SuperB R&D computing program: HTTP direct access to distributed resources

Armando Fella for the SuperB Computing Group

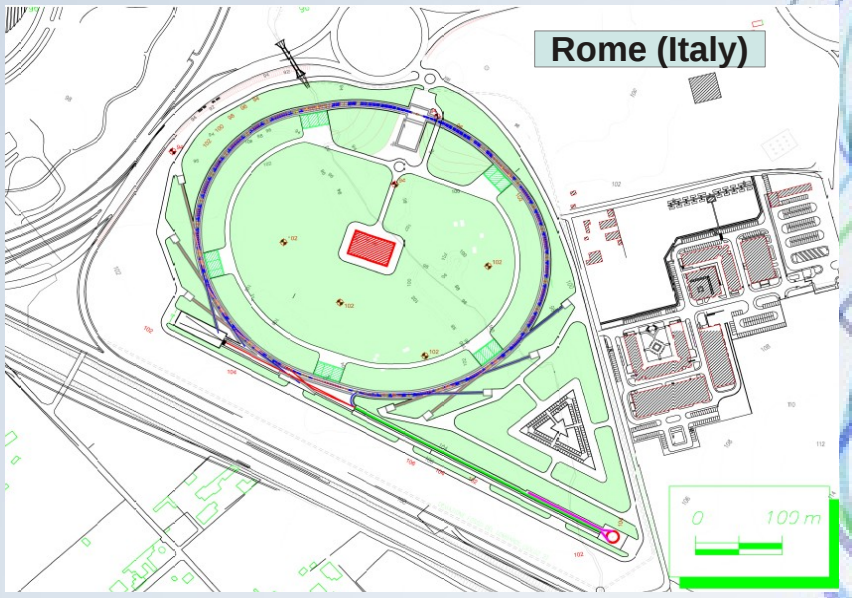
Computing for High Energy Physics, 21-25 May 2012
New York City, NY, USA

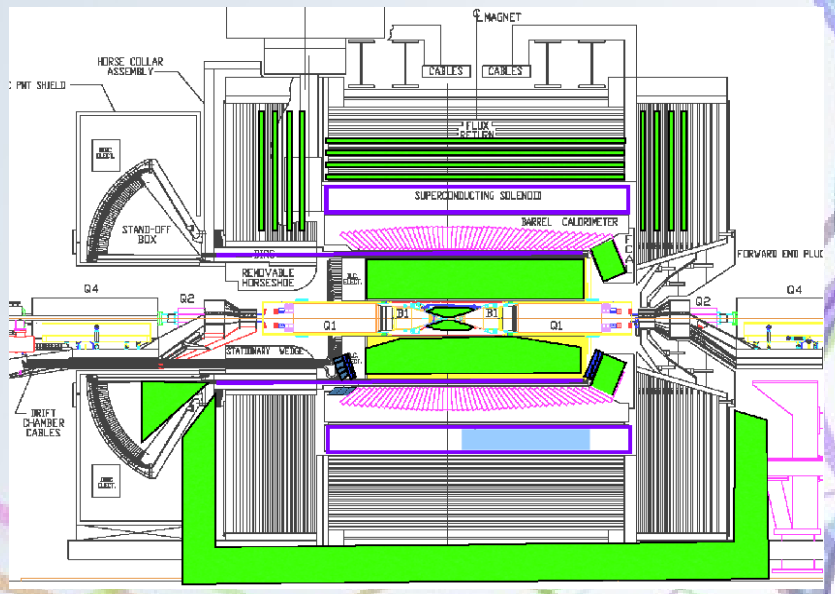
Presentation Layout

- SuperB experiment overview
- SuperB distributed resources
- Data access R&D plan
- WAN data access, work status
- Test http protocol in WAN data access
- Conclusion and future work

SuperB experiment

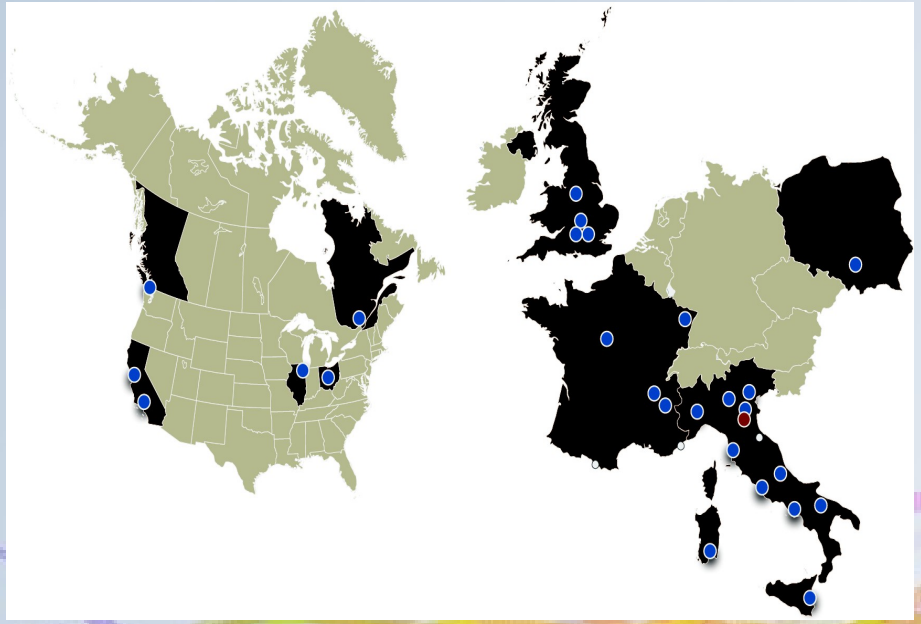
- SuperB is an asymmetric flavour factory with a two-order of magnitudes jump in luminosity with respect to present B-Factories.
- Data taking is planned to start in 2017
- Computing model definition to be frozen in TDR within one year
  - Luminosity x O(100) w.r.t. Present B-Factories (6 years of run)
    - $L_{inst} = 10^{36}\ cm^{-2}s^{-1}$
    - $L_{int} = 75\ ab^{-1}$
  - Flexible parameter choice
  - First level trigger expected rate: ~100KHz
  - Third level trigger expected rate: ~25KHz
  - Expected event size: ~200KB
- **Large international collaboration:** Canada, Italy, France, Poland, Russia, Spain, The United Kingdom and the United States.





SuperB distributed resources

- The distributed computing infrastructure, as of May 2012, includes several sites in Europe and North America
- EGI and OSG Grid flavours have been enabled
- The LHC Computing Grid architecture was adopted to provide the minimum set of services and applications upon which the SuperB distributed model could be built.
- Computing resources needed in a typical year of SuperB data taking are of the same order as corresponding ATLAS and CMS estimation for 2011



Parameter	typical Year
Luminosity (ab^{-1})	15
Storage (PB)	
Tape	113
Disk	52
CPU (KHep-Spec06)	
Event data reconstruction	210
Skimming	250
Monte Carlo	670
Physics analysis	570
Total	1700

General focus

- **Computing TDR works** include an intensive R&D program permitting the experiment to evaluate adoption or development of new solutions in data management field among the others
- **Data access** is one of the key subjects that will drive the SuperB computing activities aimed to **Data Model definition**.
- The main **areas of interest** are: WAN data access, new generation of mass data transfer system and dynamic file catalogue.
  - Activities carried on by the middleware providers like EMI project
- LHC experiments are very interested in dynamic and remote data access
  - Alice experiment is fully working in this paradigm since a couple of years, CMS and Atlas recently have implemented such a solutions in specific use cases

WAN data access opportunity

- **Remote data access** will be important together with data placement management in the following specific **use cases**:
  - **Interactive usage of SuperB data**, for example: event display and single event browsing
  - Writing and debugging **analysis code**
  - **Opportunistic analysis** executed on non SuperB-dedicated resources. Both in terms of non-SuperB grid computing centers and dynamic allocated cloud resources
  - Job execution on **site without** experiment **storage support** (Tier3 like)
  - Increasing in **reliability and availability**, recovering from temporarily/partial storage failure at SuperB sites

WAN data access requirements

- The **protocols** involved in remote and dynamic data access should provide the following **functionality**:
  - Support to **posix-like call** (open, read, seek, close)
  - Capabilities of **work through routers and firewalls**
  - **Caching and pre-fetching** features for improving performance on high latency network
  - If the protocol is natively **supported by ROOT** framework this will make the adoption much more easy
- At present time at least two protocols are good candidates that could fulfill these requirements: **xrootd and http**.
  - **Xrootd** has a high level of maturity, but it was born and used only within HEP community
  - **http** is collecting huge interest also outside the HEP environment.
- Both of these protocols are supported by ROOT framework.
- The SuperB experiment is interested in testing the remote data access with both of them to understand which one is working better from a point of view of performance and functionality.

WAN data access, past test

- First phase test: SuperB analysis execution reading data over a WAN network using both xrootd and http protocols
  - The results highlighted that the performance could rapidly decrease as soon as the network latency increases
    - We need a common software layer permitting to optimize the access to remote data, by means of data caching and pre-fetch algorithms

Data access software layer

- The development of a **general file access library** that could hide the complexity to the end user is in progress, it will implement:
  - Intelligent pre-fetching and buffering algorithms
    - Using the time spent in processing events in order to read the data from remote storage
  - Logical file name map with different physical storage url
    - Enabling the support to storage protocols not already supported by ROOT
  - Read-head buffer or caching mechanisms in order to match the performance requirements of different network, application, and storage solution
- A working Proof-of-Concept version of this library is now under development

HTTP data access test, goal

- A very preliminary set of tests have been performed in order to collect a first sample of case related results
- The **aim of the tests** are:
  - Measuring the latency due to the increase of the number of parallel read streams
  - Measuring the latency due to the increase of round trip time elapsed between source and destination
  - Support the development of data access software layer
  - Start the characterization of a concrete WAN, 'general purpose', scenario:
    - Traffic impact, typical latency, network resource overloading

HTTP data access test, layout description

- Test jobs performing data access tasks, have been submitted from CNAF site to Italian, European and extra European target sites via gLite suite
- Each job performs 1, 5, 10, 50, 100 parallel set of read streams on randomly chosen files stored on site SE
  - One read stream is a trace file ruled access to a file
  - The trace file is obtained by a real SuperB analysis job
    - An offset and a buffer size per line has been extracted by the read system calls performed by SuperB analysis
  - Each stream performs a curl access to data file per trace file line

HTTP data access test, layout description

- Curl tool:
  - curl 7.15.5 (x86_64-redhat-linux-gnu) libcurl/7.15.5 OpenSSL/0.9.8b zlib/1.2.3 libidn/0.6.5
  - VOMS proxy authentication enabled
- Source data: 250 files, ~500MB each, stored at:
  - CNAF (UI EMI 1.0, StoRM 1.8 over GPFS)
  - BARI (pure apache 2.2 over Lustre FS)
  - NAPOLI (DPM 1.8.2, not still included in this set of tests)
- No http cache mechanism implemented
- BARI SE access do not request authentication

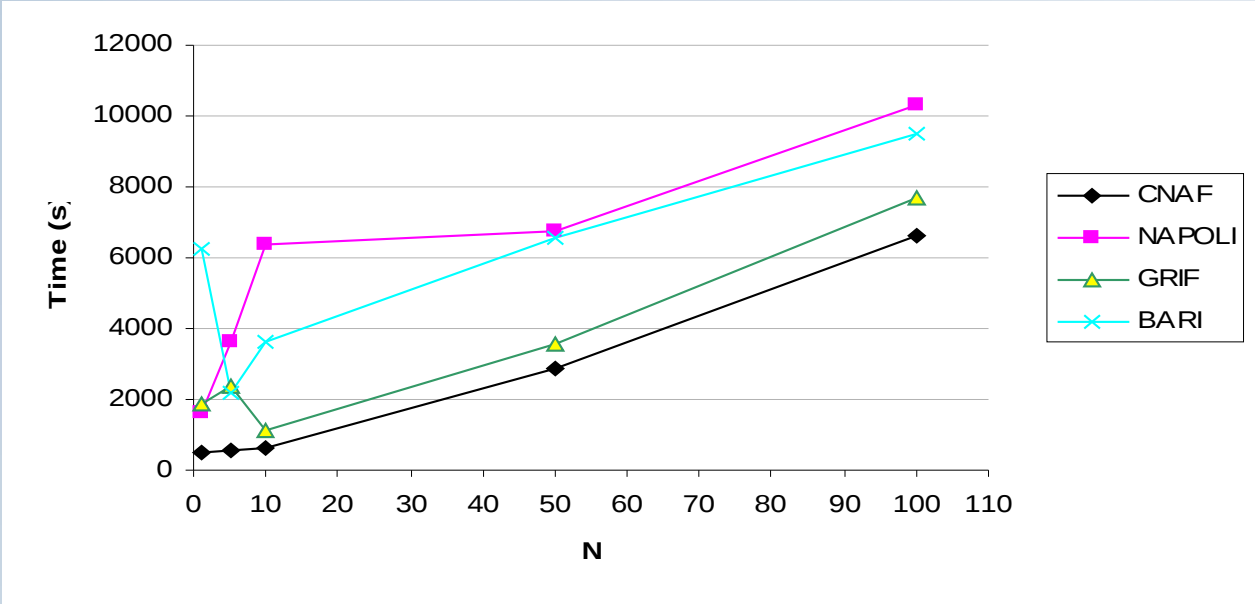
HTTP data access test, read routes

- Job submitted to following sites:
  - CNAF, NAPOLI, BARI, GRIF, CALTECH
  - accessing data file at CNAF and BARI site SE

Data sources

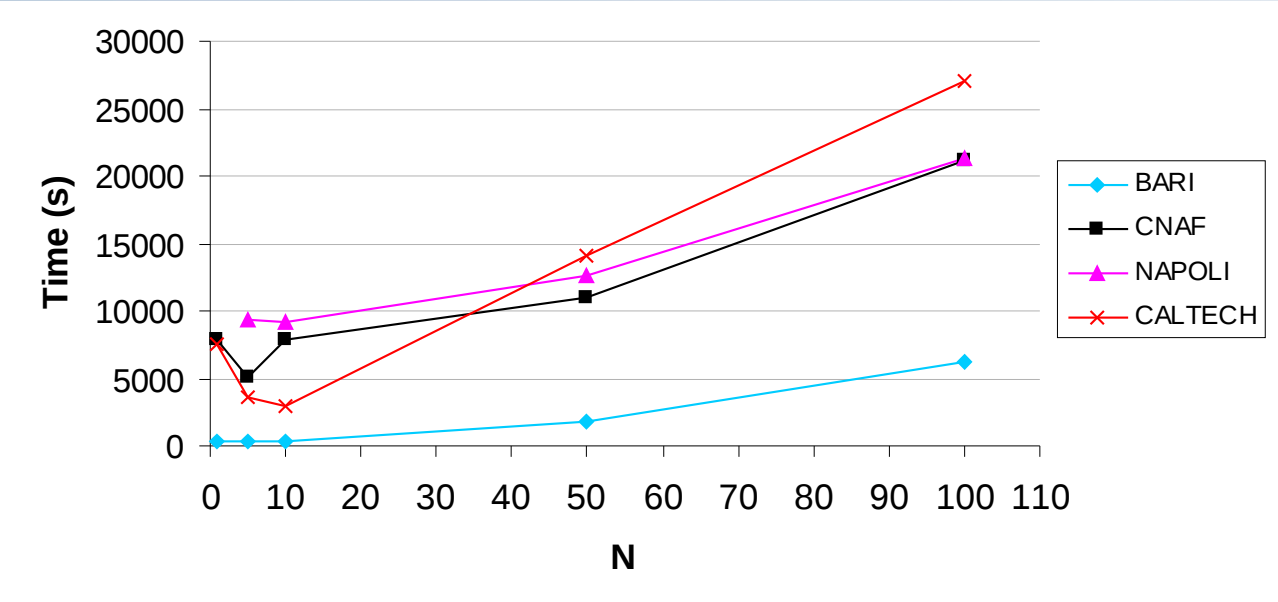
Data destinations	**CNAF**	**BARI**
CNAF	**ok**	ok
NAPOLI	ok	ok
GRIF	ok	failed
CALTECH	failed	ok
BARI	ok	**ok**

Results: CNAF storage source



	Times (seconds)			
Number of parallel accesses	**CNAF**	**NAPOLI**	**GRIF**	**BARI**
1	521	1641	1868	**6232**
5	532	3611	**2402**	2204
10	654	**6385**	1138	3651
50	**2882**	6767	3578	6550
100	6602	10341	7686	9484

Results: BARI storage source



	Times (seconds)			
Number of parallel accesses	**BARI**	**CNAF**	**NAPOLI**	**CALTECH**
1	267	**7791**		**7605**
5	336	5096	**9420**	3643
10	337	7800	9235	2885
50	1725	11012	12624	14023
100	**6159**	21105	21244	27048

Problems, results

- Unpredictable traffic load on geographycal network routes
  - 100 parallel reads overload the links in most of data source sites
  - Two cases of performance degradations also with 1 or 5 parallel streams --> temporary WAN routes saturation
  - Bari link was partially busy by intensive production activity
- Curl vs generic library performing one open, #n seek/read and one close per stream
  - Adopted design suffers of overhead due to open and close system calls per trace file line performed by curl launch
  - Future use of the data access library will fix this behaviour
- Job failures to Caltech and GRIF sites: mostly Grid infrastructure and resource availability problems

Conclusions and future plan

- No authentication layer accessing data to Bari results in a large gain in transfer time --> to be better investigated
- Bari and Napoli links will be upgraded soon to 10 GB/s network bandwidth, tests will be repeated
- Transfers times on regional routes (in Italy) are consistent with the network infrastructure
- This kind of test should be repeated to collect statistics and permit a mean, representative, measure uncorrelated with network instant load
- The tests will be repeated using the data access software layer and a test bed layout including a caching architecture: client side squid proxy