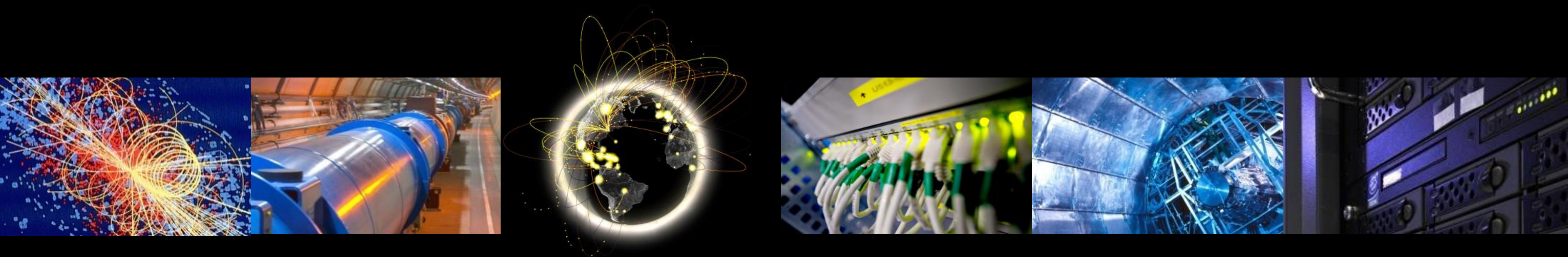
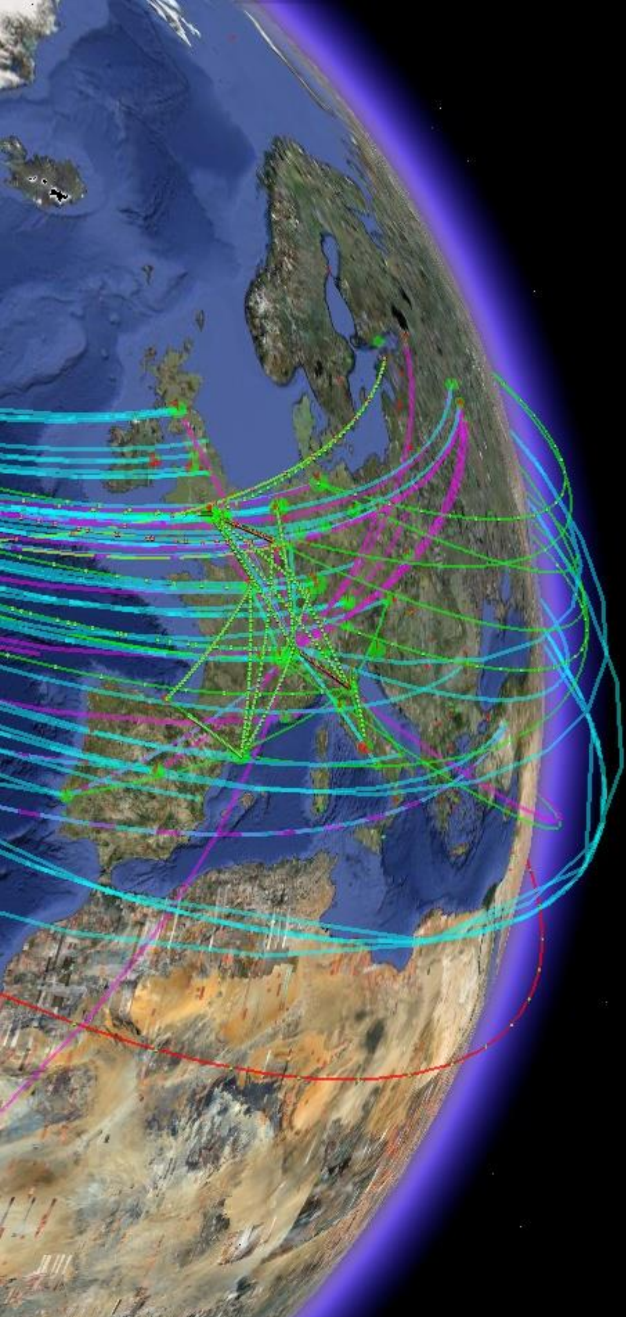
WLCG and Grid Computing Summer 2013

Fabrizio Furano

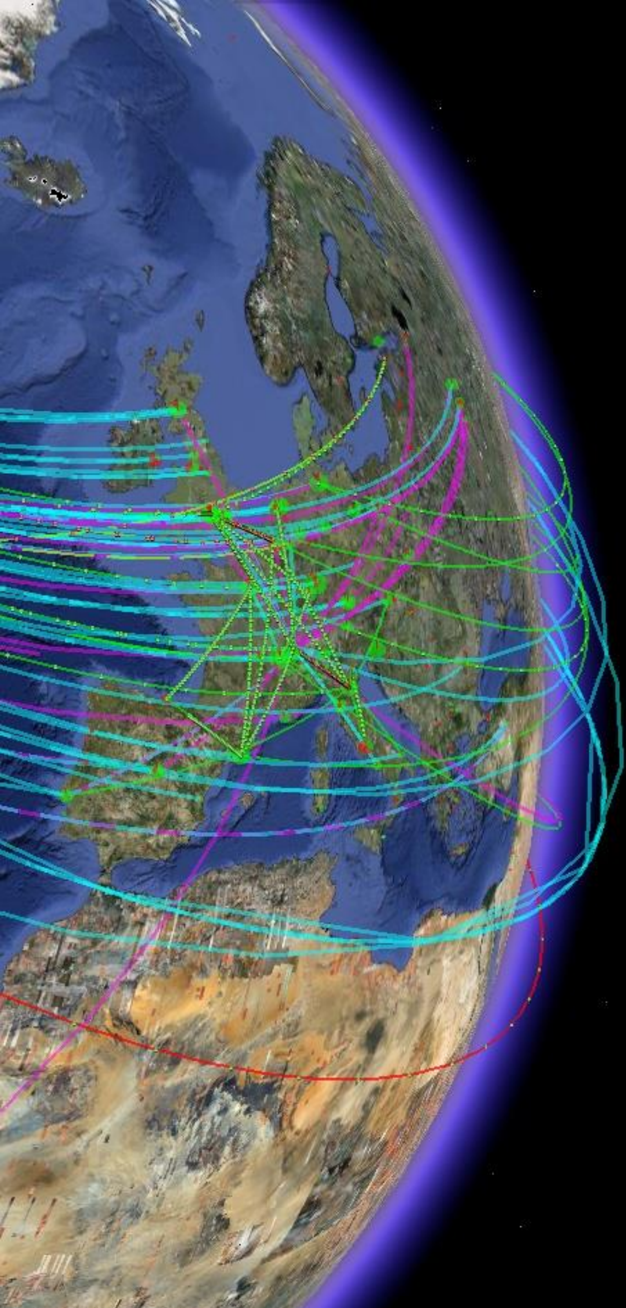
IT Support for Distributed Computing (IT-SDC) Group, CERN
WLCG
Fabrizio.Furano@cern.ch





Overview

- WLCG (today)
- Grid Computing (soon)
  - What's Next?



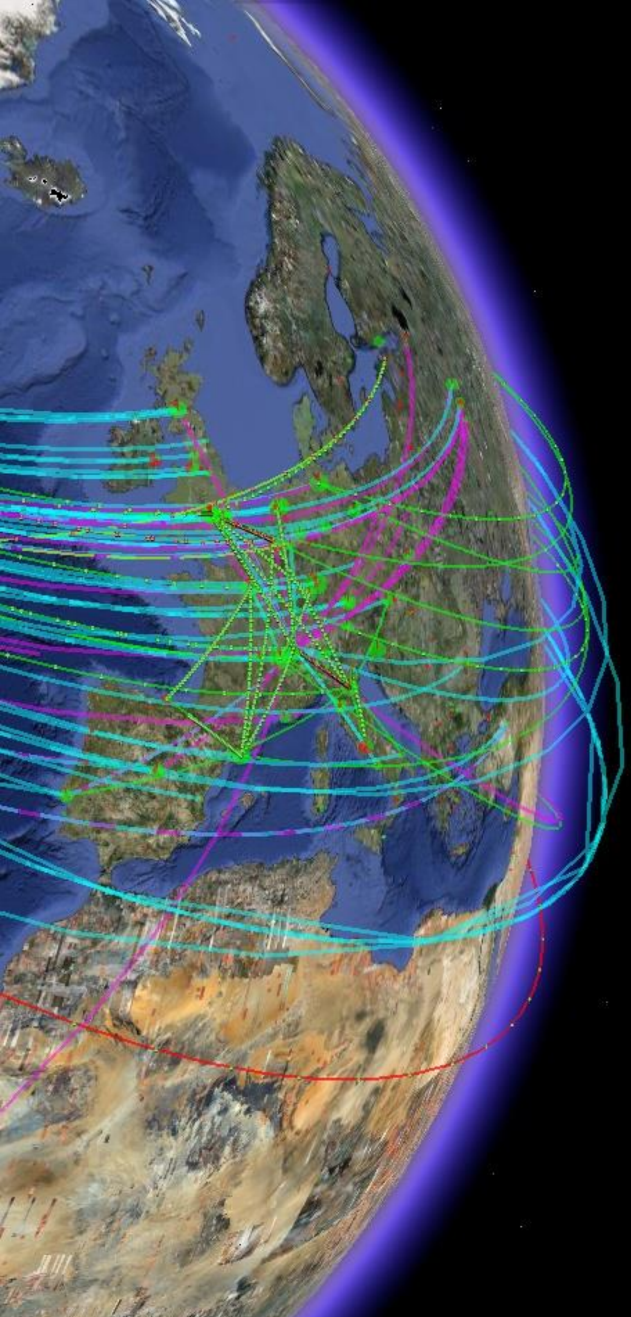
Focus

- Motivation for WLCG
- The Infrastructure
- Usage
- Next Steps
- Open questions

- Not much about the Grid technology

- Best manual and introduction:
- https://edms.cern.ch/file/722398/1.4/gLite-3-UserGuide.pdf

WLCG Worldwide LHC Computing Grid

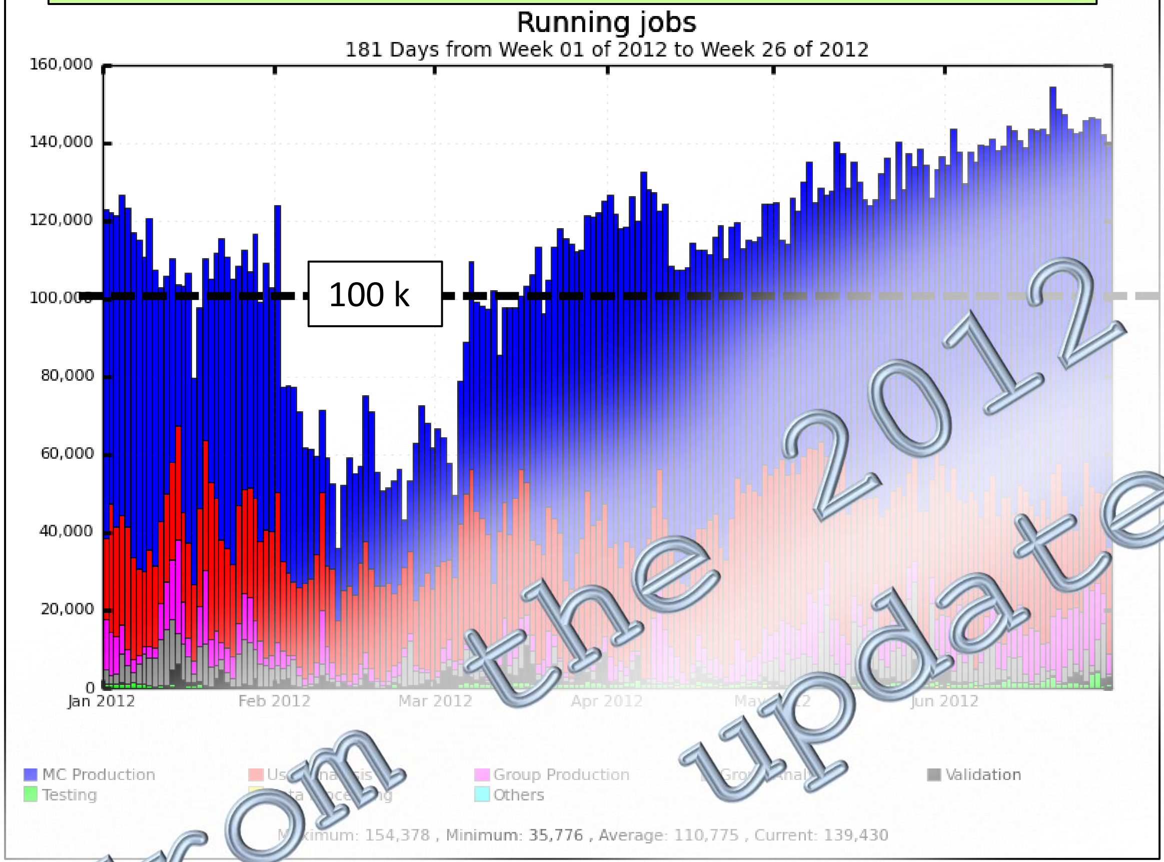


WLCG

- Why is computing for LHC a challenge?
- Why a distributed system?
- History
- Architecture
- Monitoring and Operation
- Usage

It would have been impossible to release physics results so quickly without the outstanding performance of the Grid (including the CERN Tier-0)

Number of concurrent ATLAS jobs Jan-July 2012



Includes MC production, user and group analysis at CERN, 10 Tier1-s, ~70 Tier-2 federations → > 80 sites

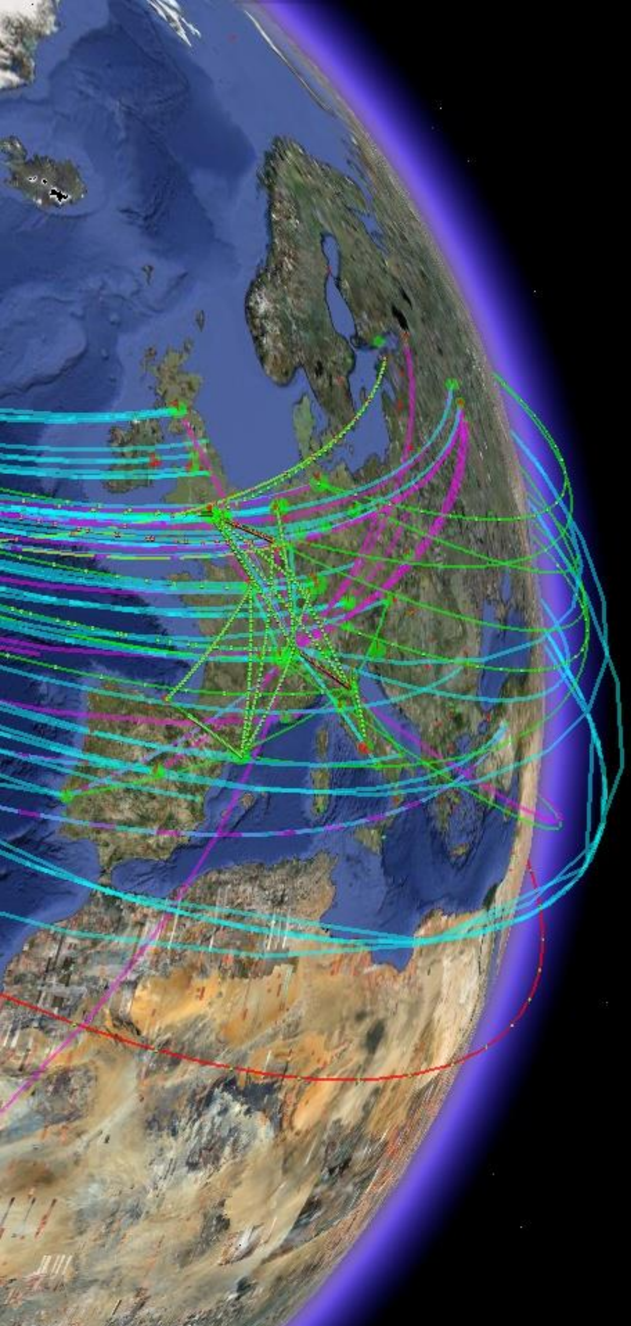
> 1500 distinct ATLAS users do analysis on the GRID

HIGGS

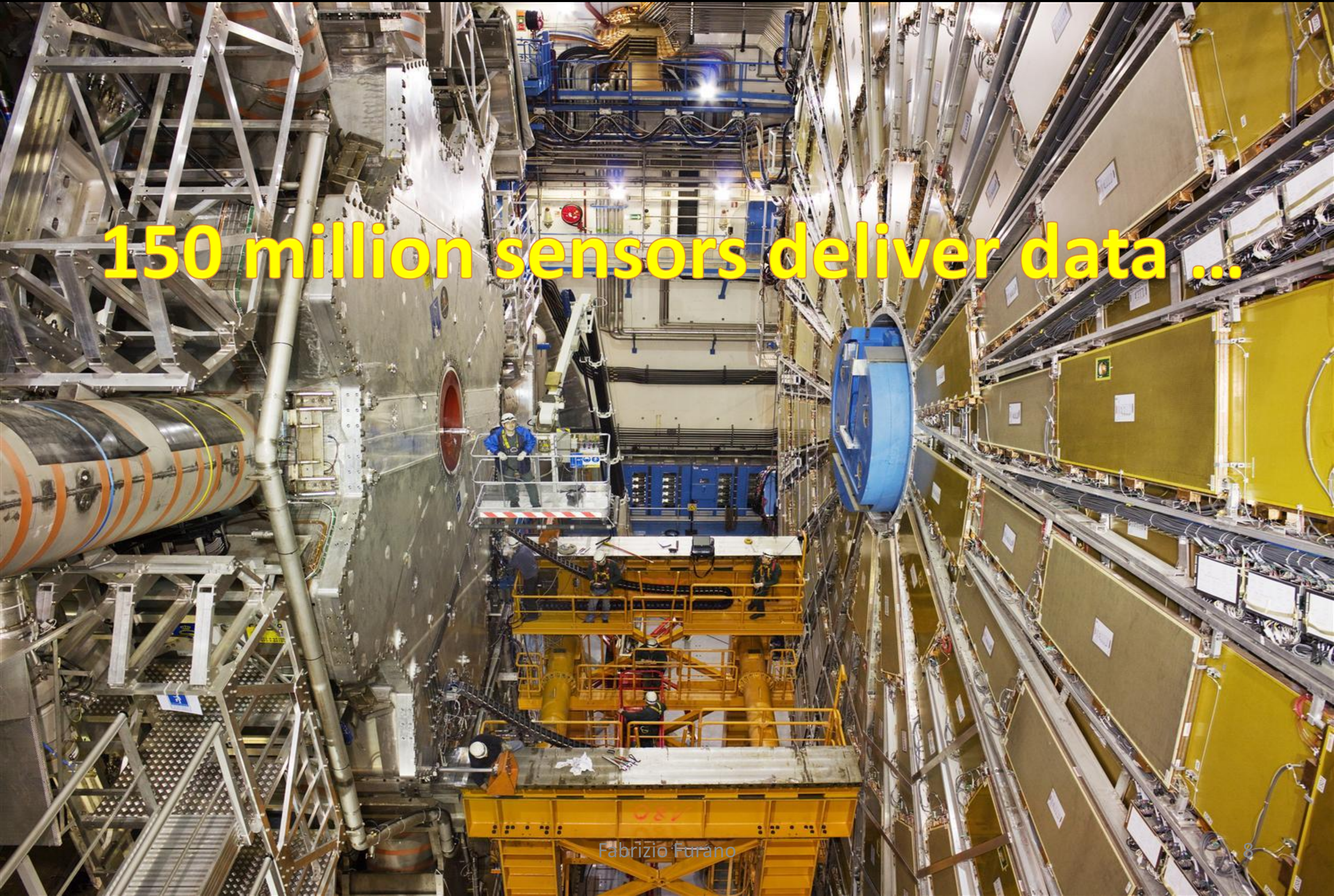
From the 2012 update

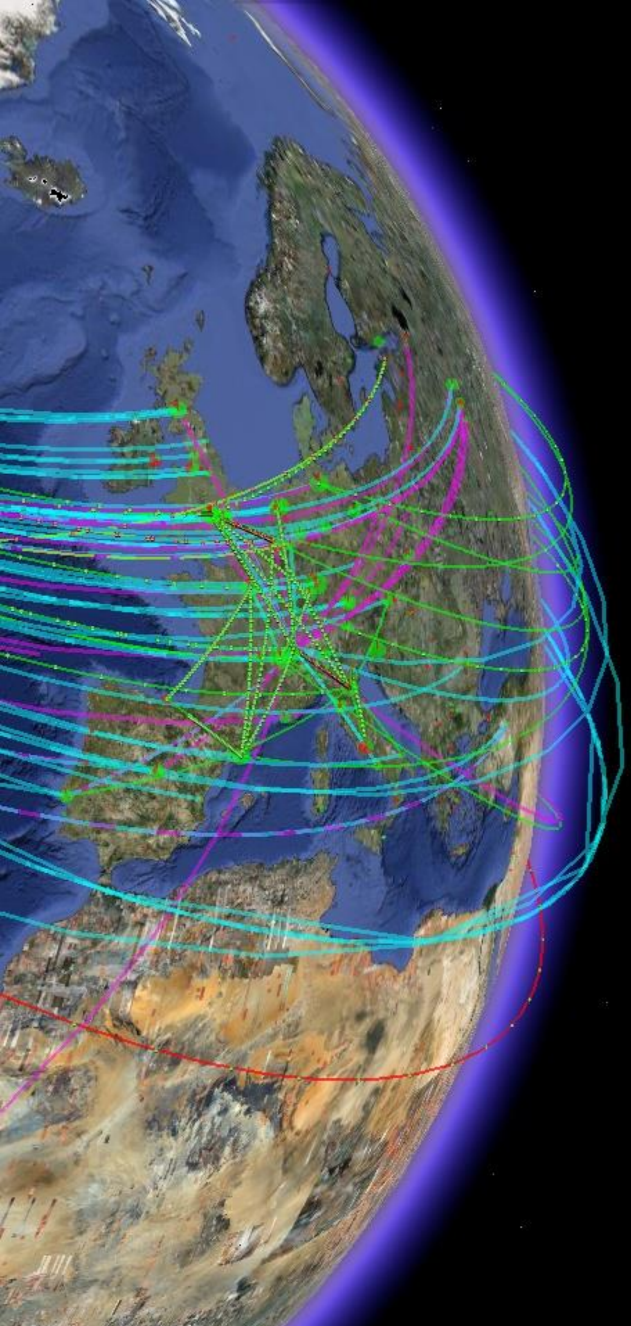
- Available resources fully used/stressed (beyond pledges in some cases)
- Massive production of 8 TeV Monte Carlo samples
- Very effective and flexible Computing Model and Operation team → accommodate high trigger rates and pile-up, intense MC simulation, analysis demands from worldwide users (through e.g. dynamic data placement)


A Large Hadron Collider
Delivering collisions up to 40 million times per second

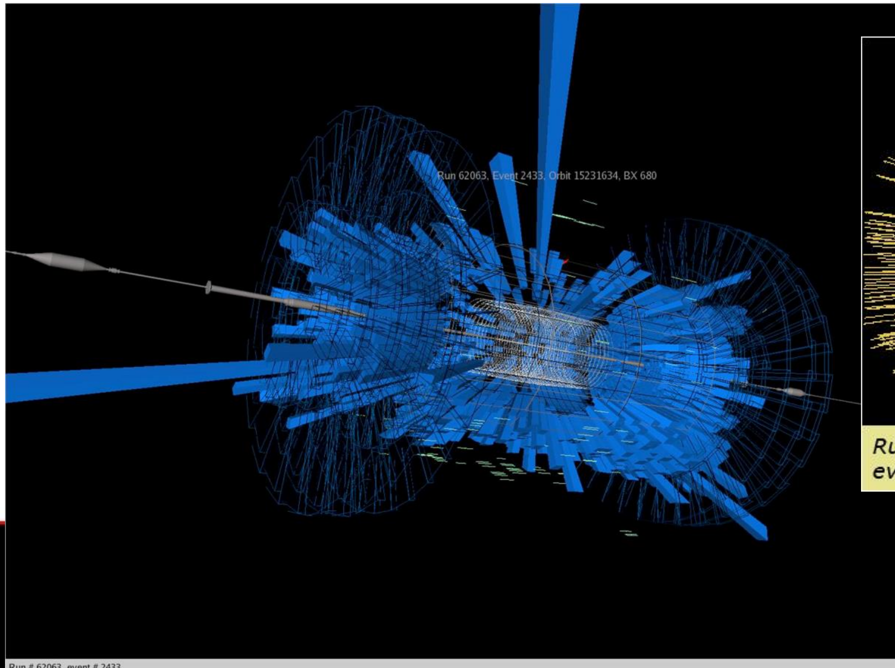


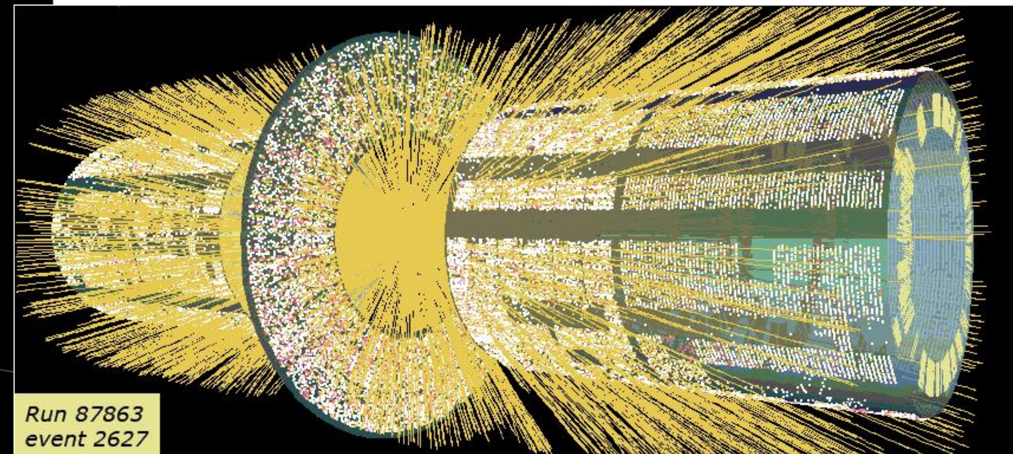
Four “Data Generators” (here : ATLAS)

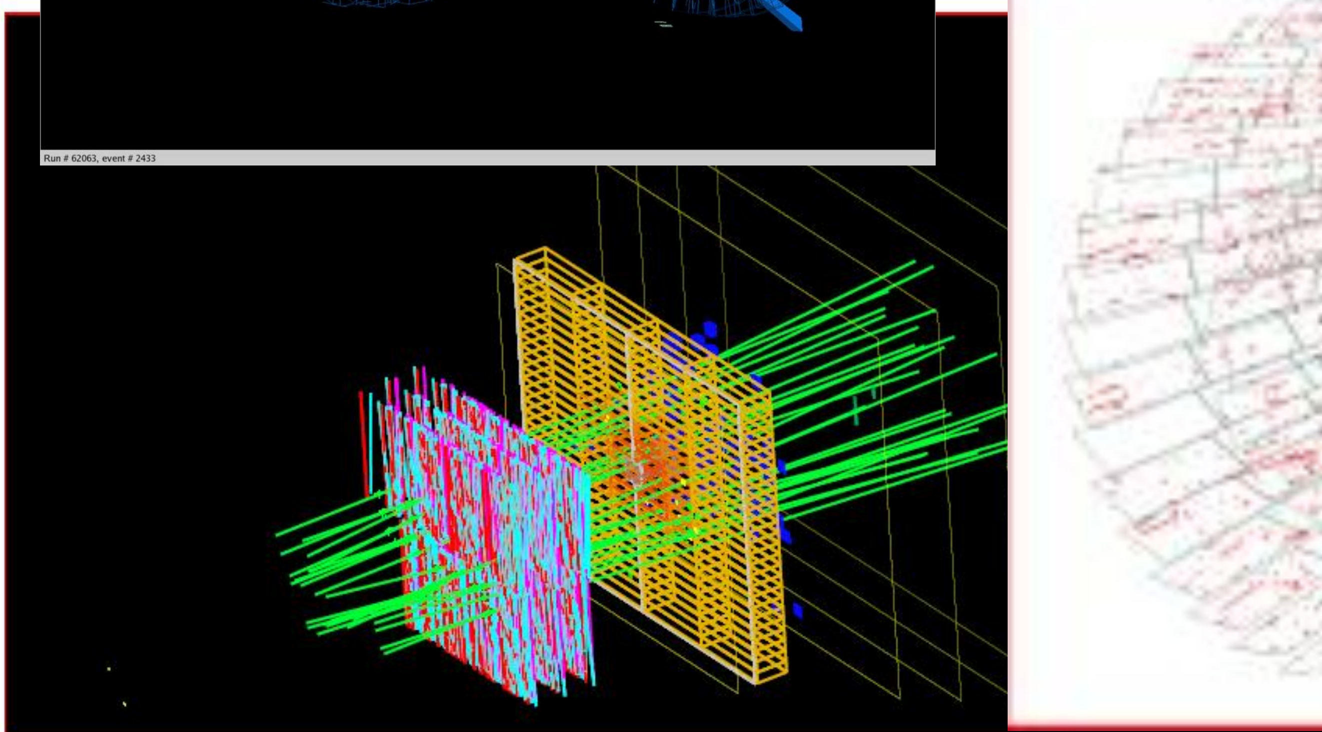


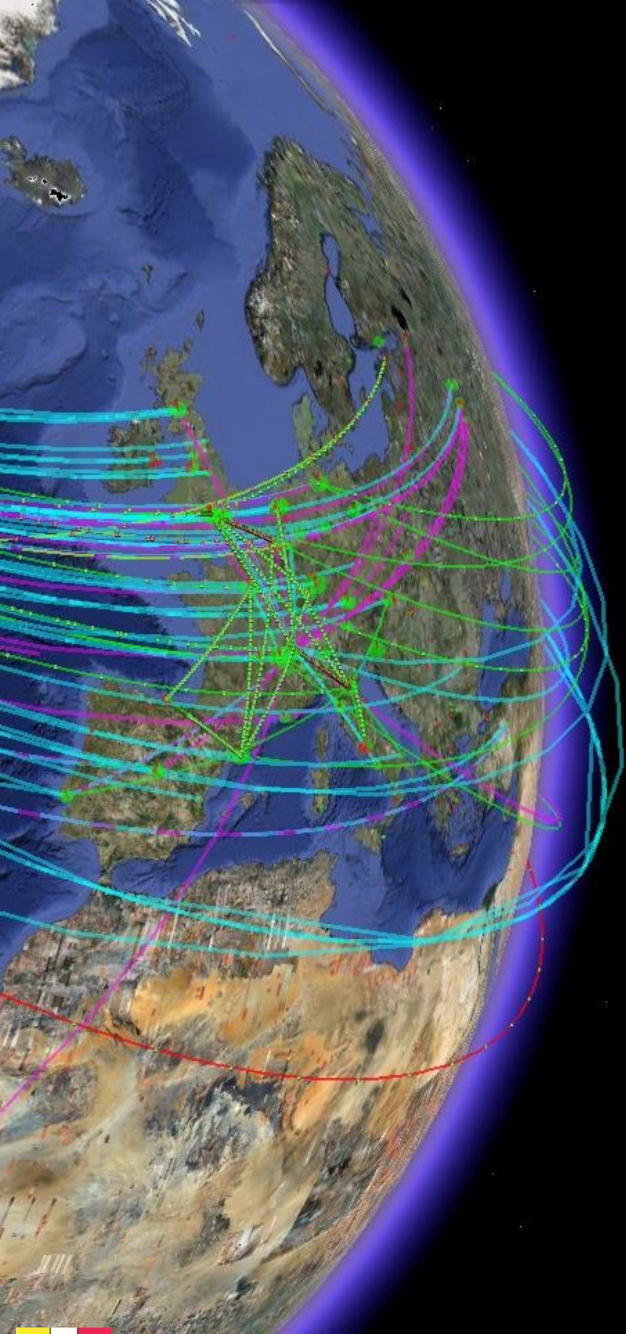


Complex events



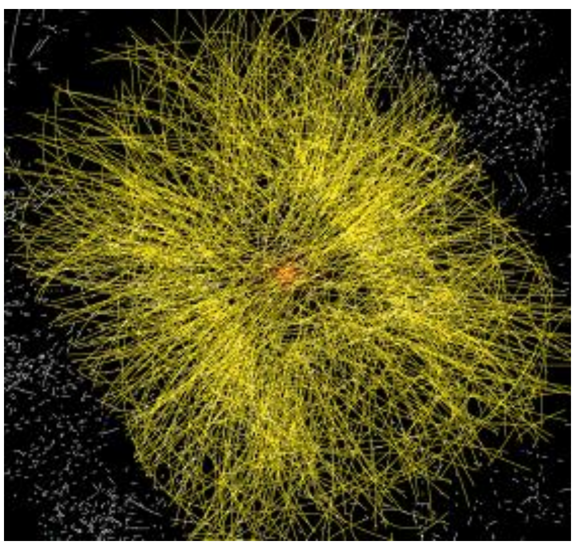


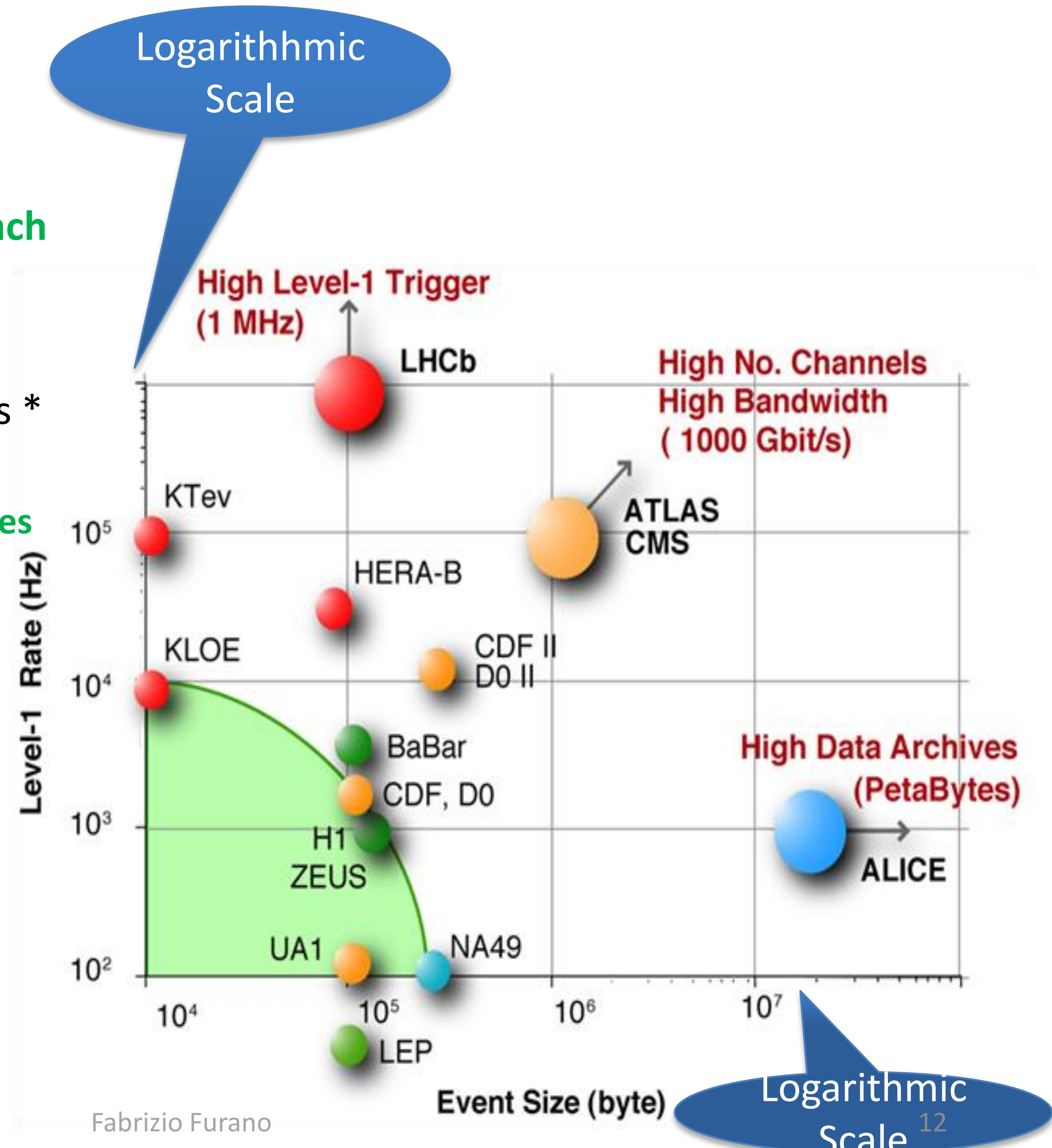


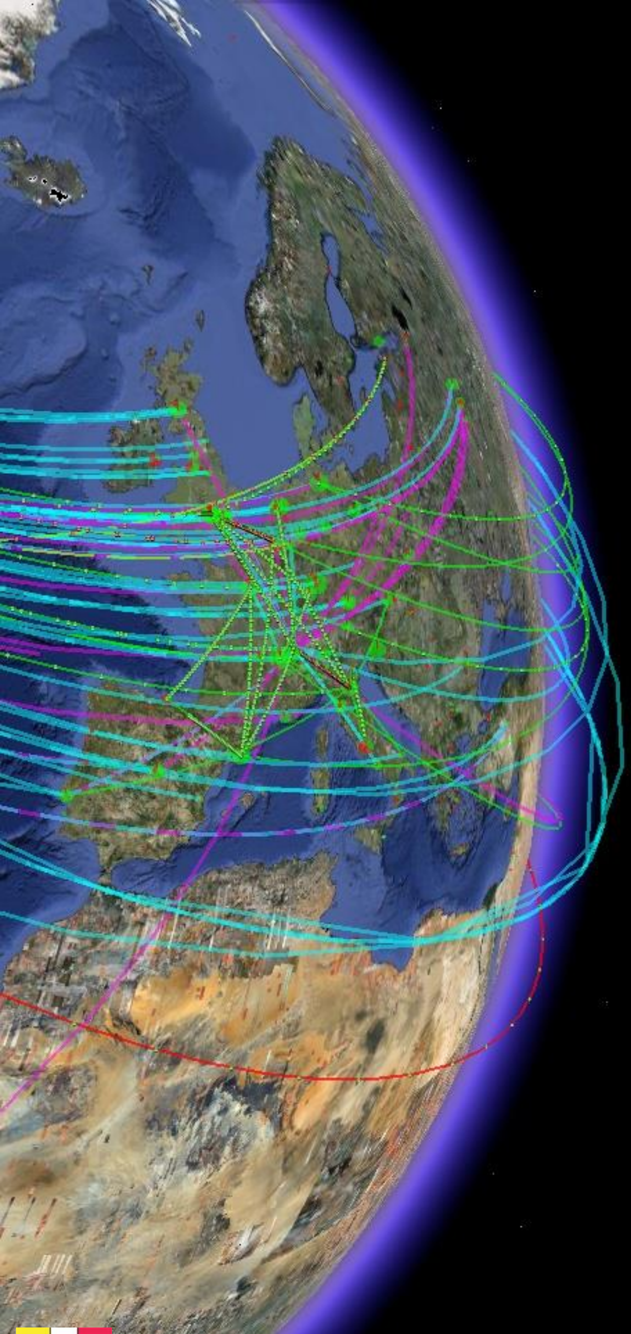


Complex Computational Tasks

- Data volume
  - High rate * large number of channels * 4 experiments
  - ➔ **15 Peta Bytes of new data each year**
- Compute power
  - Event complexity * Nb. events * thousands users
  - ➔ **340k of (today's) fastest CPU cores**
  - ➔ **45 PB of disk storage**



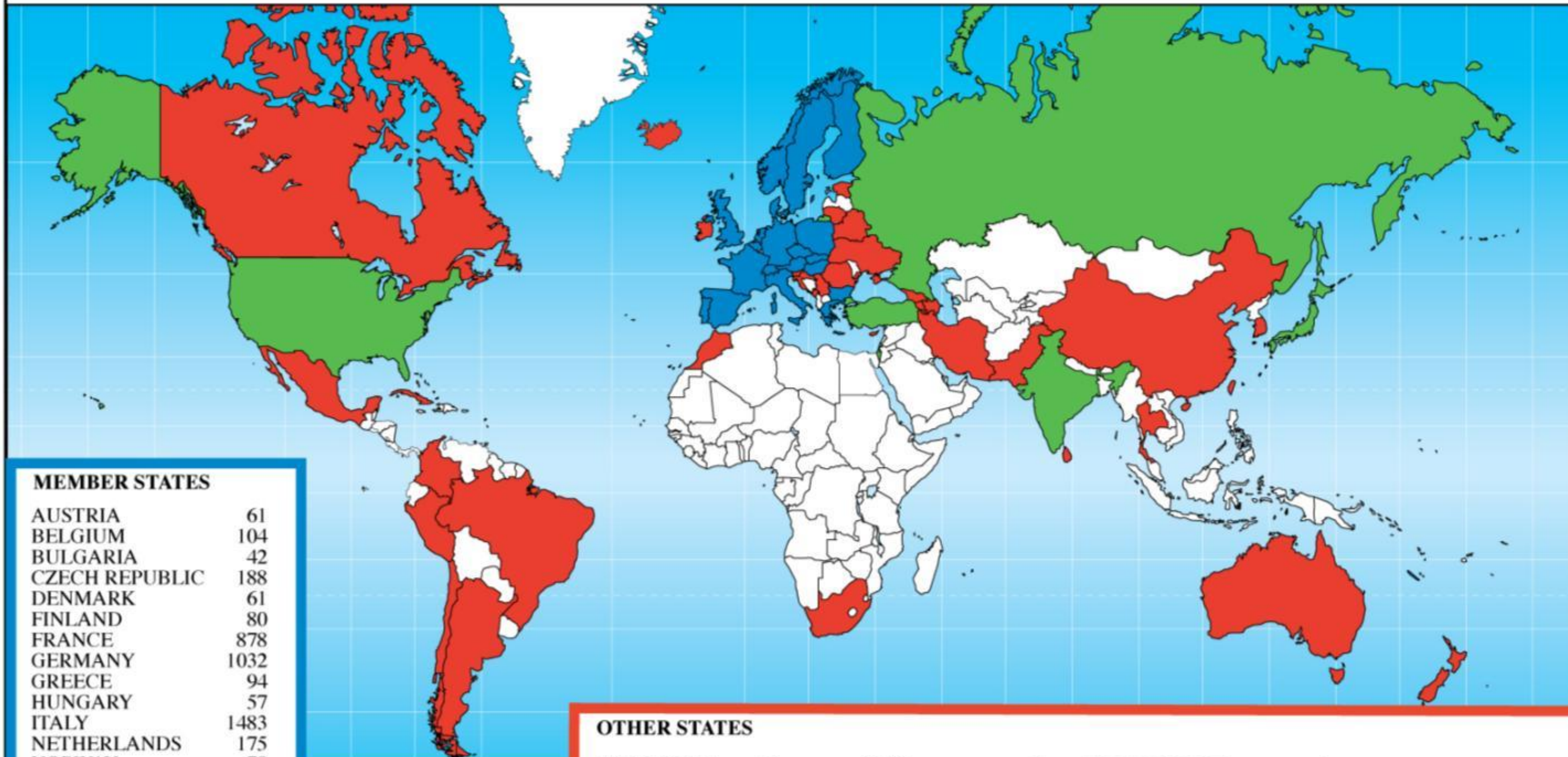




WLCG Worldwide LHC Computing Grid

Complex Large Community

Distribution of All CERN Users by Nation of Institute on 6 January 2009



MEMBER STATES

AUSTRIA	61
BELGIUM	104
BULGARIA	42
CZECH REPUBLIC	188
DENMARK	61
FINLAND	80
FRANCE	878
GERMANY	1032
GREECE	94
HUNGARY	57
ITALY	1483
NETHERLANDS	175
NORWAY	78
POLAND	174
PORTUGAL	111
SLOVAKIA	49
SPAIN	286
SWEDEN	73
SWITZERLAND	330
UNITED KINGDOM	715

6071

OBSERVER STATES

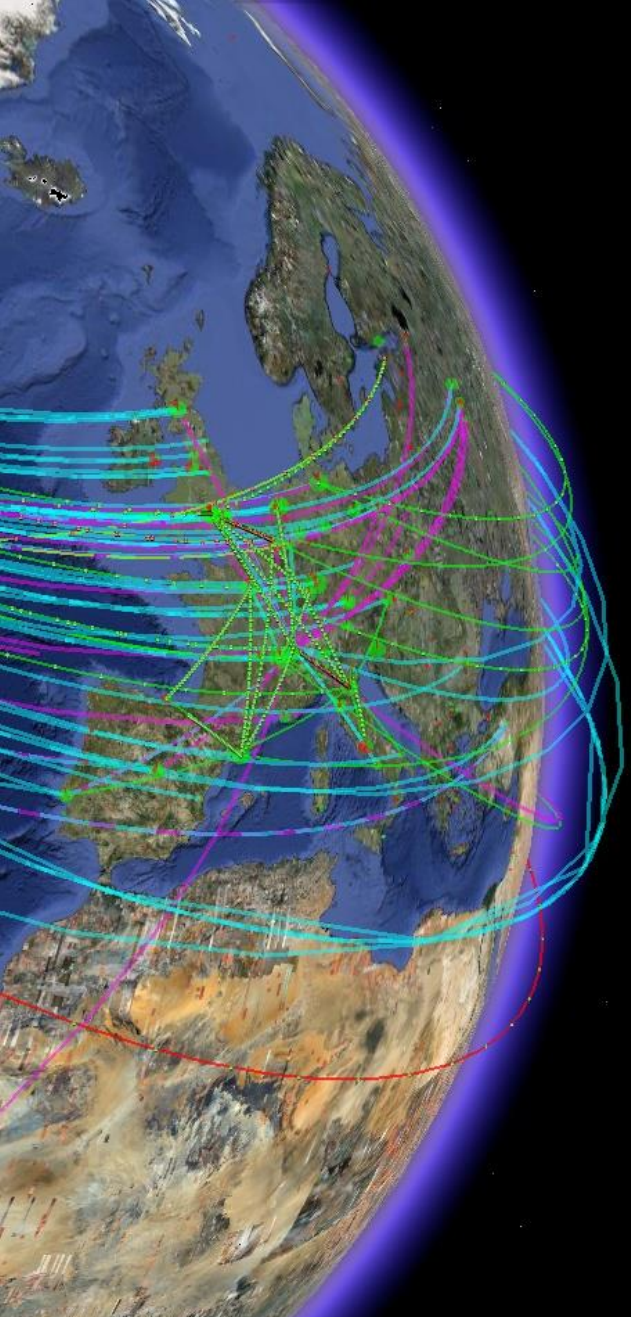
INDIA	89
ISRAEL	59
JAPAN	200
RUSSIA	883
TURKEY	52
USA	1485

2768

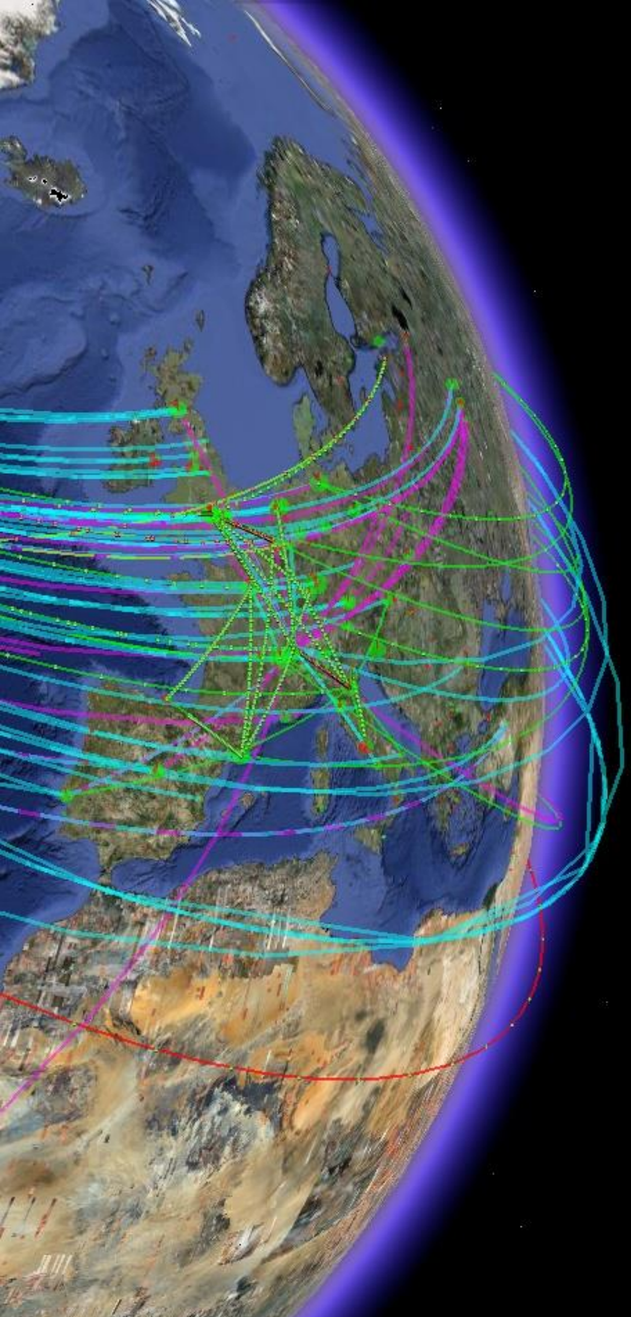
OTHER STATES

ARGENTINA	10
ARMENIA	15
AUSTRALIA	14
AZERBAIJAN	1
BELARUS	19
BRAZIL	73
CANADA	136
CHILE	4
CHINA	64
COLOMBIA	11
CROATIA	20
CUBA	3
CYPRUS	6
ESTONIA	11
GEORGIA	11
ICELAND	1
IRAN	12
IRELAND	12
KOREA	51
LITHUANIA	5
MEXICO	28
MONTENEGRO	1
MOROCCO	5
NEW ZEALAND	6
PAKISTAN	24
PERU	1
ROMANIA	49
SERBIA	17
SLOVENIA	16
SOUTH AFRICA	8
SRI LANKA	1
TAIWAN	42
THAILAND	1
UKRAINE	18

696



WLCG Challenge



WLCG

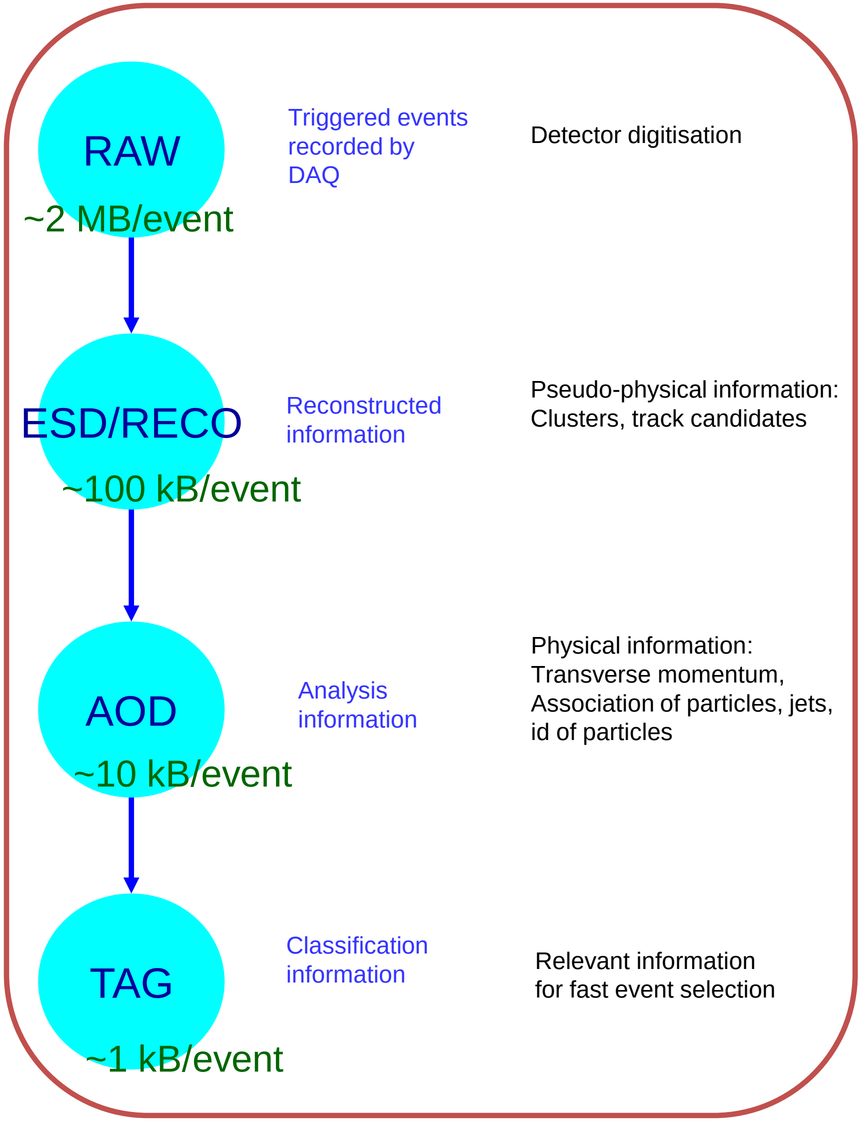
- Why is this a challenge?
- Why a distributed system?
  - and why a grid
- History
- Architecture
- Monitoring and Operation
- Usage

Data and Algorithms

- HEP data are organized as *Events* (particle collisions)
- Simulation, Reconstruction and Analysis programs process

  "one Event at a time"

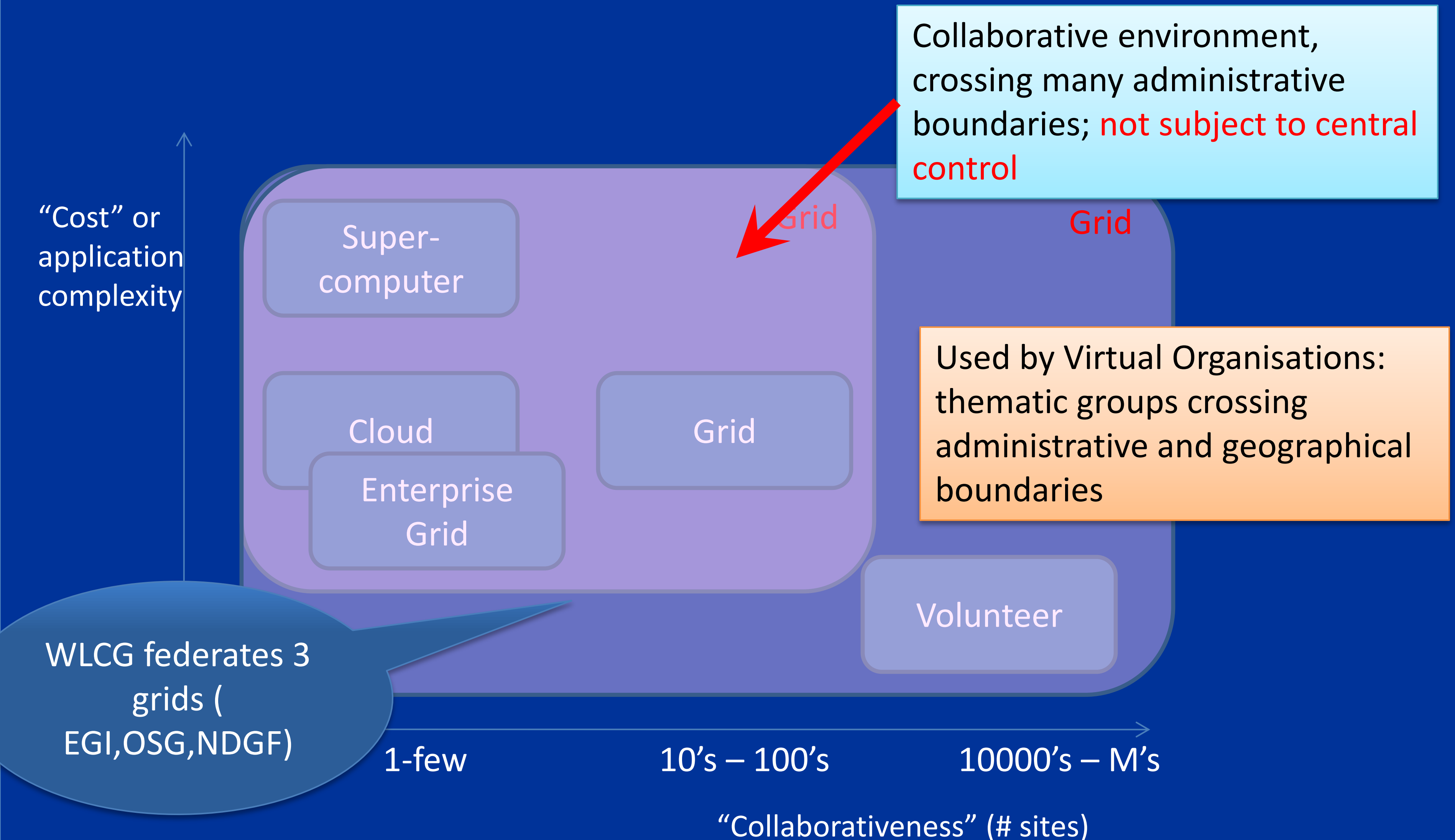
  – Events are fairly independent → Trivial parallel processing
- Event processing programs are composed of a number of Algorithms selecting and transforming "raw" Event data into "processed" (reconstructed) Event data and statistics
- **High Throughput Computing**



Why distributed, why a Grid?

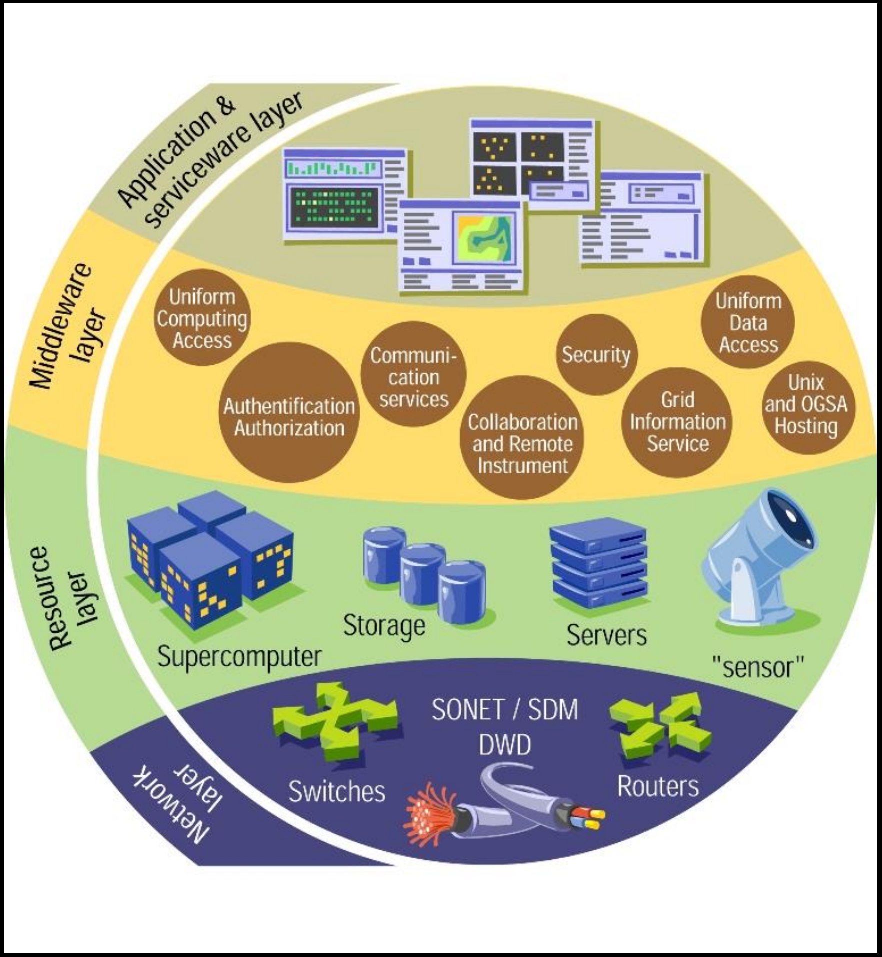
- From the start on it was clear that no center could provide ALL computing
  - Buildings, Power, Cooling, Money .....
- The HEP community is distributed and a most funding for computing is local
  - loosely coupled community
- Significant computing was available in many institutes
  - often shared with other research communities
- Several concepts of Grid computing are a good fit for our community with collaborations across many institutes
  - security/trust model (authentication, Virtual Organizations)
  - approach to heterogeneity
  - no central control
  - granularity (several hundred centers with locally managed resources)
- Both technical and political/financial reasons lead to the decision to build a grid infrastructure for LHC computing

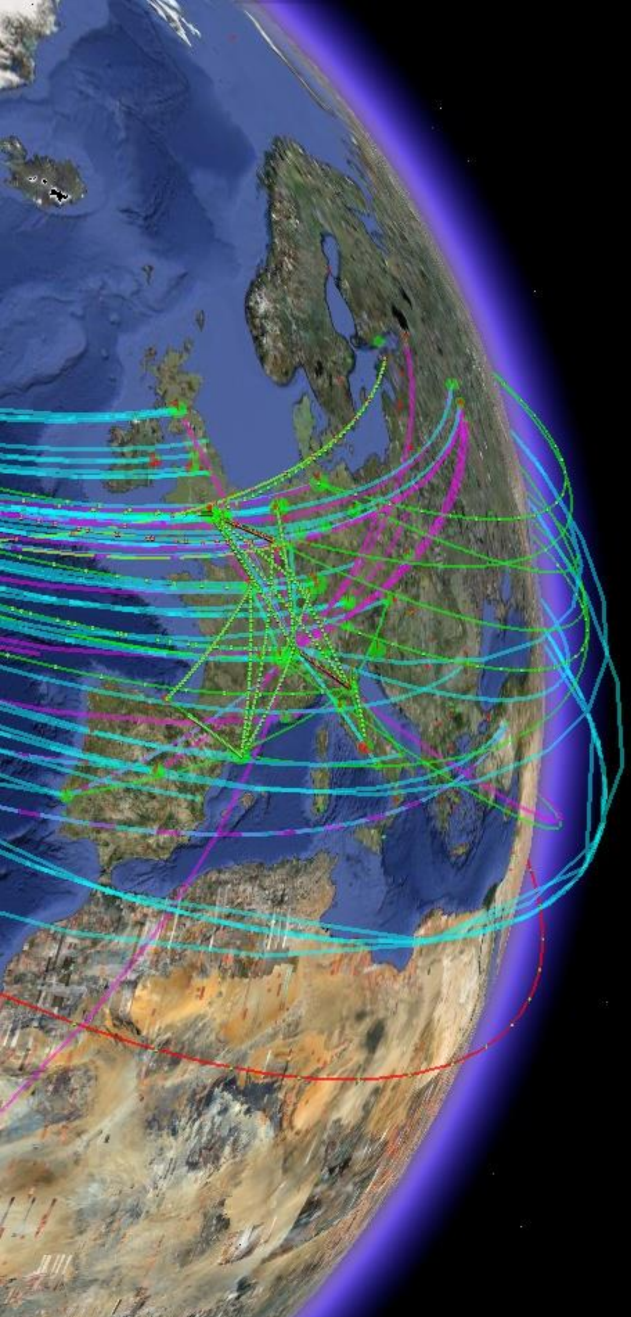
Grids and other Creatures



What is Grid Middleware?

- For today:
- The glue that creates the illusion that a distributed infrastructure is a single resource
  - without enforcing uniformity
  - without central control
- Good Grid Middleware is invisible...



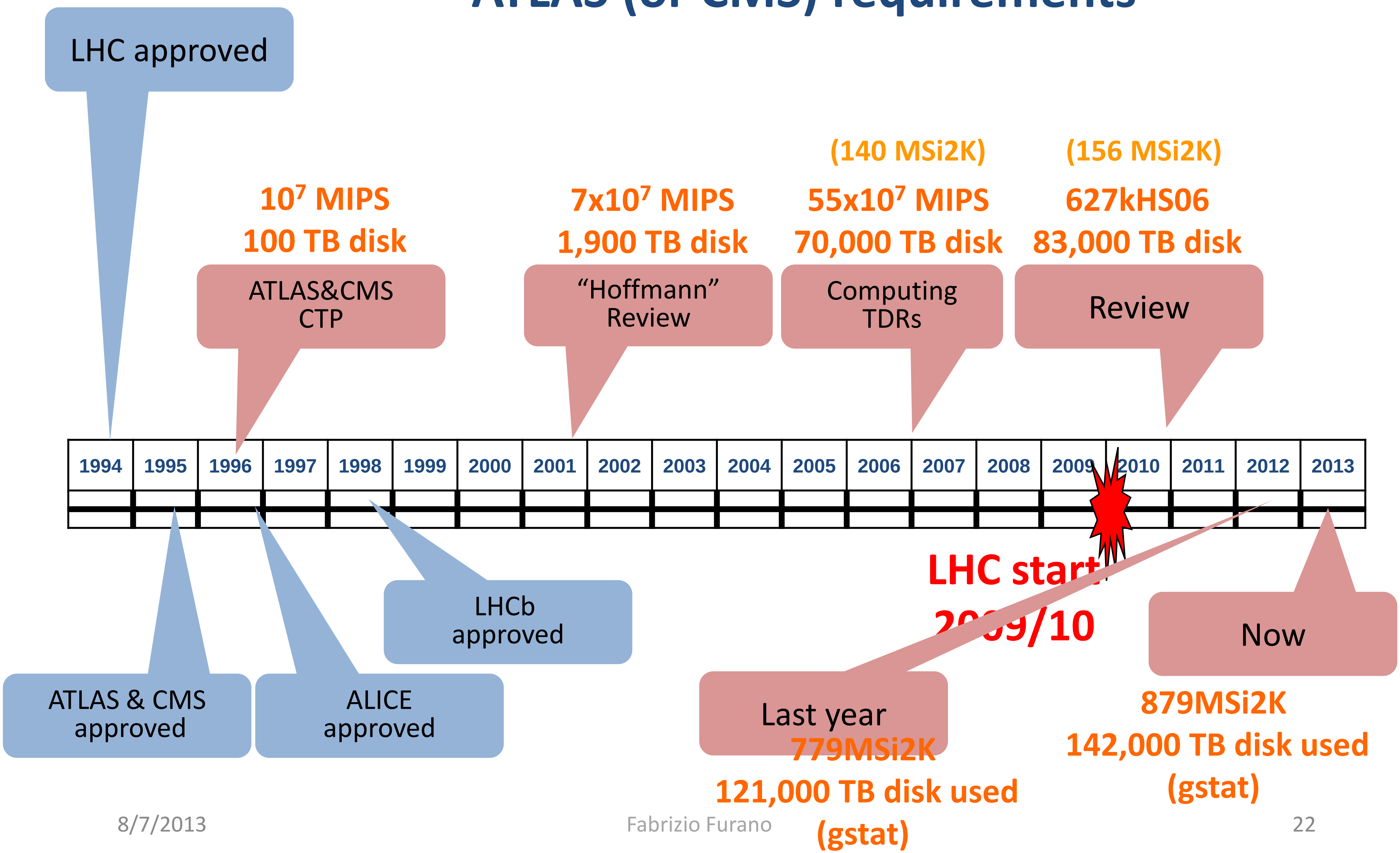


WLCG

- Why is this a challenge?
- Why a distributed system?
  - and why a grid
- **History**
- Architecture
- Monitoring and Operation
- Usage

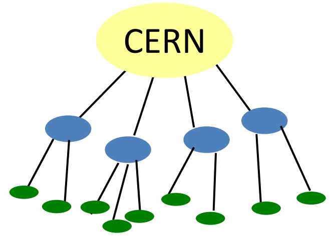
Requirements over Time

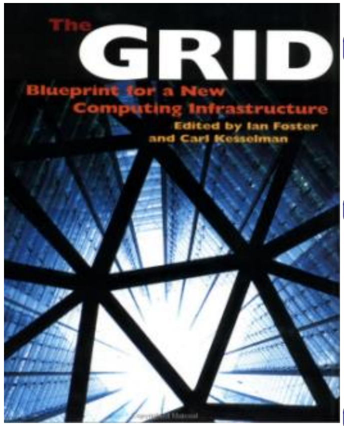
ATLAS (or CMS) requirements

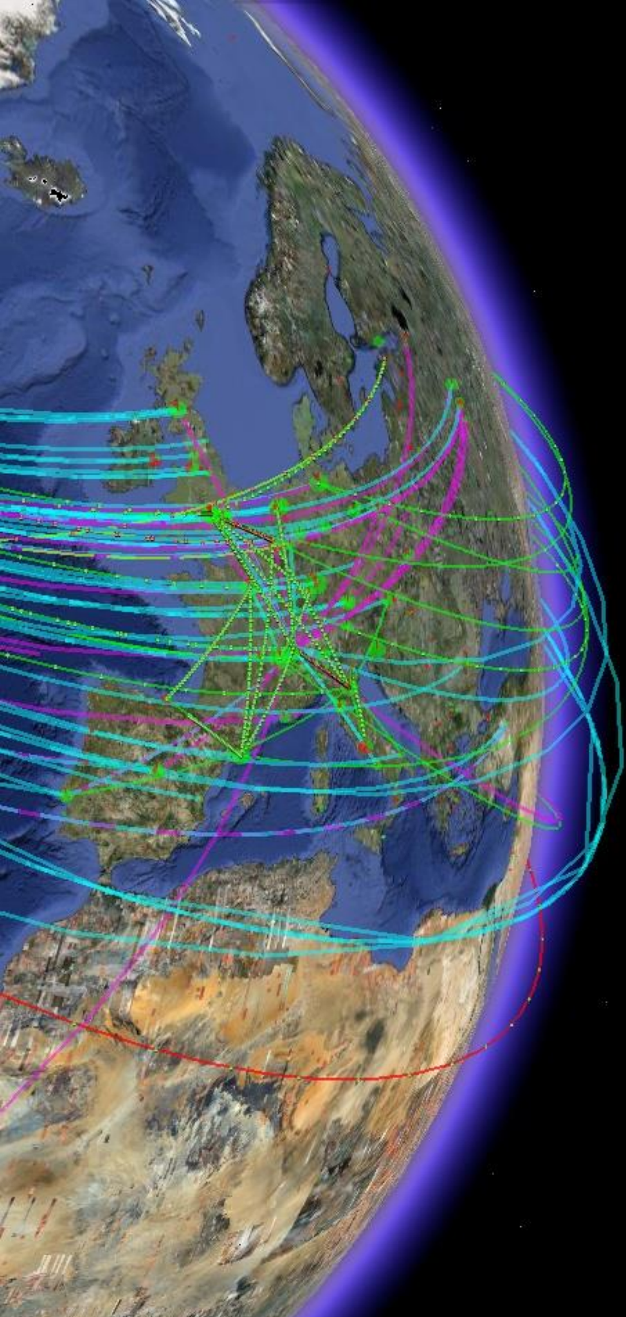


History

- 1999 - MONARC project
  - First LHC computing architecture – hierarchical distributed model, focus on network control
- 2000 – growing interest in grid technology
  - HEP community main driver in launching the DataGrid project
- 2001-2004 - EU DataGrid project
  - middleware & testbed for an operational grid
- 2002-2005 – LHC Computing Grid – LCG
  - deploying the results of DataGrid to provide a production facility for LHC experiments
- 2004-2006 – EU EGEE project phase 1
  - starts from the LCG grid
  - shared production infrastructure
  - expanding to other communities and sciences
- 2006-2008 – EU EGEE project phase 2
  - expanding to other communities and sciences
  - Scale and stability
  - Interoperations/Interoperability
- 2008-2010 – EU EGEE project phase 3
  - More communities
  - Efficient operations
  - Less central coordination
- 2010 – 201x EGI and EMI
  - Sustainable infrastructures based on National Grid Infrastructures
  - Decoupling of middleware development and infrastructure
  - Merging middleware stacks in Europe
- 2013 Post-EMI
  - Middleware maintained through collaboration
  - Infrastructure management continues



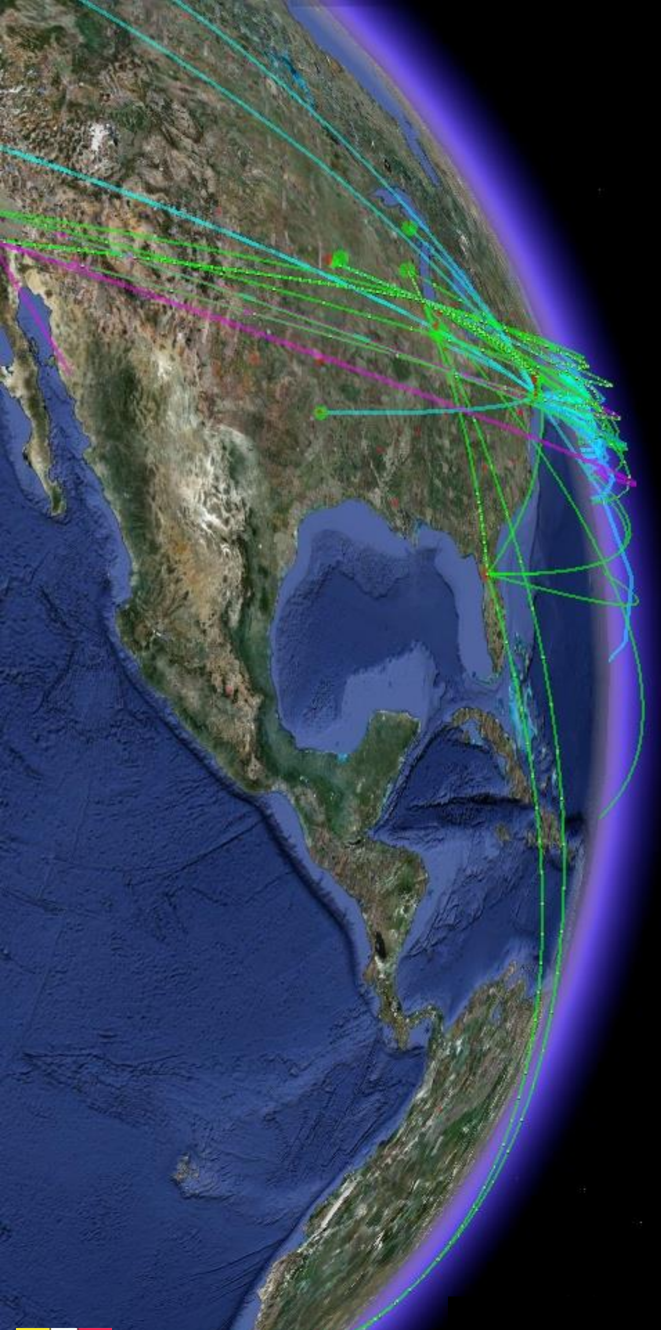




WLCG

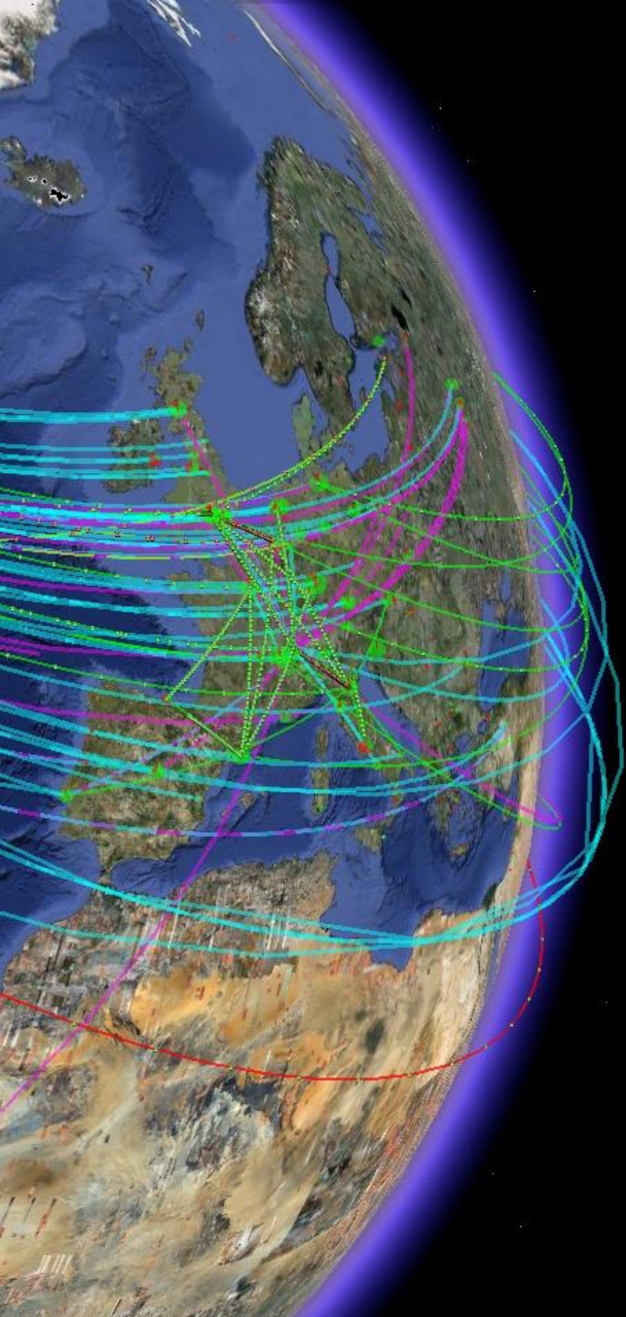
- Why is this a challenge?
- Why a distributed system?
  - and why a grid
- History
- **Architecture**
- Monitoring and Operation
- Usage

What is WLCG???



WLCG

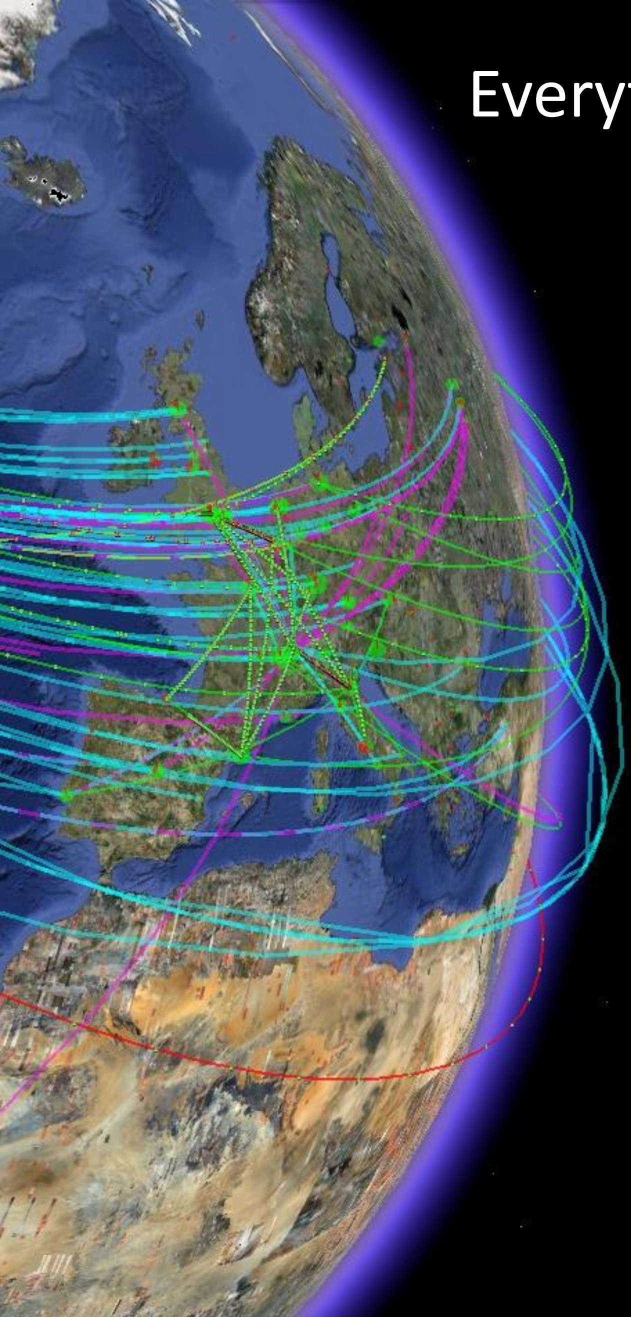
- Worldwide LHC Computing Grid
  - Distributed Computing Infrastructure for LHC experiments
    - Linking 3 distributed infrastructures
      - OSG Open Science Grid in the US
      - EGI European Grid Infrastructure
      - NDGF Nordic Data Grid Facility
    - Linking more than 300 computer centers
    - Providing > 340,000 cores
    - To more than 2000 (active) users
    - Moving ~10GB/s for each experiment
    - Archiving 15PB per year



What is needed to make it work?

- Apart from Middleware
- Apart from Computer Centers

Everything you need in a Computer Center!



- Management
- Fabric
- Networking
- Security
- Monitoring
- User Support
- Problem Tracking
- Accounting
- Service support
- SLAs…..

- But now on a global scale
  - Respecting the sites' independence
  - Linking the different infrastructures
    - NDGF, EGI, OSG

What does WLCG cover?

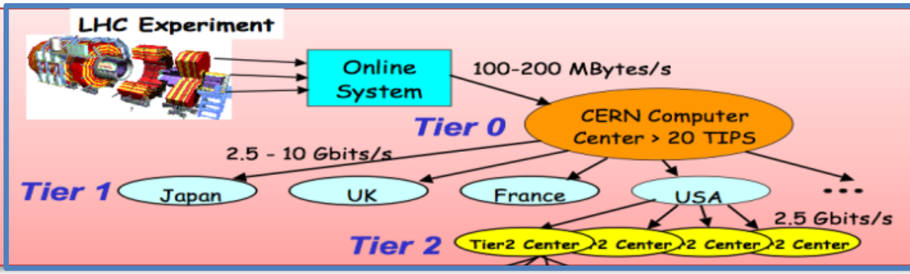
Collaboration

- Coordination & management & reporting
- Coordinate resources & funding
- Coordination with service & technology providers
- Common requirements
- Memorandum of Understanding

Framework

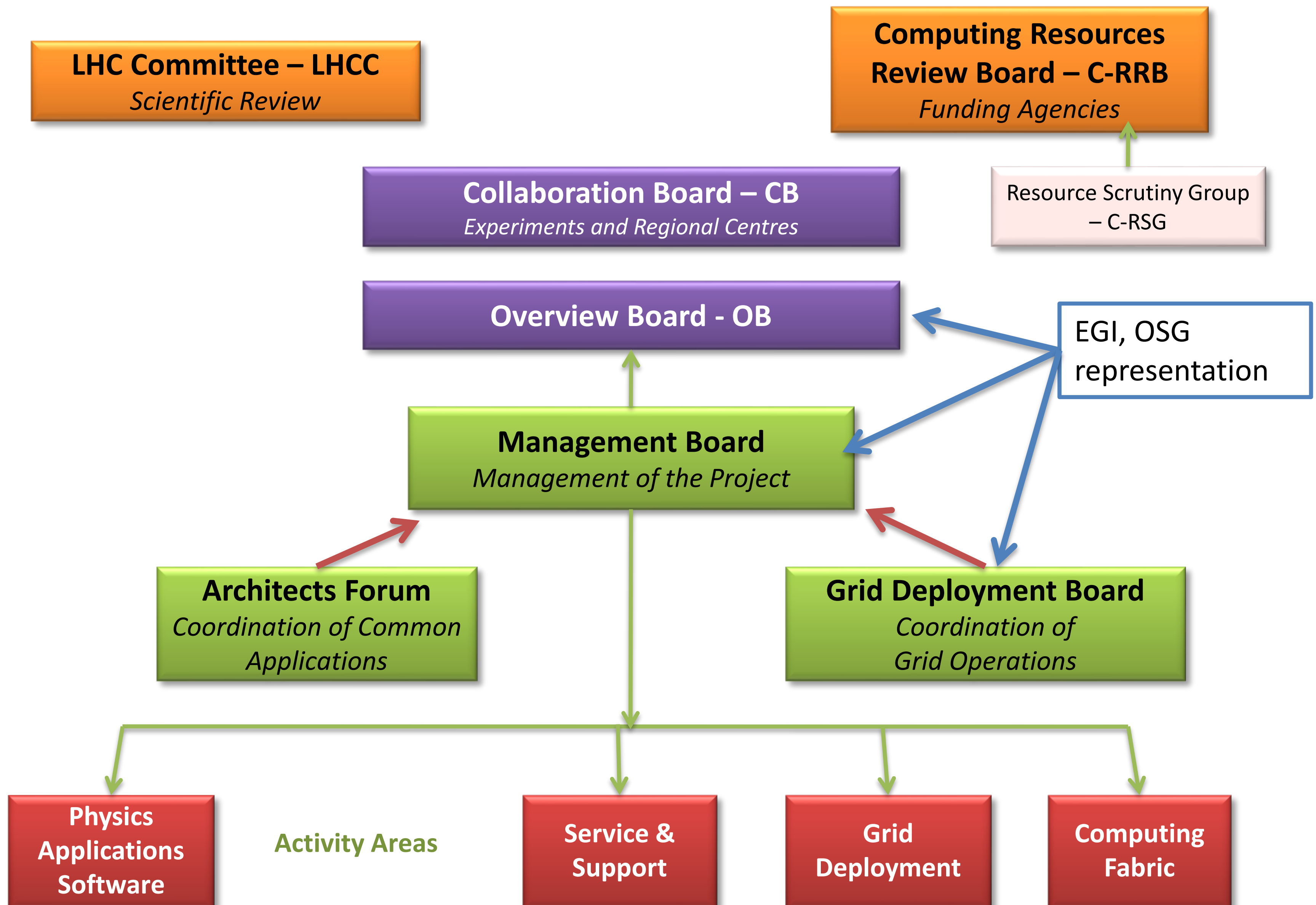
- Service management
- Service coordination
- Operational security
- Support processes & tools
- Common tools
- Monitoring & Accounting
- World-wide trust federation for CA's and VO's
- Complete Policy framework

Distributed Computing services



Physical resources: CPU, Disk, Tape, Networks

Organisation Structure



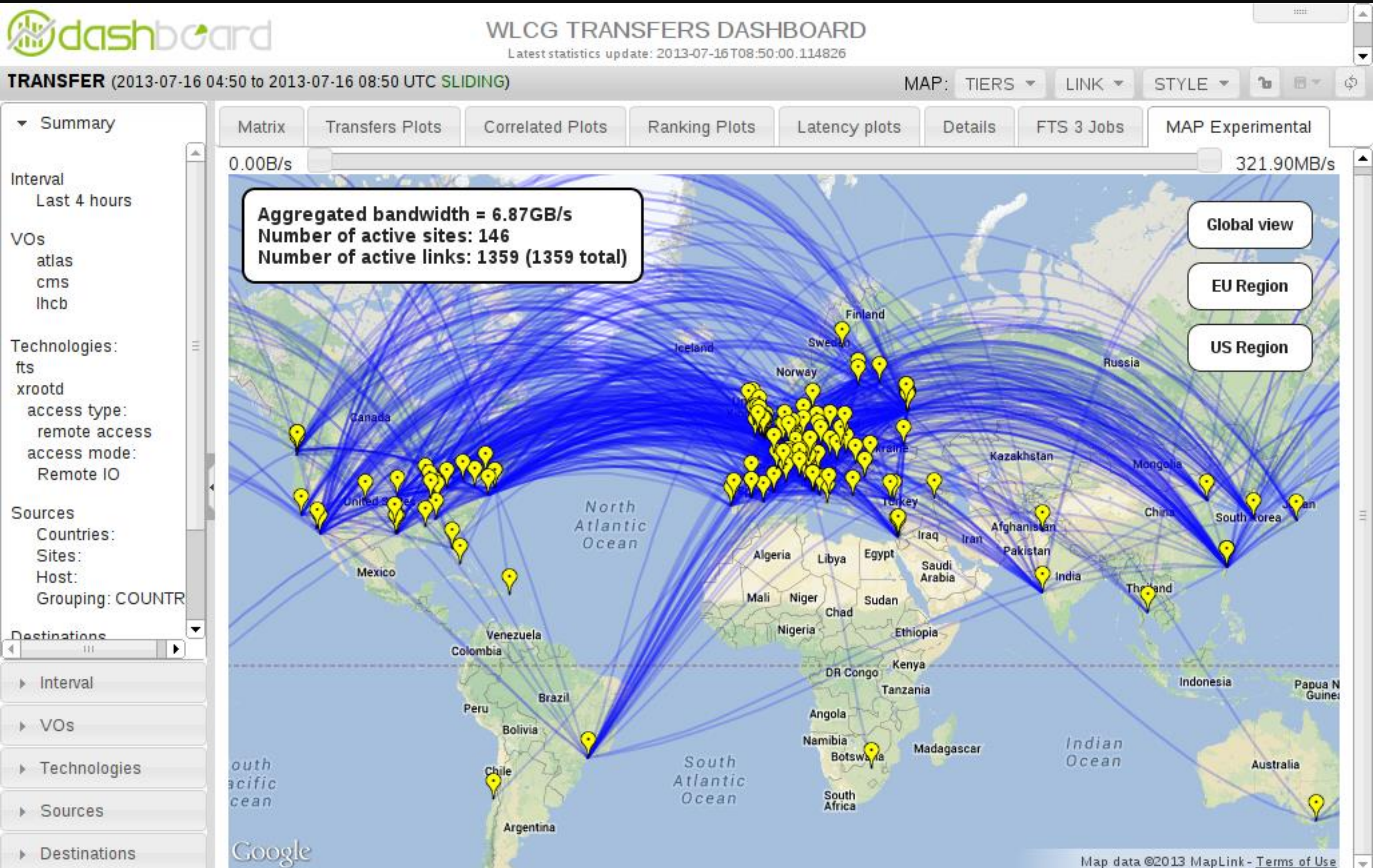
Operations

- Not all is provided by WLCG directly
- WLCG links the services provided by the underlying infrastructures
  - and ensures that they are compatible
- EGI relies on National Grid Infrastructures (NGIs)
  - +some central services
    - user support (GGUS)
    - accounting (APEL & portal)....
- Monitors the system

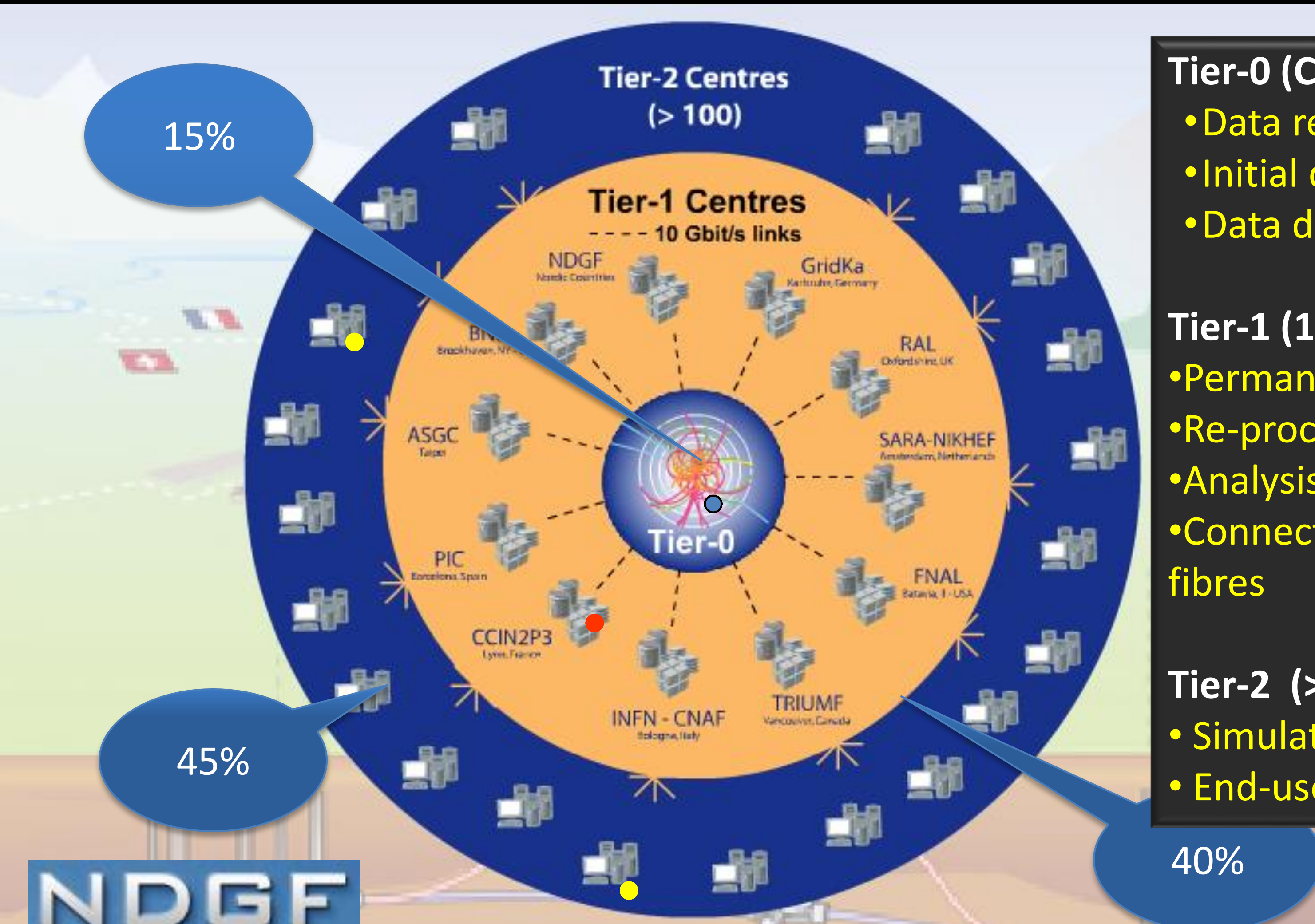
NGIs in Europe

www.eu-egi.eu

WLCG today


WLCG TRANSFERS DASHBOARD
Latest statistics update: 2013-07-16T08:50:00.114826
TRANSFER (2013-07-16 04:50 to 2013-07-16 08:50 UTC SLIDING)
MAP: TIERS
LINK
STYLE
Summary
Interval
Last 4 hours
VOs
atlas
cms
lhcb
Technologies:
fts
xrootd
access type:
remote access
access mode:
Remote IO
Sources
Countries:
Sites:
Host:
Grouping: COUNTR
Interval
VOs
Technologies
Sources
Destinations
Matrix
Transfers Plots
Correlated Plots
Ranking Plots
Latency plots
Details
FTS 3 Jobs
MAP Experimental
0.00B/s
321.90MB/s
Aggregated bandwidth = 6.87GB/s
Number of active sites: 146
Number of active links: 1359 (1359 total)
Global view
EU Region
US Region
North Atlantic Ocean
South Atlantic Ocean
Indian Ocean
Map data ©2013 MapLink - Terms of Use

Architecture



Tier-0 (CERN): (15%)
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres): (40%)
- Permanent storage
- Re-processing
- Analysis
- Connected by direct 10 Gb fibres

Tier-2 (>200 centres): (45%)
- Simulation
- End-user analysis

WLCG Collaboration Status

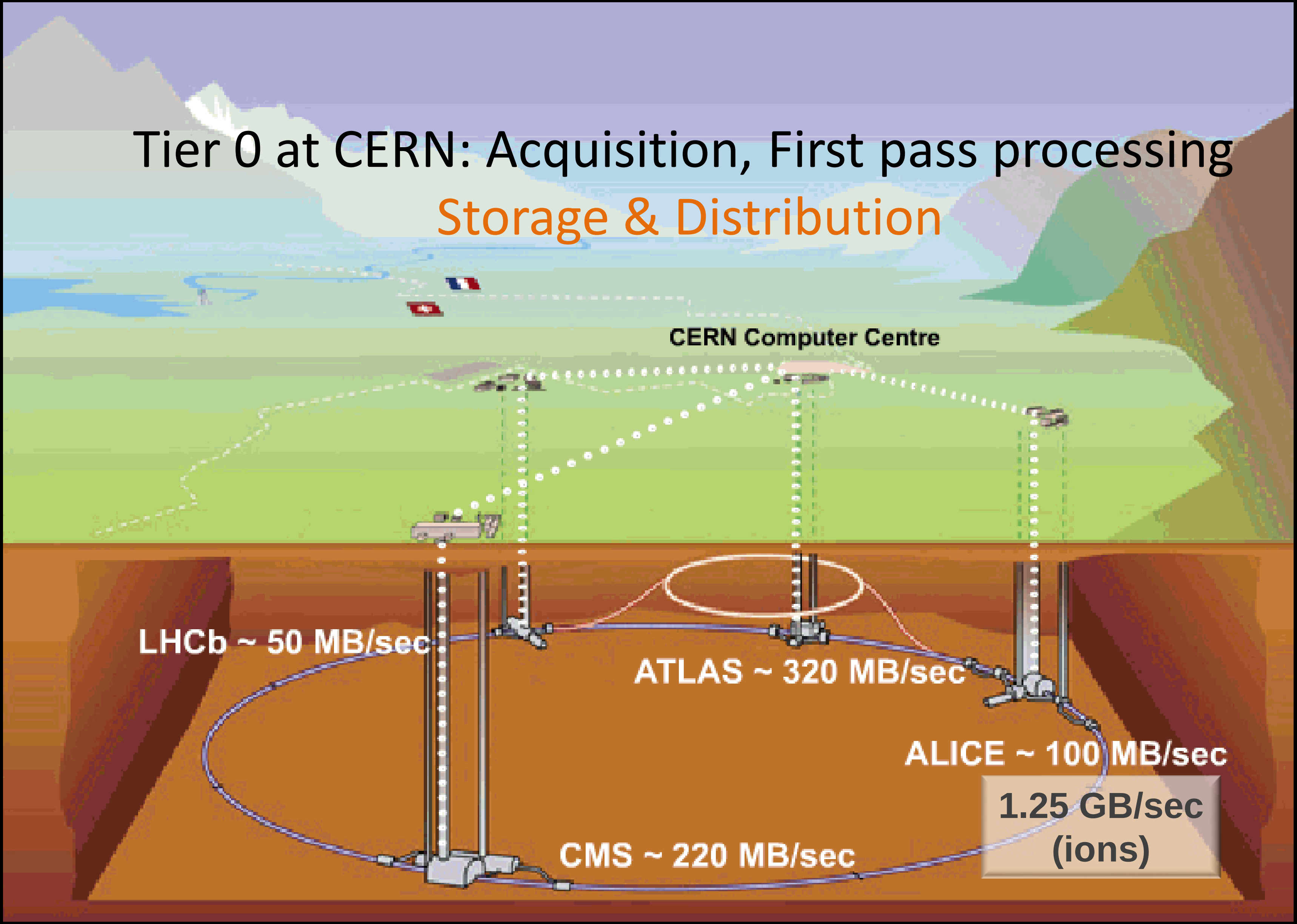
Tier 0; 11 Tier 1s; 68 Tier 2 federations

(140 Tier 2 sites) + many T3 sites

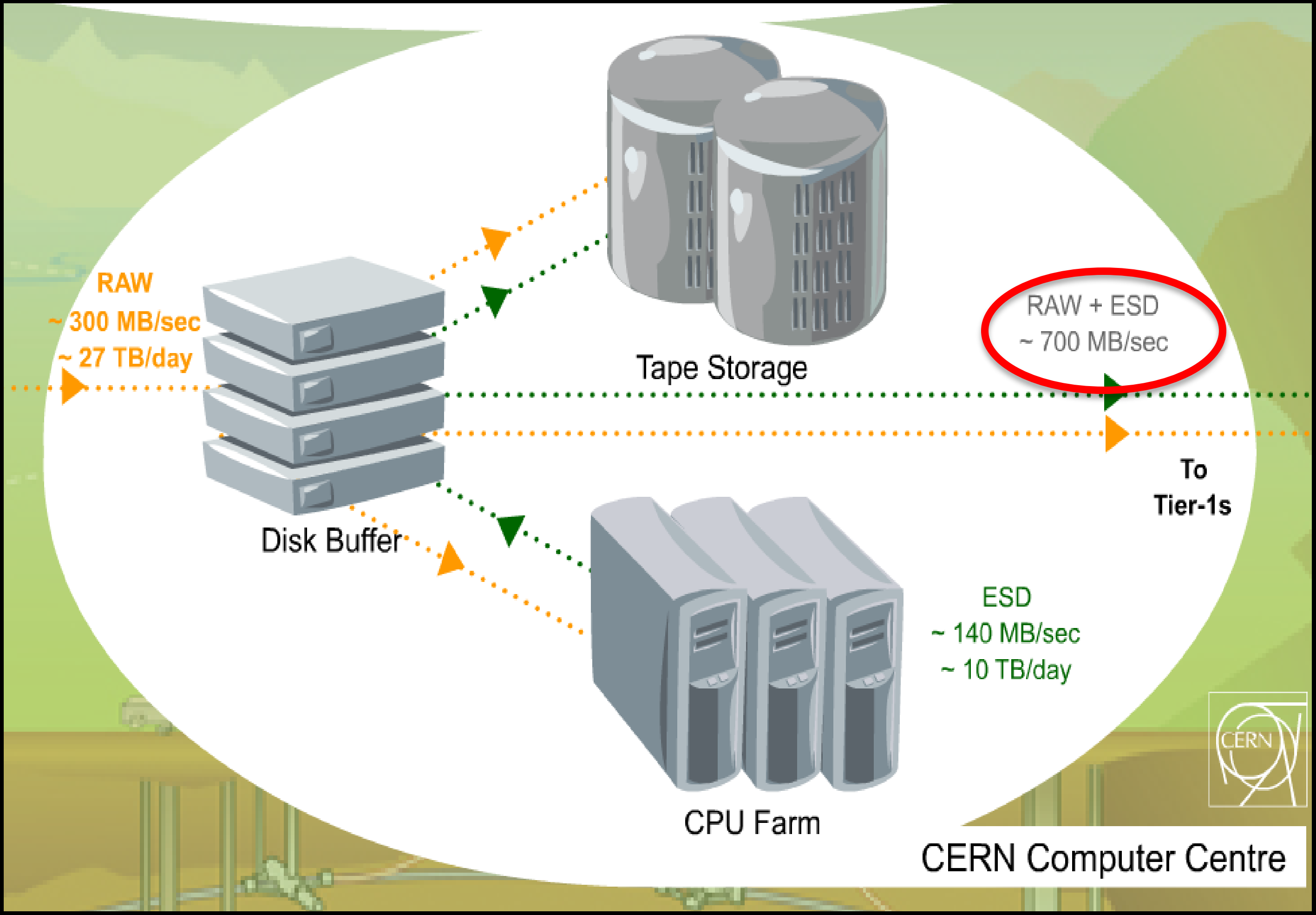
Today we have 49 MoU signatories, representing 34 countries:

Australia, Austria, Belgium, Brazil, Canada, China, Czech Rep, Denmark, Estonia, Finland, France, Germany, Hungary, Italy, India, Israel, Japan, Rep. Korea, Netherlands, Norway, Pakistan, Poland, Portugal, Romania, Russia, Slovenia, Spain, Sweden, Switzerland, Taipei, Turkey, UK, Ukraine, USA.

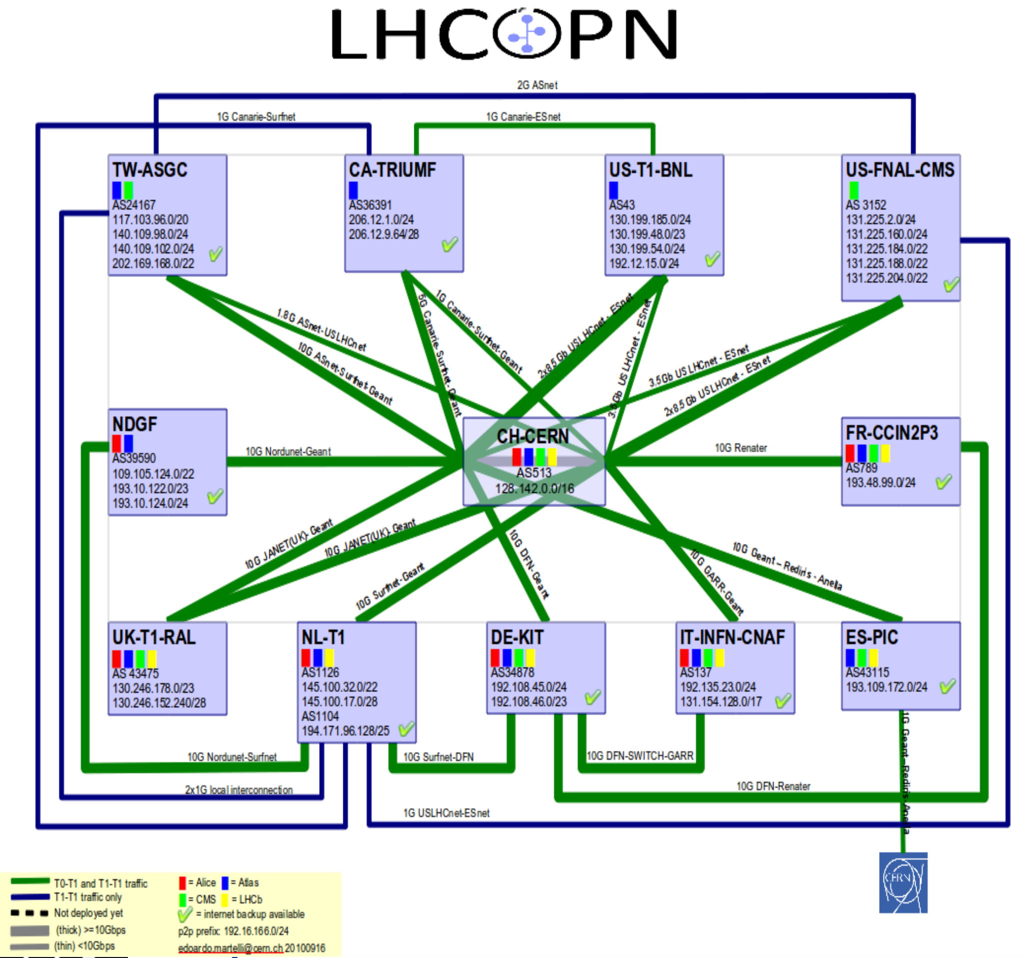
Tier 0 at CERN: Acquisition, First pass processing
Storage & Distribution

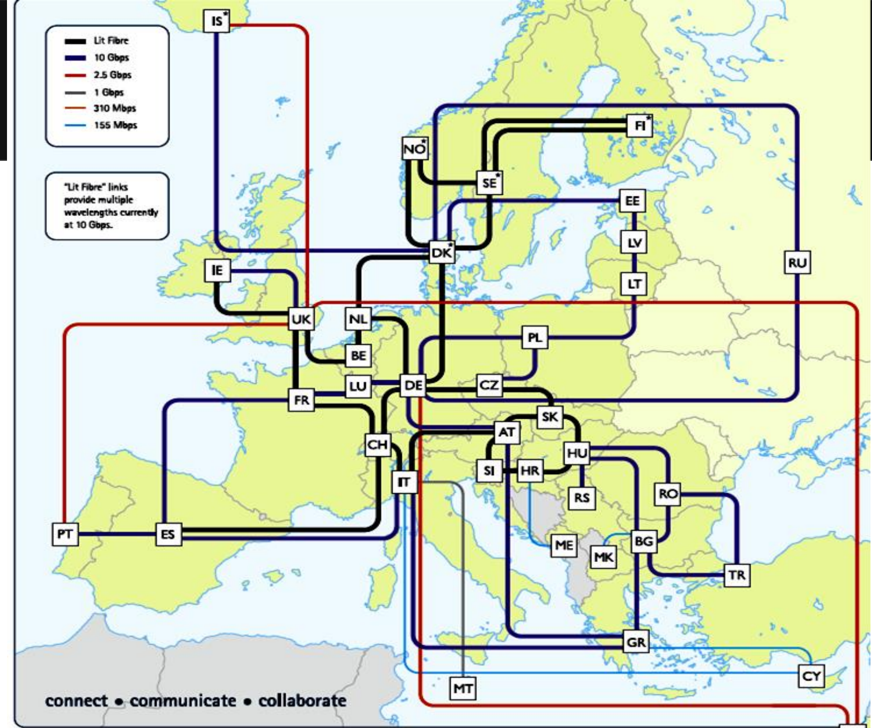


Flow in and out of the center

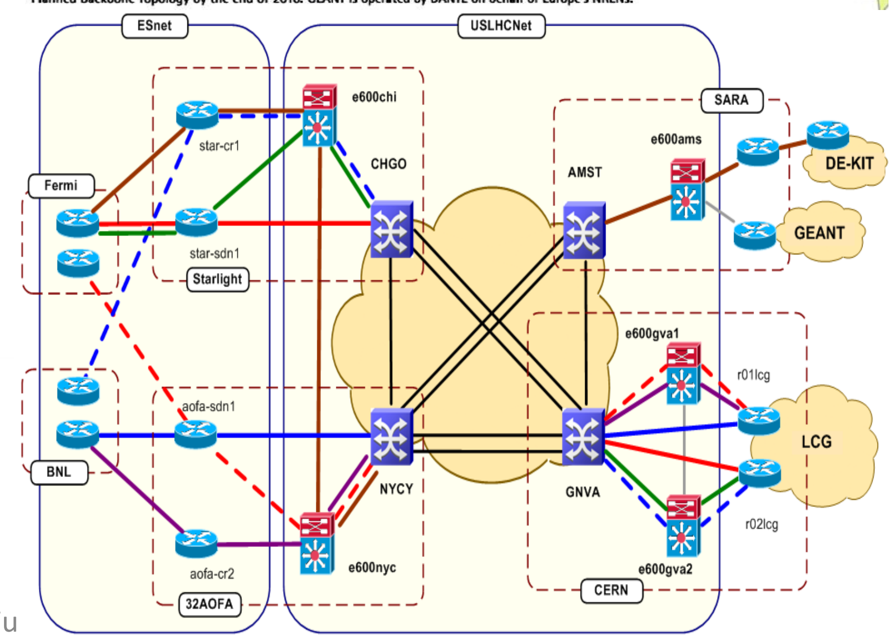


LHC Networking





Planned Backbone Topology by the end of 2010. GÉANT is operated by DANTE on behalf of Europe's NRENs.

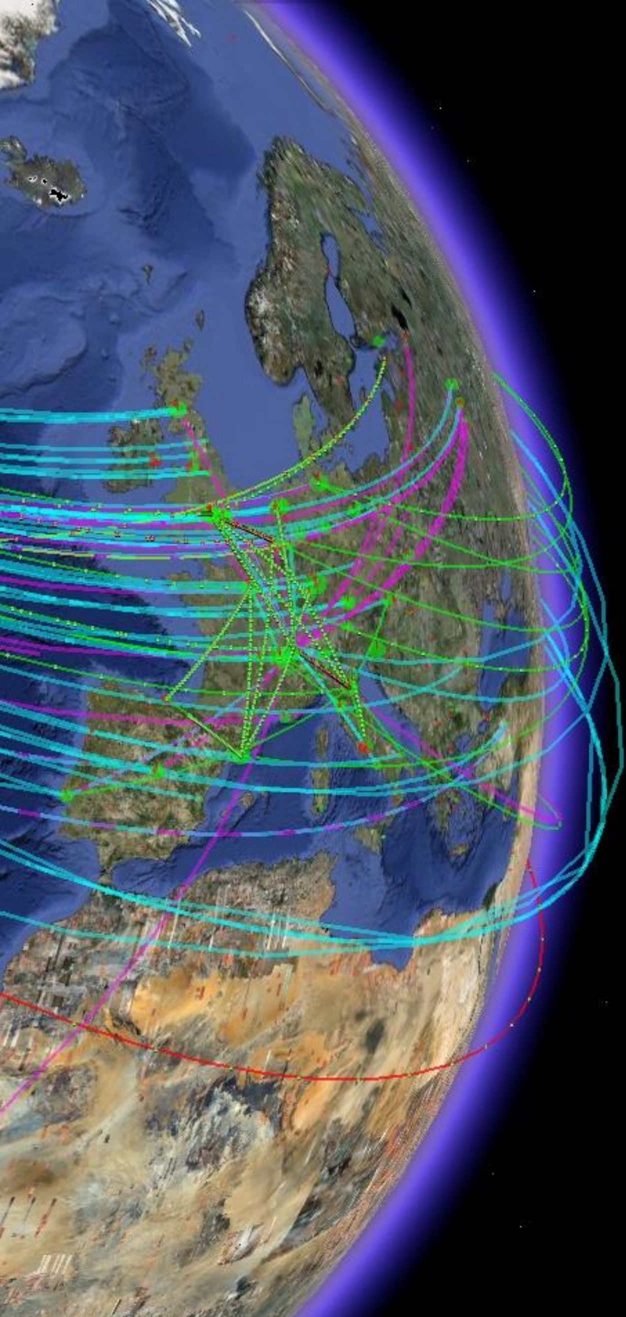


Relies on

- OPN, GEANT, US-LHCNet
- NRENs & other national & international providers

Implementation Challenges

- Managing Heterogeneity
  - Site security and usage policies
    - can't be solved by more software
  - Site management technologies
    - monitoring, accounting, installation ....
  - Local operating systems
    - all Linux based
    - but RedHat X ≠ SL X ≠ Ubuntu ≠ Debian ....
  - Local batch systems
    - SunGrid Engine, LSF, PBS, Condor, Torque&Maui ....
  - Experience and knowledge
    - SysAdmin Team: 1 part time student to 40 professionals
  - Scale:
    - >10 nodes to 20.000 nodes on a site
  - Experiments needs and visions differ



WLCG

- Why is this a challenge?
- Why a distributed system?
  - and why a grid
- History
- Architecture
- Monitoring and Operation
- **Usage**

Production Grids

- WLCG relies on a *production quality* infrastructure
  - Requires standards of:
    - Availability/reliability
    - Performance
    - Manageability
  - Used 365 days a year ... (has been for several years!)
  - Tier 1s must store the data for at least the lifetime of the LHC - ~20 years
    - Not passive – requires active migration to newer media
- Vital that we build a fault-tolerant and reliable system
  - That can deal with individual sites being down and recover
- Monitoring and operational tools and procedures are as important as the middleware

WLCG Worldwide LHC Computing Grid

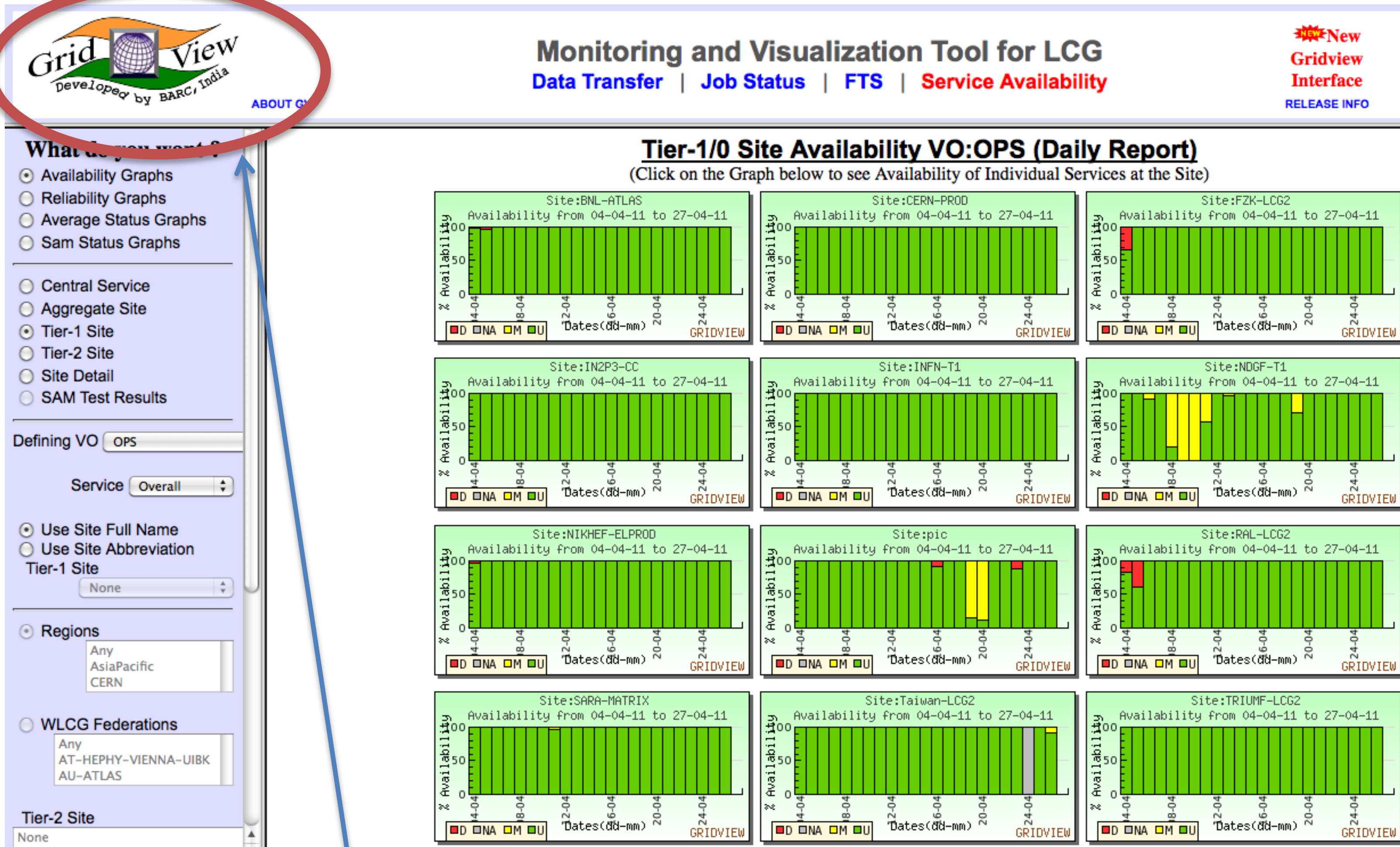
LCG

In addition to EGEE/EGI Operations

- Daily WLCG Operations Meetings
  - 30 minutes
  - Follow up on current problems
- Every two weeks WLCG T1 Service Coordination meeting
  - Operational Planning
  - Incidents followup
- Detailed monitoring of the SLAs.

SAM and availability

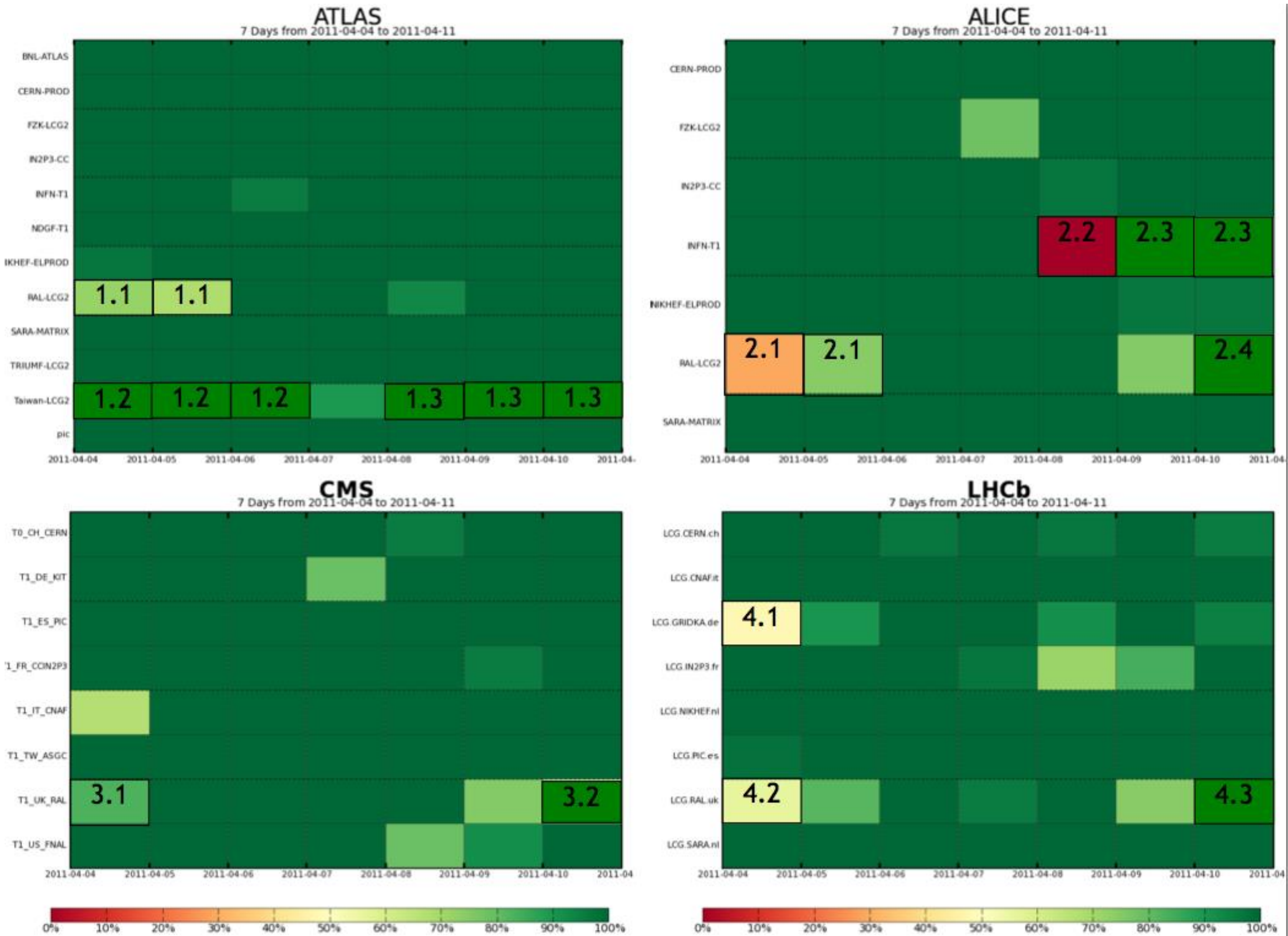
- Grid community puts a great effort into operations
- Infrastructure is continually monitored with active follow-up of issues



Huge contribution from BARC in India

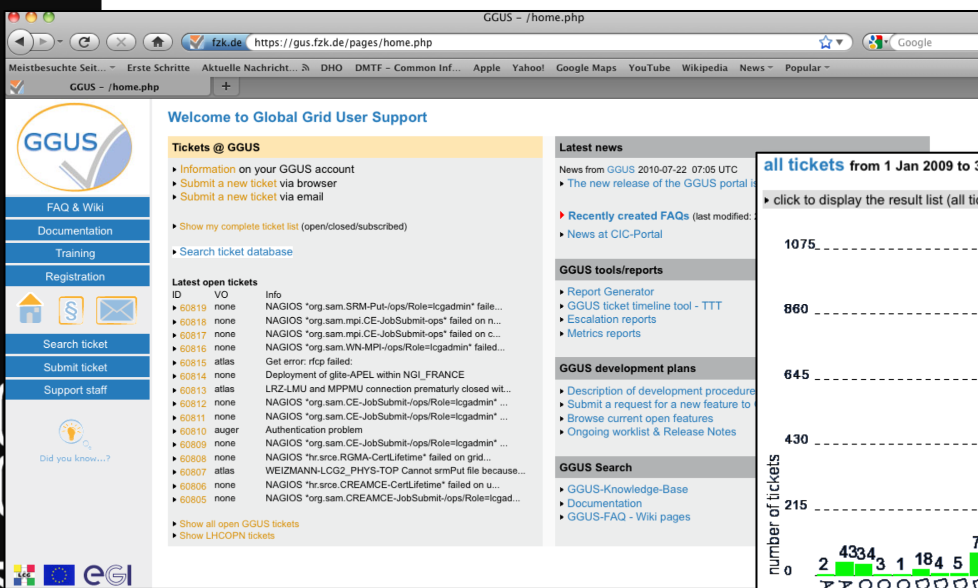
WLCG Worldwide LHC Computing Grid

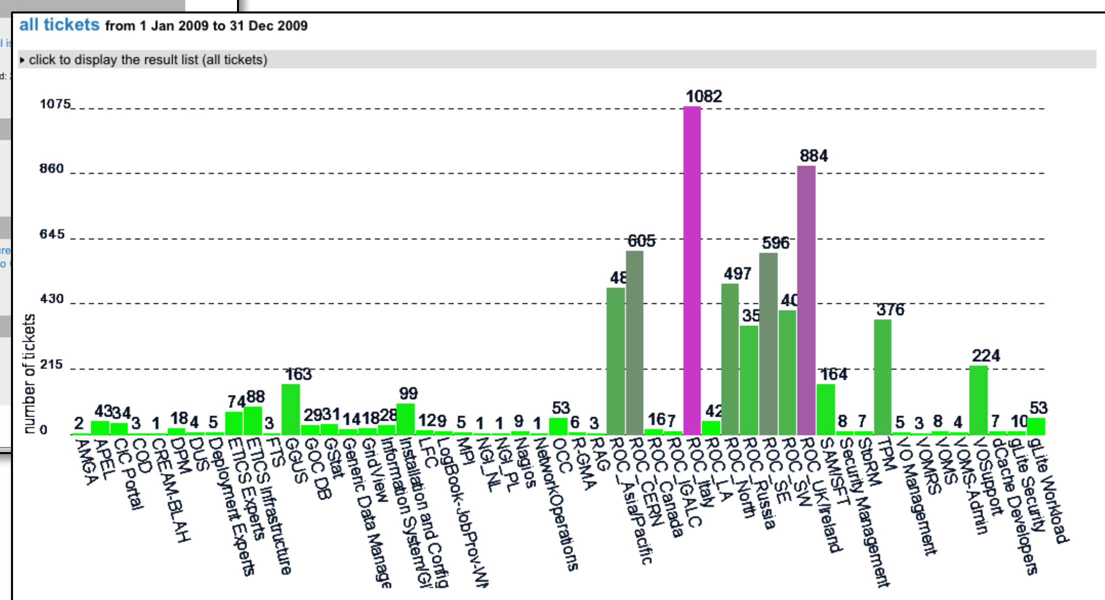
At the WLCG Management Board



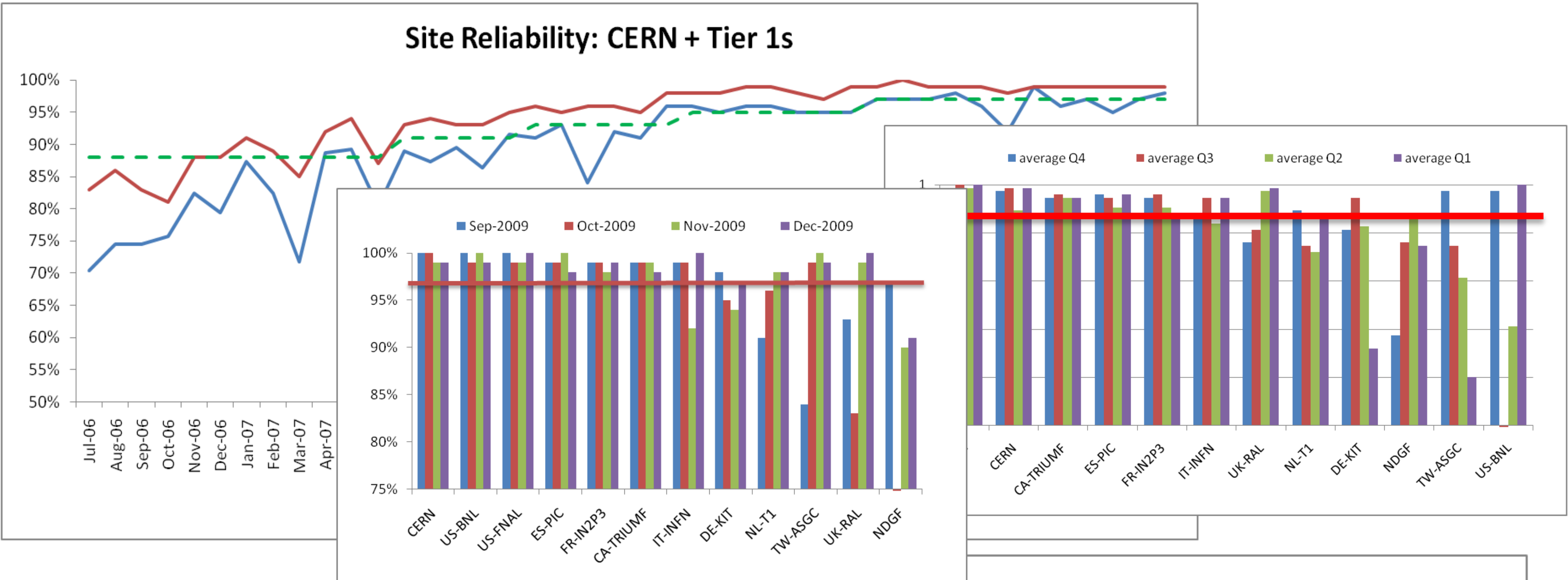
Global Grid User Support

- GGUS: Web based portal
  - about 1000 tickets per months
  - Grid security aware
  - Interfaces to regional/national support structures

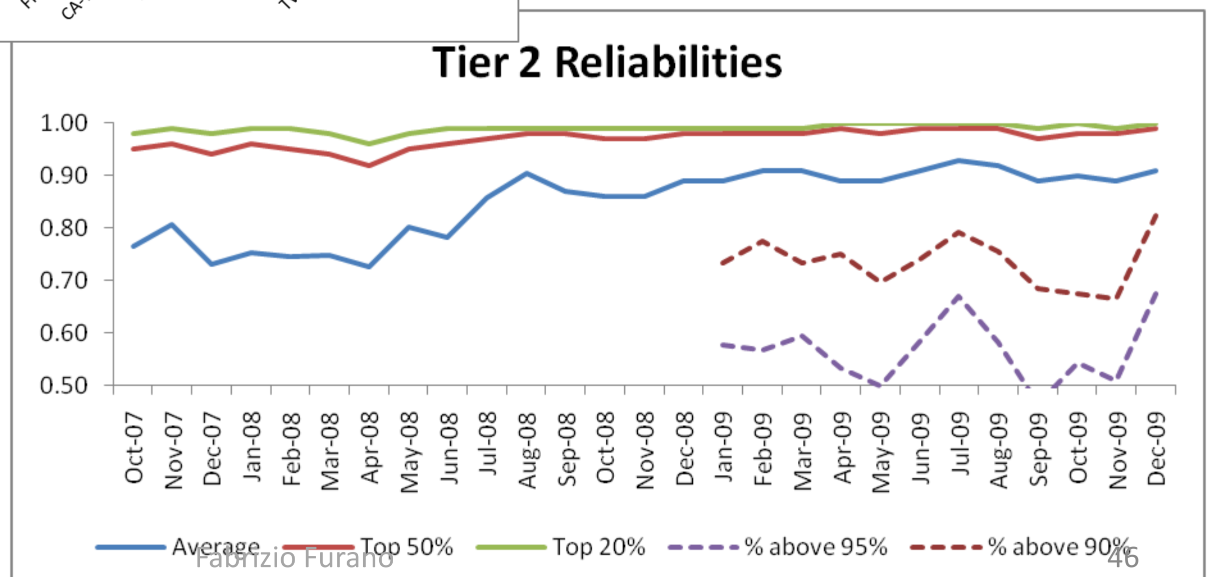




Reliabilities



- This is not the full picture:
- Experiment-specific measures give complementary view
- Need to be used together with some understanding of underlying issues



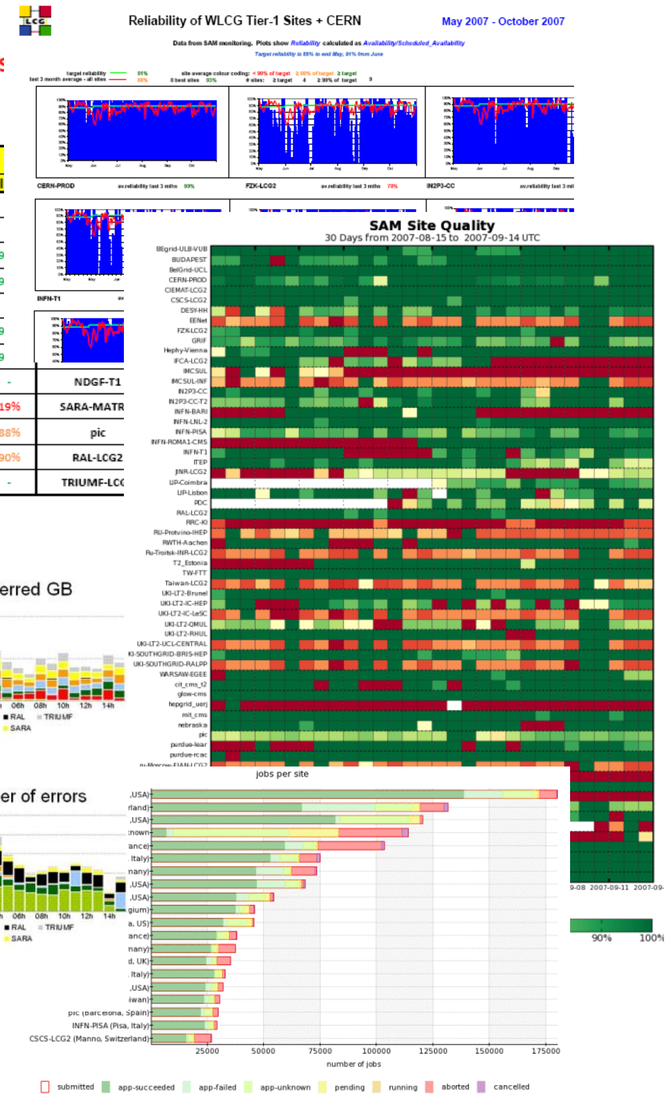
Monitoring to Improve Reliability

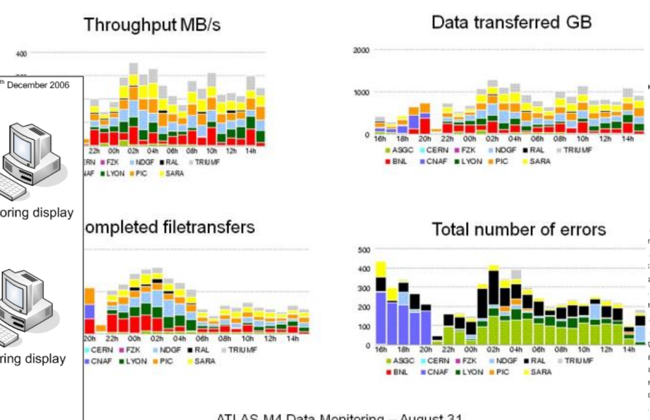
- Monitoring
- Metrics
- Workshops
- Data challenges
- Experience
- Systematic problem analysis
- Priority from software developers

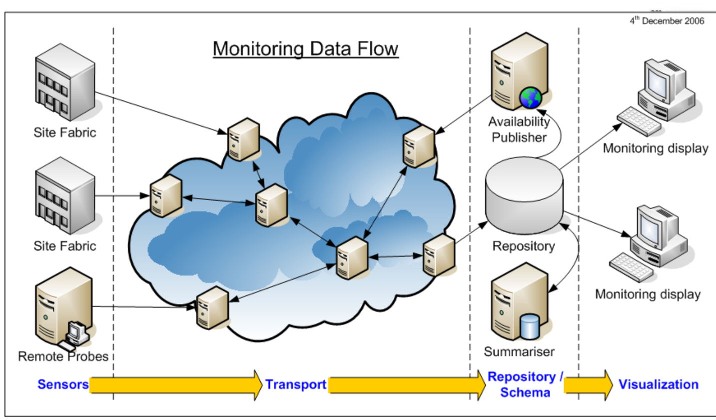
WLCG - Sites Reliability and Job Efficiency

Site	OPS	ALICE		ATLAS			CMS		LHCb	
	SAM	SAM	AGENT	SAM	GANGA	PROD	SAM	CRAB	SAM	PI
ASGC	93%	-	-	98%	22%	82%	95%	90%	-	
BNL	91%	-	-	72%	0%	0%	-	-	-	
CERN	100%	97%	99%	100%	50%	92%	100%	76%	96%	9
CNAF	80%	97%	53%	85%	52%	74%	100%	97%	66%	9
FNAL	89%	-	-	-	-	-	38%	99%	-	
FZK	91%	95%	96%	62%	73%	93%	99%	96%	91%	9
IN2P3	70%	45%	89%	26%	77%	79%	8%	99%	97%	9
NDGF	97%	0%	0%	76%	0%	84%	0%	0%	-	-
NIKHEF	92%	96%	100%	92%	45%	84%	53%	-	90%	19%
PIC	93%	-	-	100%	7%	61%	100%	100%	93%	88%
RAL	90%	96%	99%	100%	15%	93%	100%	90%	97%	90%
TRIUMF	95%	-	-	98%	4%	94%	-	-	-	-

>=91%		>=82%			<82%

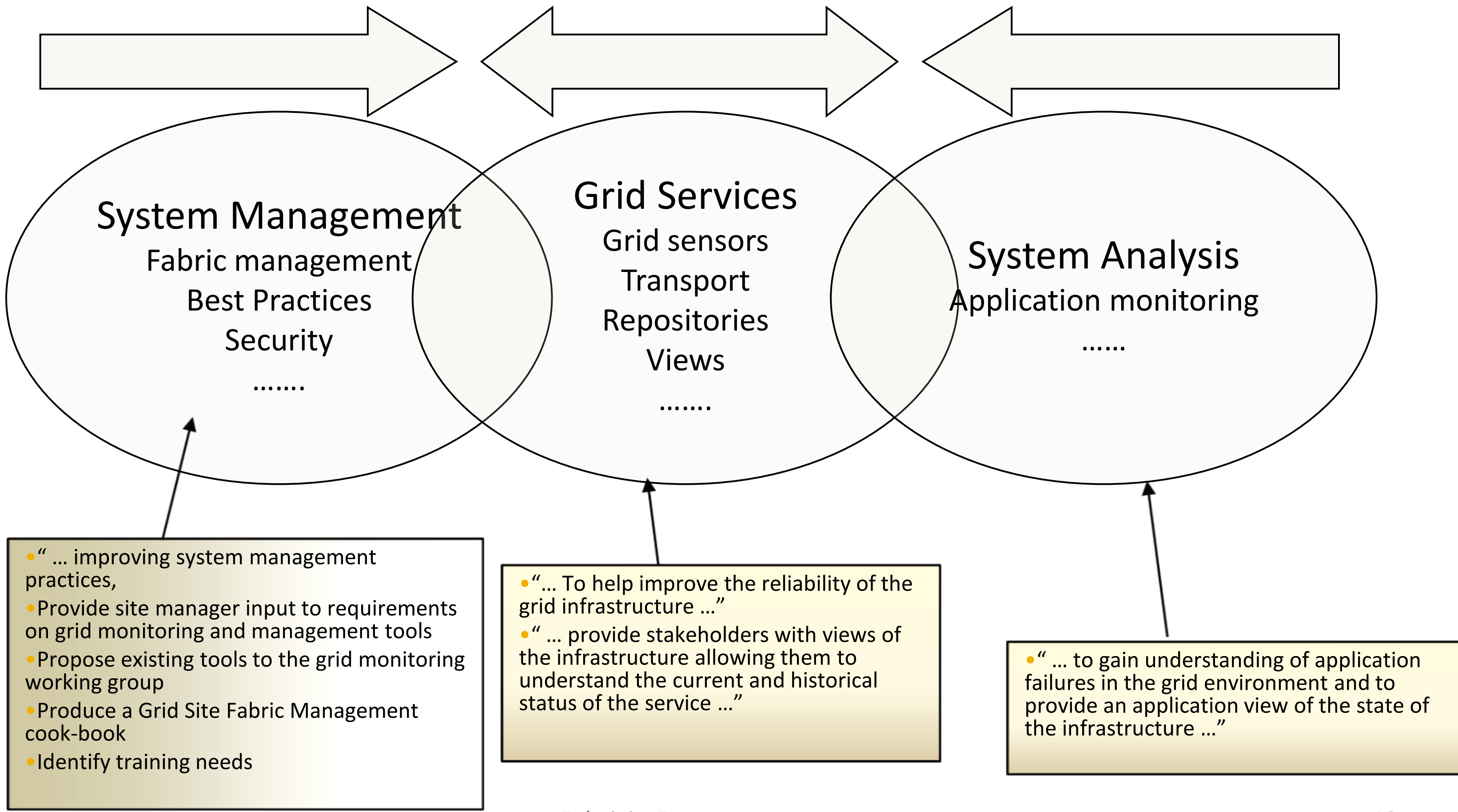






Grid Monitoring

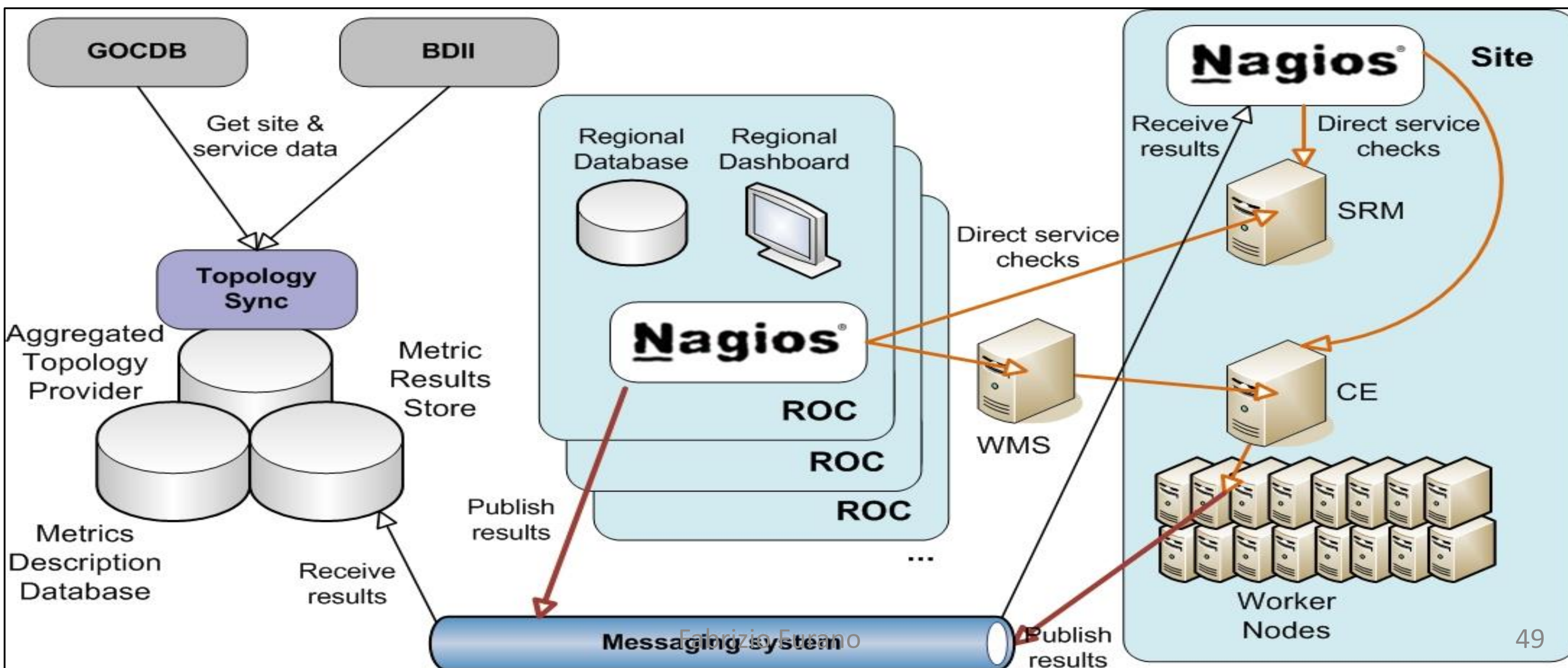
- The critical activity to achieve reliability



Monitoring

ActiveMQ

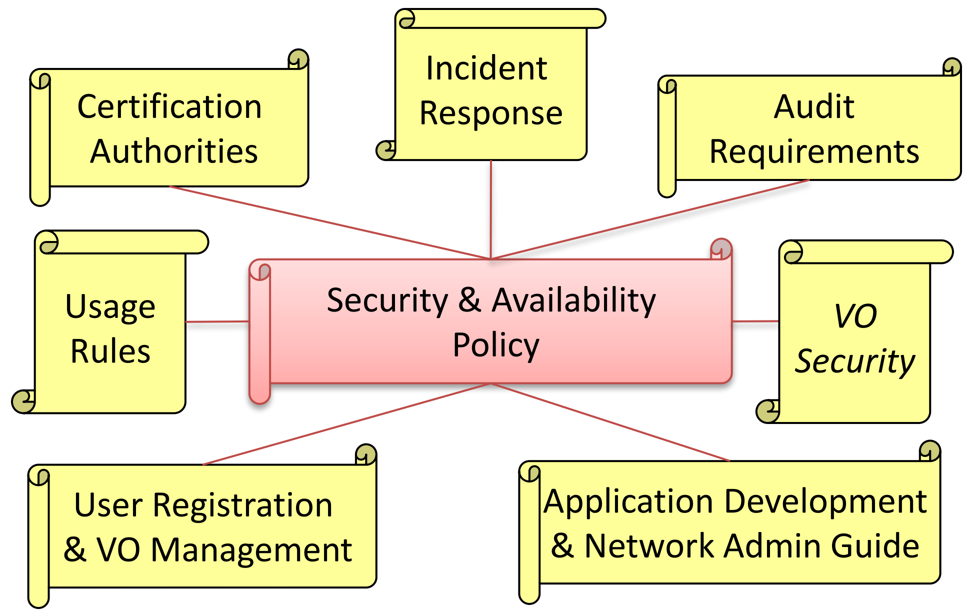
- Availability/Reliability monitoring
  - SAM tests and infrastructure
    - Now migrated to NAGIOS based system, decentralized
  - Visualization: GridView, GridMap, dashboard.cern.ch......
  - Solid foundation: Monitoring Infrastructure



Security & Policy

Collaborative policy development

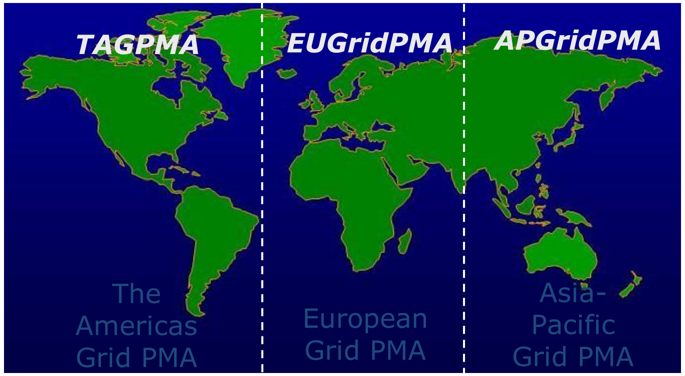
- Joint Security Policy Group
- **Certification Authorities**
  - EUGridPMA → IGTF, etc.
- **Grid Acceptable Use Policy (AUP)**
  - common, general and simple AUP
  - for all VO members using many Grid infrastructures
    - EGEE, OSG, SEE-GRID, DEISA, national Grids…
- **Incident Handling and Response**
  - defines basic communications paths
  - defines requirements (MUSTs) for IR
  - not to replace or interfere with local response plans



Security & Policy Groups

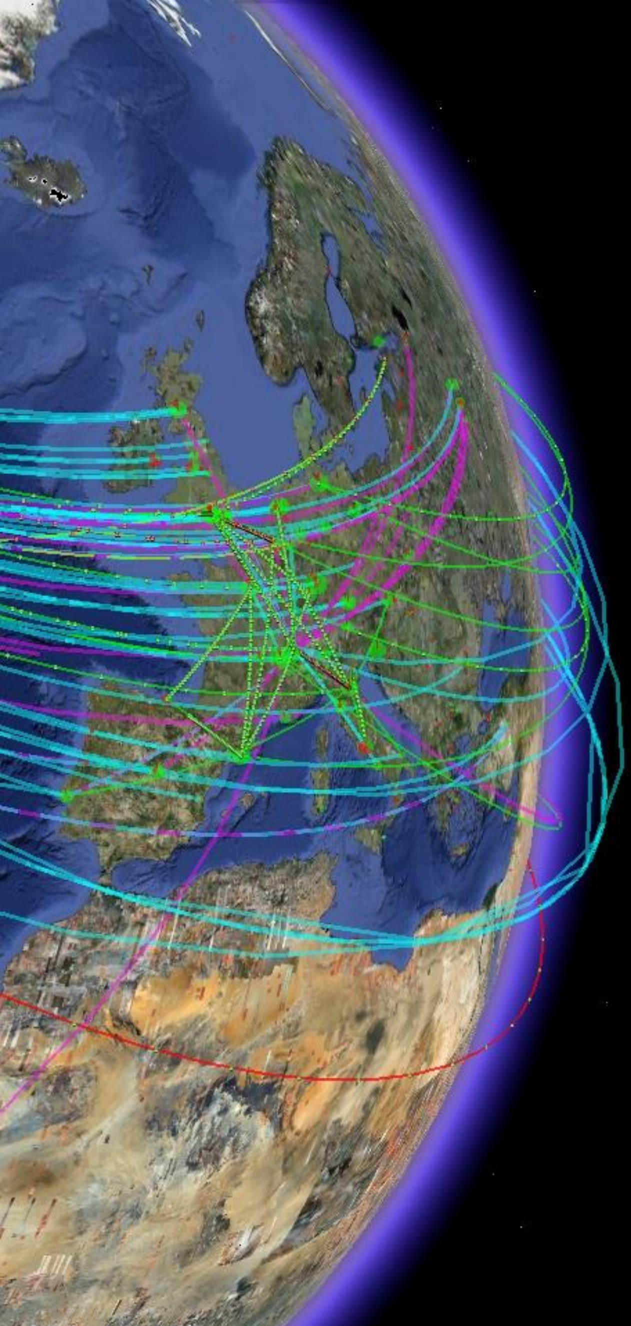
- Joint Security Policy Group
- EuGridPMA (& IGTF)
- Grid Security Vulnerability Group
- Operations Advisory Group

Security groups



- Joint Security Policy Group:
  - Joint with WLCG, OSG, and others
  - Focus on policy issues
  - Strong input to e-IRG
- EUGridPMA
  - Pan-European trust federation of CAs
  - Included in IGTF (and was model for it)
  - Success: most grid projects now subscribe to the IGTF
- Grid Security Vulnerability Group
  - Looking at how to manage vulnerabilities
  - Risk analysis is fundamental
  - Hard to balance between openness and giving away insider info
- Operational Security Coordination Team
  - Main day-to-day operational security work
  - Incident response and follow up
  - Members in all ROCs and sites
  - Frequent tests (Security Challenges)

WLCG Worldwide LHC Computing Grid

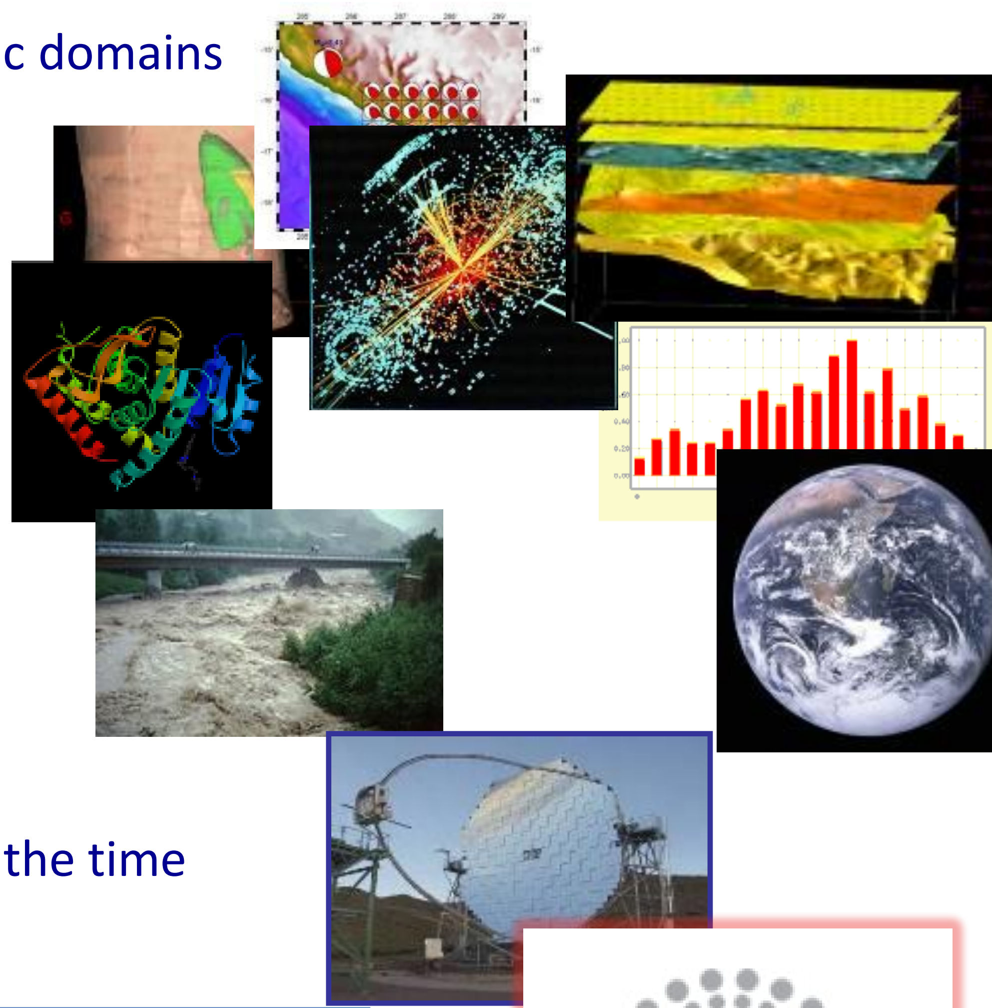


WLCG

- Why is this a challenge?
- Why a distributed system?
  - and why a grid
- History
- Architecture
- Monitoring and Operation
- **Usage**

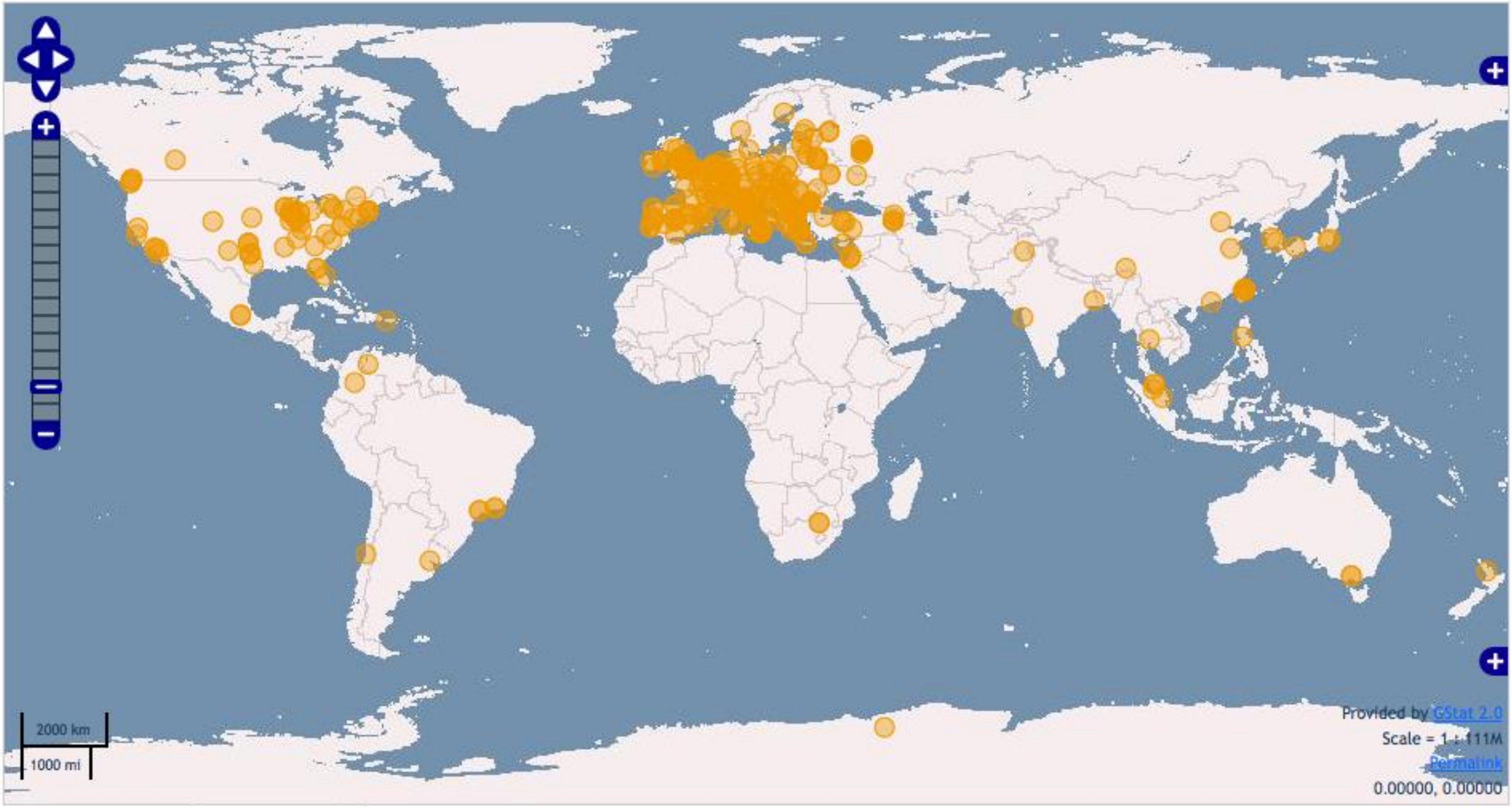
Shared Infrastructures: EGI

- >270 VOs from several scientific domains
  - Astronomy & Astrophysics
  - Civil Protection
  - Computational Chemistry
  - Comp. Fluid Dynamics
  - Computer Science/Tools
  - Condensed Matter Physics
  - Earth Sciences
  - Fusion
  - High Energy Physics
  - Life Sciences
  - ....
- Further applications joining all the time
  - Recently fishery (I-Marine)



Applications have moved from testing to routine and daily usage

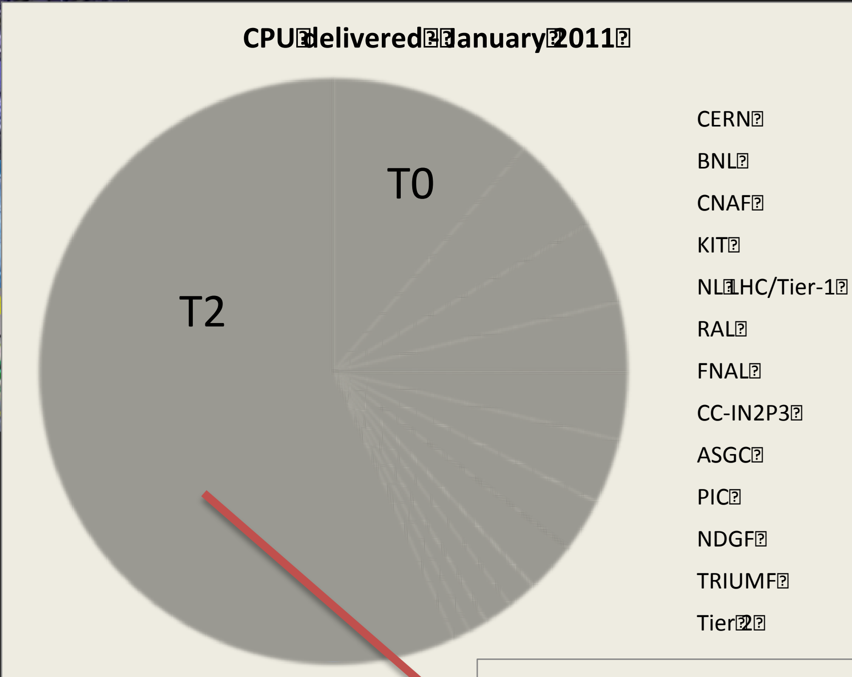
WLCG sites from Gstat Monitor



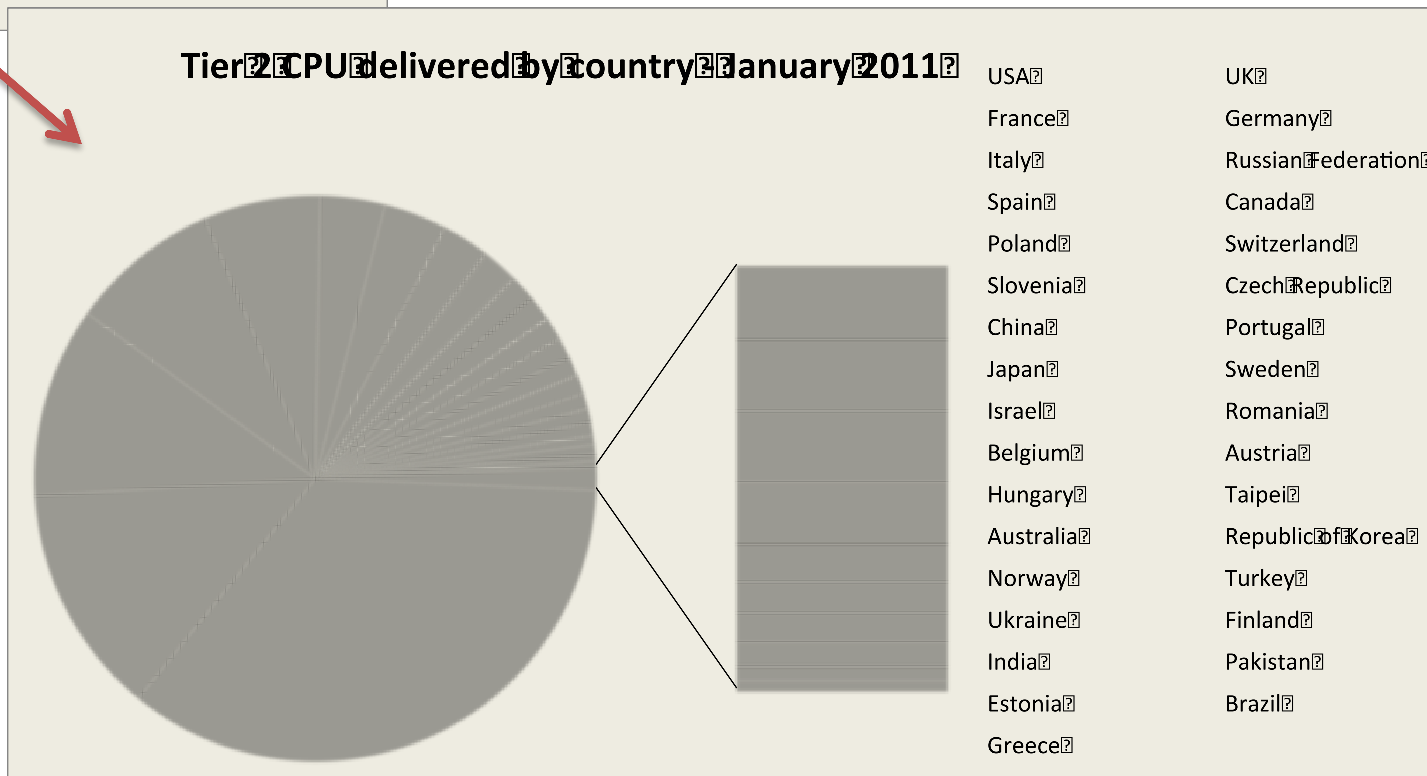
Data taken live from the LDAP based information system (BDII)

CPU – Usage at the Tiers

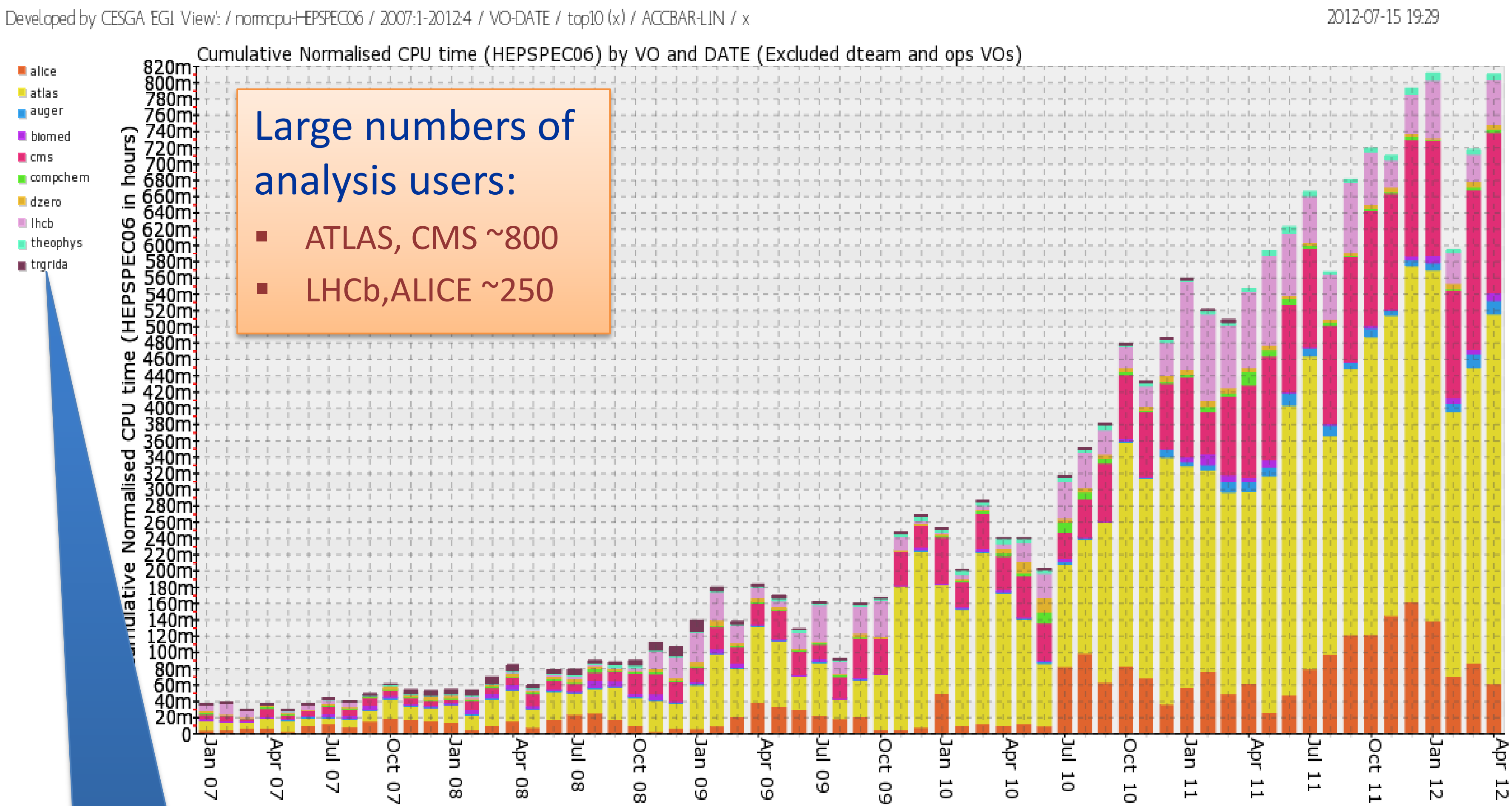
- The grid works
- All sites, large and small can contribute
  - And their contributions are needed!
- Significant use of Tier 2s for analysis
- Tier 0 usage peaks when LHC running – average is much less



Jan 2011 was highest use month ever … now exceeded



History of Grid Usage



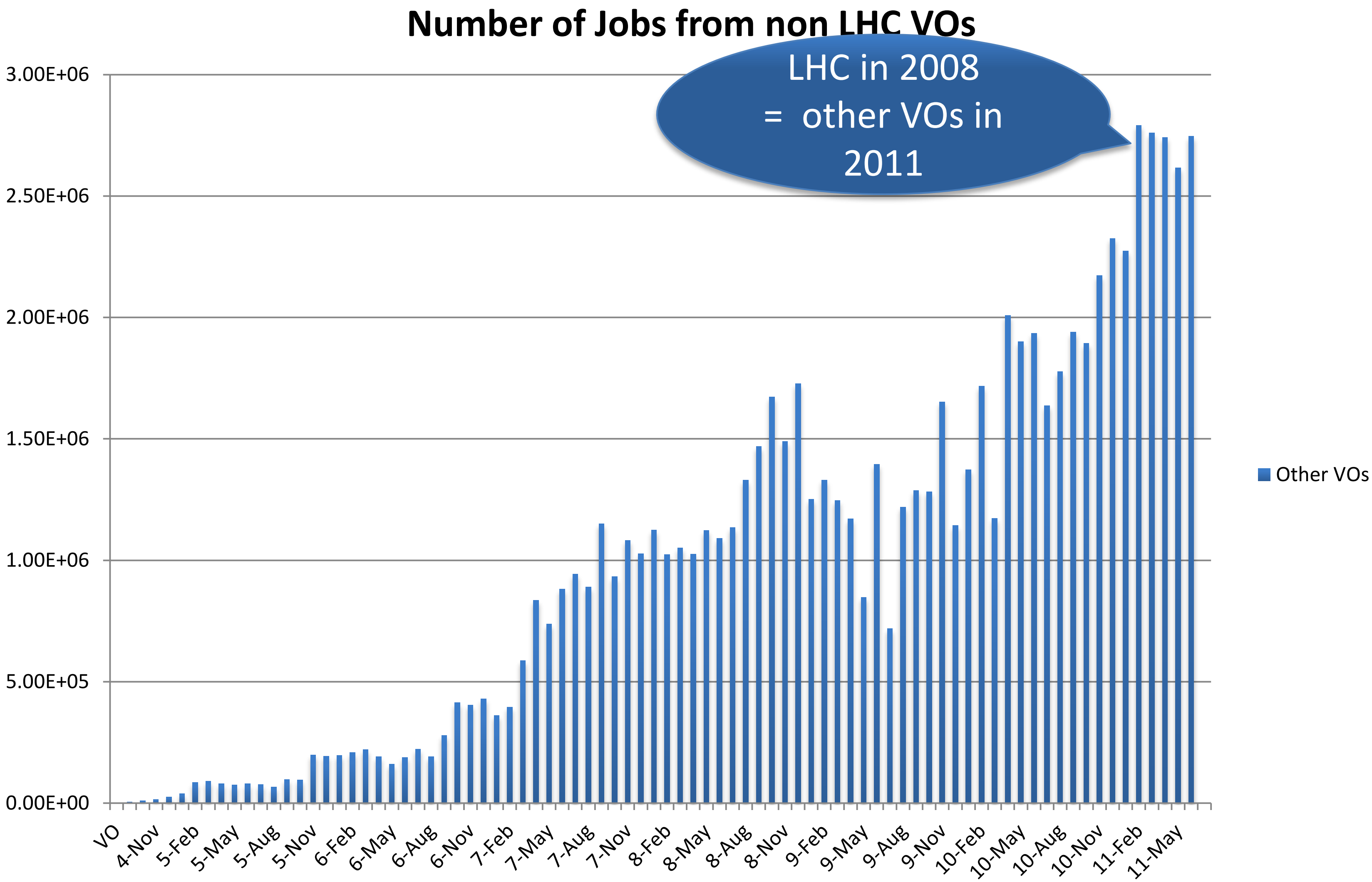
Large numbers of analysis users:

- ATLAS, CMS ~800
- LHCb,ALICE ~250

The Grid is also used by non LHC VOs

As well as LHC data, large simulation productions always on going

Non LHC VOs Usage

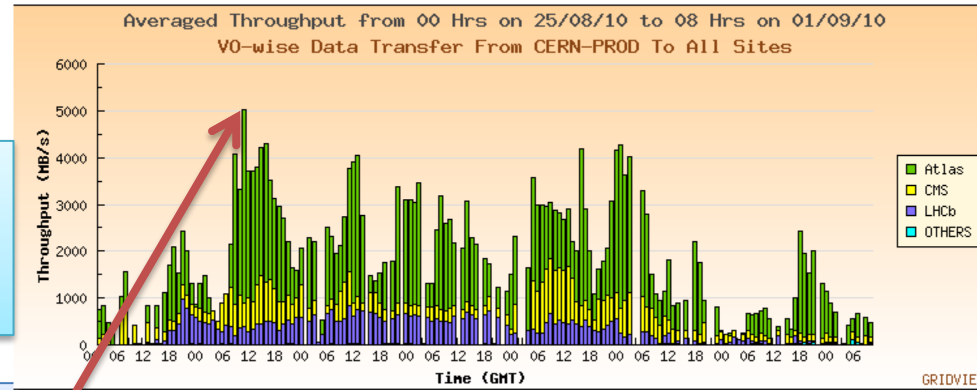


Data transfers

LHC data transfers:
April 2010 – May 2011

Rates >> higher than planned/tested
Nominal: 1.3 GB/s
Achieved: up to 5 GB/s

World-wide: ~10 GB/s per large experiment



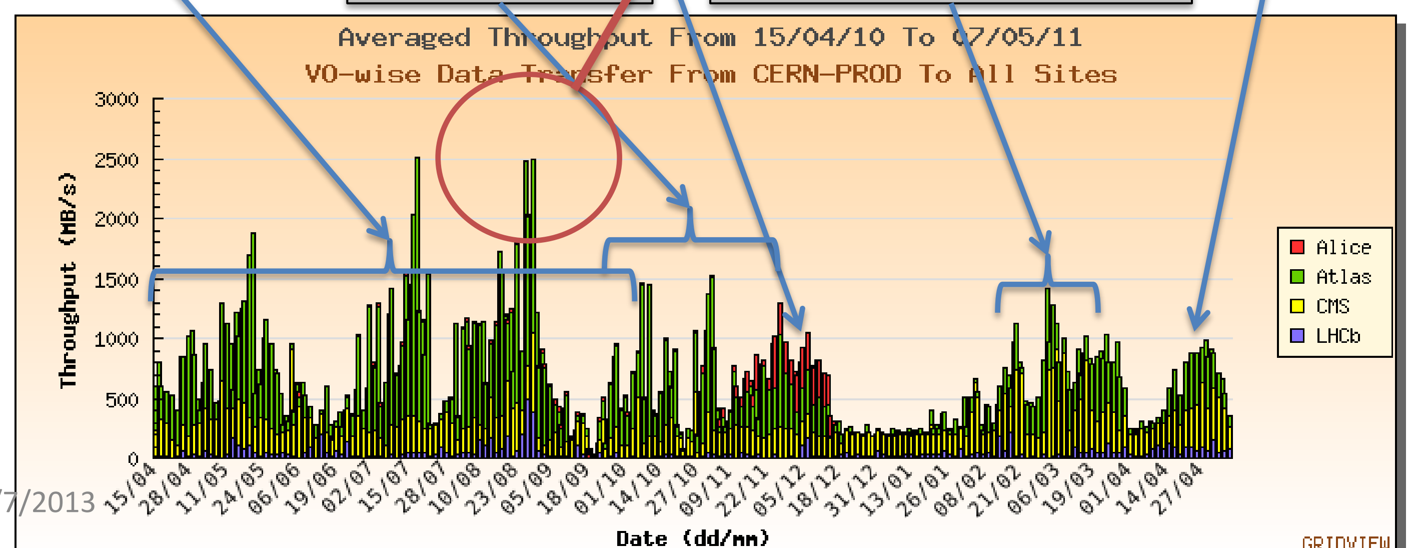
2010 pp data → Tier 1s & re-processing

ALICE HI data → Tier 1s

Re-processing 2010 data

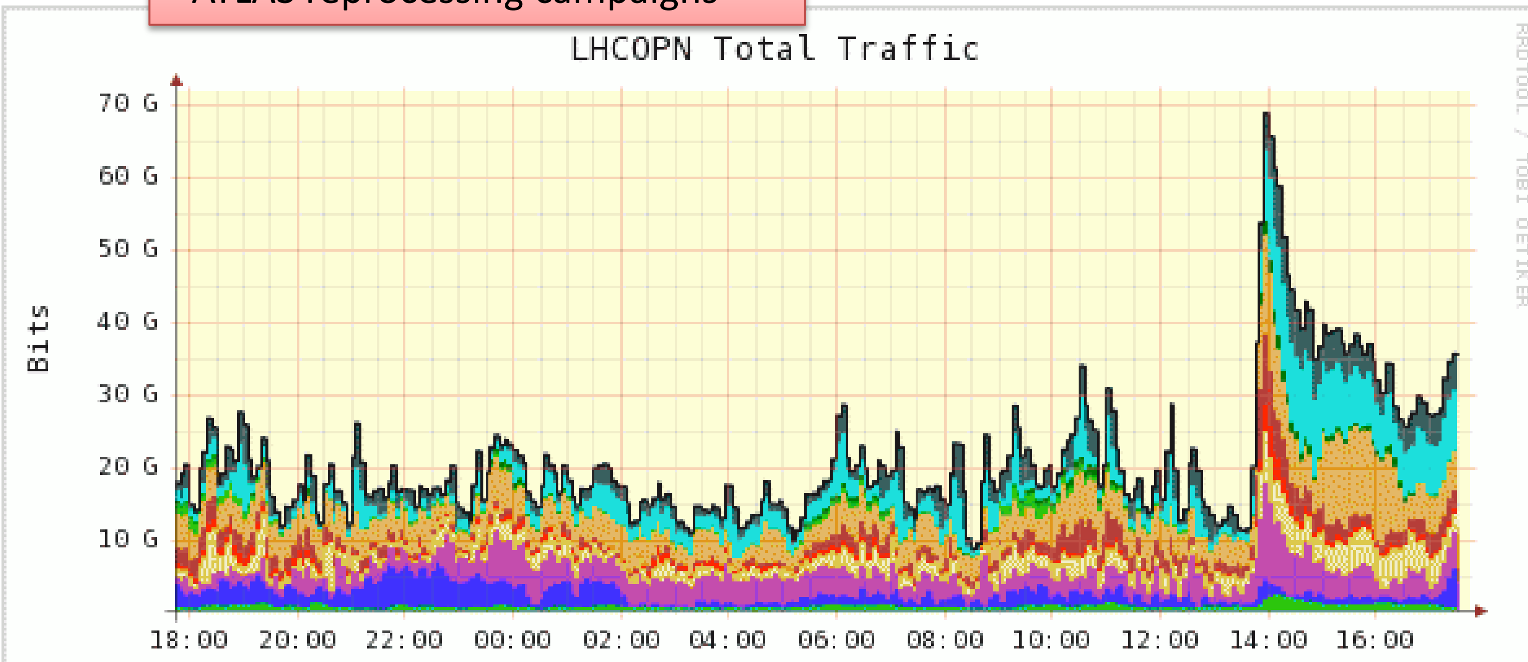
CMS HI data zero suppression & → FNAL

2011 data → Tier 1s



70 / 110 GB/s !

Traffic on OPN up to 70 Gb/s!
- ATLAS reprocessing campaigns



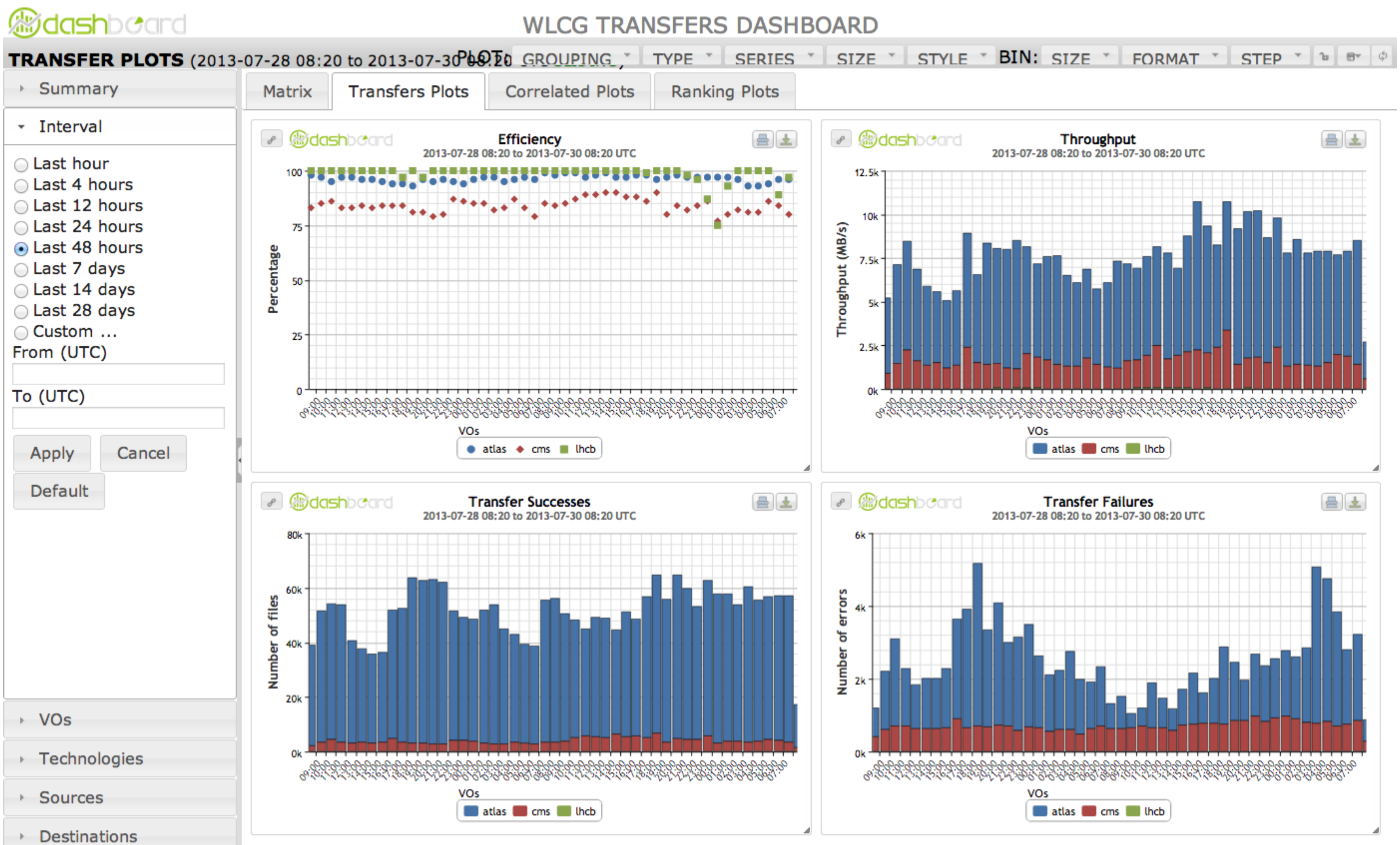
Significant levels of network traffic observed in 2010

Caused no network problems, but:

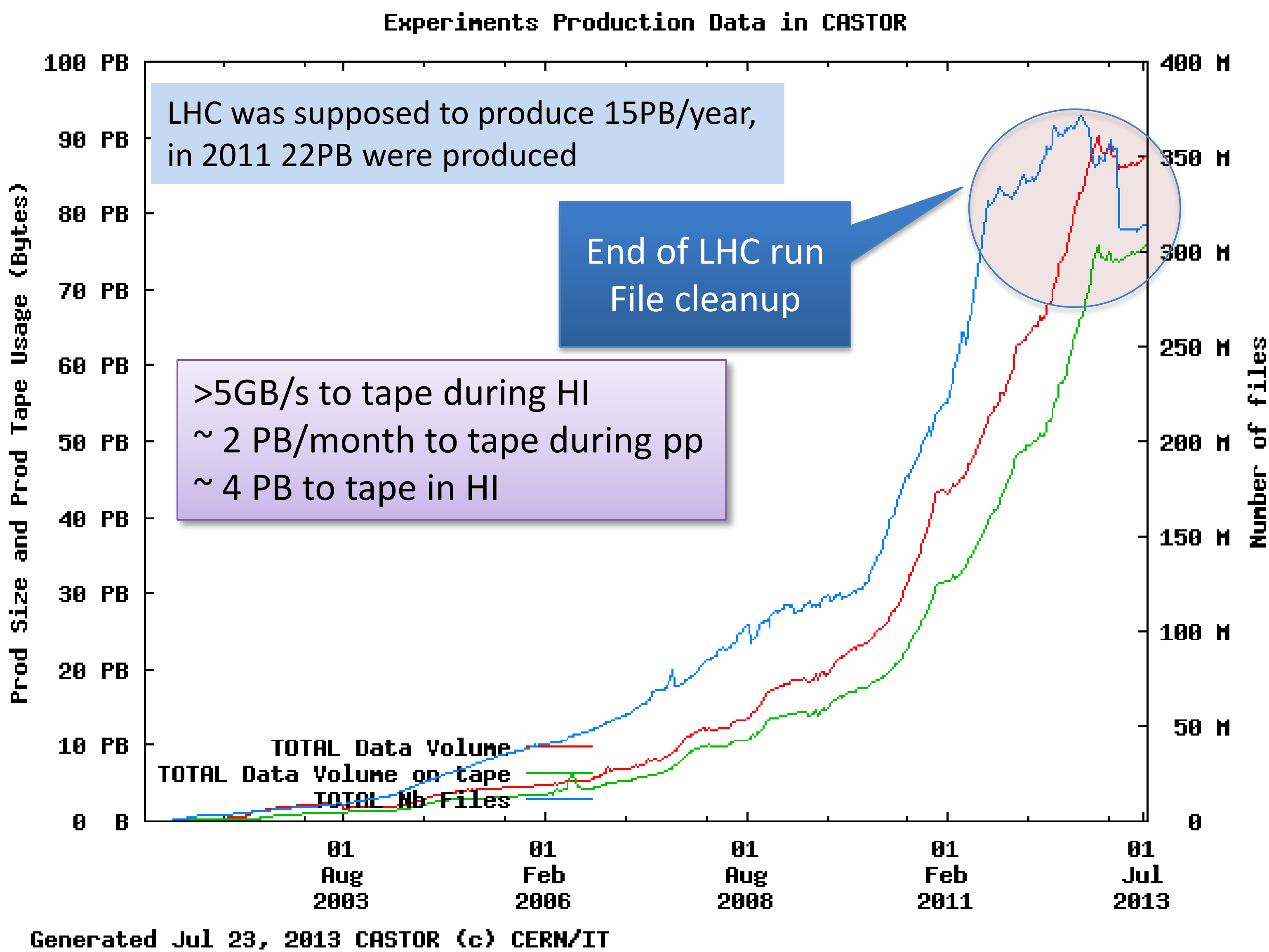
➢ Reasons understood (mostly ATLAS data management)

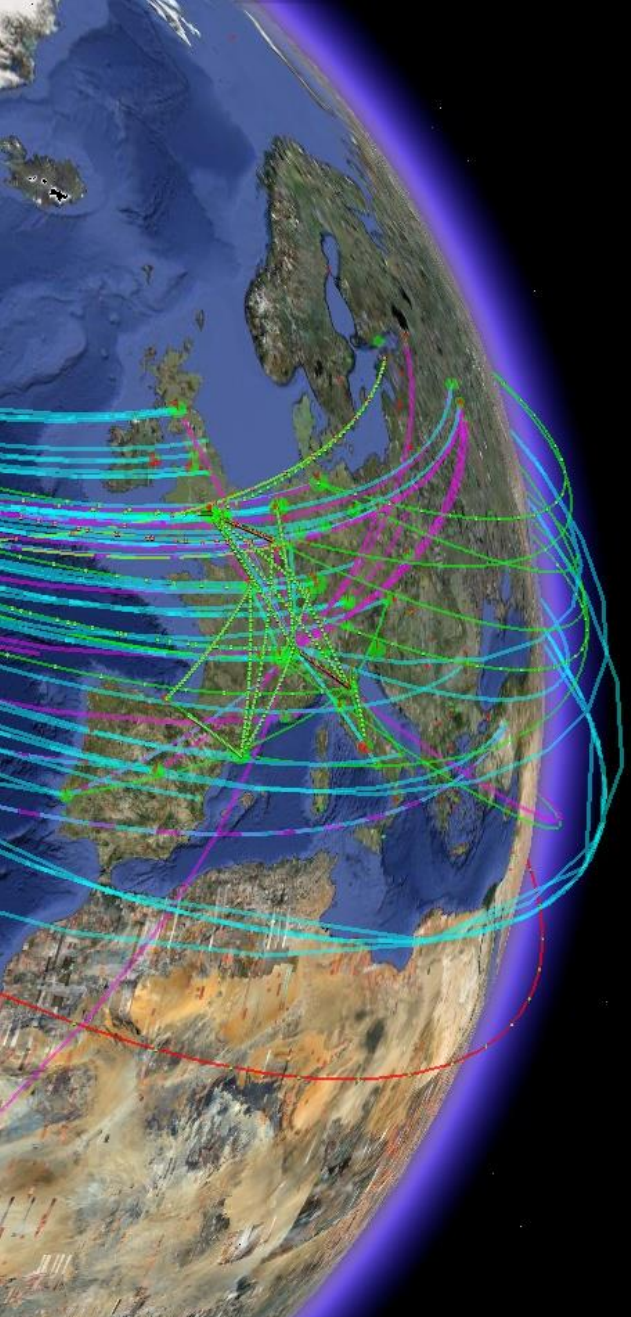
And now?

Just an example, and let's not forget that it's the end of July, during the shutdown



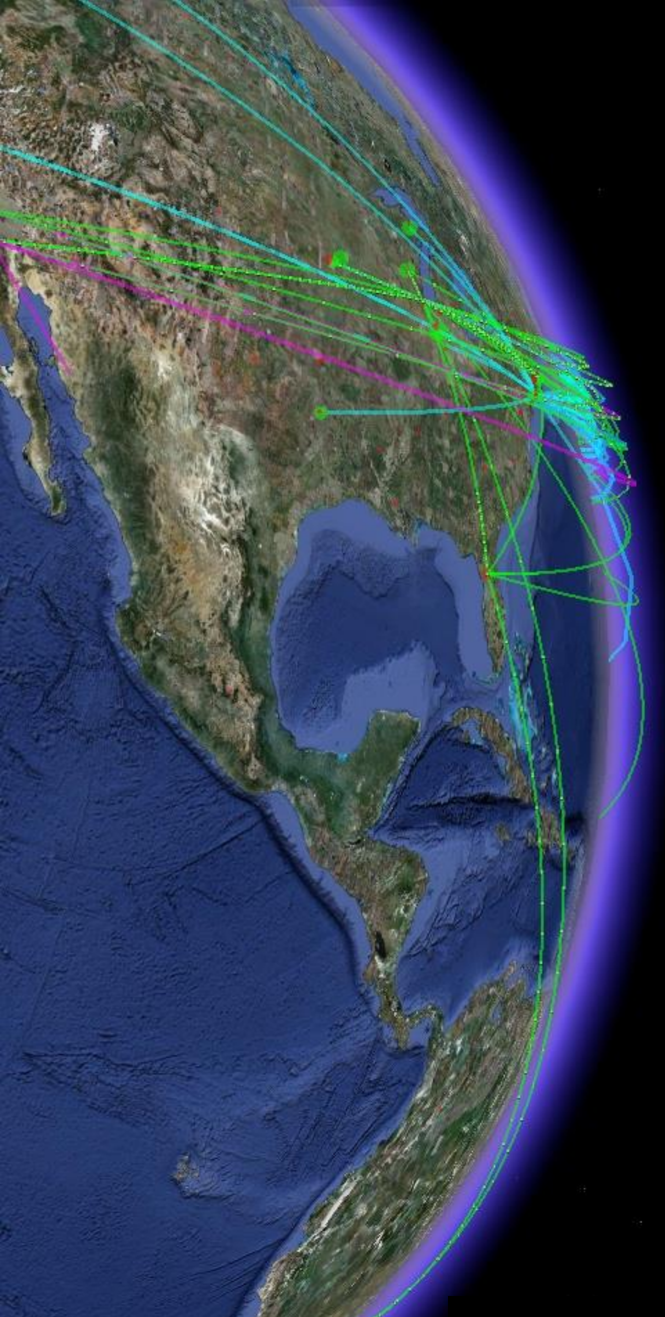
CASTOR – CERN tape storage





Summary

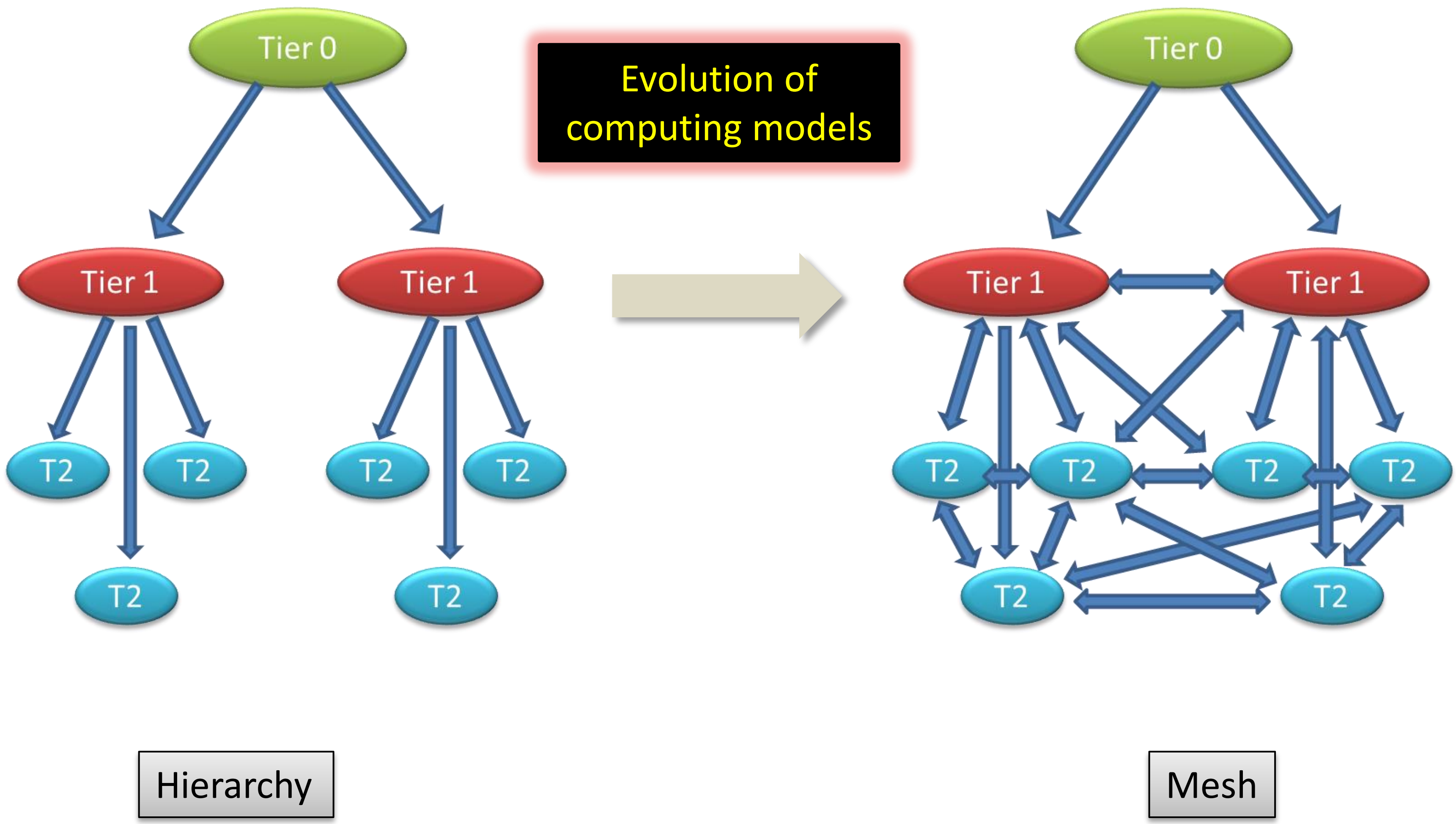
- Grid Computing and WLCG has proven itself during the first year of data-taking of LHC
- Grid computing works for our community and has a future
- Long term sustainability will be a challenge
  - Future of EGI...



Future

- WANs are now very stable and provide excellent performance
  - Move to a less hierarchical model
- Tuning for our applications
- Virtualization and Cloud Computing
- Moving towards standards
- Integrating new technology

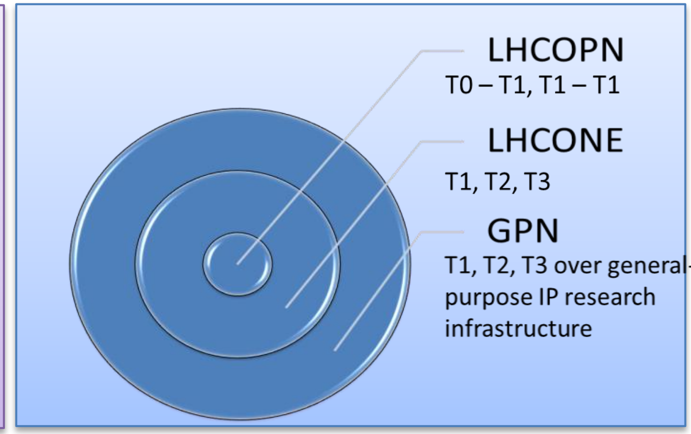
Computing Model Evolution

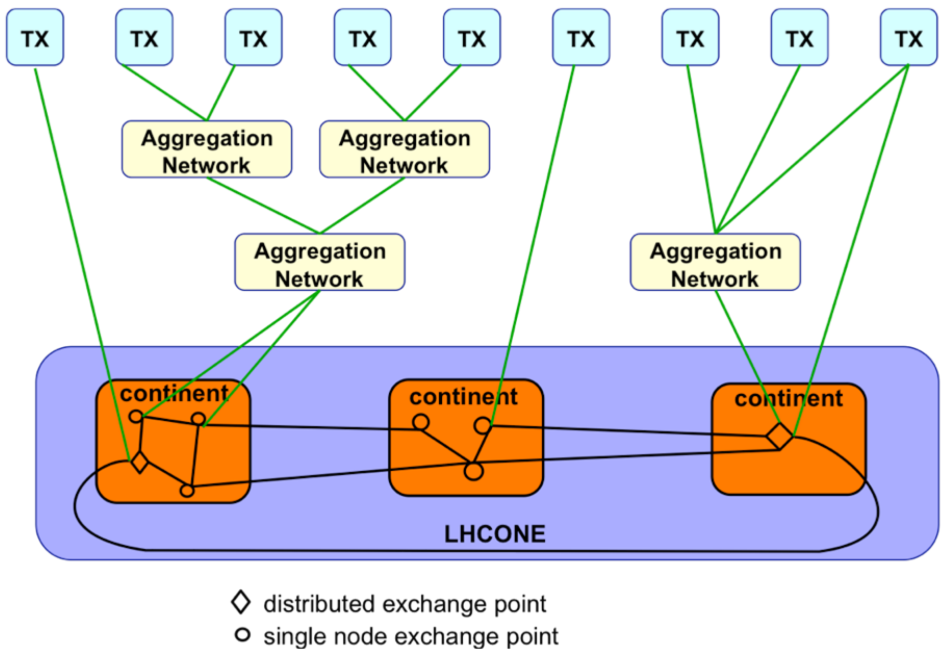


Network evolution - LHCONE

Evolution of computing models also require evolution of network infrastructure

- Enable any Tier 2, 3 to easily connect to any Tier 1 or 2

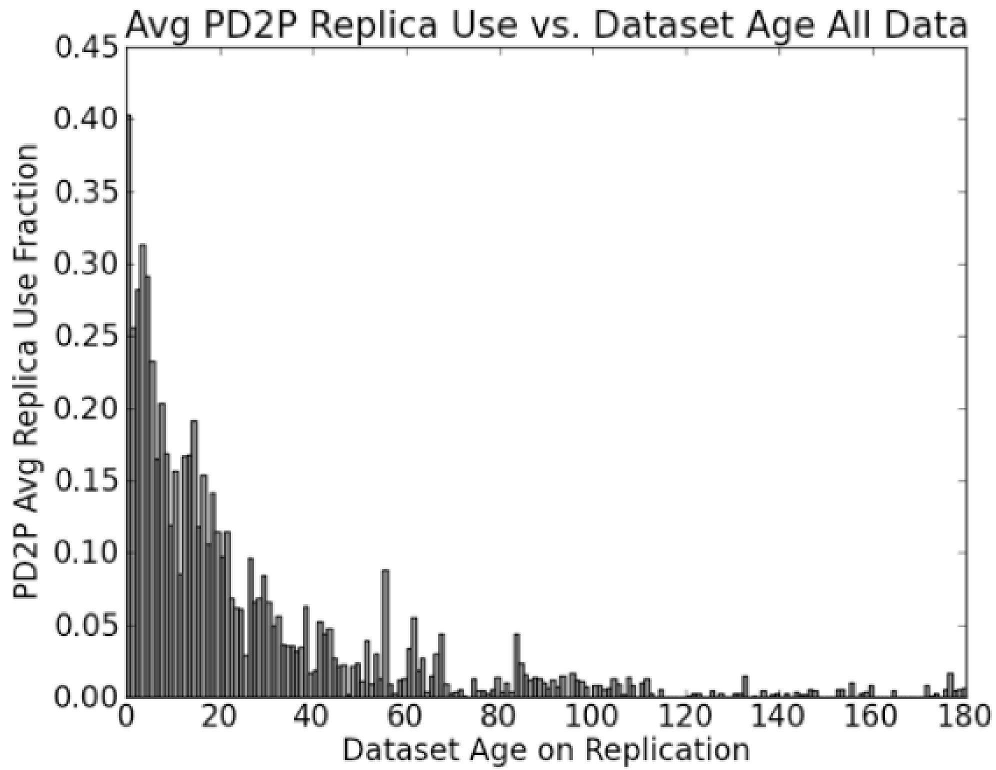


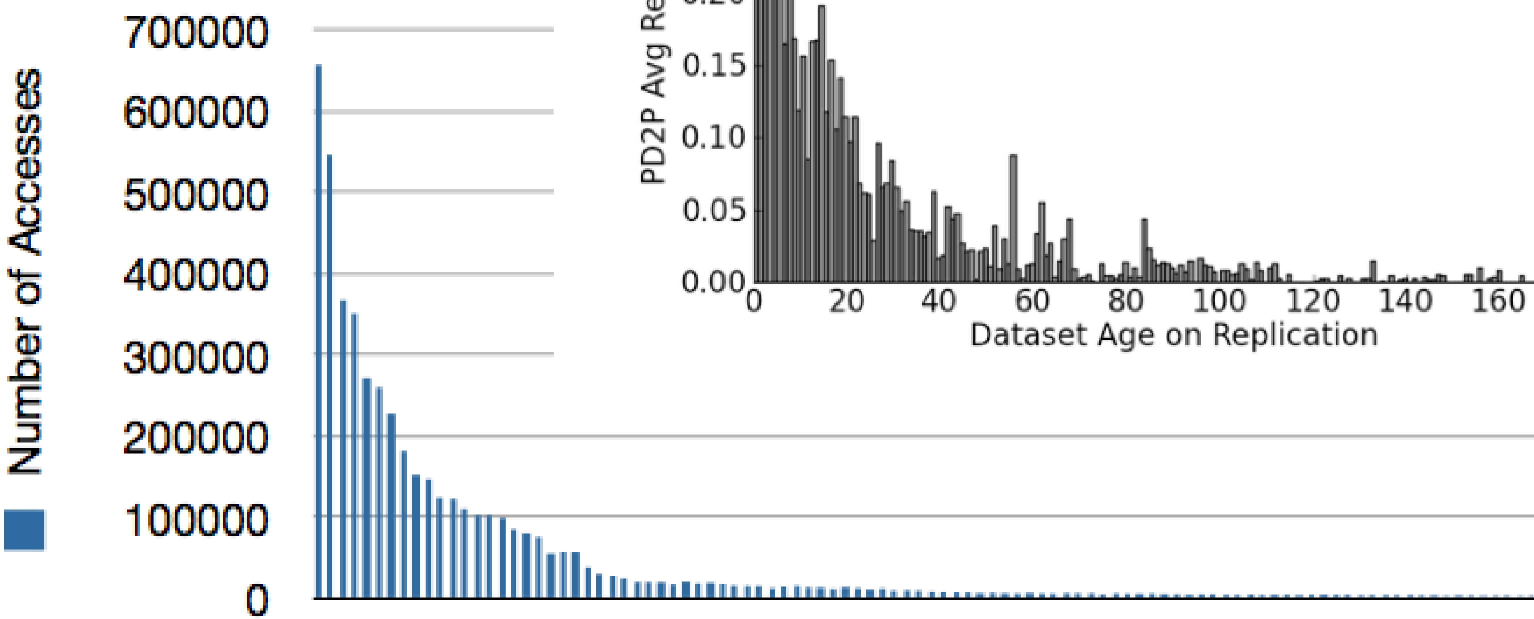


- Use of Open Exchange Points
- Do not overload the general R&E IP infrastructure with LHC data
- Connectivity to T1s, T2s, and T3s, and to aggregation networks: NRENs, GÉANT, etc.

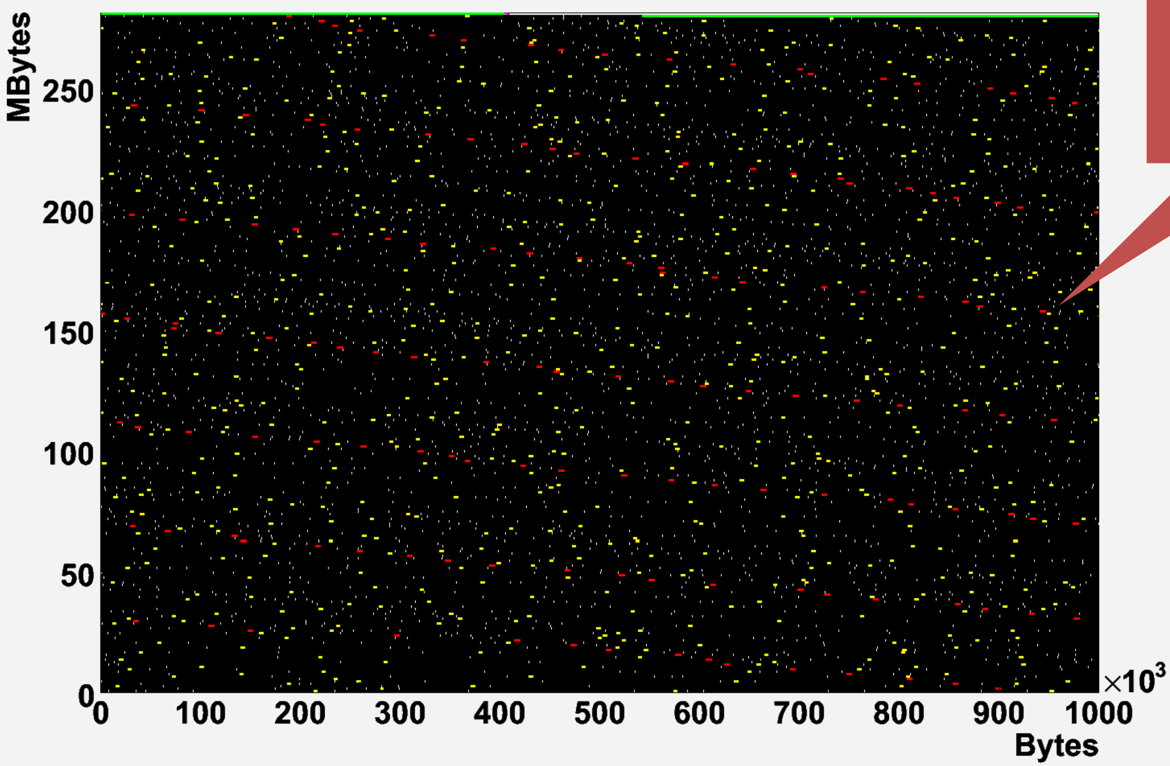
Data Popularity

- Usage of data is highly skewed
- Dynamic data placement can improve efficiency
- Data replicated to T2s at submission time (on demand)





Data access: Inside a ROOT file



3 branches have been coloured

HEP data is stored as ROOT files

- Associated information can be scattered throughout the file
- This means that file access is very sensitive to latency
- The root team have made many improvements which now open up different data management possibilities

WLCG Worldwide LHC Computing Grid

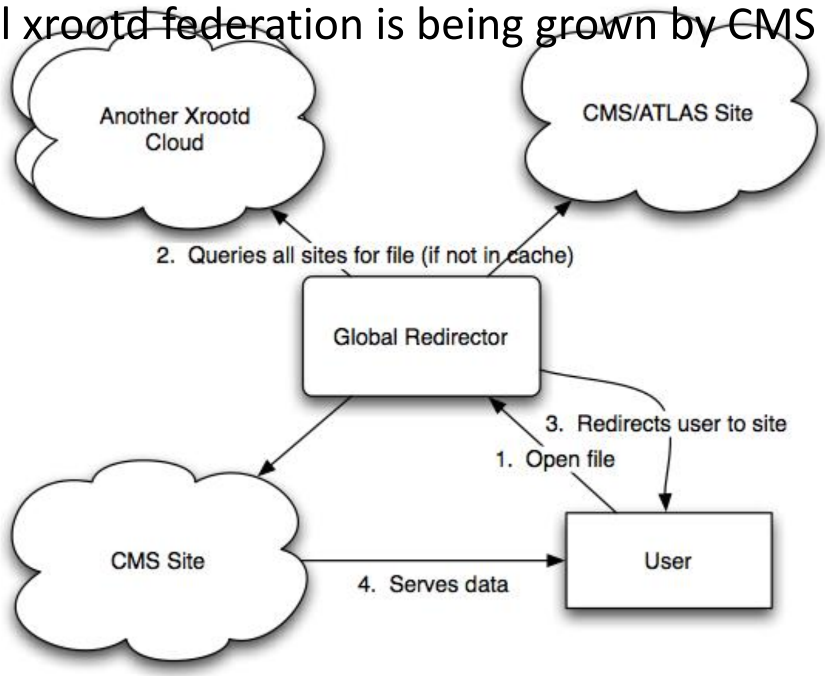
Data access: the DM problem

- Smart apps can access data directly via WAN
- The question is… where is the file that my app needs ? All I know is its name…
- What if my app needs 10 files? Where are they? Where do I request my app to run? Shall I gather them first?

- Historical solution at the base of this:
  - Write in a DB all the known locations (and keep it up to date even after HW failures)
  - The ATLAS LFC DB has now >300 million replica entries (!)
- Could be possible to apply concepts from the P2P or DNS world, in order to have always the exact information
  - and consider the catalogue as just an “indication”, to know what is supposed to be hosted in a site
- A lot of interest around the concept of “Storage Federation” (similar to “Loosely coupled storage system”)

Data access: using WANs

- Direct data access over the WAN is now a possibility for some data access frameworks/protocols
  - More efficient use of storage
  - Greater job reliability
  - Not necessarily more WAN traffic
    - Can be combined with various caching strategies
  - Can be quicker than pulling something locally from tape
  - Can be cheaper (and quicker) than creating one more replica
- XROOTD and HTTP offer this possibility (WAN optimised operations, parallelism)
- A global experimental xrootd federation is being grown by CMS and ATLAS:



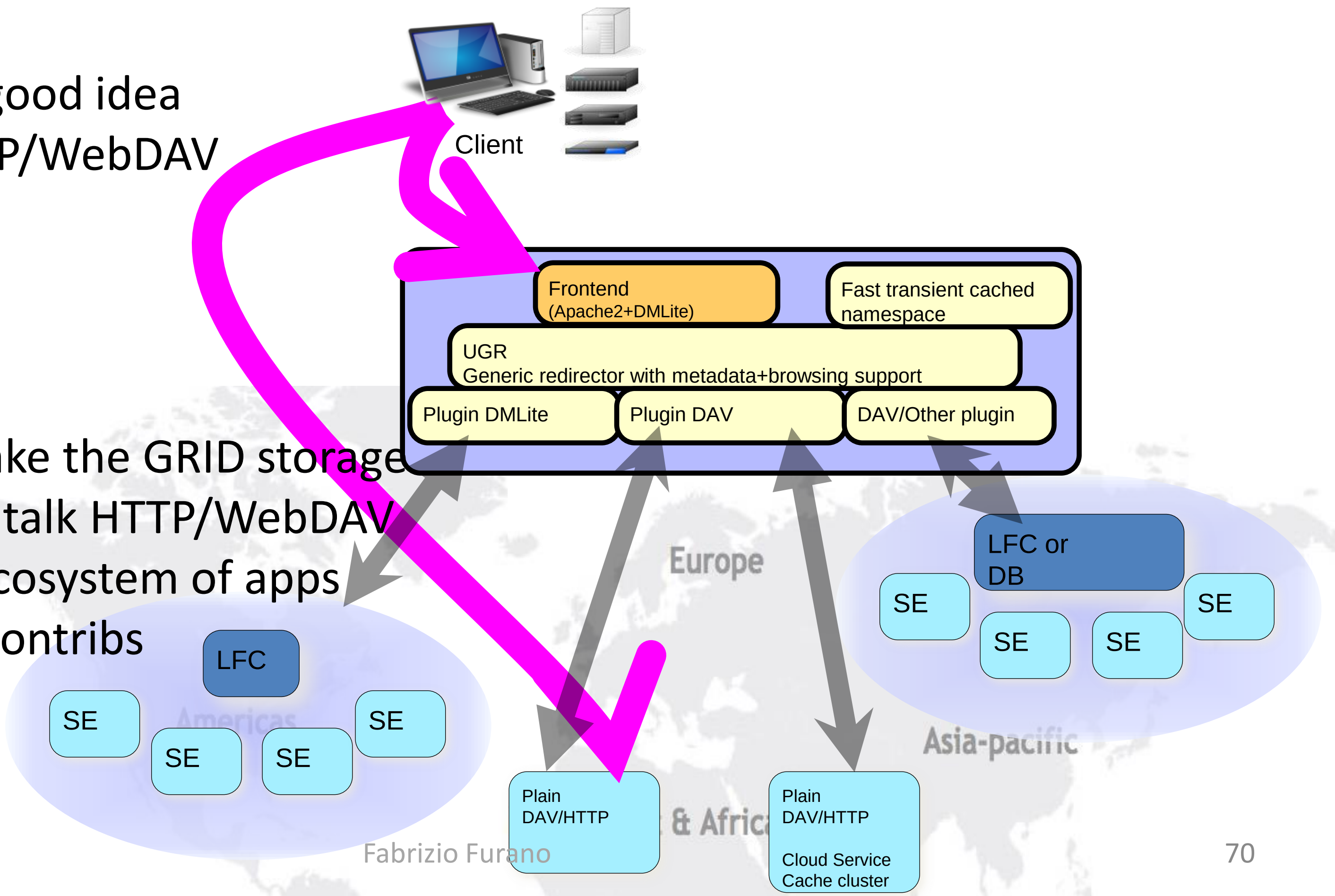
Use as a whole

A New entry, the Dynamic Federations
A high performance system able to aggregate and cache storage metadata on the fly
In practice… build a huge storage just by aggregating sites, transparently.

Can talk to any system, through plugins

We think that it's a good idea
making it speak HTTP/WebDAV

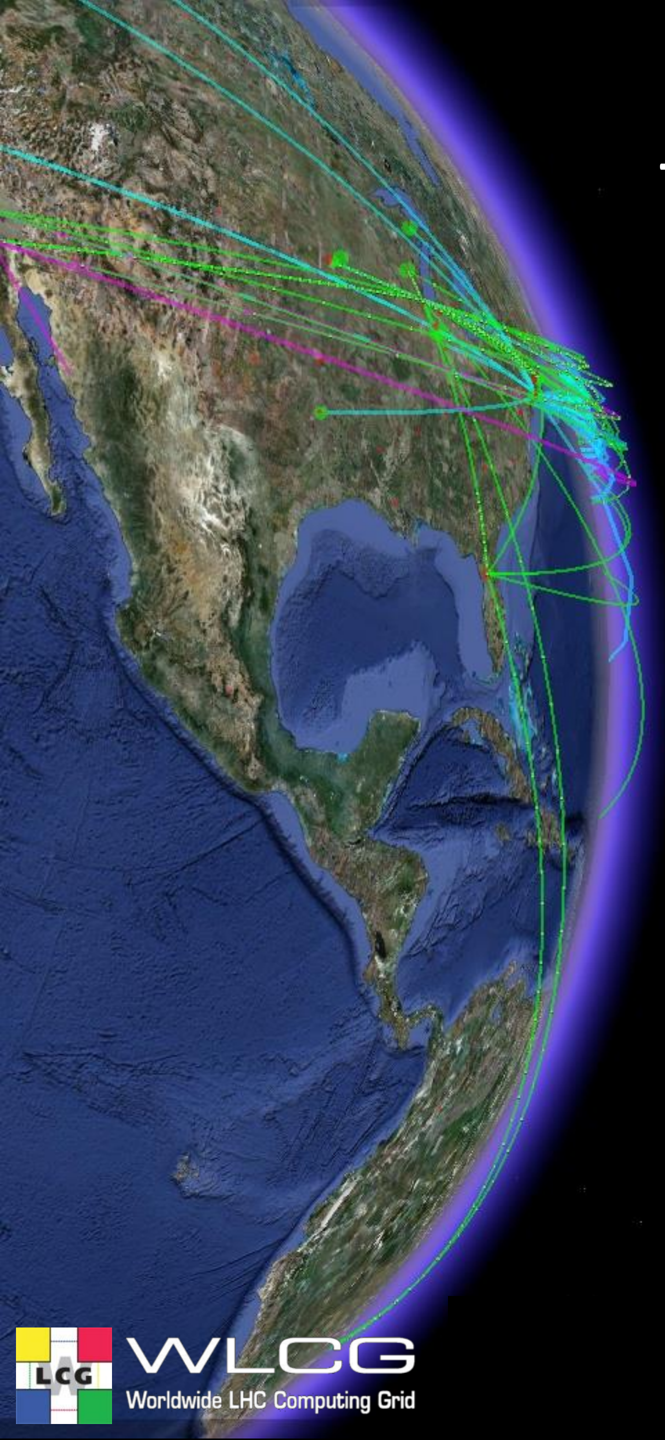
In parallel…
Very big effort to make the GRID storage
components able to talk HTTP/WebDAV
Effort to create an ecosystem of apps
Requires us to give contribs



Virtualisation & cloud

Virtualisation is interesting in a number of domains

- Application Environment
- HEP applications are platform dependent
  - Sites & laptops are varied
- Infrastructure Management
- Direct cloud use by LHC experiments
  - Simulation
  - Elasticity
  - Reprocessing & analysis
    - Data cost



To Grid or not to Grid?

Grid

- Distributed community (VO)
  - Different organizations
  - Distributed resources
- Longer term project (> 2 years)
  - With massive computing requirements (>> 100 PC nodes)
- Computing requires modest parallelization
  - MPI is available on some sites, but not easy to use in a Grid

- Don't expose middleware directly to end users
  - Link from workflow management/portals
  - Shield users from failures/complexity
  - Distributed computing requires management of failures
- Join an existing infrastructure
  - EGI is in Europe a good choice
- Use workflow management software from other Vos
  - Dirac, Panda, gCube from D4Science .....
- Get sufficient expertise.....

WLCG Worldwide LHC Computing Grid

LCG

Half Grid

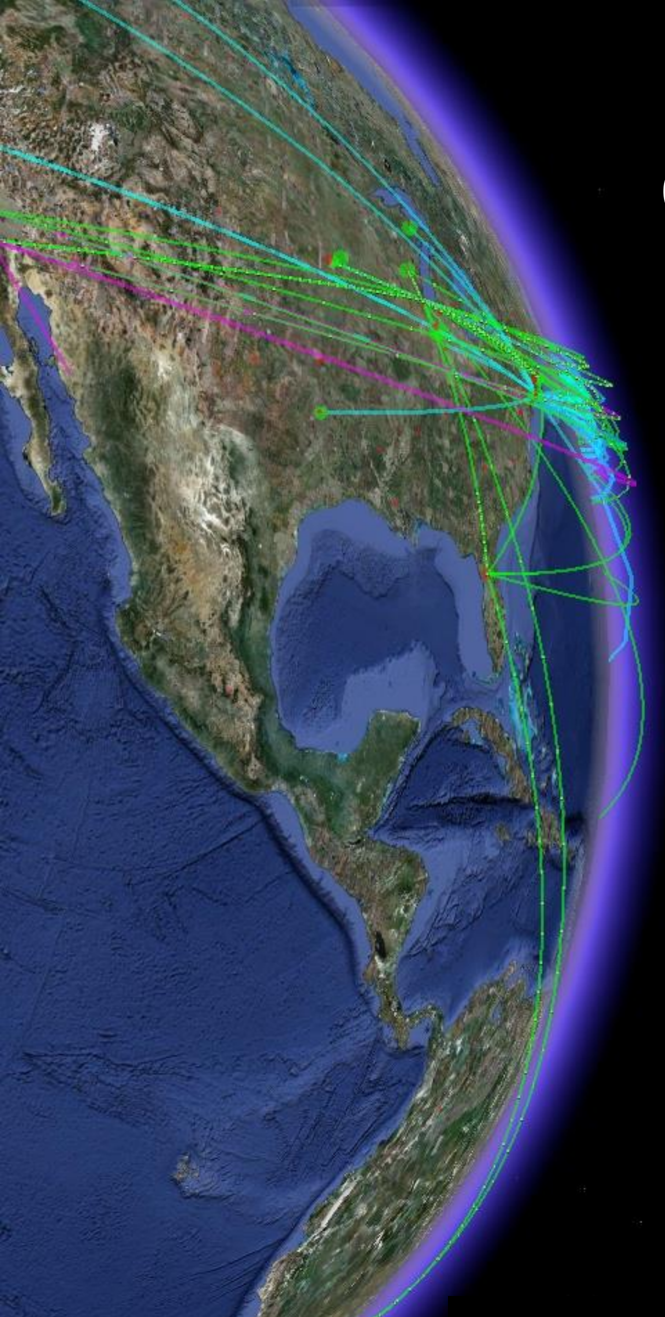
- Distributed small community (< 100)
  - Closely linked (same region or organization)
  - Distributed resources
- Medium term project (< 2 years)

- Join an existing VO (use their experience)
- Or:
  - Link your resources via Condor
    - http://www.cs.wisc.edu/condor/
- Or:
  - Use cloud computing (OpenStack, OpenNebula, Amazon EC2..)
- Or:
  - Use volunteer computing (BOINC (like Seti@home)
  - We interfaced gLite and BOINC… not much use by HEP
- You still need to invest, but you will see results faster

No Grid

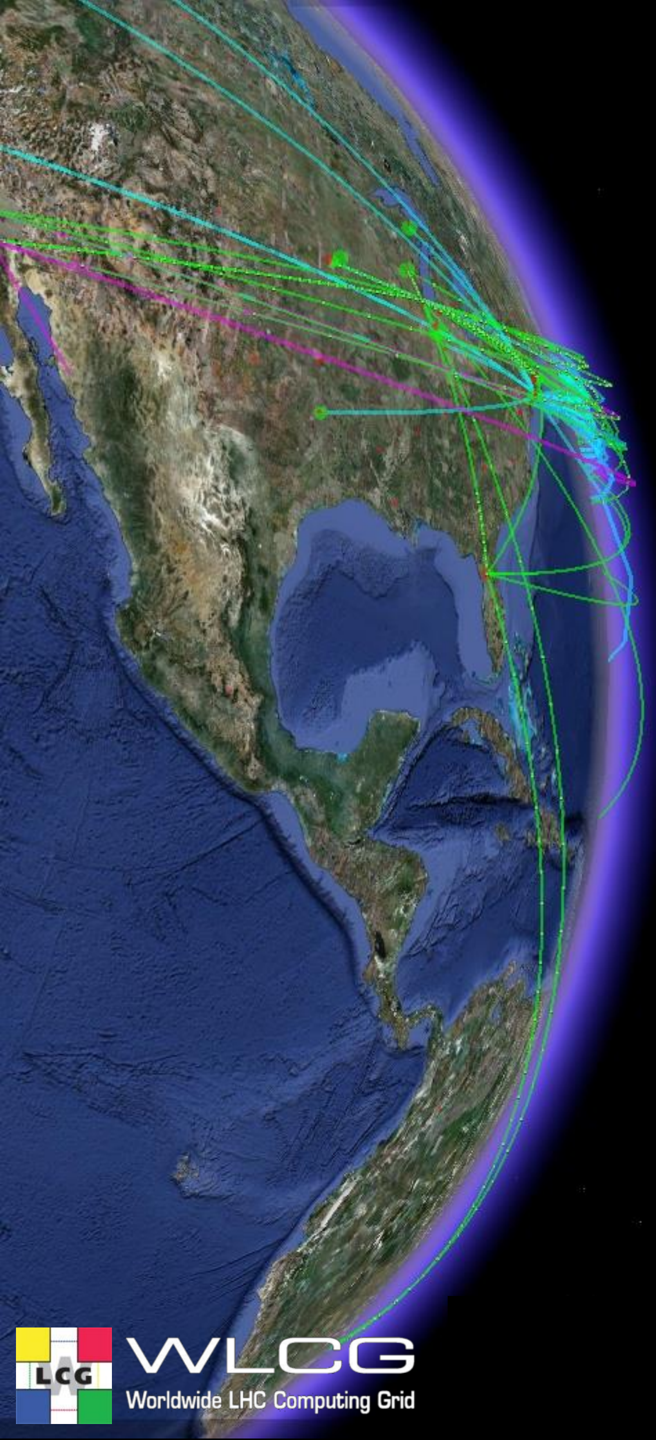
- Local team
  - Closely linked (same region or organization)
  - Distributed resources
- Short or medium term project (< 2 years)
- Massive parallel processing needed or HPC needed

- If you choose using the grid nevertheless…
  - Understand the startup costs



Credits

- Slides have been re-used from countless individuals
  - too many to name them individually.....



Thank you