The Worldwide LHC Computing Grid

its impact and evolution

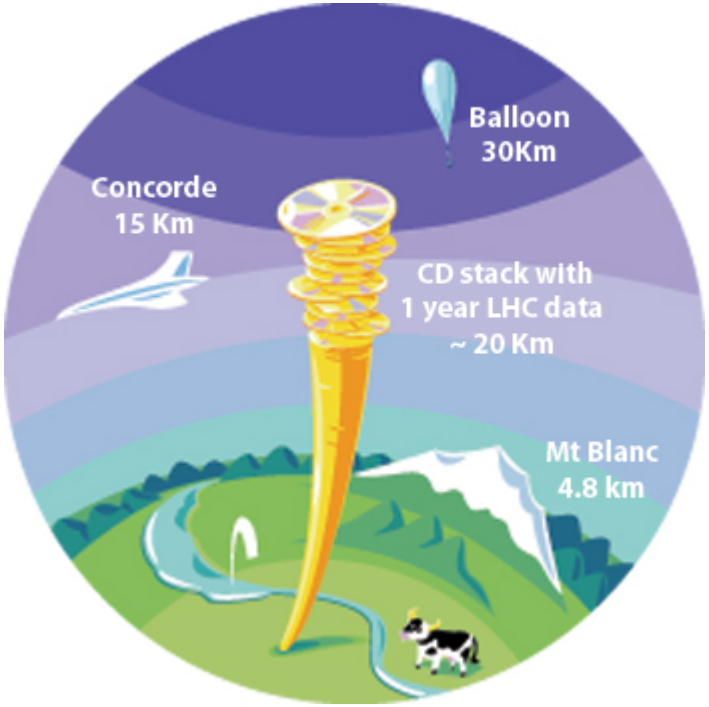
Contents

- Grids in Particle Physics
- Use of grids for other research communities
- Grids and Business
- Moving from grids to clouds
- Conclusions

The LHC Data Challenge

CERN IT Department

- The accelerator will run for 20 years
- Experiments are producing about **20 Million Gigabytes** of data each year (about 3 million DVDs – 700 years of movies!)
- LHC data analysis requires a computing power equivalent to **~100,000 of today's fastest PC processors**
- Requires many cooperating computer centres, as CERN can only provide ~20% of the capacity



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Solution: the Grid

CERN IT Department

- Use the Grid to unite computing resources of particle physics institutes around the world

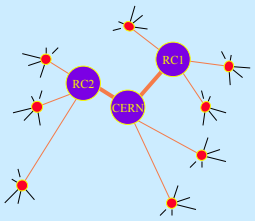
The **World Wide Web** provides seamless access to information that is stored in many millions of different geographical locations

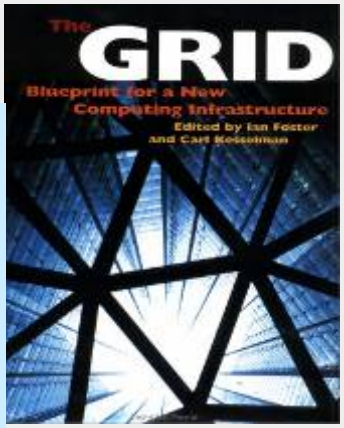
The **Grid** is an infrastructure that provides seamless access to computing power and data storage capacity distributed over the globe

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

How CERN got into grids

- Partially decentralized model
  - replicate the event data at about five regional centres
  - data transfer via network or movable media





GriPhyN, iVDGL, PPDG → GRID 3 → OSG → WLCG

EU DataGrid → EGEE 1 → EGEE 2 → EGEE 3

LCG 1 → LCG 2

1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 20

Service Challenges

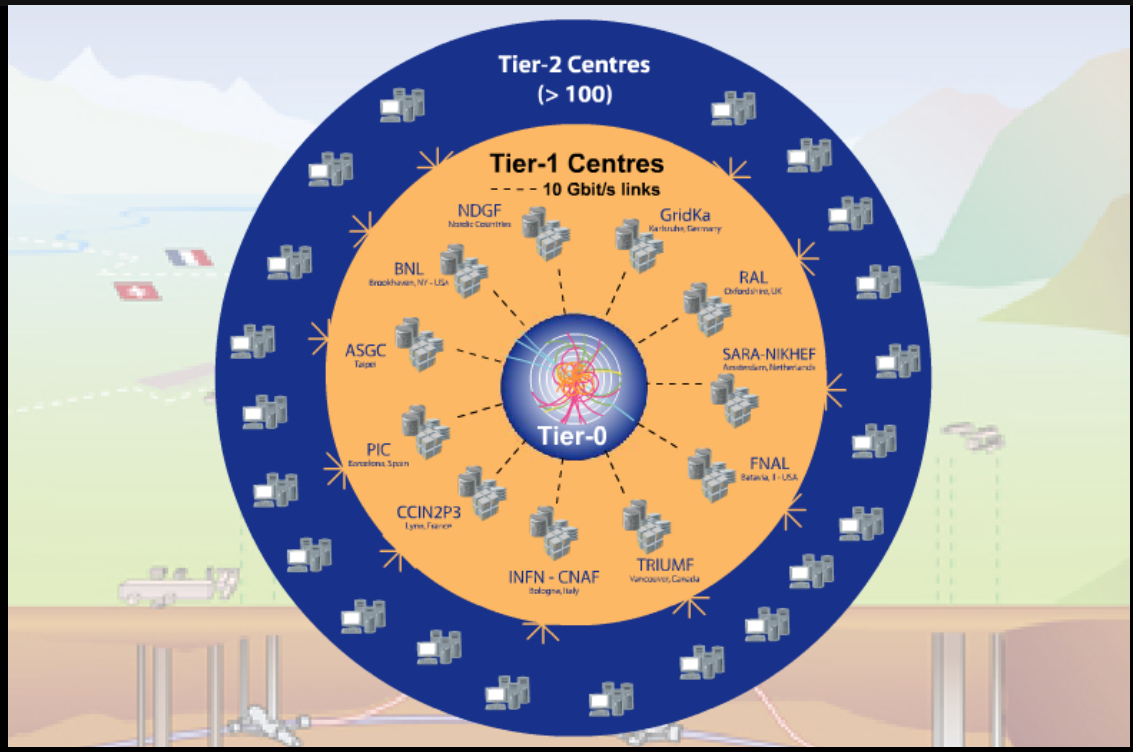
Cosmics

First physics

Data Challenges

WLCG – what and why?

- A distributed computing infrastructure to provide the production and analysis environments for the LHC experiments
- Managed and operated by a worldwide collaboration between the experiments and the participating computer centres
- The resources are distributed – for funding and sociological reasons
- Our task was to make use of the resources available to us – no matter where they are located

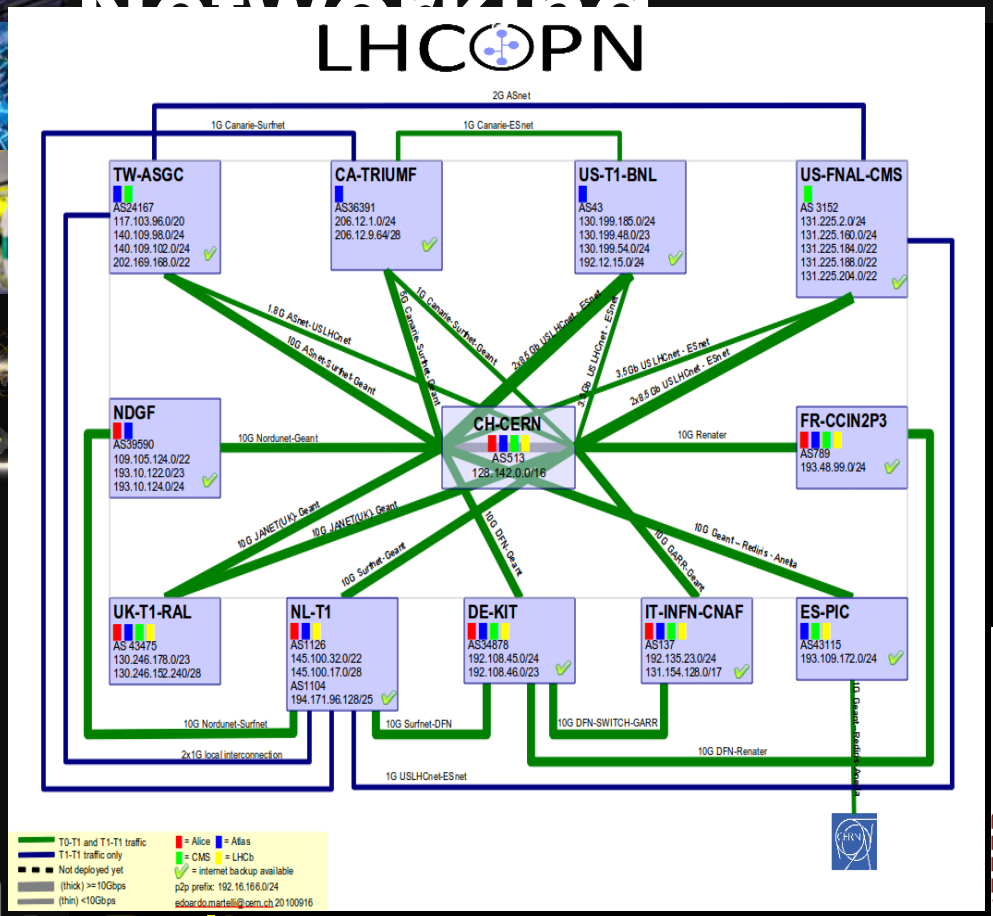


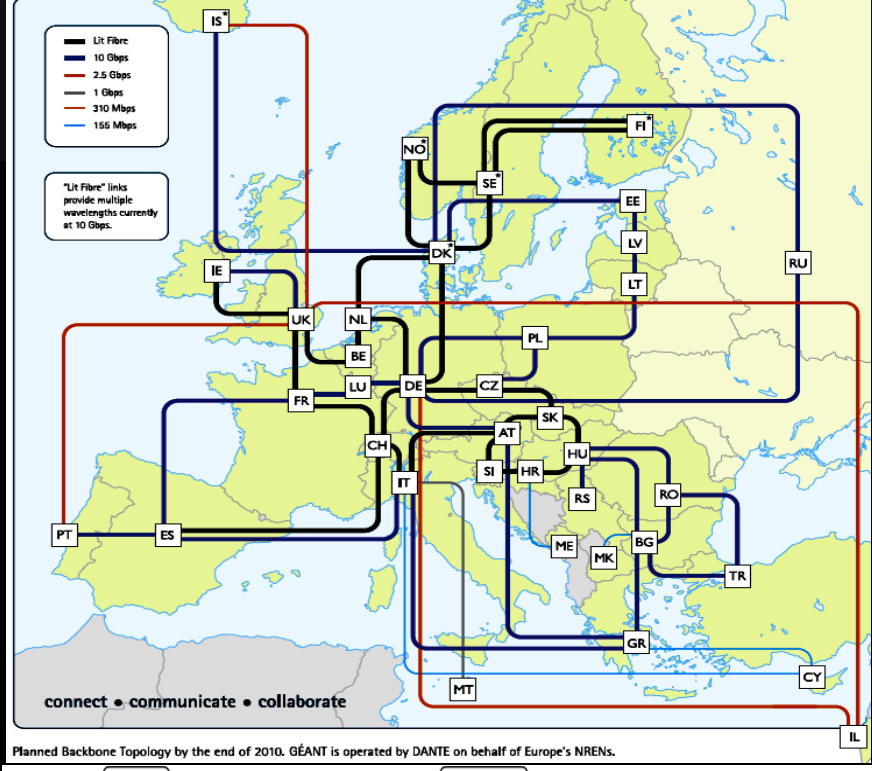
Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

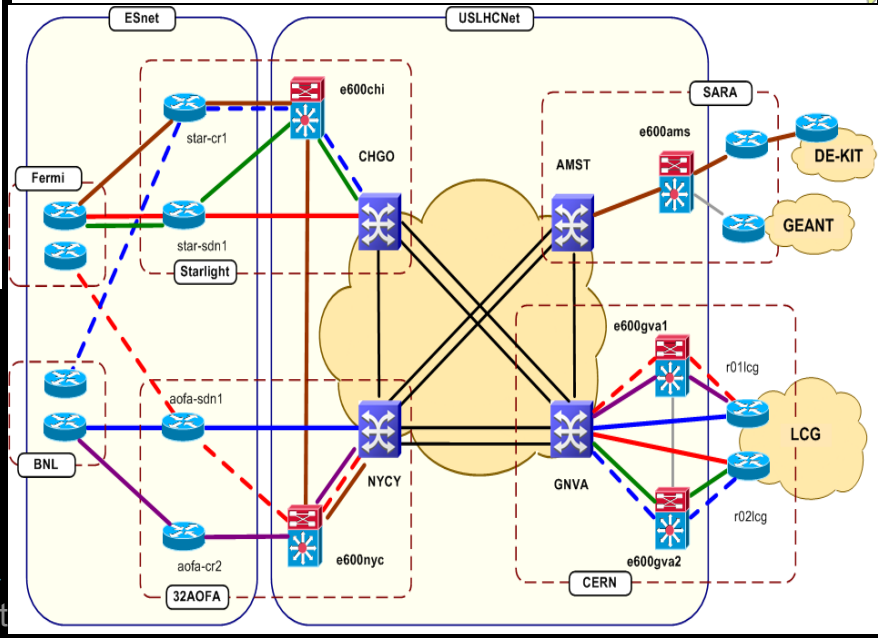
Tier-2 (~130 centres):
- Simulation
- End-user analysis

LHC Networking





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

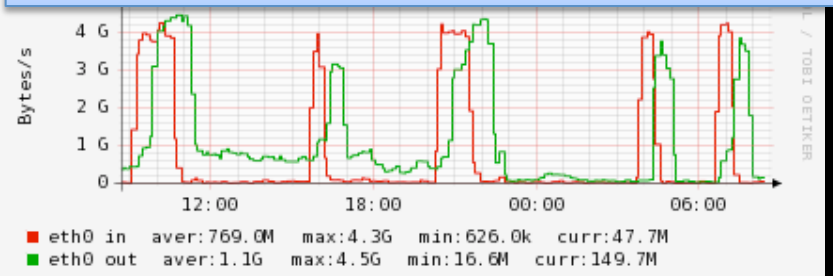


- Relies on
  - OPN, GEANT, US-LHCNet
  - NRENs & other national & international providers

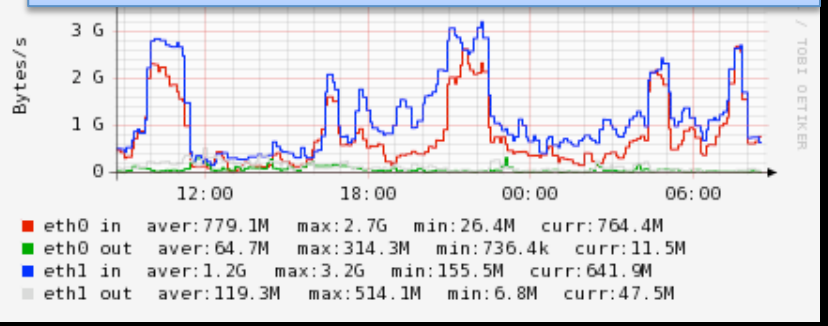
WLCG: Data Taking

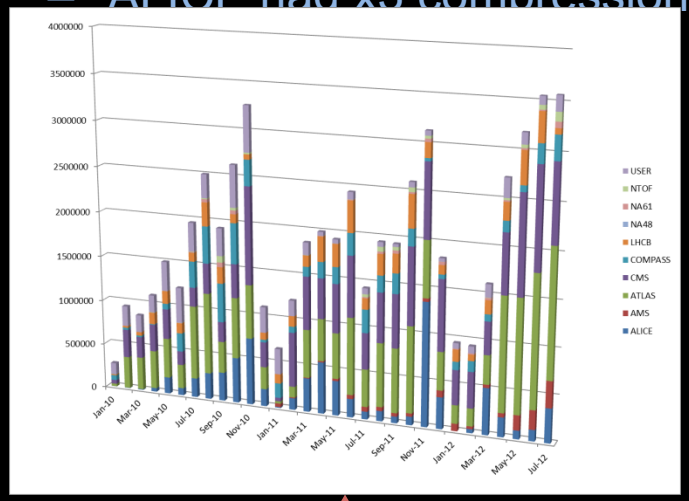
- Castor service at Tier 0 well adapted to the load:
  - Heavy Ions: more than 6 GB/s to tape (tests show that Castor can easily support >12 GB/s); Actual limit now is network from experiment to CC
  - Major improvements in tape efficiencies – tape writing at ~native drive speeds. Fewer drives needed
  - ALICE had x3 compression for

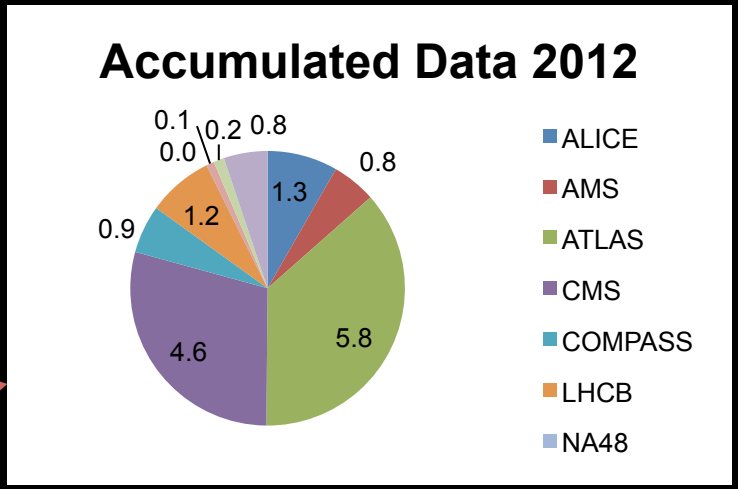
HI: ALICE data into Castor > 4 GB/s (r



HI: Overall rates to tape > 6 GB/s (r+

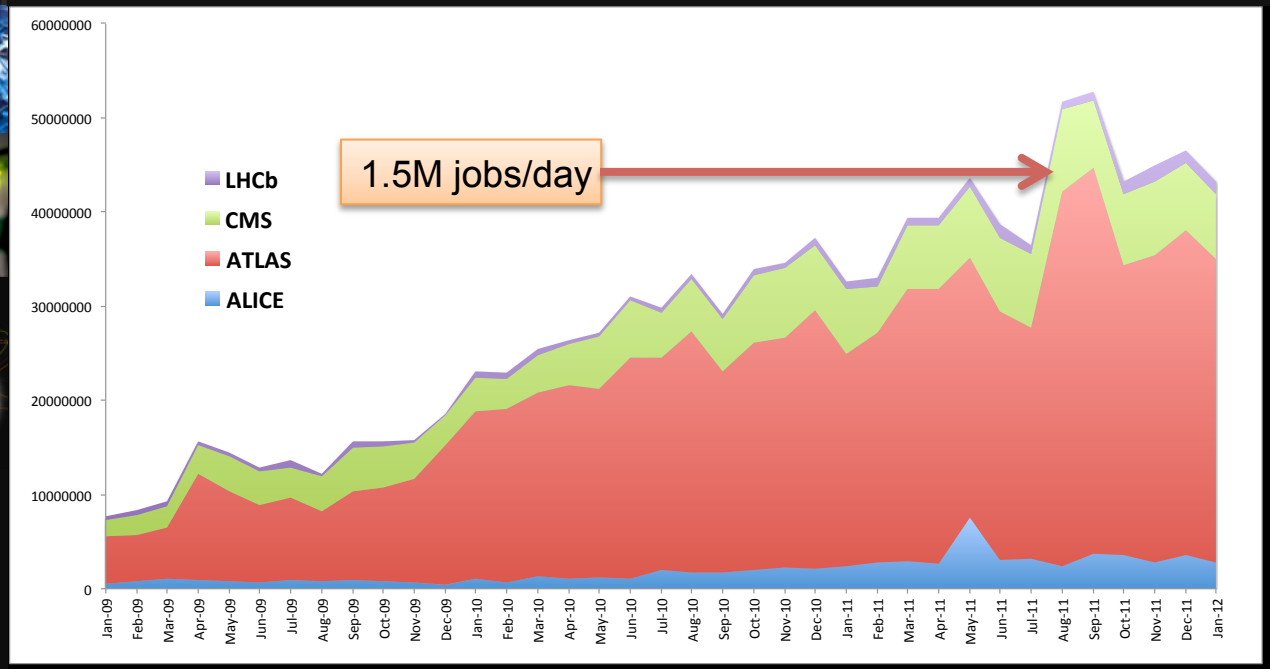






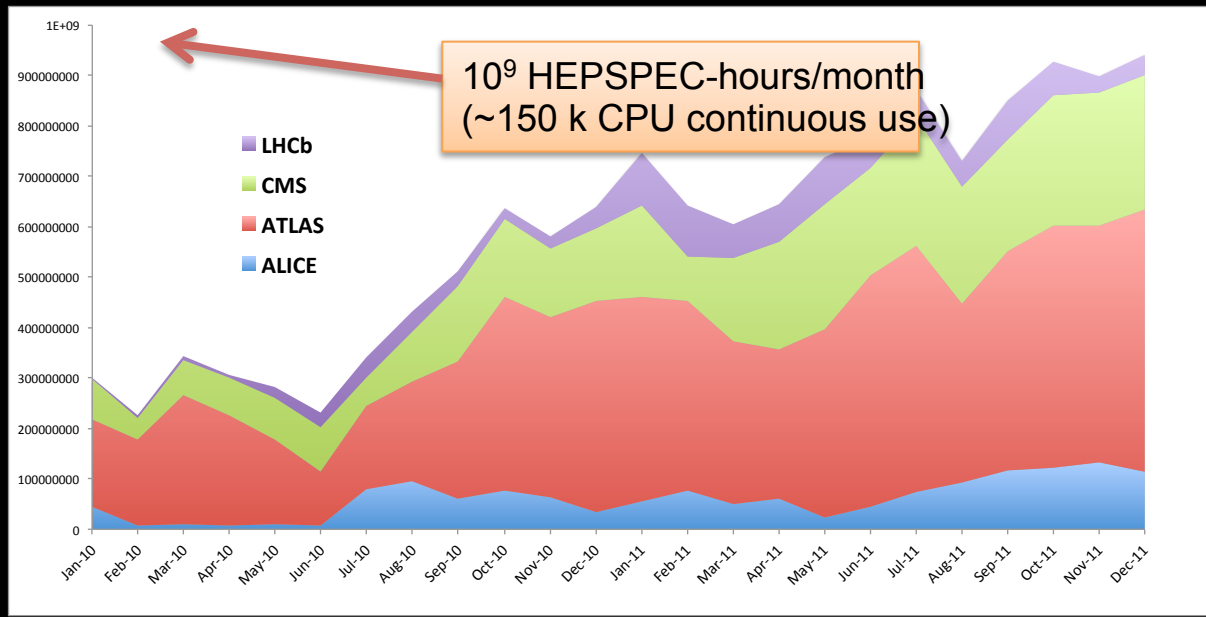
23 PB data written in 201
16 PB in 2012 !

Overall use of WLCG



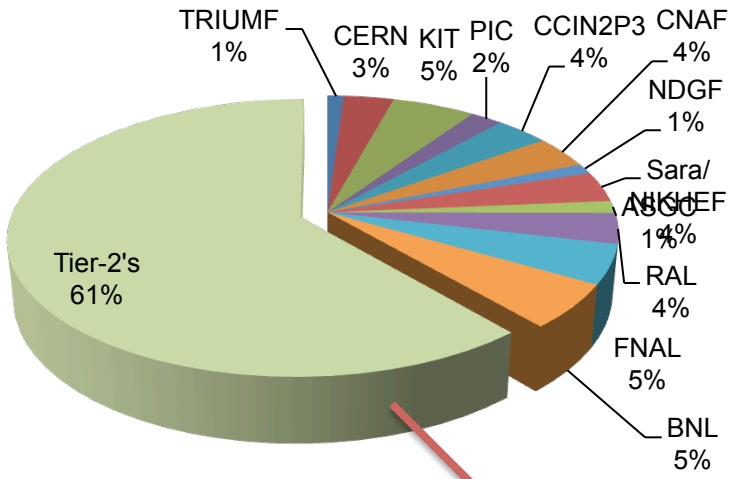
Usage continues to grow even over end of year technical stop

- # jobs/day
- CPU usage

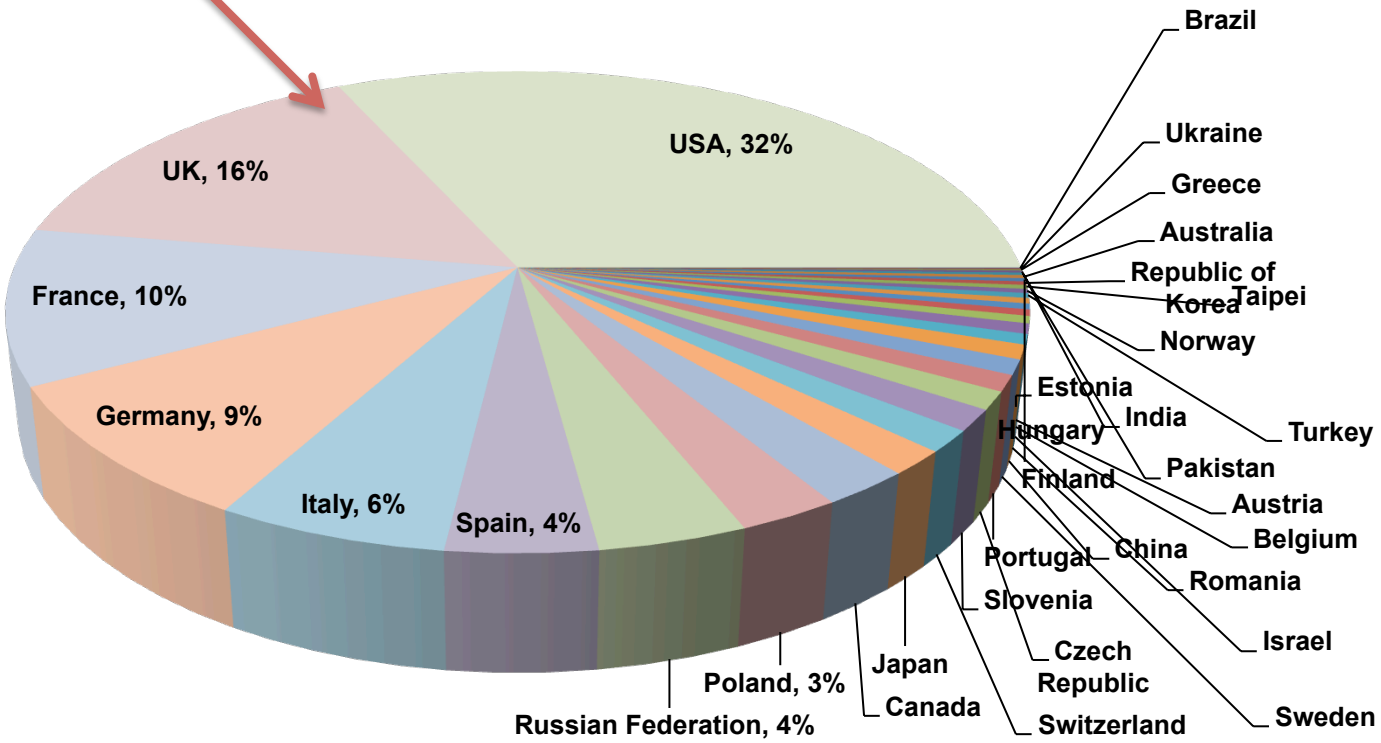


WLCG Worldwide LHC Computing Grid

CPU – 11.2010-10.2011

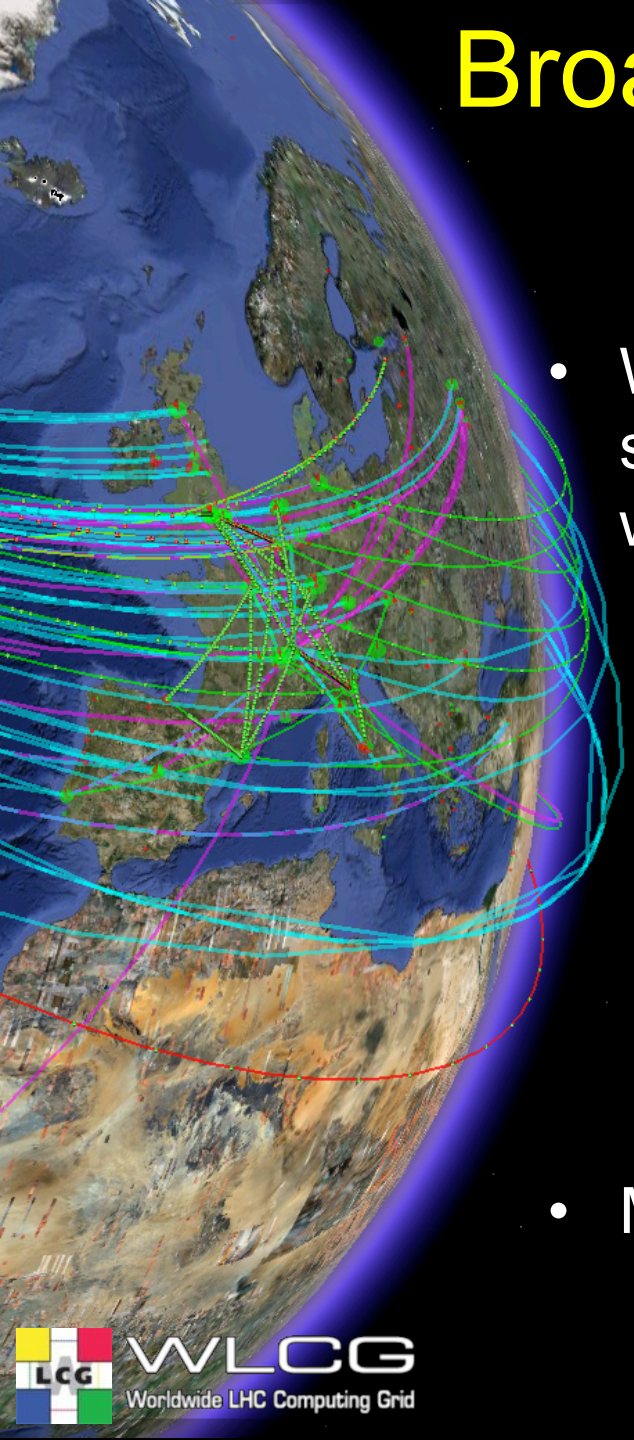


Significant use of Tier 2s for analysis



WLCG Worldwide LHC Computing Grid

Broader Impact of the LHC Computing Grid



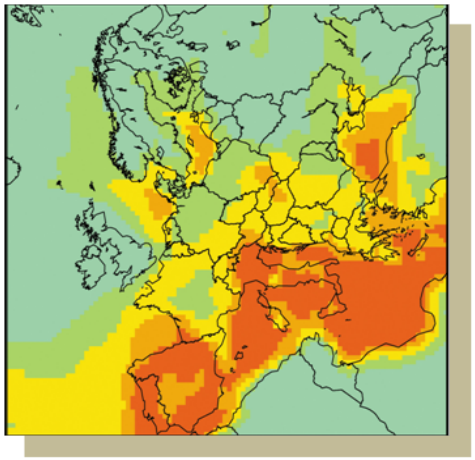
- WLCG has been leveraged on both sides of the Atlantic, to benefit the wider scientific community
  - Europe (EC FP7):
    - Enabling Grids for E-sciencE (EGEE) 2004-2010
    - European Grid Infrastructure (EGI) 2010--
  - USA (NSF):
    - Open Science Grid (OSG) 2006-2012 (+ extension?)
- Many scientific applications

Archeology
Astronomy
Astrophysics
Civil Protection
Comp. Chemistry
Earth Sciences
Finance
Fusion
Geophysics
High Energy Physics
Life Sciences
Multimedia
Material Sciences
…

WLCG Worldwide LHC Computing Grid

Earth Sciences

- Seismology
- Atmospheric modeling
- Meteorology
- Flood forecasting
- Pollution
- …



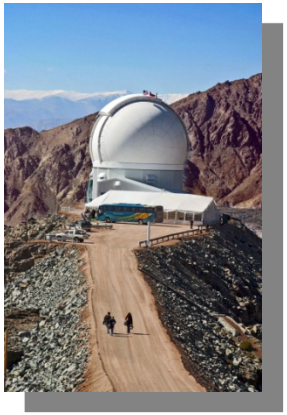
EGEODE the "Expanding GEosciences On-DEmand" Open Virtual Organization provides a suite of Grid computing services designed to meet the needs of Oil & Gas research

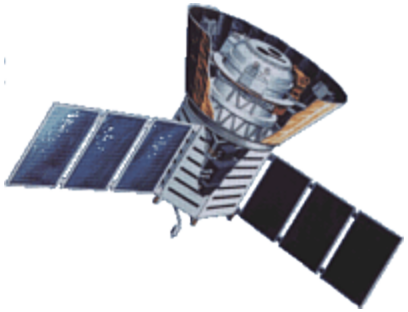
FOOTPRINT (EU project): identify the dominant pathways and sources of pesticide contamination in the agricultural landscape.

Astronomy & Astrophysics

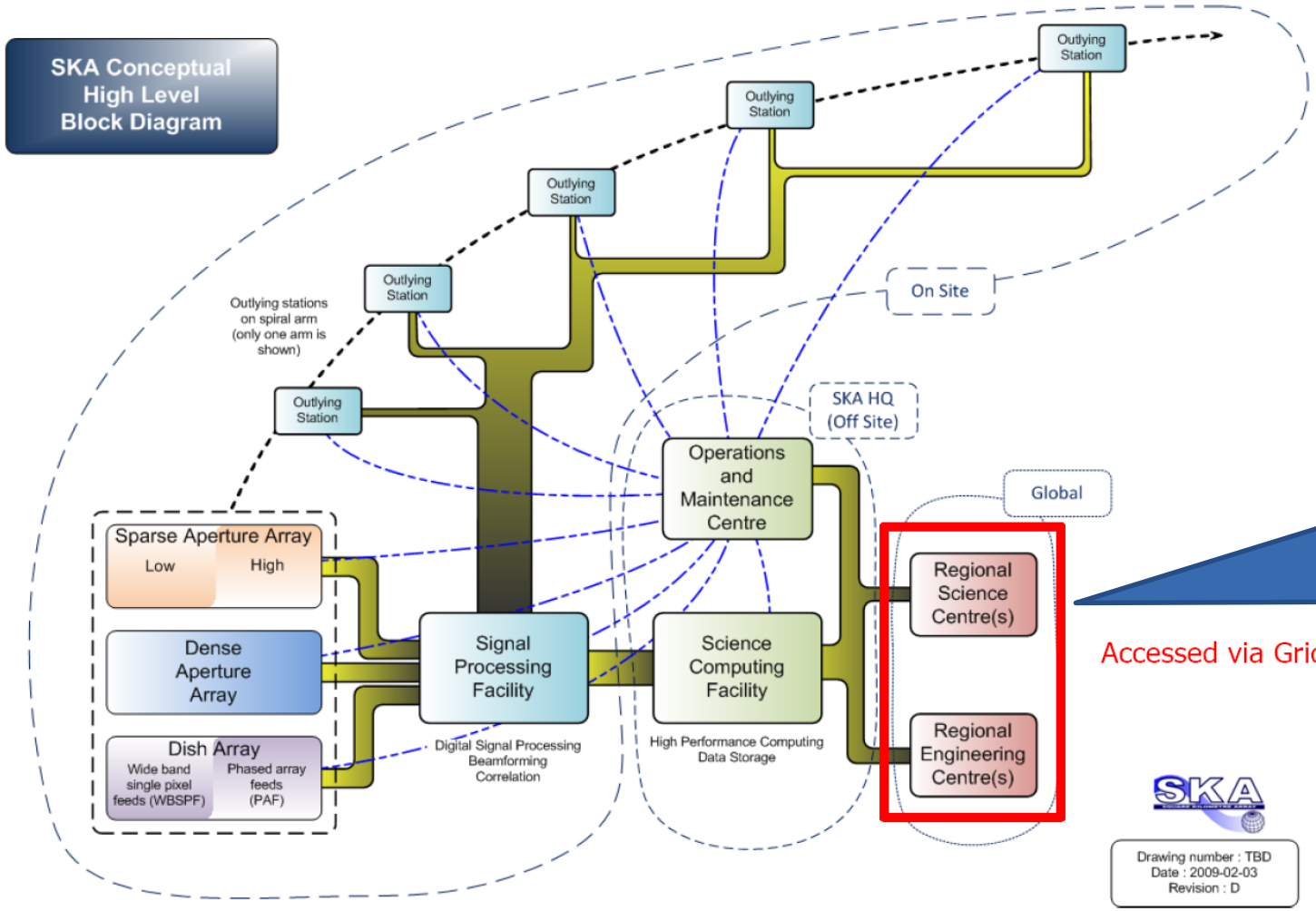
LOFAR large distributed radio telescope

AUGER & ARGO Cosmic Ray Observatories





Square Kilometer Array and grids

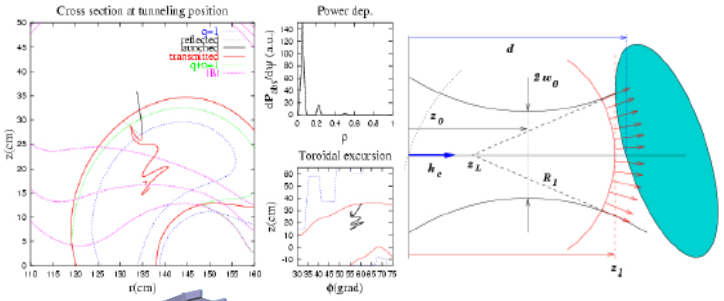


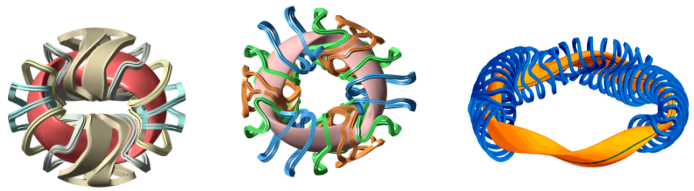
EGEE style grids most relevant here

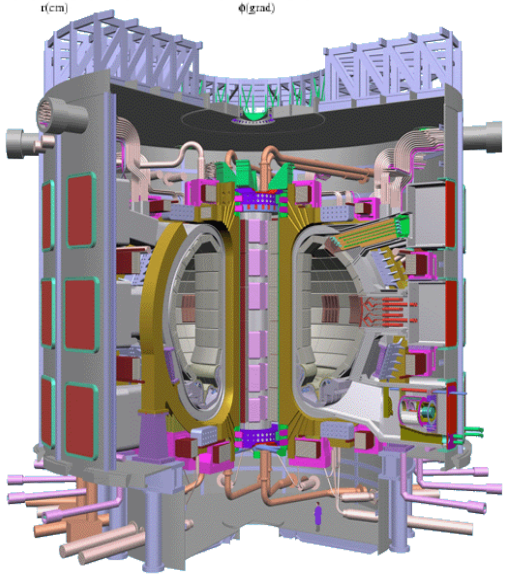
Fusion

Commercial exploitation of fusion energy still needs to solve several outstanding problems requiring exceptional computing facilities including supercomputers and cluster-based grids

- Ion Kinetic Transport
- Massive Ray Tracing
- Stellarator Optimization



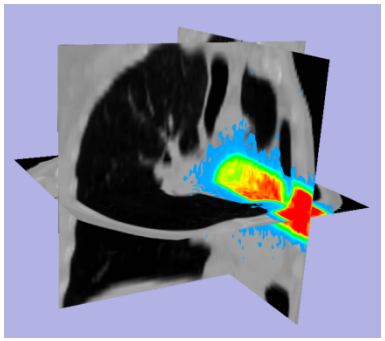




Life Sciences

Medical Imaging - ThIS (Therapeutic Irradiation Simulator)

- Monte-Carlo simulation of irradiations of living tissues with photons, protons or light ions beams for cancer therapy



Bioinformatics - Grid Protein Sequence Analysis

In silico drug discovery

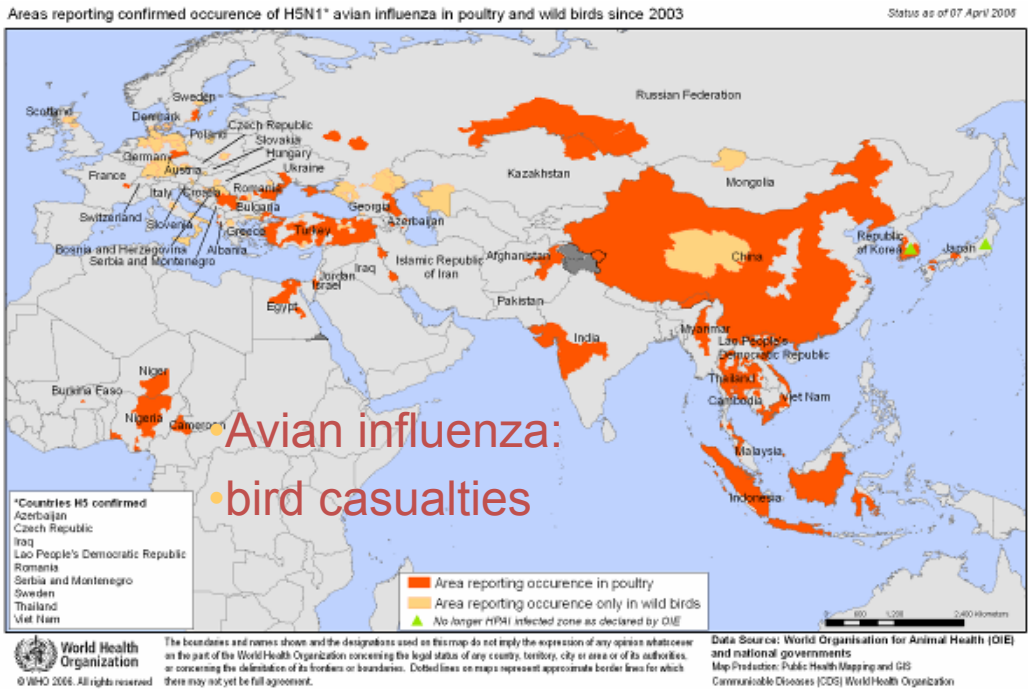
- Diseases such as HIV/AIDS, SRAS, Bird Flu, Malaria etc. are a threat to public health due to world wide exchanges and circulation of persons

- Grids open new perspectives to *in silico* drug discovery
  - Reduced cost and adding an accelerating factor in the search for new drugs

International collaboration is required for:

- Early detection
- Epidemiological watch
- Prevention
- Search for new drugs
- Search for vaccines

WISDOM
Initiative for grid-enabled drug discovery
against neglected and emergent diseases



Information Society Technologies

Similarity Search

Measurement of Pulmonary Trunk

Temporal Modelling

RV and LV Automatic Modelling

Visual Data Mining

Surgery Planning

Genetics Profiling

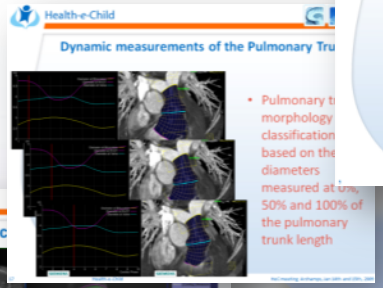
Personalised Simulation

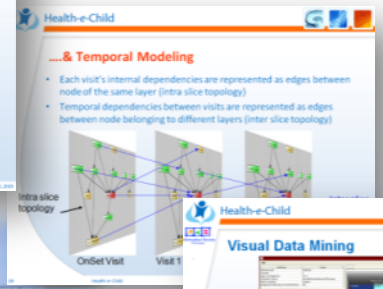
Treatment Response

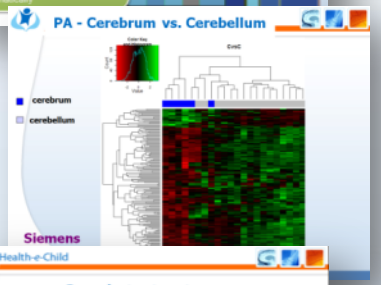
Semantic Browsing

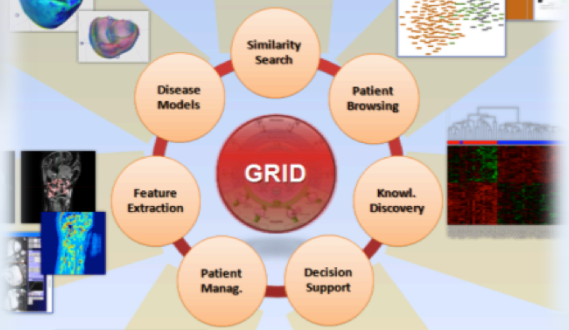
Inferring Outcome

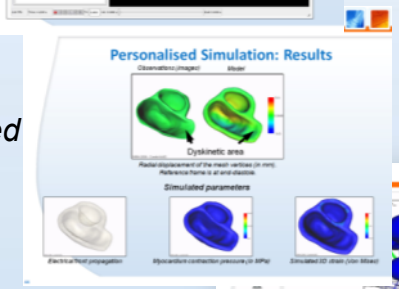
Biomechanical Models
Tumor Growth Modelling

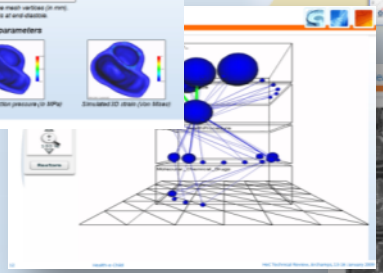


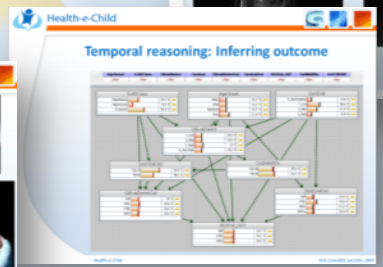


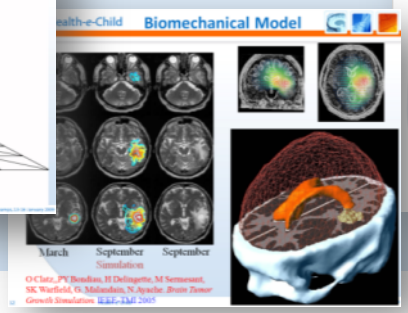












Contents

- Grids in Particle Physics
- Use of grids for other research communities
- **Grids and Business**
- Moving from grids to clouds
- Conclusions

EGEE – What do we deliver?

- **Infrastructure operation**
  - **Sites distributed across many countries**
    - **Large quantity of CPUs and storage**
    - **Continuous monitoring of grid services & automated site configuration/management**
    - **Support multiple Virtual Organisations from diverse research disciplines**

- **Middleware**
  - **Production quality middleware distributed under business friendly open source licence**
    - Implements a service-oriented architecture that virtualises resources
    - Adheres to recommendations on web service inter-operability and evolving towards emerging standards

gLite Lightweight Middleware for Grid Computing

- **User Support - *Managed process from first contact through to production usage***
  - **Training**
  - **Expertise in grid-enabling applications**
  - **Online helpdesk**
  - **Networking events (User Forum, Conferences etc.)**

GGUS Global Grid User Support

Sample of Business Applications

- **SMEs**
  - NICE (Italy) & GridWisetech (Poland): develop services on open source middleware for deployment on customer in-house IT infrastructure
  - OpenPlast project – (France) Develop and deploy Grid platform for plastics industry
  - Imense Ltd (UK) - Ported gLite application and GridPP sites
- **Energy**
  - TOTAL, UK - Ported application using GILDA testbed
  - CGGVeritas (France) – manages in-house IT infrastructures and sells services to petrochemical industry
- **Automotive**
  - DataMat (Italy) – Provides grid services to automotive industry

LATEST NEWS

CellProfiler on the Cloud
Secure Gene Sequencing Portal
Constellation Technologies
Model-based Analysis of Tiling-arrays for ChIP-Chip

POPULAR

About Constellation Technologies
Constellation Technologies
Model-based Analysis of Tiling-arrays for ChIP-Chip
Constellation Technologies and Cambridge University Explore the Cloud Together
Middleware Technology

Technology

Middleware Technology

Middleware is a "glue" software that pools together various computing resources to create the Grid. It allows users to securely access the integrated computing and storage resources of in a way similar to accessing an enourmous virtual computer. EGEE develops and deploys a middleware distribution called gLite (pronounced "gee-lite"). gLite is a result of collaborative efforts of more than 80 people in 12 different academic and industrial research centers as part of the EGEE Project. gLite provides a framework for building grid applications tapping into the power of distributed computing and storage resources across the Internet.

The gLite middleware stack provides the user both with foundation level and higher level services. Foundation level services ensure security, resource access and systems to monitor grid activity. These provide the basis for a consistent and dependable production infrastructure. Higher level services provided by gLite include job management, data catalogues and data replication, providing applications with the tools to build end-to-end solutions. Other third party projects complete a rich ecosystem built on the gLite foundation services.

Constellation Technologies will be developing cloud computing solutions for the next generation of Internet based on gLite.

Contents

- Grids in Particle Physics
- Use of grids for other research communities
- Grids and Business
- **Moving from grids to clouds**
- Conclusions

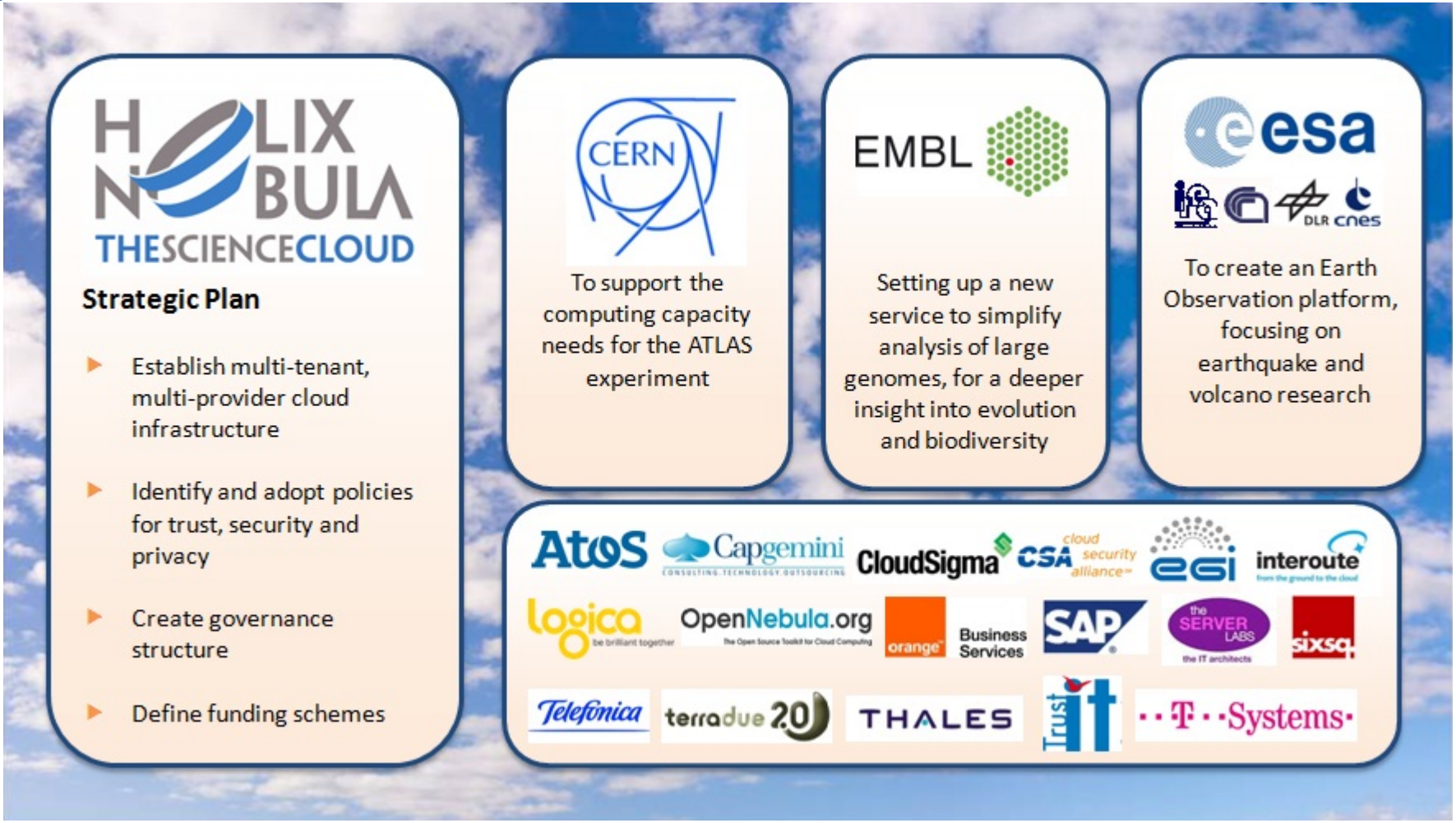
How to evolve WLCG?

A distributed computing infrastructure to provide the production and analysis environments for the LHC experiments

- **Collaboration** - *The resources are distributed and provided "in-kind"*
- **Service** - *Managed and operated by a worldwide collaboration between the experiments and the participating computer centres*
- **Implementation** - *Today general grid technology with high-energy physics specific higher-level services*

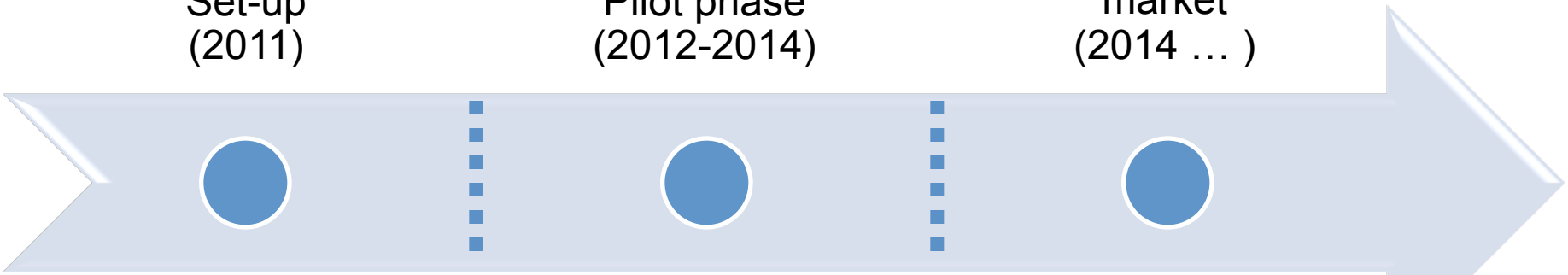
Evolve the **Implementation** while preserving the **collaboration** & **service**

A European Cloud Computing Partnership big science teams up with big business



Email:contact@helix-nebula.eu Twitter: HelixNebulaSC Website: http://www.helix-nebula.eu/

Timeline



- Agree strategic plan
- Select flagships use cases
- Identify service providers
- Define governance model

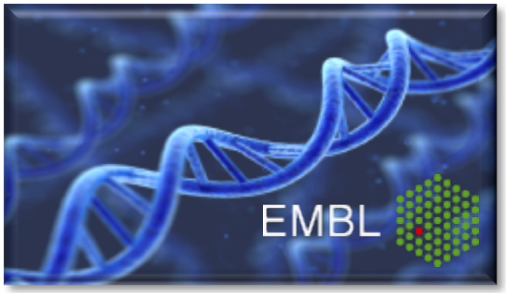
- Deploy flagships
- Analysis of functionality, performance & financial model
- Success Stories

- More applications
- More services
- More users
- More service providers

under grant 312301 with 1.8M

Initial flagship use cases

HELIX NEBULA THESCIENCECLOUD

ATLAS High Energy Physics Cloud Use	Genomic Assembly in the Cloud	SuperSites Exploitation Platform
		
To support the computing capacity needs for the ATLAS experiment	A new service to simplify large scale genome analysis; for a deeper insight into evolution and biodiversity	To create an Earth Observation platform, focusing on earthquake and volcano research

- **Scientific challenges with societal impact**
- **Sponsored by user organisations**
- ***Stretch* what is possible with the cloud today**

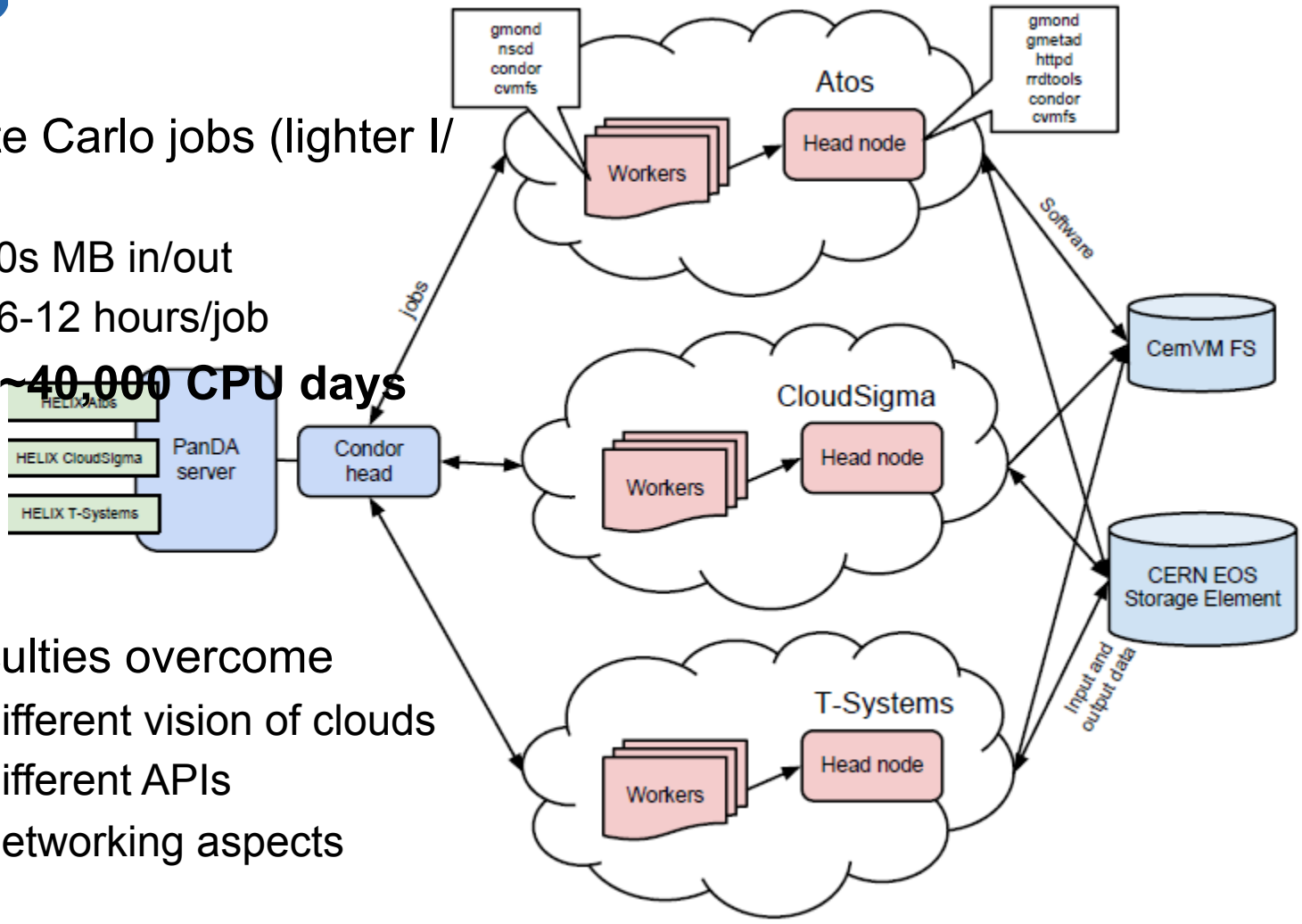
Flagship deployments
First results

- Proof of Concept deployments within the Pilot Phase started January 2012
- Each flagship has been deployed with a series of providers independently

CERN, EMBL and ESA succeeded in eploying scientific applications each involving tens of thousands of jobs running at data entres operated by Atos, CloudSigma and T-Systems

CERN-ATLAS flagship configuration

- Monte Carlo jobs (lighter I/O)
  - 10s MB in/out
  - ~6-12 hours/job
- **Ran ~40,000 CPU days**
- Difficulties overcome
  - Different vision of clouds
  - Different APIs
  - Networking aspects

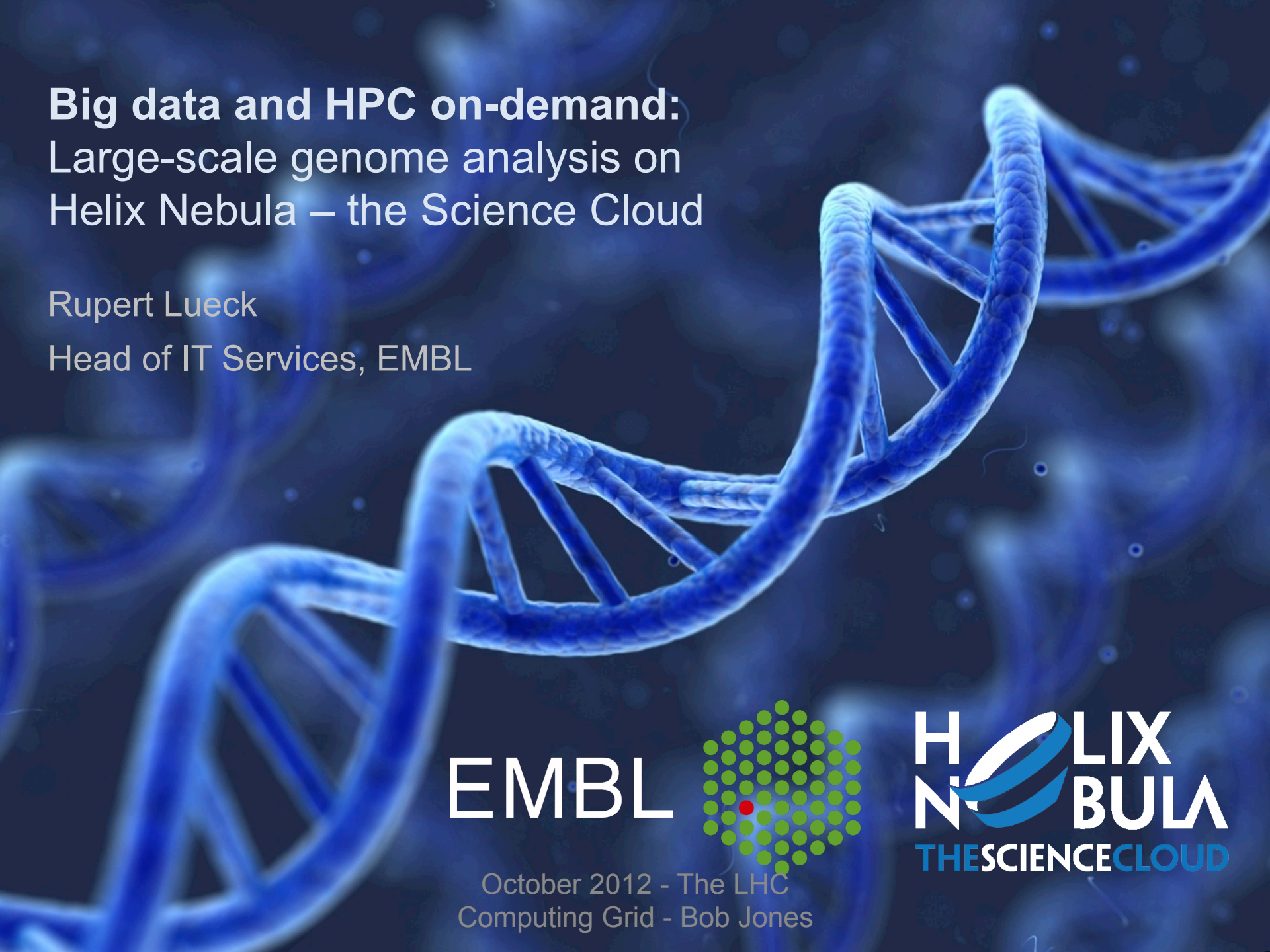


Ramón Medrano Llamas,Fernando Barreiro, Dan van der Ster (CERN IT), Rodney Walker (LMU Munich)

Big data and HPC on-demand: Large-scale genome analysis on Helix Nebula – the Science Cloud

Rupert Lueck

Head of IT Services, EMBL

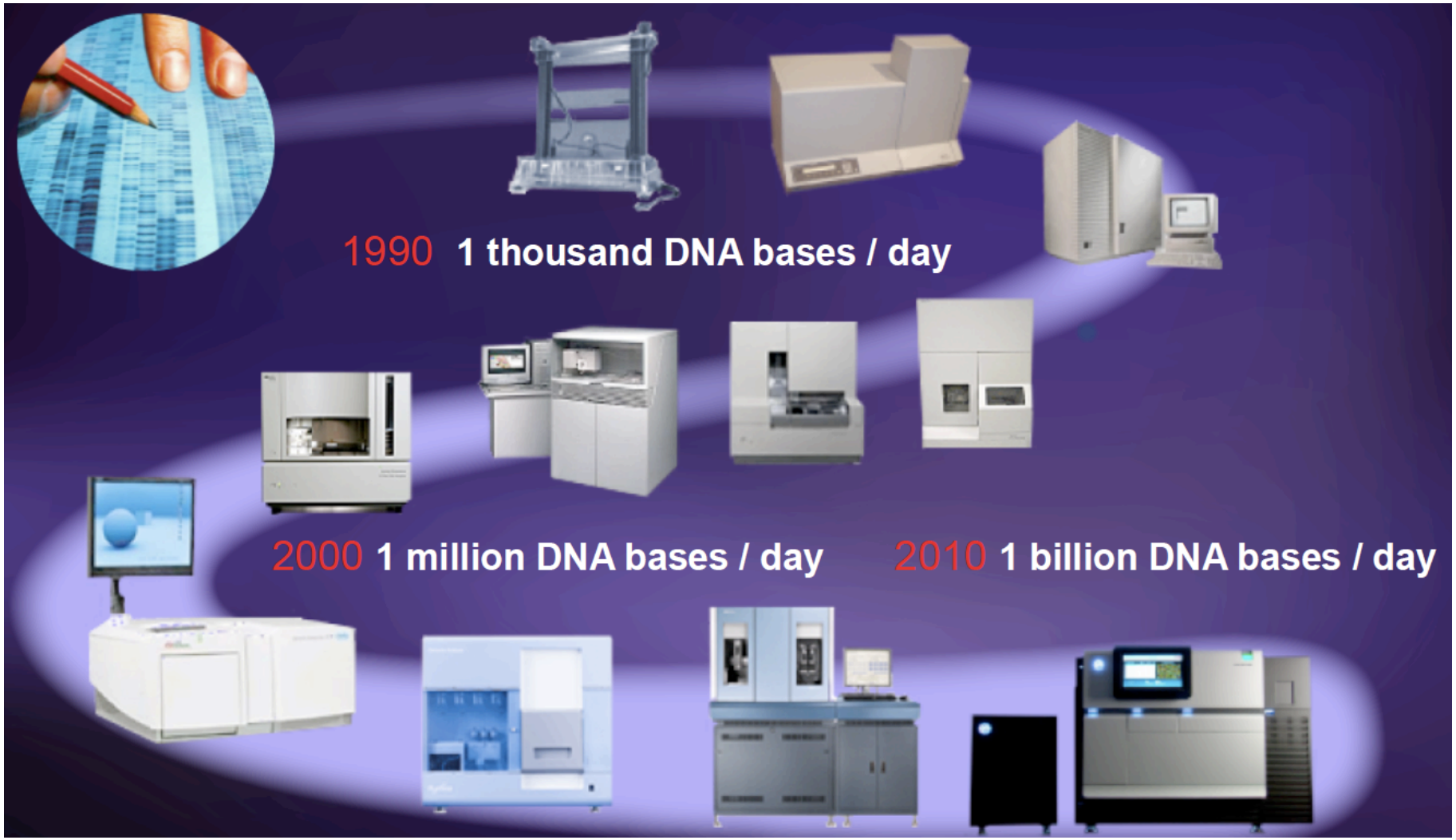


EMBL

HELIX NEBULA THE SCIENCE CLOUD

October 2012 - The LHC Computing Grid - Bob Jones

Next Generation Sequencing (NGS) Revolution



EMBL

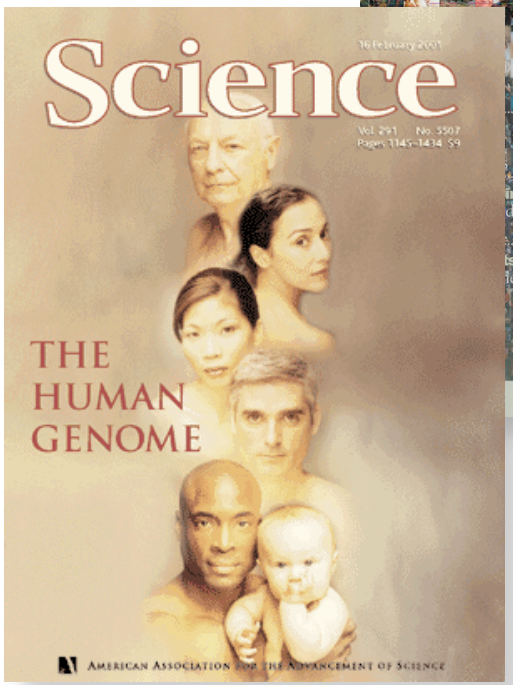
NGS Impact on Human Genome Sequencing

- Human genome project
  - 10 years
  - Large International Consortium
  - Thousands of Sequencers
  - $3,000,000,000

- Sequencing today
  - < $10,000
  - A few hours
  - One machine

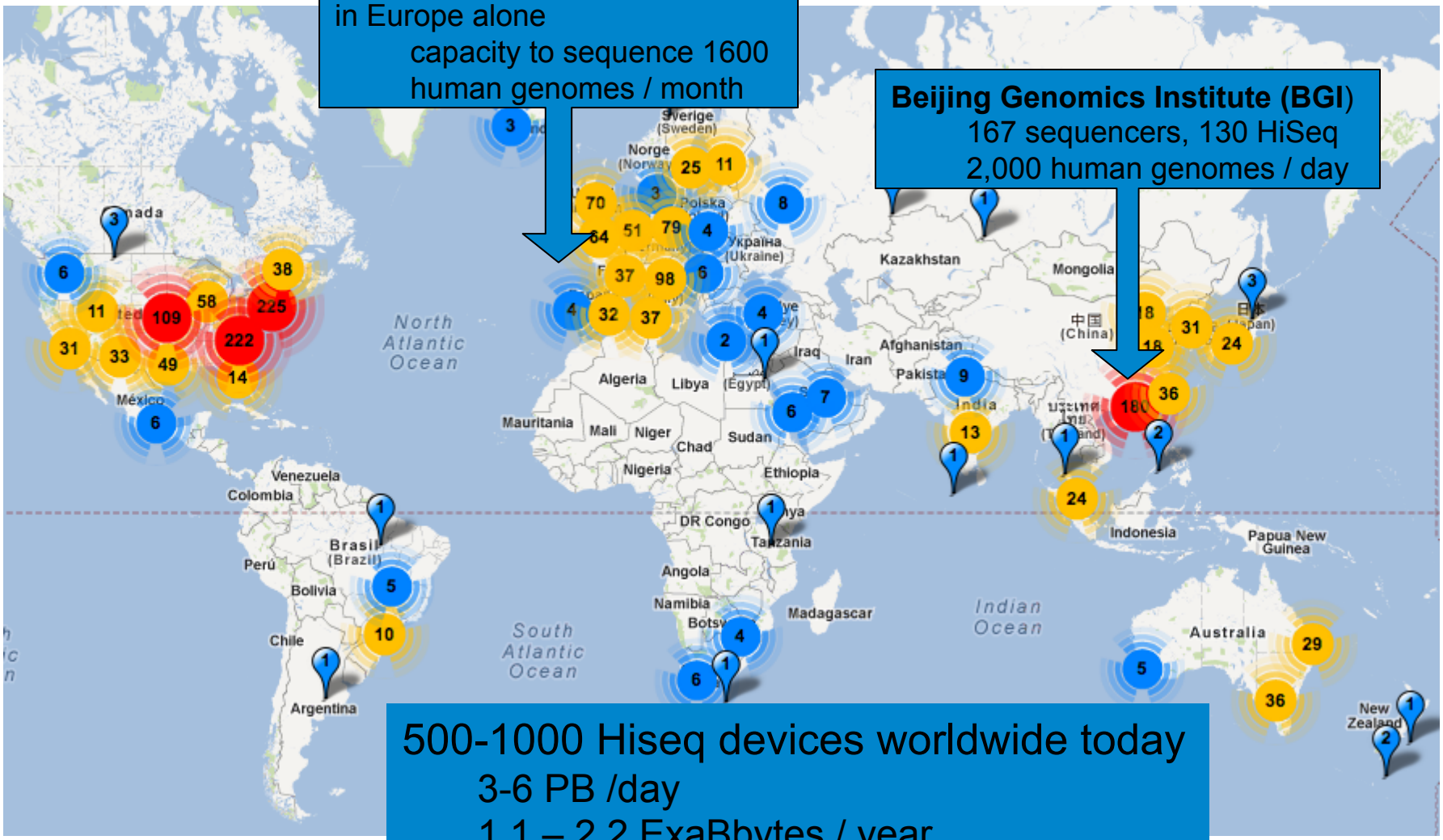
2000

2010



EMBL

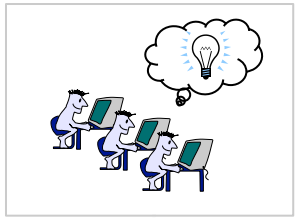
World Map of High-throughput Sequencers



EMBL

EMBL Flagship project: Whole-Genome Assembly

Access



Scientists

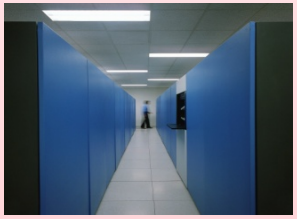
Data acquisition

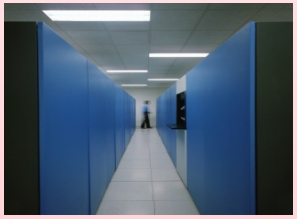
NGS Labs

Cloud Storage

HELIX NEBULA
THESCIENCECLOUD

On-demand processing





Cloud Service

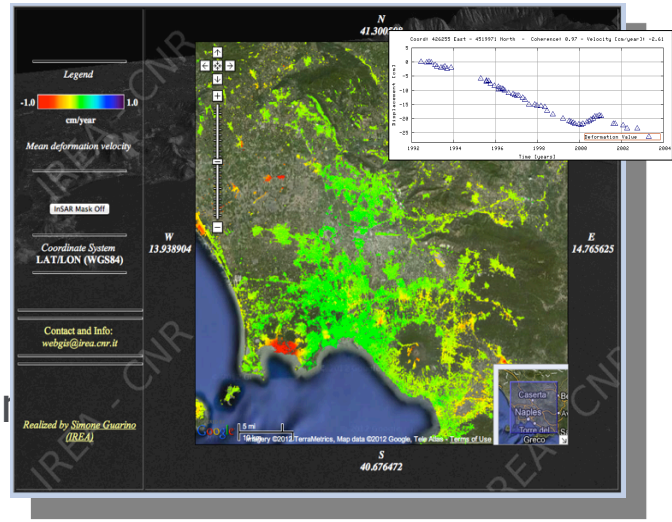
Integration with other cloud services / Archiving

ESA's experience with Helix Nebula and outlook

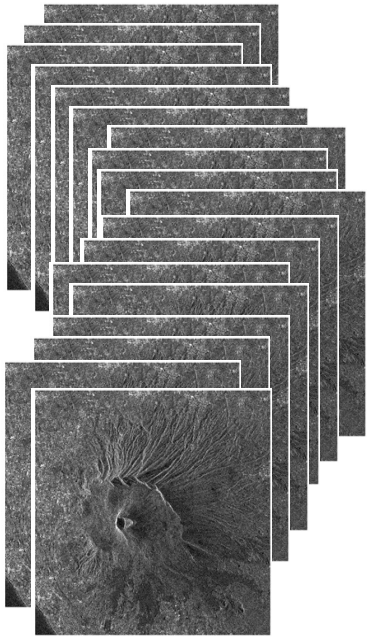
Wolfgang Lengert,
ERS and ADM-Aeolus Mission Manager

Earth Observation Application Platform exploiting 20 years of satellite data

- EO Application Platform
  - OpenNebula
  - Data Catalogue and Access
  - Map-Reduce computing model
  - Software repository
  - Utilities for sw development and testi
- Cloudification of application
  - CNR / IREA (Italian Research Council in Naples) developed an application (SBAS) measuring the vertical movement of ground in sub cm from space.
  - **SBAS** targets
    - Time series over **20 years** with ESA archive
    - Points of Interest are at **world scale**
    - **TBytes** of data to process



CNR SBAS Processing

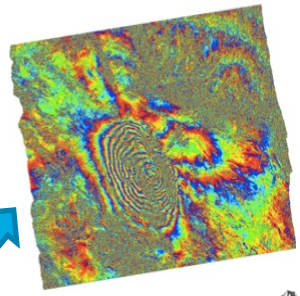


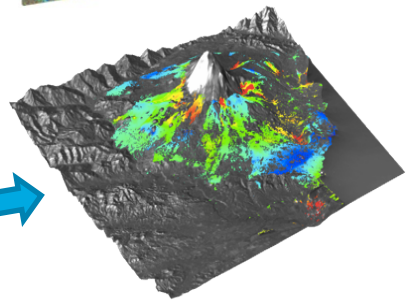
~150 Satellite images: **1.5TB**

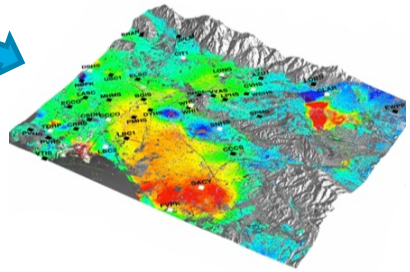
HELIX NEBULA THESCIENCECLOUD

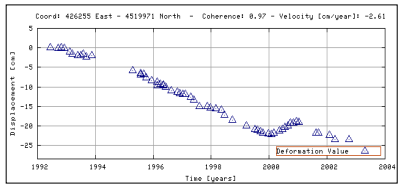
CNR

Time Processing: **150h**









Earthquakes

Volcanoes

Oil & Gas

Water Resources

Supersite Exploitation Platform: potential actors benefits

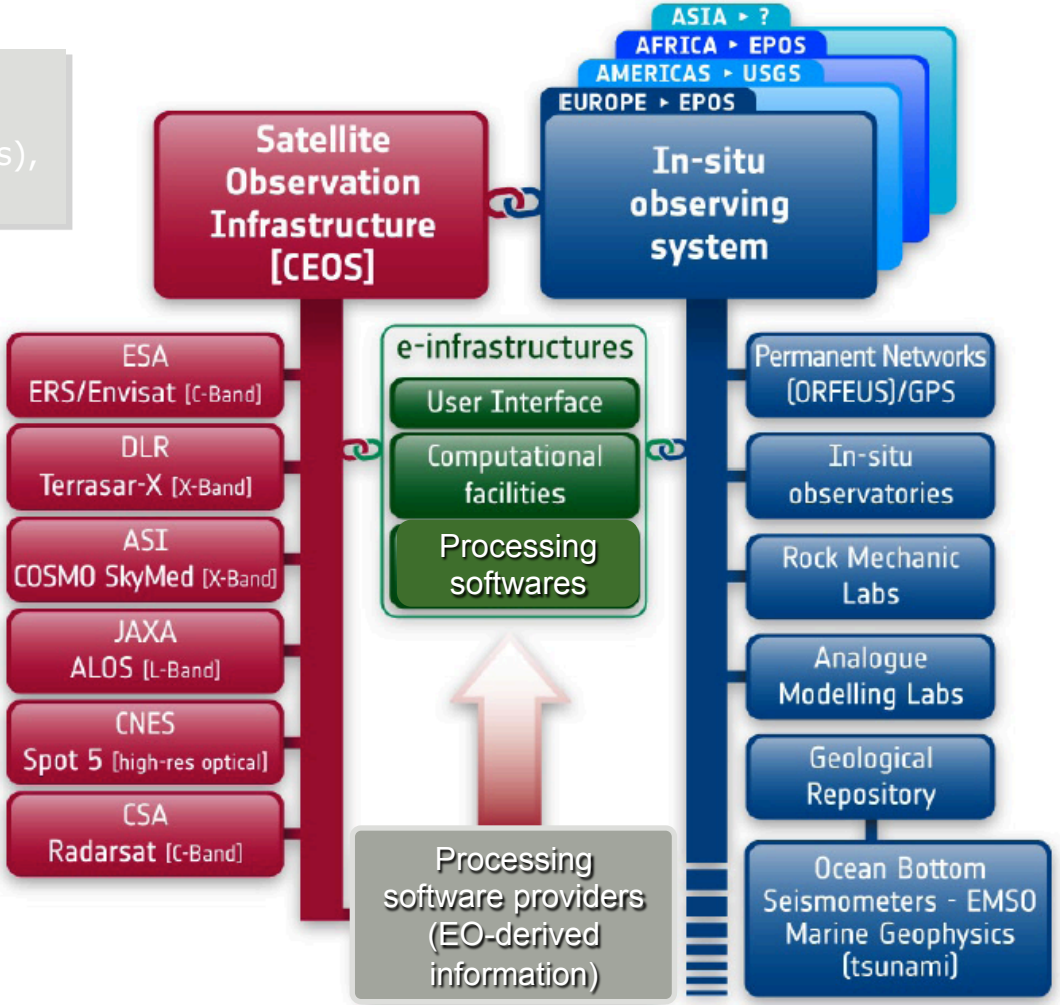
EO data provider benefits:
- Enlarge EO data exploitation (space agencies)
- Increase EO data sales (commercial distributors), in particular EO data archives

End-user benefits:
- More data, either free or at low cost
- Processing capabilities free or at low cost
- Processing softwares free or at low cost
- Forum for discussing/exchanging results

→ More science

Processing software provider benefits:
- Low investment
- Increase sales
- Increase software visibility

IT companies (computational facilities) benefits:
- New business
- Access to a global user community
- Contribution to science



Conclusions

- The Physics community took the concept of a grid and turned into a global production quality service aggregating massive resources to meet the needs of the LHC collaborations

- The results of this development serve a wide range of research communities;
  have helped industry understand how it can use distributed computing;
  have launched a number of start-up companies and
  provided the IT service industry with new tools to support their customers

- Open source licenses encourage the uptake of the technology by other research communities and industry while ensuring the Physics community contribution is acknowledged

- Providing access to computing infrastructures by industry and research communities for prototyping purposes reduces the investment and risk for the adoption of new technologies

Conclusions

- Many research communities and business sectors are now facing an unprecedented ***data deluge***. The Physics community with its LHC programme has unique experience in handling data at this scale

- Cloud computing uses many of the advances pioneered by grid computing and makes it accessible to business sectors, governments and the general public

- The on-going work to evolve the LHC computing infrastructure to make use of cloud computing technology can serve as an excellent test ground for the adoption of cloud computing in many research communities, business sectors and government agencies

- The Physics community is exploring how commercial cloud services can serve the research infrastructures of the future and provide new markets for European industry