Distributed Storage work status

Giacinto DONVITO – INFN/IGI
Claudio GRANDI – INFN-Bologna
Armando FELLA – INFN-Pisa
on behalf of distributed storage group

Outline

- People involved/interested
- List of activities
- Few updates on:
  - Data Model
  - HTTP remote data access
  - HADOOP testing
- To-do and future works

People involved

- **Giacinto Donvito** – INFN-Bari:
  - Data Model
  - HADOOP testing
  - http & xrootd remote access
  - Distributed Tier1 testing
- **Silvio Pardi, Domenico del Prete, Guido Russo** – INFN Napoli:
  - Cluster set-up
  - Distributed Tier1 testing
  - Gluster testing
  - SRM testing
- **Gianni Marzulli** – INFN-Bari:
  - Cluster set-up
  - HADOOP testing
- **Armando Fella** – INFN-Pisa:
  - http remote access
  - NFSv4.1 testing
  - Data Model
- **Elisa Manoni** – INFN-Perugia:
  - Developing application code for testing http & xrootd data access
- **Paolo Franchini** – INFN-CNAF:
  - http remote access
- **Claudio Grandi** – INFN-Bologna:
  - Data Model
- **Stefano Bagnasco** – INFN-Torino:
  - GlusterFS testing

List of activities

- Data model
- HTTP remote data access
- Storage technologies tracking:
  - HADOOP testing
  - GlusterFS testing
  - EOS testing
  - NFSv4.1
- Distributed Tier1 studies

No major updates on these items mainly due to lack of Man Power

SuperB Data Model

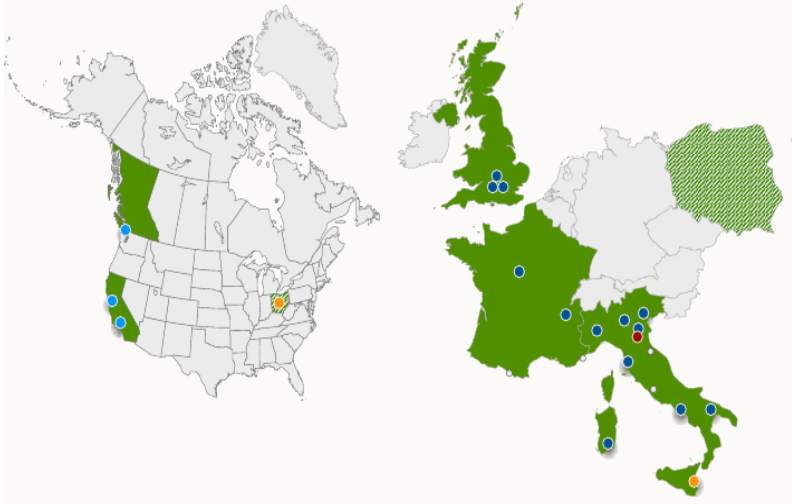
- LHC experiment are taking advantage of:
  - A fully redundant LHCOPN for CERN-T1 (and T1-T1)
  - And soon: “LHCONE” for T2/3
- Full-mesh data routes
- Hierarchy, data routes and workflows imposed by
  - Use case peculiarities:
    - Reconstruction, Reprocessing, Simulation, Production, Analysis
  - Installed facilities, infrastructure:
    - Presence of MSS, How much CPU/Disk, network connectivity, etc
  - Human factors:
    - Physics group location, technical expertise on site, politic choices

SuperB Distributed Sites

- SuperB will be a fully-distributed experiment
- With site from at least 2 grid flavors (OSG, EGI)
- Specific categories of sites associated to specific use cases?
  - Analysis and MC production everywhere?
  - Reprocessing are community or resource driven?
- Which use-cases will be fulfilled by South-Italy's computing centers?

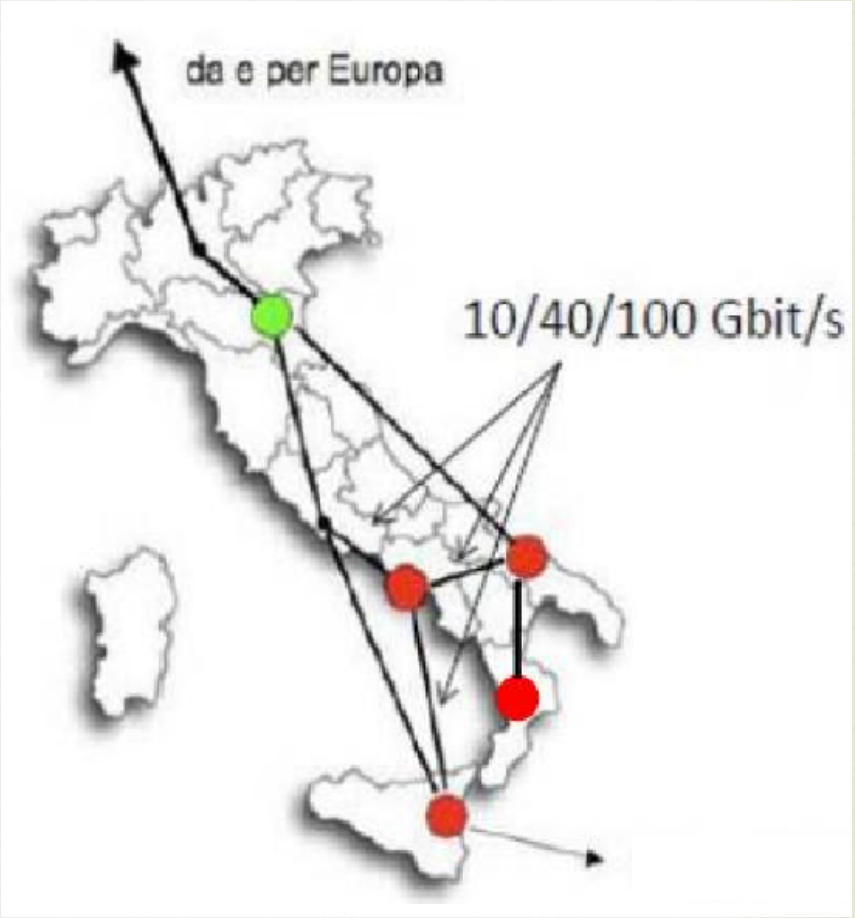
- **LHC Tier-1s** (3)
  - infn-t1, in2p3-cc, ral-lcg2
- **LHC Tier-2** (16)
  - uki-lt2-qmul, uki-southgrid-ralpp, uki-southgrid-ox-hep, grif, in2p3-lpsc, wt2(slac), cit-cms-t2b, victoria-lcg2, cyfronet-lcg2, infn-bari, infn-catania, infn-lnl-2, infn-milano, infn-napoli-atlas, infn-pisa, infn-torino
- **Other** (8)
  - infn-ferrara, infn-perugia, infn-cagliari, napoli-grisu, napoli-unina, in2p3-ires, cit-hep-ce, osc

Green: EGI sites
Red: OSG sites



Data Route: SuperB use case

- Tier0 at experiment location Roma Tor Vergata
- Four sites with dedicated computing resources
  - Plan: three sites forming a main distributed computing center with LHC Tier1s like duties, capabilities, core business
  - Network upgrade plan involving south of Italy sites
  - Full-mesh data route profiting of LHC network infrastructures



Computing Model survey

- The storage group prepared a survey to help defining the SuperB Data and Computing Model
  - http://mailman.fe.infn.it/superbwiki/index.php/Distributed_Computing/Distributed_storage_portal
- Why do we need to start defining the Computing Model now?
  - Some of the choices affect the functionalities needed from the computing tools currently in development/adoption
  - Same of the choices may also affect the topology of the computing infrastructure (services needed at sites)
- Do we pretend to define the Computing/Data Model now?
  - No. Nothing is carved in stone.
- So why a survey now?
  - Because some of the questions the computing group has may already have obvious answers.
  - Because even if some answers will change with time we may get anyhow an overall direction to follow.

Computing Model survey

- Data taking
  - Currently assuming choices typical of lepton colliders, but the high luminosity may suggest choices closer to those of hadron colliders (e.g. multiple physics streams, express stream, etc...)
- Data formats
  - Currently assuming the same of Babar. Is this still valid?
    - Will Event Directories (indexes of individual events in different files/datasets) be used?
    - What kind of skims will be used (filters, data reduction, ...)?
    - What are the exact flows of MC full/fast simulation? MC data formats?
    - How is a dataset defined? Is it "open" or "closed"?
    - How are Conditions Data organized (RDB, flat files, ...)?

Computing Model survey

- Organized Processing
  - Frequency of reprocessing, IO definition
  - Organized physics groups productions
- Calibrations
  - Are there dedicated calibration/alignment samples?
  - Frequency? Latency?
- Analysis
  - Accessing any possible data format?
  - Is “Sparse” data access possible?
- Quantitative information
    - What is the amount of MC to be produced?
    - Do the following (by Steffen) need to be reviewed? http://agenda.infn.it/getFile.py/access?resId=0&materialId=0&confId=4678

HTTP remote data access

- The work is going on:
  - Please look at Paolo Franchini talk Thursday Morning “Computing - Overflow” session

HadoopFS testing

- The activity was started in Bari few months ago but since January we have an FTE fully dedicated to those test
- We are mainly focusing on resilience to failures
  - NameNode failure
  - Disk & DataNode failures
  - Racks failures
  - Data Centers failures
- At the moment the test are using different networks on the same computing center
  - Testing problems on using it through firewall
- Deep testing on FUSE
- HDFS + WebDav Testing

HadoopFS first results

- Resilience to failures
  - NameNode failure ➔ OK
  - Disk & DataNode failures ➔ OK
  - Racks failures ➔ OK
  - Data Centers failures ➔ Still not FULLY OK
    - The built-in replica algorithms are not fully compatible with data distribution among geographically distributed computing centers

HadoopFS first results

- Firewall tests ➔ OK
- FUSE testing ➔ OK
  - Still not supporting complex write operations
- Apache WebDav + FUSE ➔ OK
  - HDFS Native WebDav implementation ➔ Need further development
- We are currently testing 3 different HDFS testing:
  - 0.20.203
    - Quite stable and complete version
  - OSG version
    - Based on a older 0.20.xxx version + few interesting patches from CMS peoples
  - 1.0.x version
    - Will be the next “stable” release

Future works and ToDo

- To interact with Online and Offline experts to try to answer to “open questions” in order to complete the Data&Computing Model
- To go on with HTTP test working both on the storage technology and on the application software tuning
  - Looking to the EMI development in terms of HTTP dynamic catalogue
- To finalize HDFS testing on local farm
  - To start testing a geographically distributed environment
- To start doing deep HDFS performance and scalability tests

- To start writing the Data and Computing Model

Final thought and conclusions

- Several activities still pending due to a endemic lack of Men Power
  - We really need new people joining the group with significant effort dedicated
- Technology is evolving out-there and we need to catch all the new possibility that could help the SuperB community
- LHC and others experiments are testing new solution and models, so we have to carefully look at what is happening:
  - We are participating to both Storage and DataManagement TEG (Technology Evolution Group) born within WLCG in order to understand how the computing model of LHC experiment will evolve in the future
    - The outcome of those groups will be of help in choosing both technical solution and general design option

Final thought and conclusions

- Now we are focusing on Data Model and Computing Model, in order to write Computing TDR
  - This is the right time to begin thinking to these issues
  - Answers collected now could be revised in the future, but are important to drive the Computing group in building the general design
  - It should be a good opportunity to start a positive interaction between Computing and Physics people
    - For example new emerging technologies could help the physics community to find new solutions to "old" problems