1st Synergy LNF-OAR Workshop

LHC CM From GRID to Cloud: the ATLAS Italian cloud and LNF Tier-2

Elisabetta Vilucchi
INFN-LNF 16/4/2014

Outline

- The LHC experince: the Grid
- LHC CM evolution
  - Network upgrades
  - Next LHC CM evolution: from Grid to Cloud
- The ATLAS Italian computing infrastructure
- The ATLAS LNF Tier-2

LHC: Large Hadron Collider

Many, many data

~15 PB/year now

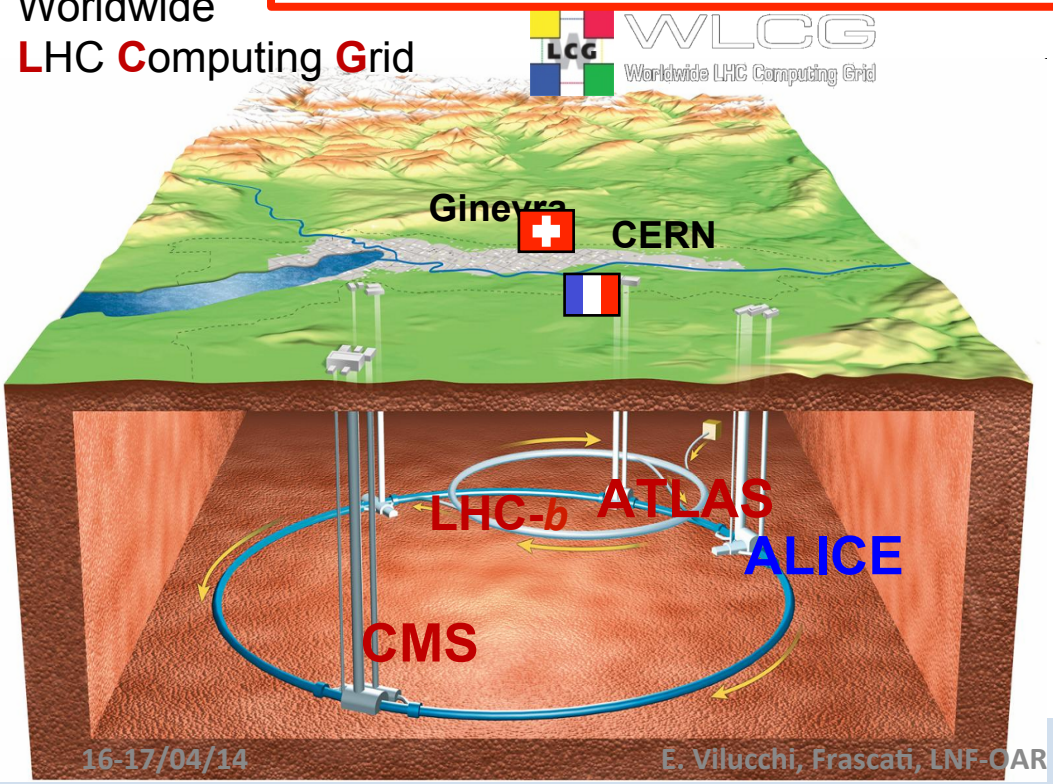
(~130 PB/year in 2021)

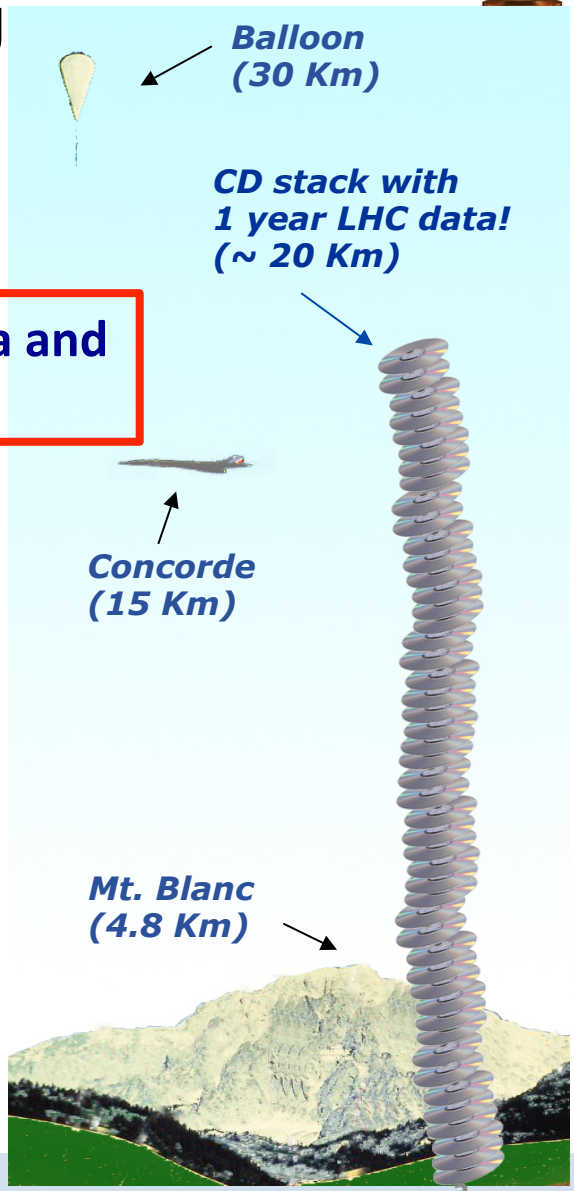
1 Petabyte (1PB) = 1000 TB *= 10 times the text content of the World Wide Web*

Very large computing power distributed worldwide

The GRID to manage this huge amount of data and distributed computing power

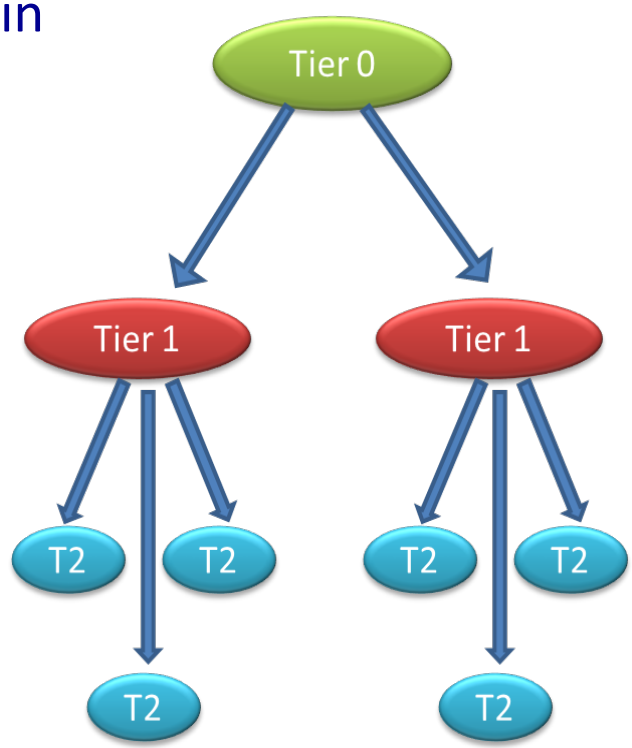
Worldwide
LHC Computing Grid





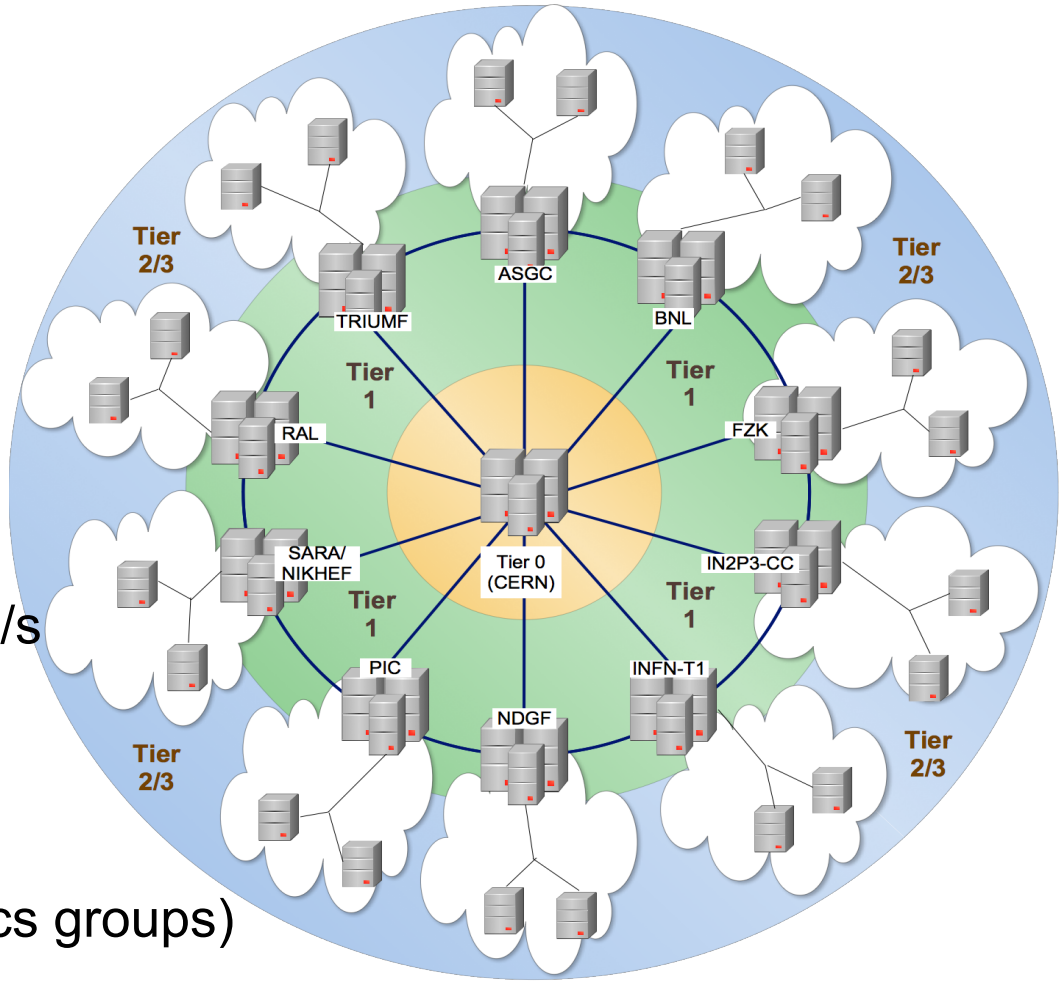
The first LHC Computing Model (CM): MONARC

- Static Hierarchical Computing Model based on a multi-tier distributed architecture deployed within the Grid paradigm: LHC Computing Grid
- One Tier-0 (Cern), one Tier-1 per country, some Tier-2s per associated Tier-1, many Tier-3s
- Any site (Tier) with its specific tasks, where network costs favour regional data access (performing network connections only between Tier-0 and Tier-1s)
- Due to network limitations, static data pre-placement,
  - Data from Tier-0 to Tier-1s and from Tier-1s to their Tier-2s: jobs-go-to-data



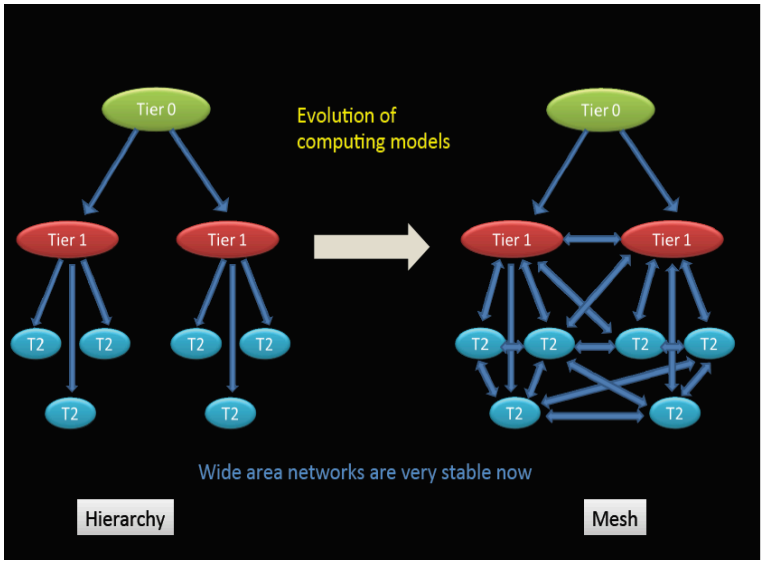
The multi-tier architecture

- More than 200 centres in ~40 countries
- Tier-0 at CERN
  - Data recording
  - Initial data reconstruction
  - Data distribution
- 11 Tier-1s
  - Permanent storage
  - Re-processing
  - Simulation
  - Analysis
  - Connected by direct 10 Gb/s network links
- ~200 Tier-2s
  - Simulation
  - Analysis (end-user, physiscs groups)



Network evolution brings to CMs evolution

- Network is as important as site infrastructure
  - key point to optimize storage usage and jobs brokering to sites
- At the beginning network was the bottleneck. The hierarchical model was based on the assumption of a rather limited connectivity between computing centres.
  - Only links between well connected sites (Tier-0 and Tier-1s) were dedicated to cover fundamental roles.
- Network capacity improved very fast
- WAN is very stable and performance is good
  - It allows to relax MONARC model:
  - migration from hierarchy to full mesh model: sites are all directly interconnected and independent of the Tier1s
- Data management based on popularity concept
  - Dynamic storage usage
  - Reduction of data replicas. Only data really needed is sent (and cached)
- Network awareness
  - Workload management systems and data transfers will use networking status/performance metrics to send jobs/data to sites



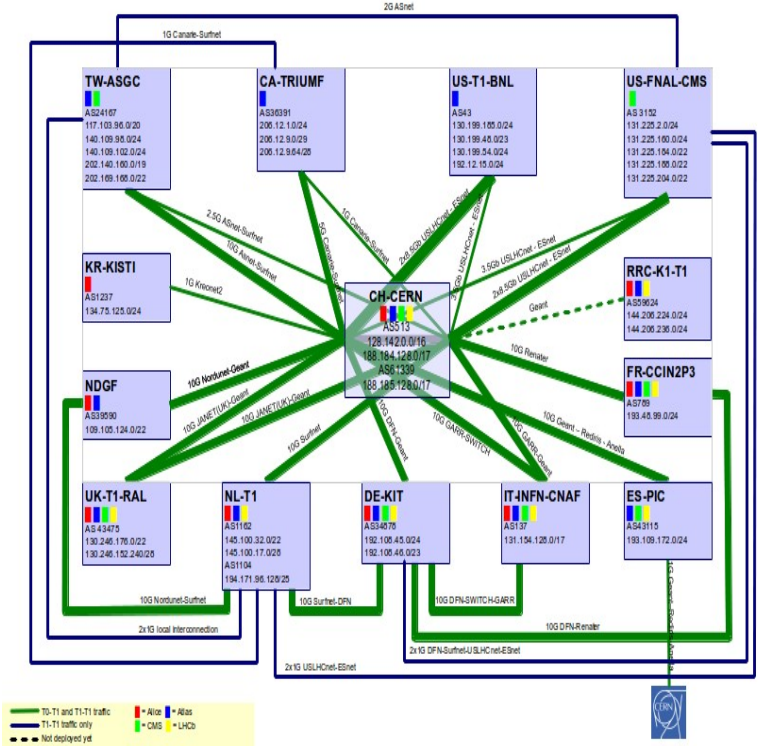
It's cheaper to transport data than to store it

The LHC backbone

LHCOPN

LHCOPN collaboration with CERN & GEANT

- Private, closed, infrastructure with primary purpose to provide **dedicated** connectivity between Tier-0 and Tier-1s
- Tier-1s connected directly to CERN
  - 18 10G Links
  - 2 x 100G links between CERN and Wigner data center (Budapest)
  - Capacity used also for Tier1-Tier1 traffic, via CERN or additional links, and transit to Tier0 via other Tier1s

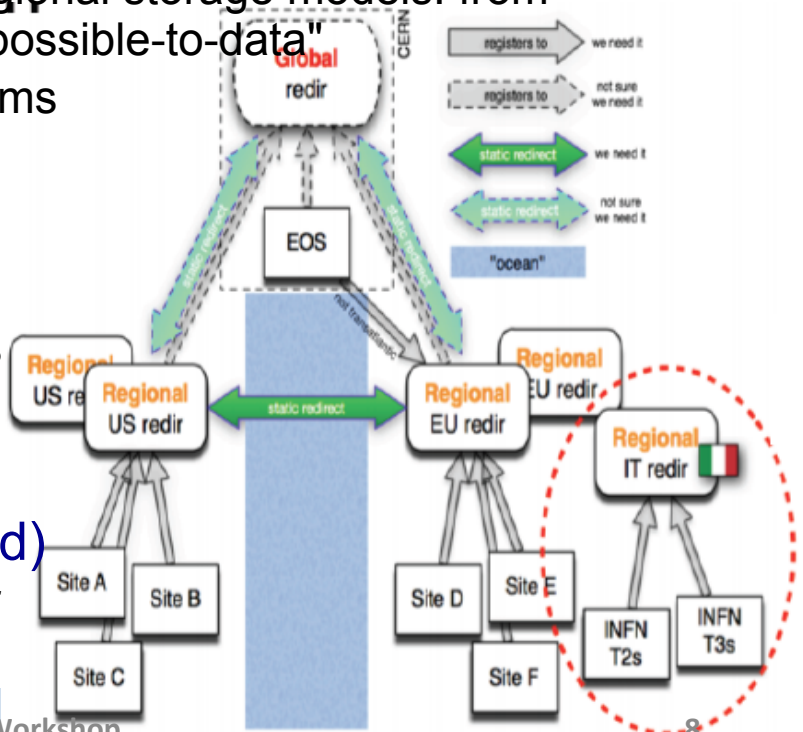


LHCONE an initiative for Tier-2 Network

- Network providers, jointly working with the experiments, have proposed a new network model for supporting the LHC experiments, known as the LHC Open Network Environment (LHCONE)
- LHCONE is a reserved network to LHC Tier1/2/3 sites, the goal is to provide some guarantees performance protecting the infrastructures against potential “threats” of very large data flows
- LHCONE will complement LHCOPN.

How to exploit such a performing network

- Remote access via WAN is a reality: Storage Federation
- The LHC experiments are deploying federated storage infrastructures based on xrootd or http protocols
  - Provide new access modes & redundancy
    - Jobs access data on shared storage resources via WAN
    - Relaxes CPU-data locality opening up regional storage models: from "jobs-go-to-data" to "jobs-go-as-close-as-possible-to-data"
    - Failover capability for local storage problems
    - Dynamic data caching based on access
- A data solution for computing sites without storage: opportunistic, cloud, Tier3
  - Disk can be concentrated only in large sites
  - Reduction of operational load at sites
  - Lower disk storage demands
- An example: *FAX (*Federated ATLAS Xrootd)
  - a way to unify direct access to a diversity of storage services used by ATLAS



CM evolution: from Grid to Cloud

- The present LHC CM implementation is the Grid
- But the transition from Grid to Cloud (and virtualization) is also leading to an important infrastructural change and a new way for the procurement of the resources
- General-purpose infrastructure, used to give people uniformity and transparency in using the resources
  - Same aim of the Grid, but much easier: access to the cloud is less complex
- Cloud resources: free opportunistic cloud resources
  - HLT farms accessible through cloud interface during shutdowns or LHC inter-fill
  - Academic facilities
  - Commercial cloud infrastructure: Amazon EC2, Google.

LHC CM implementation: the Grid

- Grid gives selected user communities uniform access to distributed resources with independent administrations
  - Computing, data storage, devices, …
- **Ian Foster** Grid Checklist (*"What is the Grid? A Three Point Checklist"*, 2002)
  - *Grid is a system that*: *coordinates resources that are not subject to centralized control ...*
  - *... using standard, open, general-purpose protocols and interfaces ...*
    - A Grid is built from multi-purpose protocols and interfaces that address such fundamental issues as authentication, authorization, resource discovery, and resource access.
  - *... to deliver nontrivial qualities of service*.
    - A Grid allows its constituent resources to be used in a coordinated fashion to deliver various qualities of service, relating for example to response time, throughput, availability, security,…

Many communities using the Grid

- Grid computing paradigm was adopted by many communities:
  - High-energy physics
  - Astrophysics
  - Fusion
  - Computational chemistry
  - Biomed – biological and medical research
  - Earth sciences
  - UNOSAT – satellite image analysis for the UN
  - Digital libraries
  - E-learning
  - Industrial partners in EGEE

I. Foster Cloud definition

- **Ian Foster, Yong Zhao, Ioan Raicu, Shiyong Lu**
  "Cloud Computing and Grid Computing 360-Degree Compared", 2008:
  - *A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet.*
- Cloud Computing is a specialized distributed computing paradigm that differs from traditional ones in that:
  - it is massively scalable,
  - can be encapsulated as an abstract entity that delivers different levels of services to customers outside the Cloud,
  - it is driven by economies of scale,
  - the services can be dynamically configured (via virtualization or other approaches) and delivered on demand.

Grid vs Cloud 1

- Business model:
  - Cloud: a customer will pay the provider on a consumption basis (such as electricity)
  - Grids: project-oriented, the VO provide to the users or the community a share of computing and storage resources they can access.
- Resources:
  - Grids: heterogeneous and dynamic existing resources from multiple geographically distributed institutions (Virtual Organization)
    - resources have their their hardware, operating systems, local resource management, and security infrastructure
      - Resources integration implemented with the middleware: a set of standard protocols, middleware, toolkits, and services built on top of these protocols.
      - Resources are pre-reserved
  - Clouds are usually referred to as a large pool of computing and/or storage resources, which can be accessed via standard protocols via an abstract interface from the Internet.
    - cloud computing systems are demand driven (consumers' actual needs)
- Cloud computing solutions give enterprises significantly more flexibility.
  - They can dispense with IT infrastructures of their own and only have to pay for the resources and services they actually use ("pay-per-use"/ "pay as you go").
  - These can be dynamically adapted to changed business requirements and processes with the help of virtualization technologies and service oriented, distributed software systems.

Grid to Cloud for LHC CM

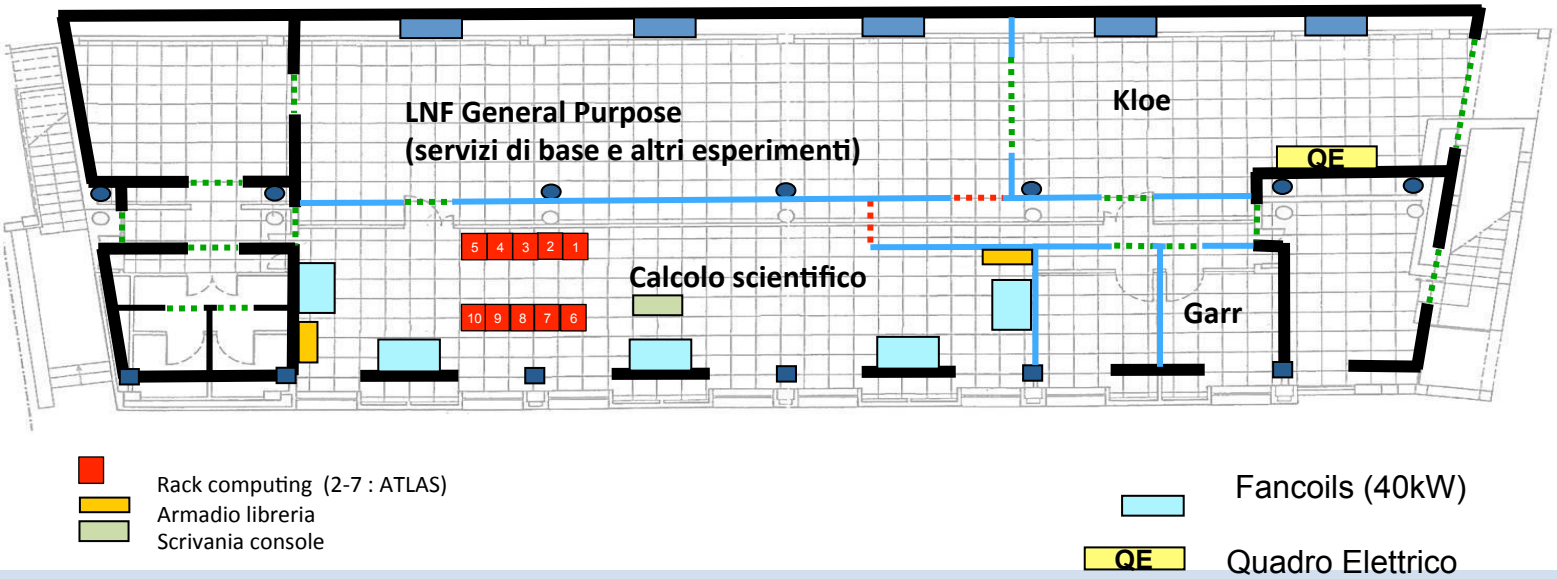
- From the infrastructure administration point of view, administration of cloud sites is easier than traditional grid centres.
- Load from sites to central services:
  - All the experiment software delivered by central services to sites
  - Servers configurations supplied as images directly from central experiment support
  - Very few services still at the level of the site
- Not a solution for everything (so far at least): not yet addressing completely the needs of I/O intensive tasks from the storage point of view
- **GRID is still the baseline, anyway Network & Cloud are pushing towards a simplification of the computing infrastructures**
- **With cloud infrastructure the integration of new community (non LHC) will be easier.**

The ATLAS Italian "cloud"

- The INFN Computing Infrastructure is made of:
  - 1 Tier1 at CNAF (Bologna)
  - 10 Tier-2s serving the 4 LHC experiments
  - LHC Tier3s in almost all groups
  - many experiment farms in all the universities
- Global amount of available Italian resources:
  - CNAF: main Italian computing centre for LHC experiments and several others
    - ~140 kHS06, ~13 PB of disk, ~16 PB of tape
  - Tier-2s: Global amount of resources: ~125 kHS06 CPU and ~10 PB disk
- Network connection provided by GARR (GARR-X):
  - 10 Gbps WAN connection for all the Tier2s
  - CNAF 3x10 Gbps WAN connection
  - 100 Gbps transition starting from south sites
- ATLAS Italian Tier-2s:
  - **Frascati**
  - Milano
  - Napoli
  - Roma1

LNF Tier-2: the infrastructure

- Brand new data center available at LNF.
- ~90m^2 surface.
- 160 kW electric power (UPS 66kW).
- Conditioning system: >160 kW .
- Redundant conditioning system.
- Recycling of waste heat (PUE ~1,24 during the winter months).



LNF "green" Computing Infrastructure

- Data Center waste heat recovery

About 400 kW of waste heat, coming from the cooling system of the data centre and other technological equipment, will be used to heat some buildings in the winter season.

This activity aims at saving 55 k€ per year of natural gas, and to avoid expensive extraordinary maintenance activities on a very old heating plant.

With this action, in the winter season, Frascati DC will have: **PUE = 1,24**.

This work will be carried out during the year 2014 with ordinary funds

- Fuel cell Integrated power and cooling system for a data centre

LNF has also submitted a more ambitious R&D project, in cooperation with some firm, whose aim is at delivering an innovative powering, cooling and continuity system, to serve the data center, based on the usage of molten carbonate fuel cells.

The idea would allow the maximum exploitation of the primary energy (methane) for the powering of the DC equipment, their cooling and the power supply continuity.

Also under evaluation is the possibility of carrying out the project with available financial incentives.

LNF Tier-2

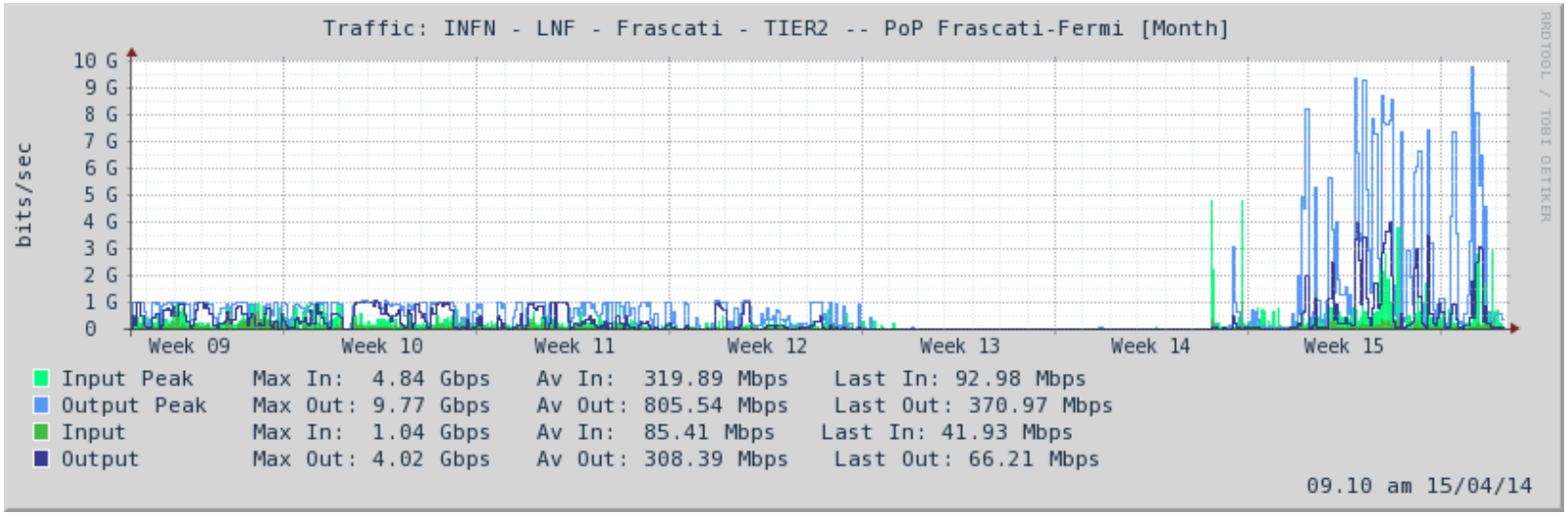
- CPU: 6340 HEPSPEC (~1150 job slots)
- Storage: 580TBn (~700TBr)
- Network: 10Gbps WAN connection, 10Gbps LAN (disk servers and rack switches)
- EMI-3 midlleware, Disk Pool Manager (DPM) as SRM, Torque/Maui batch system
- High availibility and reliability in the last years:
  - Availability 2011-13 = time_site_is_available/total_time = 95%
  - Reliability 2011-13 =

time_site_is_available/(total_time-time_site_is_sched_down) = 97%

- More than 90% of efficiency of ATLAS analysis and production jobs.
- LNF Tier-2 is dedicated to ATLAS jobs, but also the other LHC VOs and Belle 2 VO jobs are supported
- Collaboration outside INFN: Megalab (regional project) in collaboration with CNR, ESA/Esrin and Regione Lazio

LNF Tier-2

- Two weeks ago we connected the farm with 10Gbps link to LHCONE, traffic suddenly increased.
- More than 2 millions of ATLAS completed jobs last year.



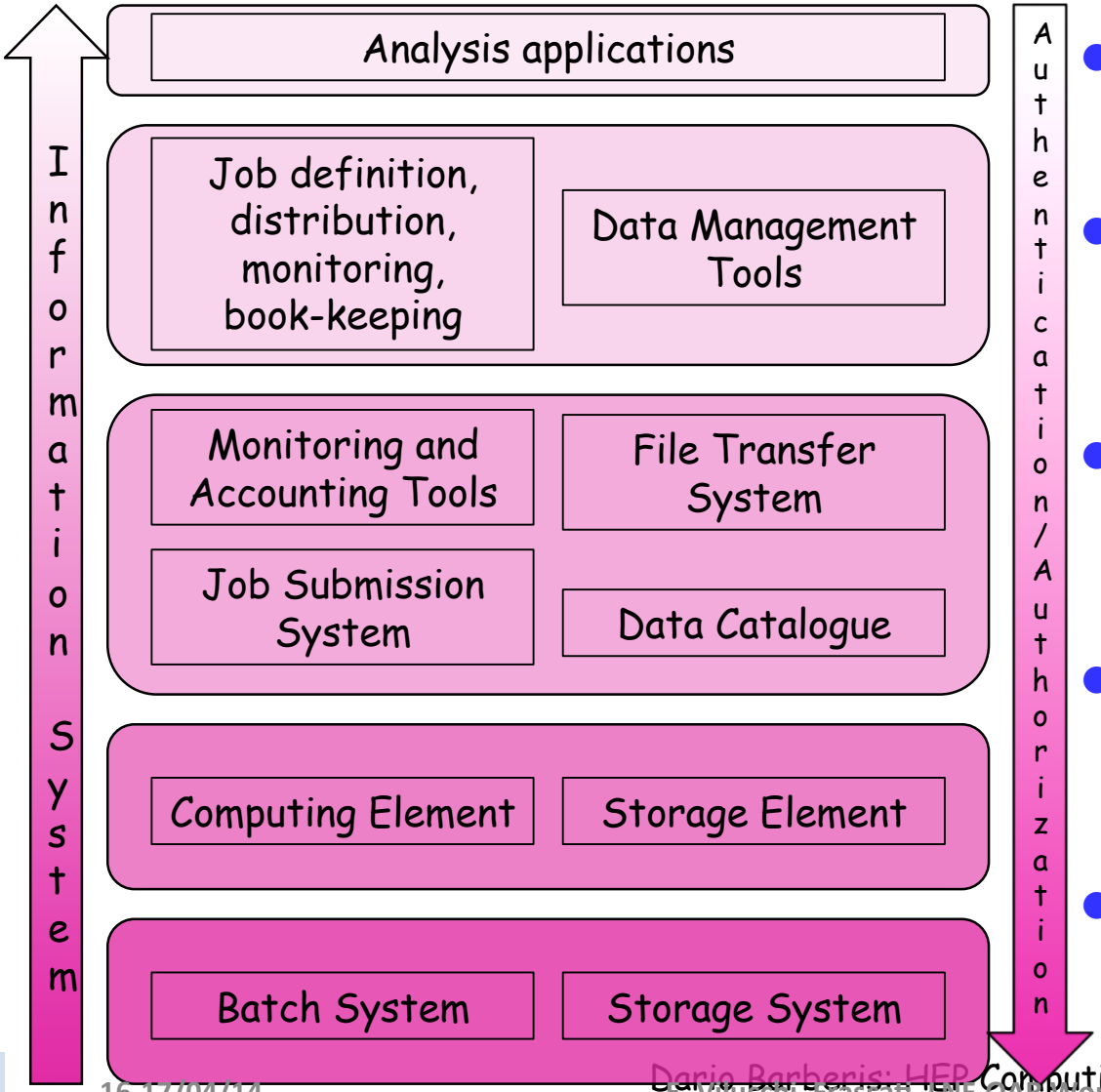
- INFN Tier-2s review few months ago
  - Frascati resulted one of the most performing and efficient Tier-2

Conclusions

- LHC results showed that Grid computing is efficient for massive distributed computing and the transition towards a simplified computing model, based on the cloud paradigm, is already started.
- Frascati Tier-2 just moved in a new computing centre, large enough to host many other computing resources
- Frascati reached an high level of maturity to play a significant role in any new computing infrastructure in the next years.

Backup slides

Layered Service Structure



- Experiment applications (running jobs)
  - Written by physicists
- Experiment production and data management infrastructure
  - Written by computing experts in the experiments
- Common HEP application layer
  - Written by computing specialists at CERN and other laboratories
- Grid middleware
  - Funded by EU FP6/7, US DoE/NSF, national projects and academic institutions
- Local resources
  - Storage, batch, operating system

Computing Model: main operations

- Tier-0:
  - Copy RAW data to CERN Castor Mass Storage System tape for archival
  - Copy RAW data to Tier-1s for storage and subsequent reprocessing
  - Run first-pass calibration/alignment (within 24 hrs)
  - Run first-pass reconstruction (within 48 hrs)
  - Distribute reconstruction output (ESDs, AODs & TAGS) to Tier-1s
- Tier-1s:
  - Store and take care of a fraction of RAW data (forever)
  - Run "slow" calibration/alignment procedures
  - Rerun reconstruction with better calib/align and/or algorithms
  - Distribute reconstruction output to Tier-2s
  - Keep current versions of ESDs and AODs on disk for analysis
  - Run large-scale event selection and analysis jobs
- Tier-2s:
  - Run simulation (and calibration/alignment when/where appropriate)
  - Keep current versions of AODs on disk for analysis
  - Run analysis jobs
- Tier-3s:
  - Provide access to Grid resources and local storage for end-user data
  - Contribute CPU cycles for simulation and analysis

Web access to software and databases

- Placing web servers and a sequence of cascading caches in front of major services is a very cost-effective way to provide robust and distributed read access to popular information
- Frontier shields Oracle servers from overloads due to repeated identical queries from jobs accessing the same conditions data (same runs or time intervals)
  - Data are cached in the Frontier launchpad (server) and the site Squid before getting to the worker node
  - Access times for conditions data decreased from several minutes to a few seconds for Tier-2s with large latency times to the nearest Tier-1
- CVMFS is a web-based file system with Squid caches at each site
  - No need to pre-install all software releases on each site; software is pulled and cached locally when used
  - Conditions data files are also available through CVMFS, saving space on the SE
  - Deployment for ATLAS in rapid progress (3/4 of sites now, aim for completion in 2012)

Databases

- Not all data is part of event records stored in sequential or structured files
- Oracle databases are used for several purposes:
  - Online:
    - Detector configuration parameters
    - Detector condition monitoring (temperatures, pressures, voltages and currents, gas mixtures)
    - Calibration parameters to be loaded in front-end processors
  - Offline:
    - Detector calibrations and alignment for offline event reconstruction and analysis
      - a few GB/day of new data
    - Event, file and dataset catalogues
      - a few TB/year of new data
    - Grid site parameters and topology
    - Task production book-keeping and logging
- Databases are accessed directly (online) or through web services (offline, from any site)