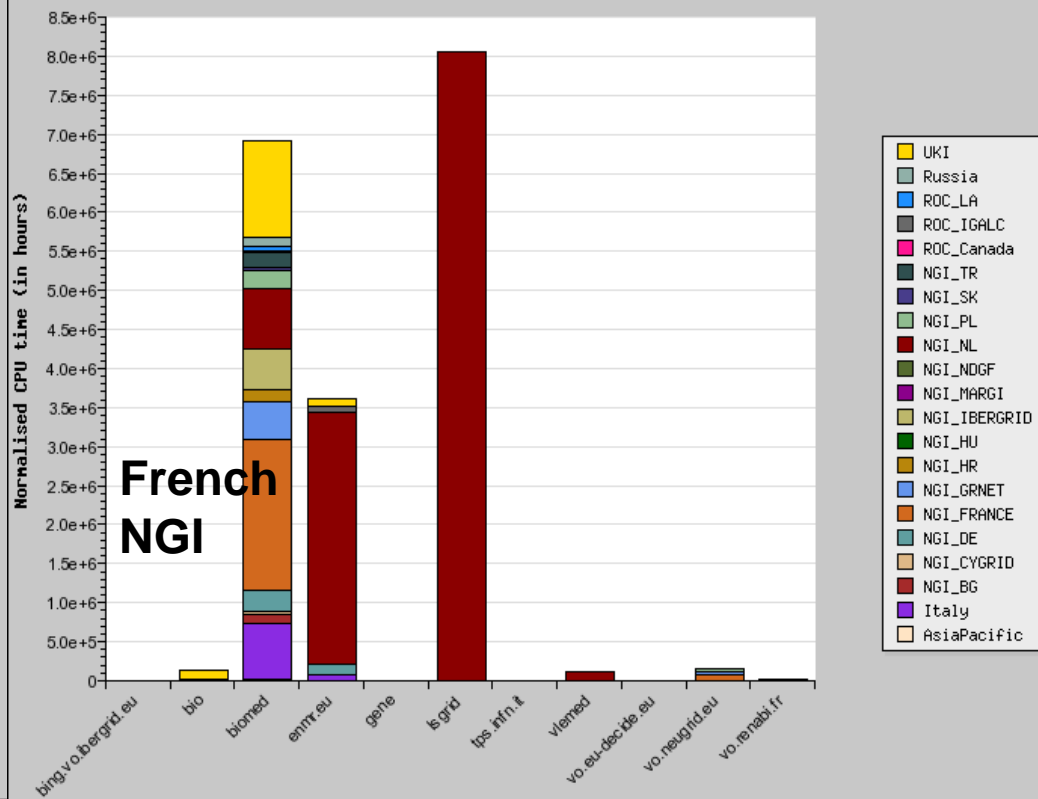
Life Sciences Normalised CPU time (kSI2K) by REGION and VO
Production Sites. January 2009 - January 2010



(C) CESGA 'EGI View': Life Sciences / normcpu / 2009-1-2010:1 / REGION-VO / Production / ACCBAR-LIN

Year 2009

Life Sciences Normalised CPU time (kSI2K) by REGION and VO
Production Sites. January 2010 - January 2011



(C) CESGA 'EGI View': Life Sciences / normcpu / 2010-1-2011:1 / REGION-VO / Production / ACCBAR-LIN

2011-01-17 20:12 UTC

Year 2010



RENABI GRISBI: large-scale bioinformatics



www.grisbio.fr

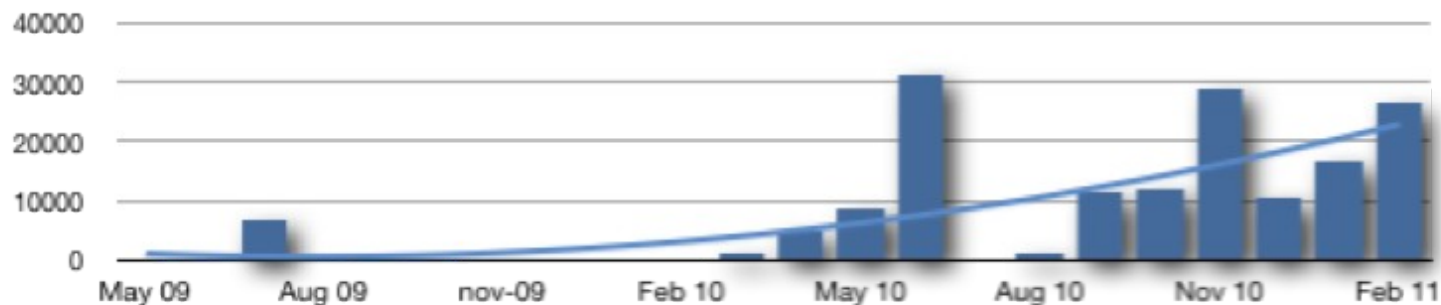
- Distributed Research Infrastructure supported by RENABI (French Bioinformatics coordination), the French NGI and the Biology funding agency IBISA
- 6 centers from RENABI Bioinformatics collaboration
 - 51 users
 - 1700 cores, 220 TB storage



Bioinformatics achievements

- National Bioinformatics VO: vo.renabi.fr
- Pre-deployed bioinformatics resources (data and tools), reuse of RENABI bioinformatics platform
- Yearly training events and workshops
- Scientific applications
 - Structural biology (conformational sampling in NMR structure calculation)
 - Genome-wide haplotype association study
 - High-Throughput Sequencing data analysis

Accounting jobs of vo.renabi.fr





Sharing data

➤ Files

- LFC: File catalog
- Hydra: Encryption service

➤ Metadata

- AMGA: grid-enabled RDBMS interface
- GRelC: databases mediation service

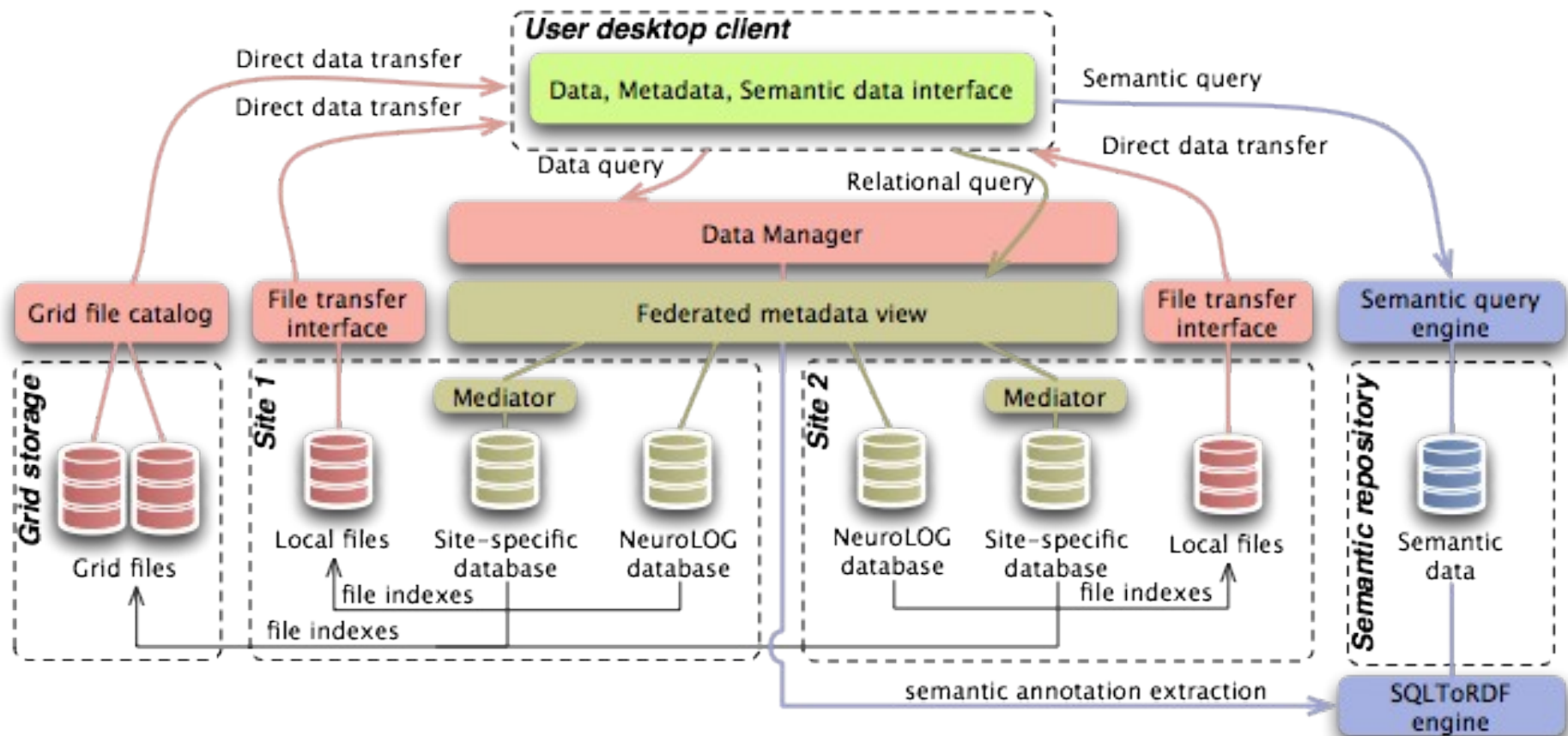
➤ Community data representation

- NeuroLOG: heterogeneous neuroradiology data
- VIP: anatomo-physiological models and medical image modalities
- GINGSENG: biosignals, epidemiological data



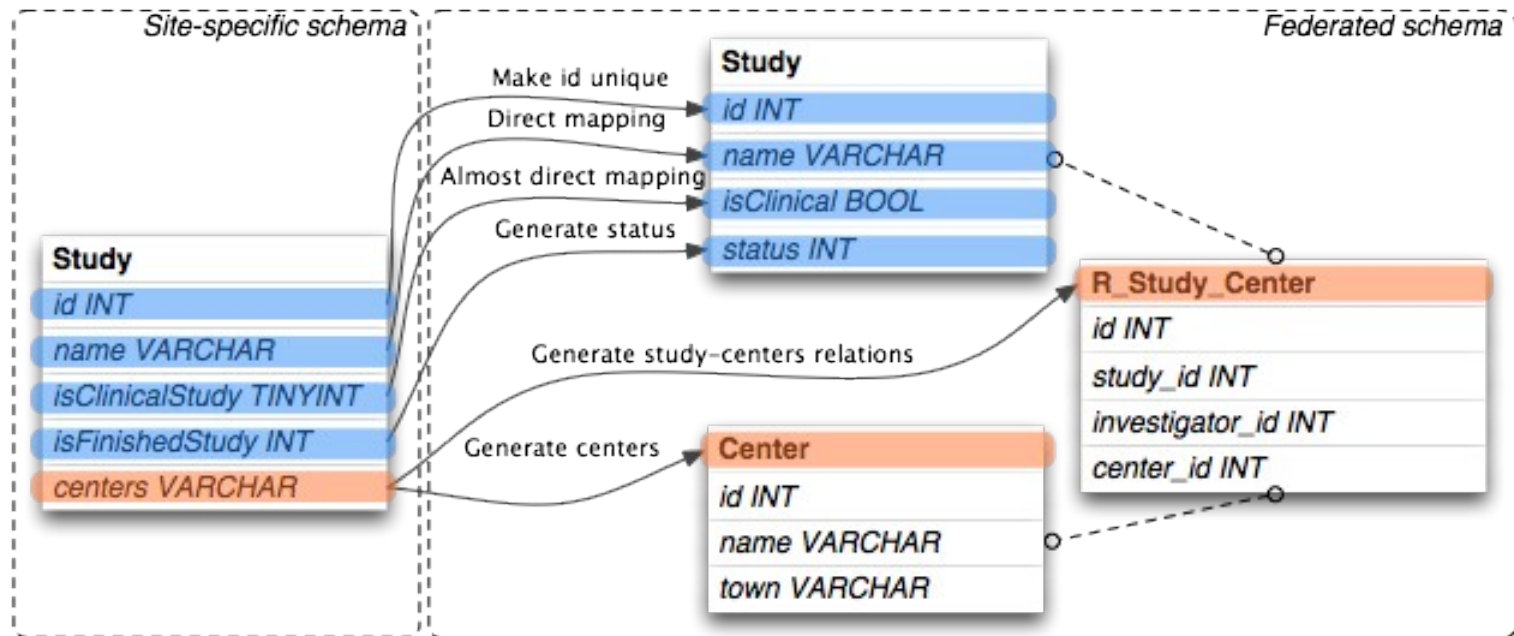
NeuroLOG neurosciences Data Management Layer

- Federation of neuroscience data stores
- Files + relational data + semantic data

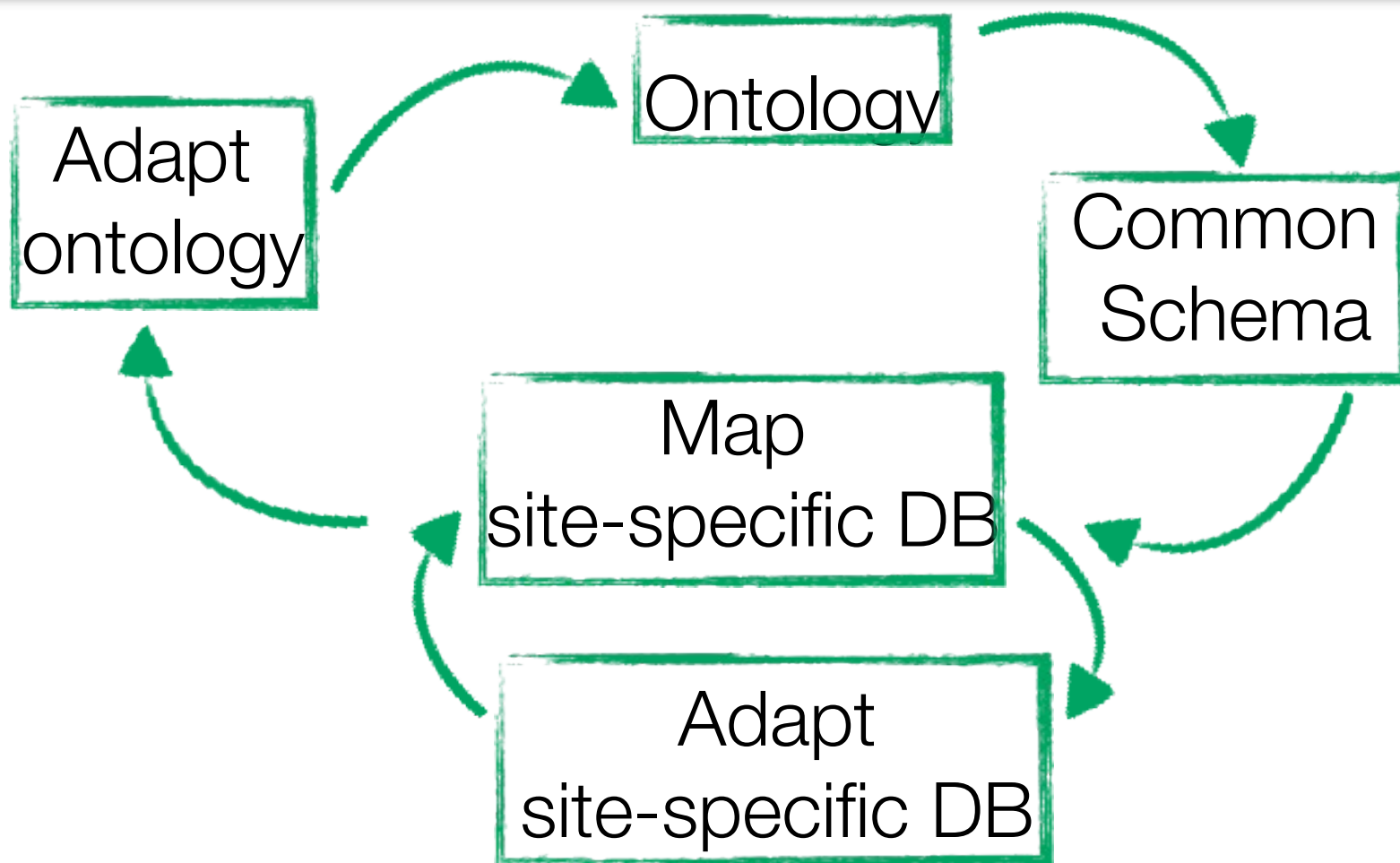


Neurosciences data mediation

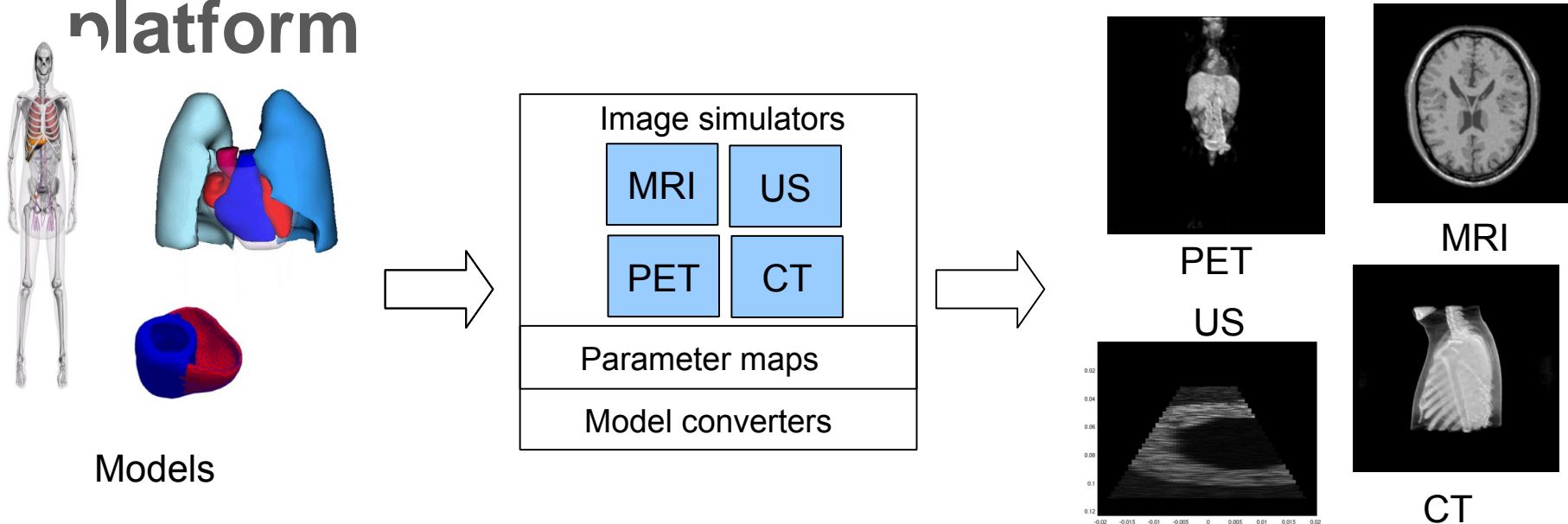
- Alignment of site-specific data schemas to a common one
- Schema mapping needed for each site



Ontology-driven approach

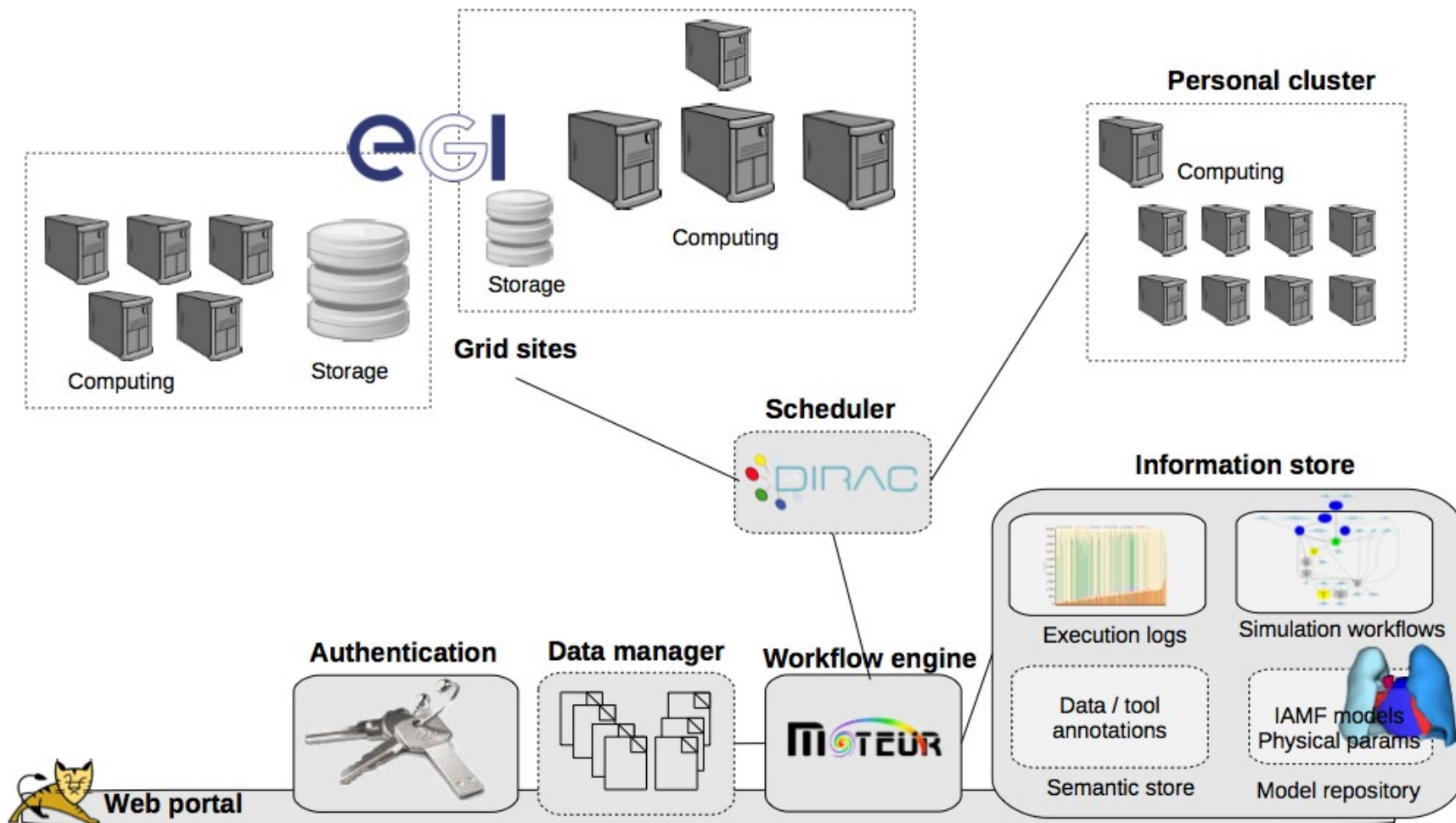


- Integrate various models and image simulation tools in a grid-enabled platform



- Exploitation to test and validate image analysis tools

VIP data integration





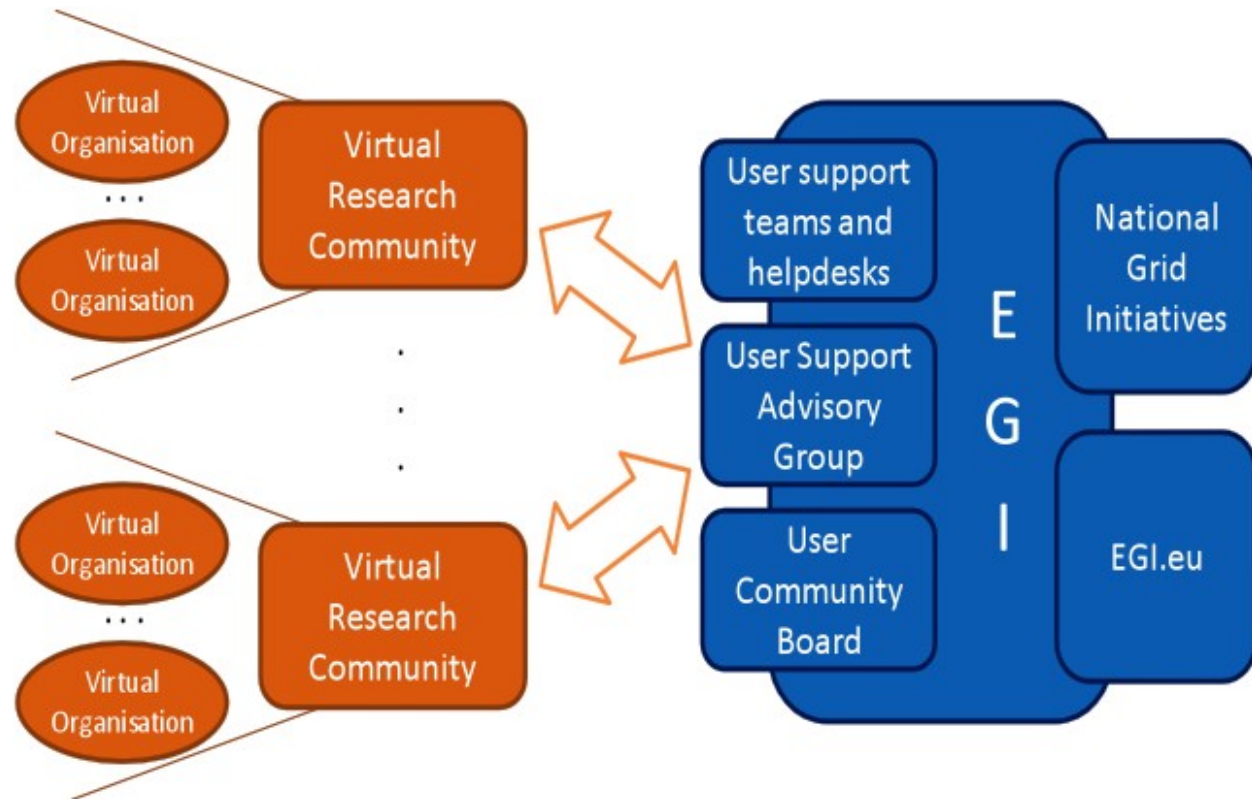
Scientific publications

- **LCG (80 papers over the last year)**
 - <http://lpc.web.cern.ch/LPCC/index.php?page=lhc-articles>
- **Grid Observatory (21 papers, since 2008)**
 - <http://grid-observatory.org/index.php?id=72>
- **Reported in the French communities since July 2010 (partial)**
 - Medical imaging: 12
 - Bioinformatics: 3
 - Complex systems: 2
 - Earth Sciences: 2

Virtual Research Communities

➤ EGI model

- But VRCs expand beyond EGI user groups



➤ EU view: ESFRIs

What is a VRC?

➤ Being representative of a community

- Self-established, self-organized (no guideline)
- Grouping existing Vos
- Unlike VOs, no technical tooling / representation

➤ Why?

- Formalize interactions with resource providers
- Capitalize experience and efforts, reuse tools
- Number is strength: promote requirements, liase with EGI

➤ How?

- Prove legitimacy
- Fund people (signif cant workload involved)



LS Grid Community VRC

➤ Participating user groups

- VOs: biomed, vlemmed, Isgrid, pneumogrid
- NGIs: Dutch, French, Italian, Spanish, Swiss
- Projects: EGI InsPIRE, LifeWatch ESFRI
- Hosting body: HealthGrid (non-profit association)

➤ VRC set up

- Started June 2010
- SLAs / MoUs with resource providers on-going
- User management tooling on-going



LS Grid Community services

➤ Provide technical services

- Operate VOs and shared services
- Improve users experience
 - Technical team for monitoring and troubleshooting (duty shifts)
 - Provide targeted user support and application porting
 - Report VO issues, promote best practices

➤ Training and induction

- Organize community training events (e.g. EGI UF)

➤ Dissemination

- Transfer knowledge among VRC partners
- Advertise actions, liaise with other groups



Difficulties in setting up VRCs

- **There is a “hole” in the EGI / EU model**
 - VRCs hardly exist today
 - Limited funding for **grid user** community organization
 - ESFRIs have agenda of their own and often not grid-addicts. They may prefer infrastructures under their control.
- **Hardly comparable status among different user communities**



Conclusions

➤ **Community building: towards VRCs**

- Very heterogeneous user communities
- Different levels of community building progress
- Little tooling available to handle VRCs

➤ **Grid experience**

- Different use of grid resources
 - Sharing / accessing community data is often a key point
- Grid investment is still high for users