# Cluster Setup and Distributed File System

## R&D Storage

S. Pardi for the R&D Storage Group

# People Involved

Gaetano Capasso - INFN-Naples

Domenico Del Prete – INFN-Naples

Diacono Domenico – INFN-Bari

Donvito Giacinto – INFN-Bari

Armando Fella – INFN-Pisa

Silvio Pardi – INFN-Naples

Spinoso Vincenzo – INFN-Bari

Guido Russo – INFN-Naples

**Incoming people are welcome!**

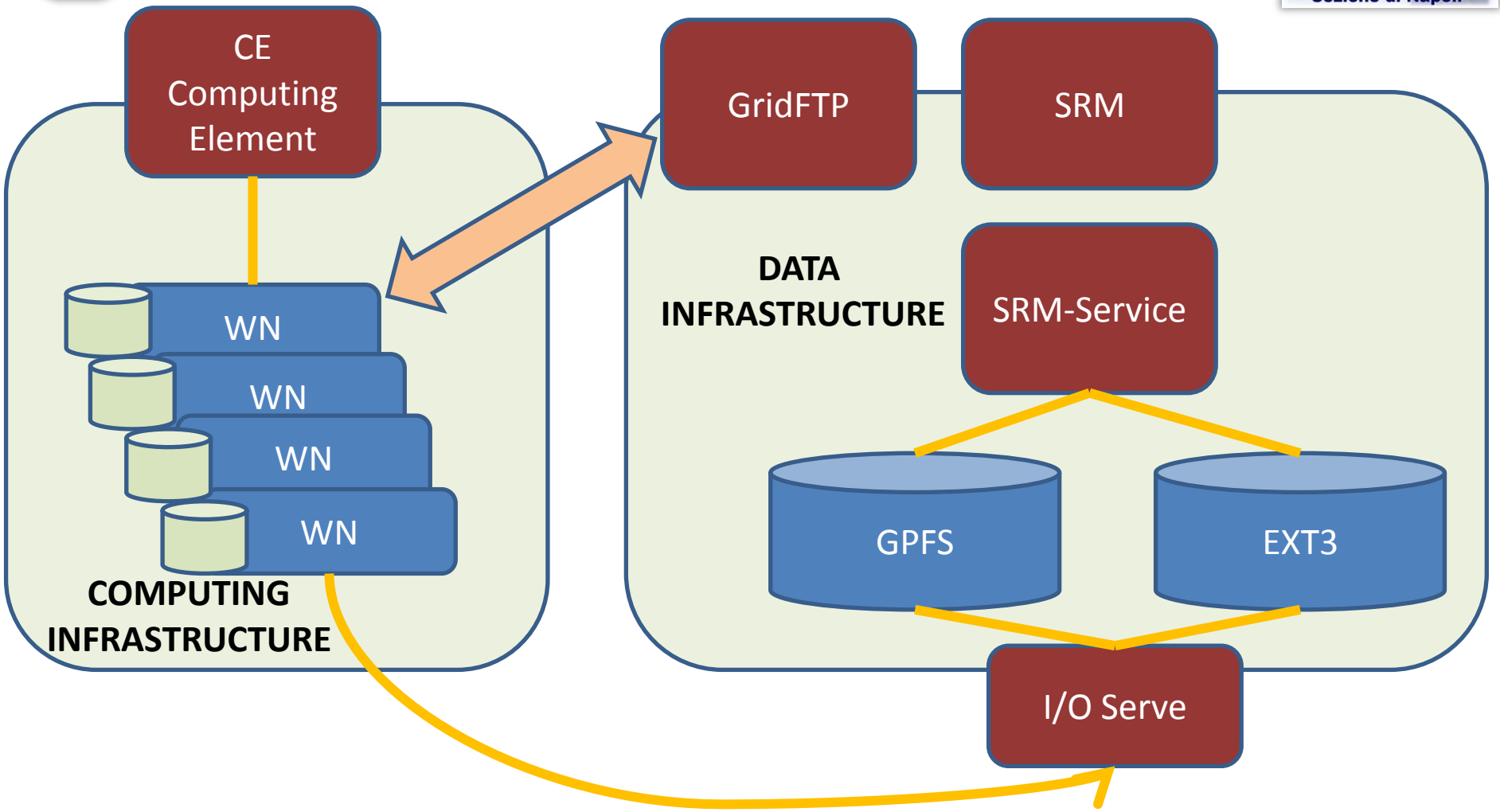
# Storage R&D

- Investigate usage of WN disks performance and applicability to apps/frameworks Hadoop/Proof/caching.
- Interactivity + usage of personal resources for personal analysis tasks
- File/Replica location. Investigate on storage systems able to do it natively
- Investigate simulation-Based approach to define the minimum requirements for storage sites

# Work on Storage R&d

- Work in progress:
  - Testing both functionalities and performance
    - File system under test:
      - Lustre, Xrootd, HDFS, GlusterFS
  - Testing new file systems with analysis jobs:
    - Jobs coming from SuperB itself and CMS
  - Testing several linux distribution in order to find out a usable NFSv4.1
    - Still no definitive solution found

# COMMON STORAGE ARCHITECTURE



# TIER2 ATLAS IN NAPLES

**108** Nodes for a total of **1.056 cores**

Whereof

**83** Worker Node for a total of **882 cores**

**19** Storage Node for a total of **144 cores**

**6** Services Nodes for a total of **30 core**

**~14%** of Total Core for the storage infrastructure

+

Complex Network infrastructure to support traffic  WN to SE.

+

19 Server and Storage System in fiber optic

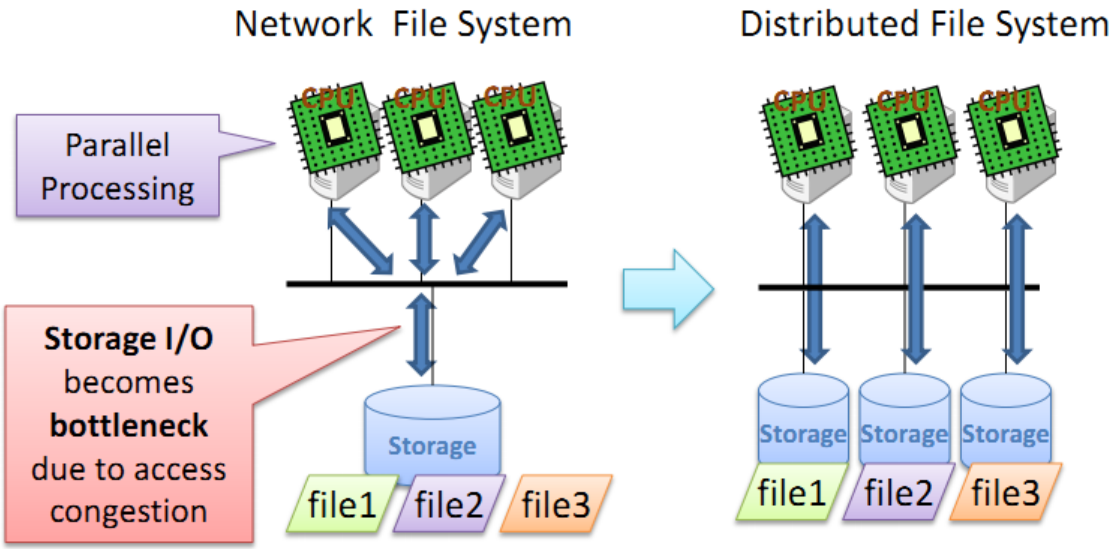
# TIER2 ALICE-CMS IN BARI

- 135 Nodes for a total of 1350 cores
- 20 Disk Server => ~700TB of storage
  - Each server has multiple Gbps connection
- 1 x 10 Gbps storage server dedicated for tests
- The network infrastructure is able to provide 1Gbps to each WN

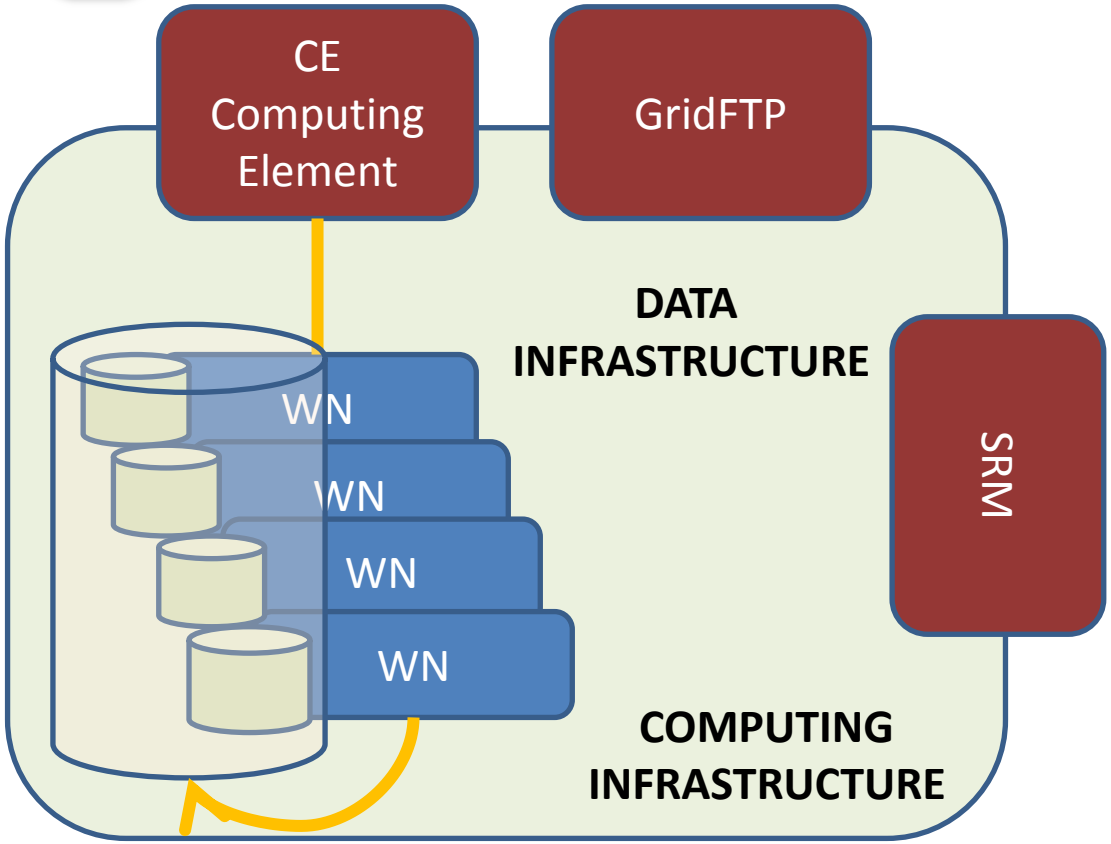
**~12%** of Total Core for the storage infrastructure

# Masahiro Tanaka
# SuperB Ferrara Meeting 2010

## Needs for Distributed File System



# NEW POSSIBLE SCENARIOUS?

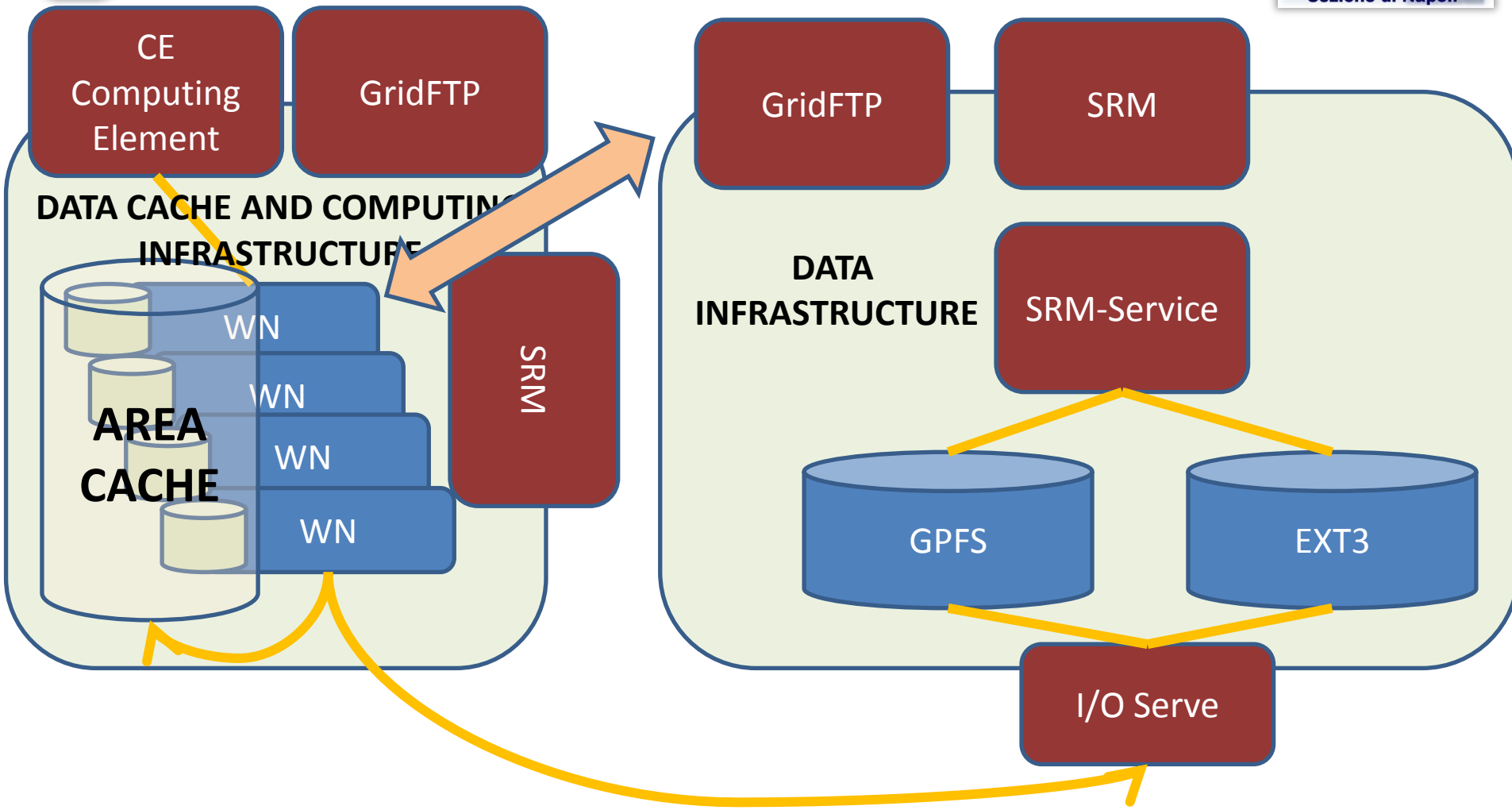



We can design a cluster setup to integrate data storage and computing services in the same nodes.

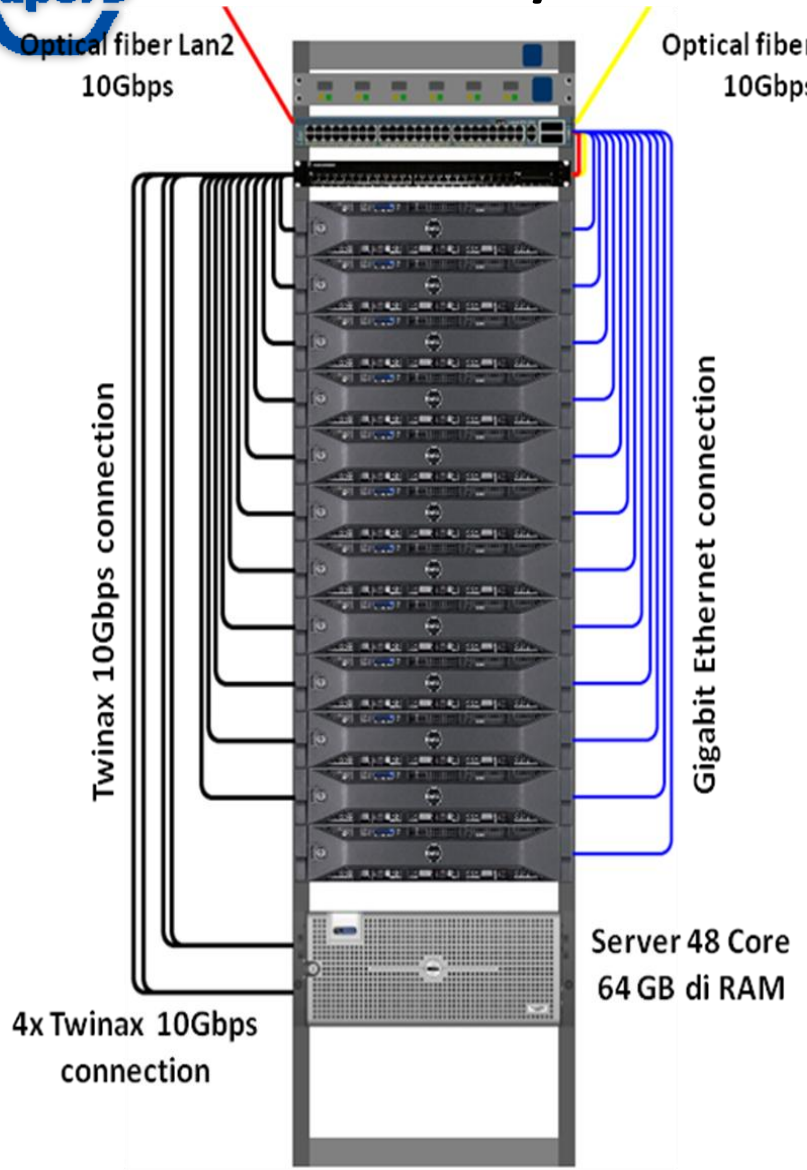
Computer facility as a farm of high density Box in term of core and disk

To day in a 2 Unit server we can have more than 10TB of disk and 48 core

# HIBRID SOLUTION



# 10Gbit/s Farm Setup in Naples




- DELL Server PowerEdge R510 biprocessor quadcore Intel(R) Xeon(R)
- Memory 32Gbyte RAM, a *PERC H200 controller RAID and* four local disk of 500GB in Raid0 configuration.
- Broadcom NetXtreme I 57711 double port SFP+ 10GbE NIC with ofload TOE e iSCSI, PCIe x8.
- OS SL 5.5

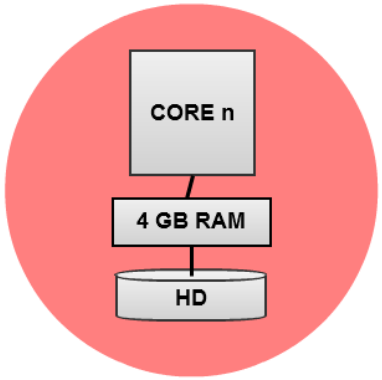
**Interested Peoples can use the cluster for Test**

# The configuration of each node

2x CPU 2,13GHz 4Core, 4,80GT/s + 4,80GT/s QPI
8x 4GB Dual Rank 1066MHz
8x HDD 500GB 7200rpm SATA Hot-Plug 3,5”

10 Gbps
10 Gbps
Twinax

1 Gbps
1 Gbps
UTP




. . . 8x one node (12 nodes in this test)

96 core (8x 12), 384 GB Ram (32x 12), 48 TB HD (4x 12)

# FILE SYSTEM IN EVALUATION

In this presentation we show first benchmark obtained with two Open source solution currently in investigation.

- Hadoop File System
- GlusterFS

# HDSF CARACTERISTICS

HDSF is the Hadoop File system.

- HDFS is designed to be highly fault-tolerant
- An HDFS instance may consist of hundreds or thousands of server machines, each storing part of the file system's data
- Automatic data Replication between the Cluster Nodes with a replica factor configured by the system administrator
- The detection of faults and quick, automatic recovery is a core architectural goal of HDFS.
- HDFS doesn't provide a native POSIX interface, but we have a fuse module
- HDFS doesn't provide multi-streaming writing and random-writing but just append.
- HDFS give best performance with map/reduce framework

# GLUSTERFS CARACTERISTICS

GlusterFS

- GlusterFS is an open source, clustered file system capable of scaling to several petabytes and handling thousands of clients. It aggregates various storage bricks over Infiniband RDMA or TCP/IP network.

- GlusterFS is based on a stackable user space design and can deliver exceptional performance for diverse workloads.

- GlusterFS is a standard posix file system. It works as a supervisor on already existing posix filesystem (i.e. Ext3) with the functionality of data distribution and data replica of complete files among the Brick nodes. GlusterFS does not provide namenode, it use distributed hash and the posix extention to manage the file replication between the nodes.

- GlusterFS Geo-replication provides a continuous, asynchronous, and incremental replication service from one site to another over Local Area Networks (LANs), Wide Area Network (WANs), and across the Internet.

# FILE SYSTEM IN EVALUATION

| Feature | HADOOP FILE SYSTEM (HDFS) | GLUSTER-FS |
|---|---|---|
| Open Source | YES – Java | Yes in C |
| Metadata Architecture | Single Name Node for HDSF infrastructure | No Name Node |
| Posix Interface | Non-native | Native |
| Data Distribution feature | Data are split in block of 64MB and distributed on the DiskNode | Tree FS: Distributed, Replica and Striped. |
| Data replica | Automatic at Block level | Automatic at file level or at block level in the Striped mode. |

# DOUBLE FILE SYSTEM CONFIGURATION

CURRENT SETUP:

OS SL 5.5 with gLite 3.2 Middleware. The Nodes are ready to be include in the Atlas TIER2 and available for SuperB FastSim Program (to do in the next week)

IMPLEMENTATION OF THE **HADOOP** DISTRIBUTED FILE SYSTEM ON 8 NODE

IMPLEMENTATION OF **GLUSTERFS** DISTRIBUTED FILE SYSTEM ON 9 NODE

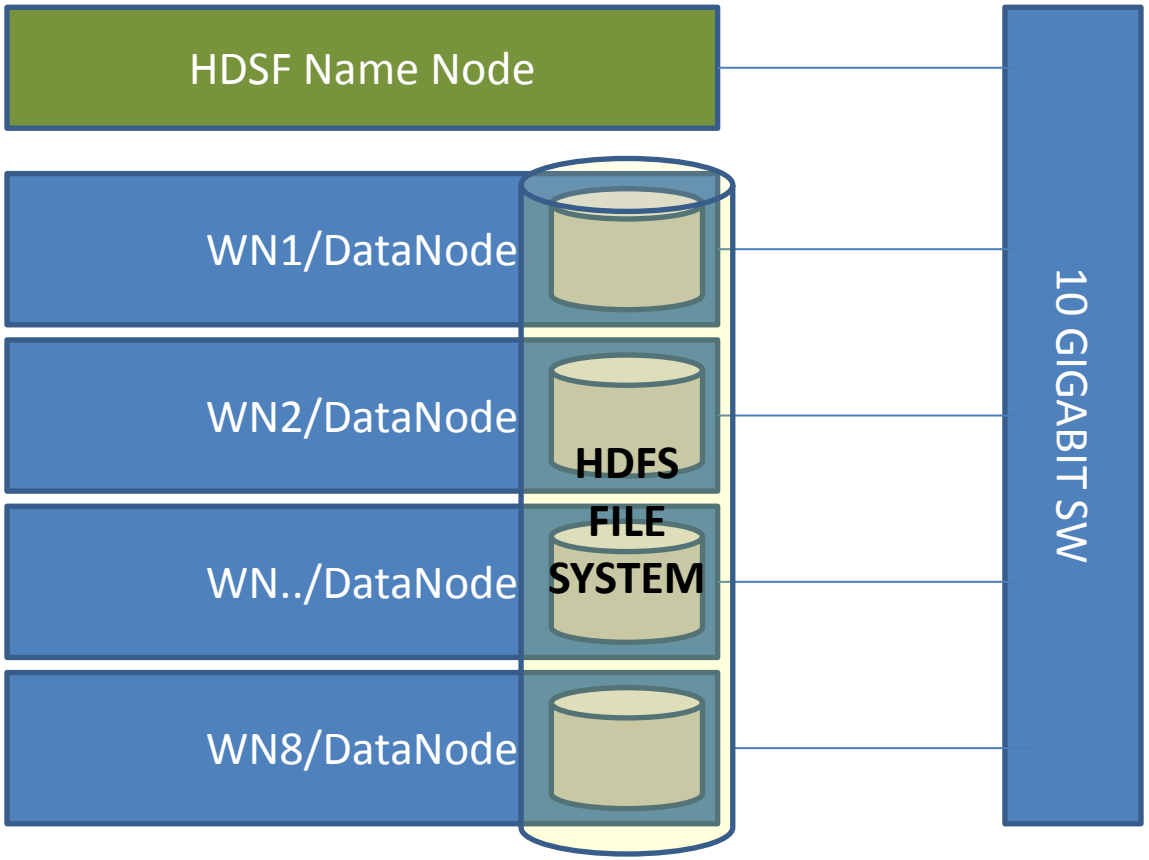
A HDFS area composed by 8 DataNode and one Name node, the fuse module. Each Worker Node of the cluster, mount and use the filesystem through the posix interface.
A GlusterFS area of 4.5TB row configured as distributed filesystem with replica factor 3 for a total of 1.5TB

# HDSF-HADOOP IMPLEMENTATION

1 NAME NODE

8 WORKER NODE / DATANODES that share the local disk

All the DataNode mount HDSF Posix Inferface through
Fuse module



# GLUSTER IMPLEMENTATION

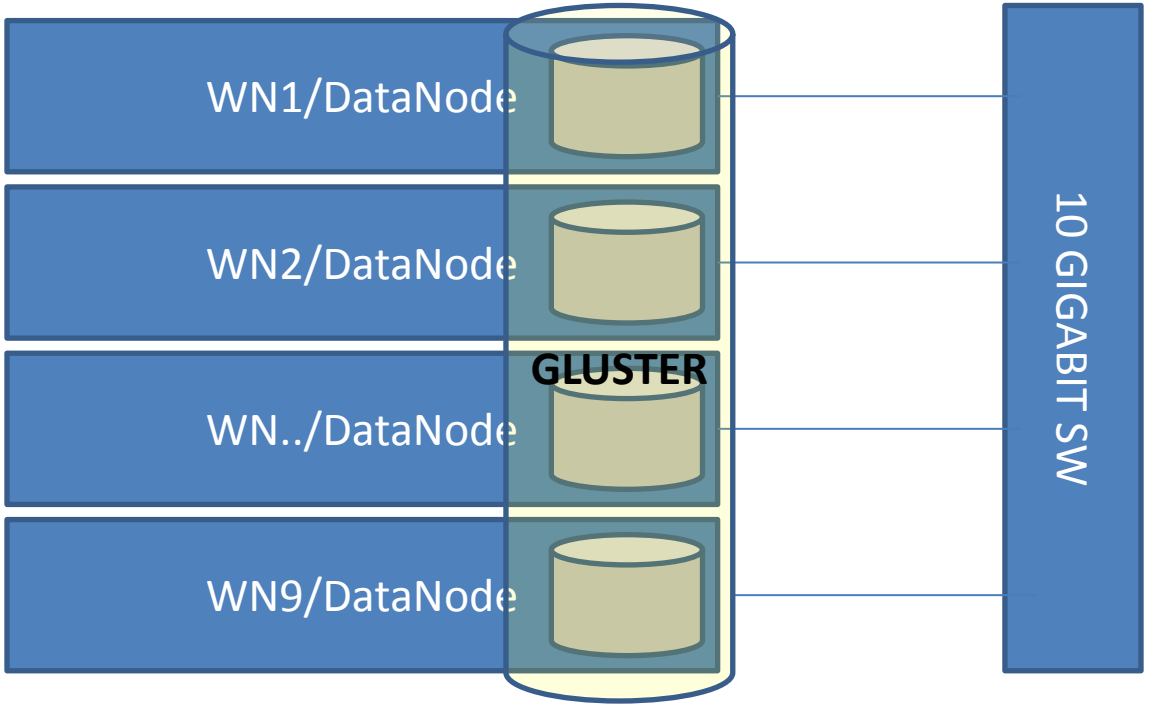
GLUSTERFS IS A FILE CLUSTER FILE SYSTEM THAT WORK WITHOUT DATA-NODES. IT CAN WORK IN 3 DIFFERENT CONFIGURATION: DISTRBUTED, STRIPED AND REPLICA. IT USE THE POSIX EXENSION TO GUARANTEED THE CONSISTENCE OF FS.

**NO NAME NODE**

9 WORKER NODE / DATANODES that share the local disk

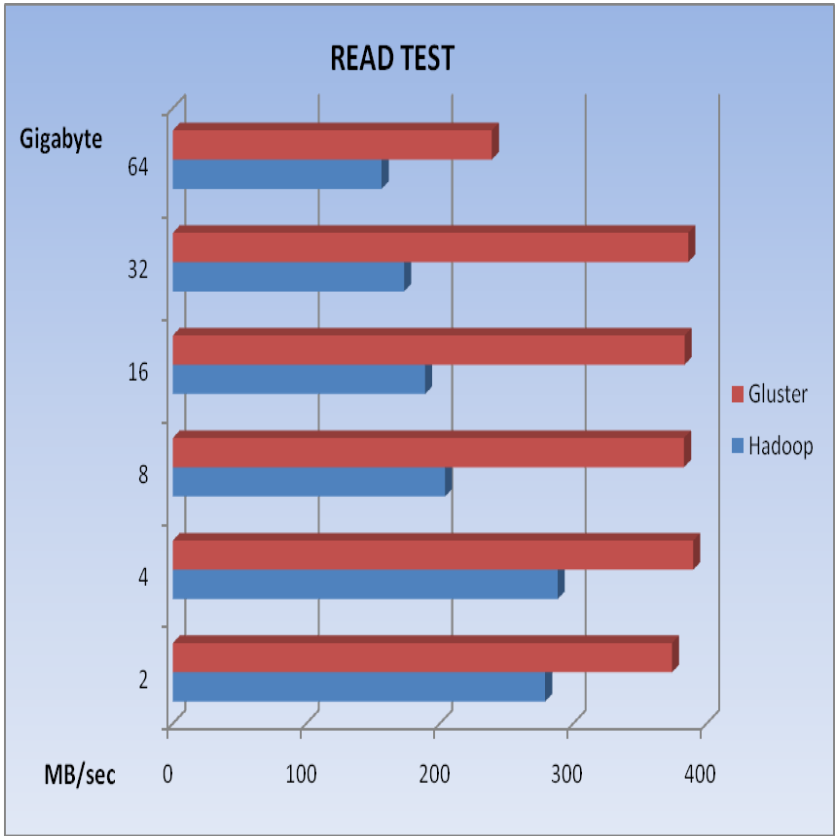
All the DataNode are client and server.

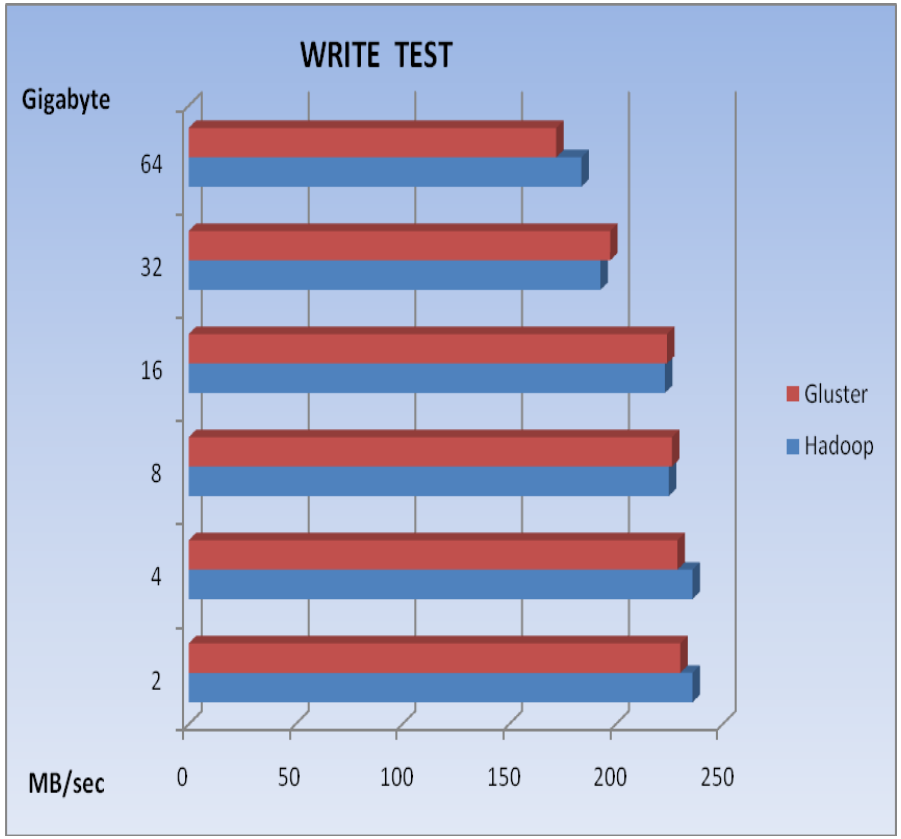
The Gluster FS is configured with replica 3



# GlusterFS Read/Write Test

Basic Benchmark performed from a single nodes.
We used: IO-Zone for GlusterFS and Benchmark provide by HDFS.
HDFS shows strong limits when used ad General purpose FS without Map/Reduce Framework



# Job Analysis Benchmark

Performance test executed through a SuperB Analysis job provided by Alejandro Perez.

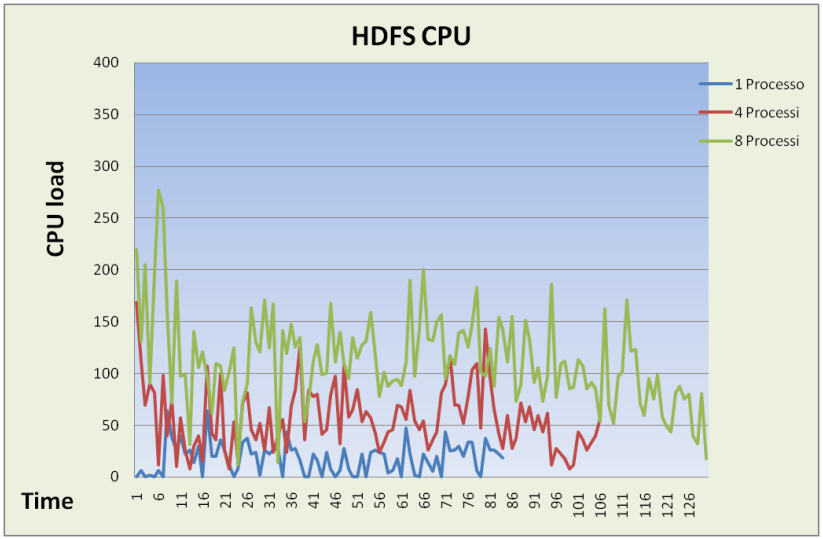
It executes an analysis test on a set of simulated root files previously replicated in the GlusterFS and Hadoop volumes.

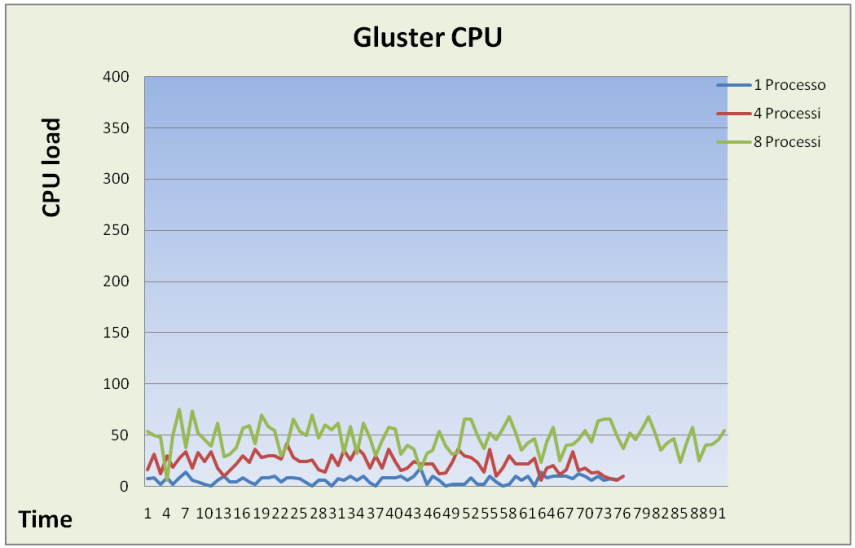
The following compare the execution time and the bandwidth occupation during  a set of different job execution setup

# Jobs on a Single Node

As Benchmark we send test job that analyze 10 Files for a total of about 3GB

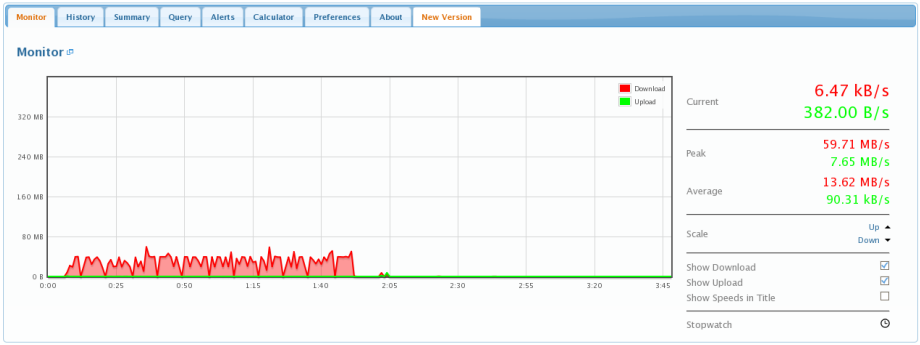
| File System | 1 Job | 2 Jobs | 4 Jobs | 8 Jobs |
|---|---|---|---|---|
| GlusterFS | 115s | 116s | 114s | 135s |
| HDFS | 141s | 145s | 148s | 198s |



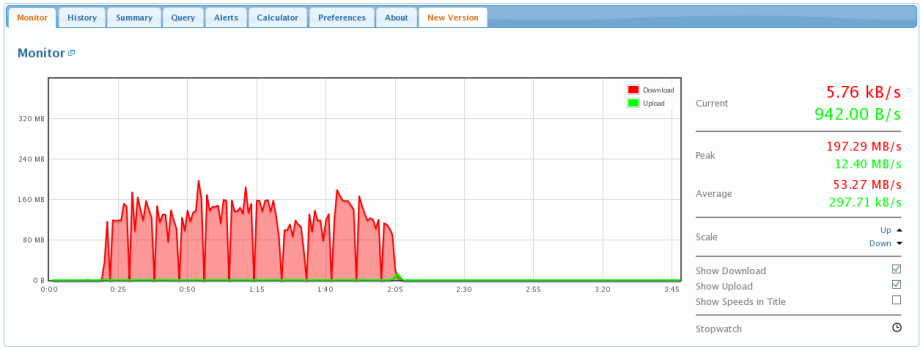



# Network traffic

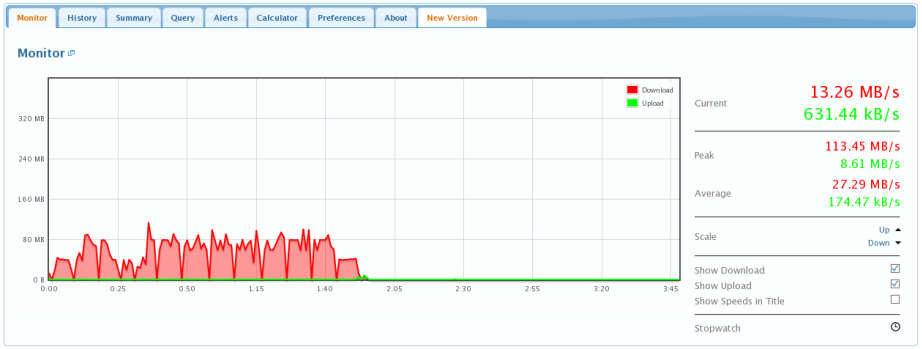
## Single Job max rate 59MB/s



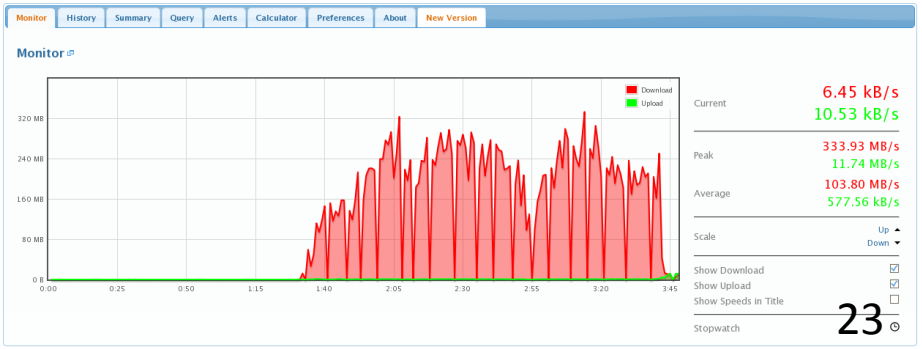

## 4 Jobs max rate 197MB/s




## 2 Jobs max rate 113MB/s

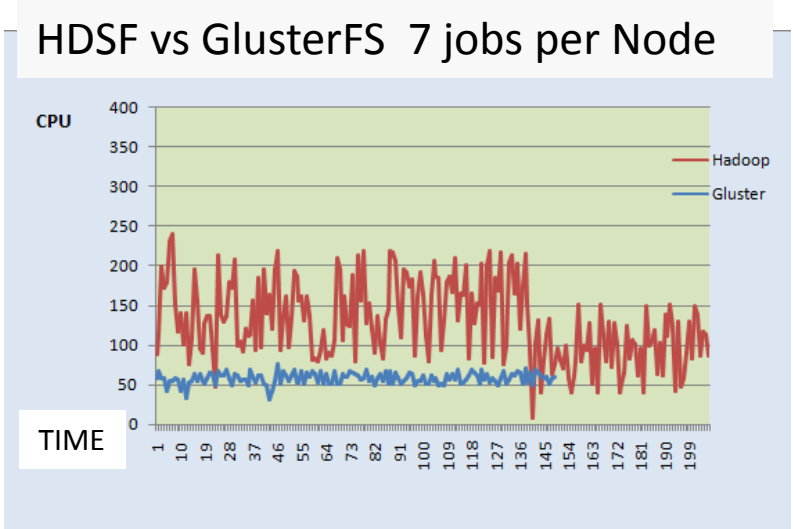



## 8 Jobs (max box size) max rate 333MB/s

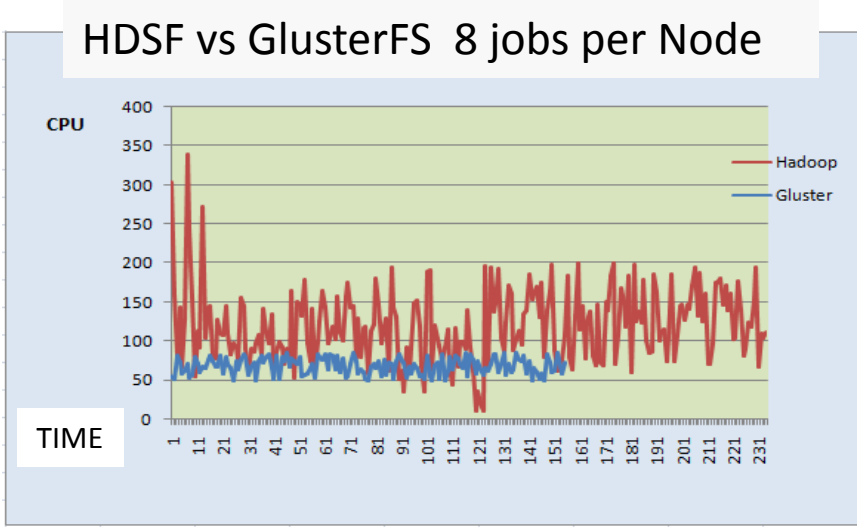


# 7/8 Job on 9 nodes

| File System | 7 Jobs | 8 Jobs |
|---|---|---|
| GlusterFS | 135s | 155s |
| HDFS | 205s | 225s |

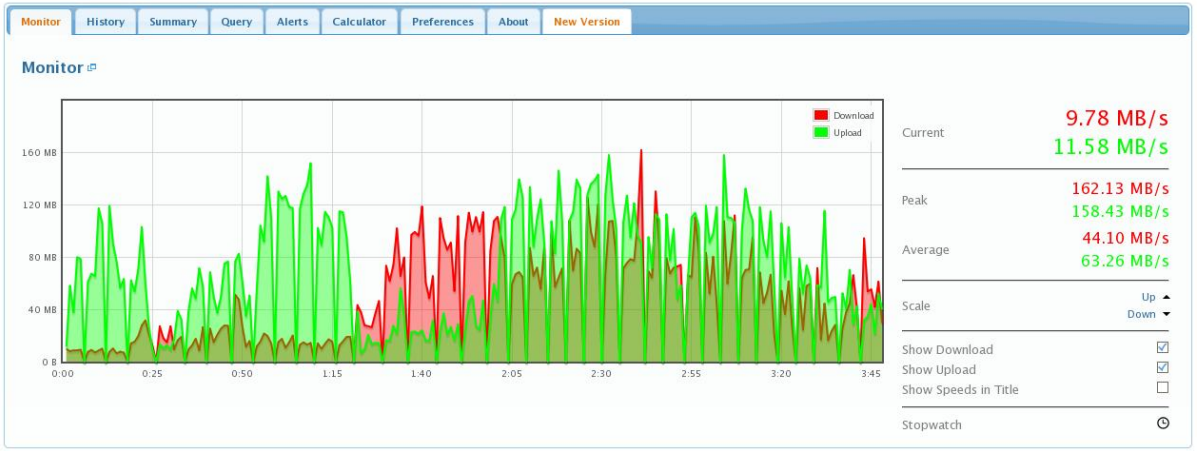


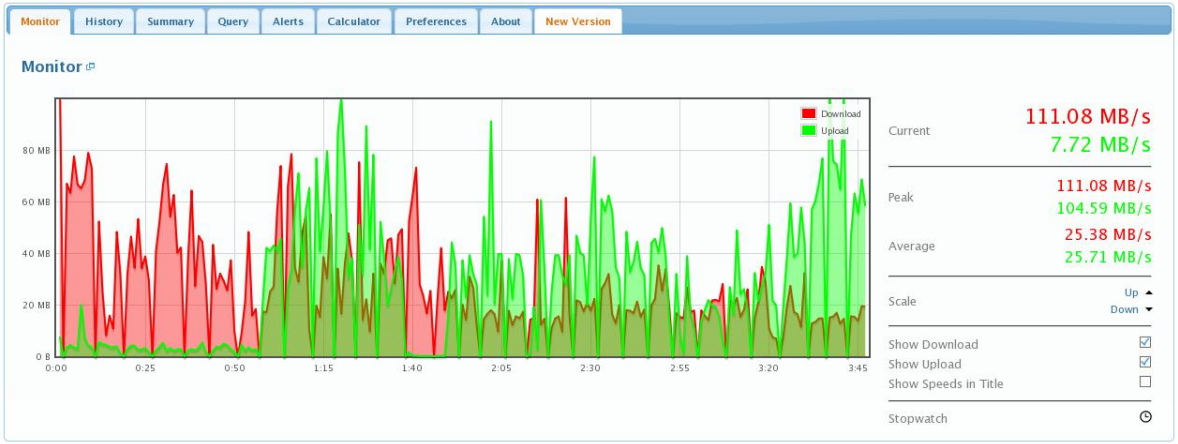

12% of core dedicate for storage




No core dedicate for storage

# Network Traffic




In this test we sent 8 analysis jobs for node

We achieve an Aggregate Network occupation of about 10Gbit/s In the Cluster .

# work on storage R&d

| | *GlusterFS* | *Hadoopfs* |
|---|---|---|
| *Resilient to failure* | ★★★ | ★★★★ |
| *Posix compliance* | ★★★★ | ★★ |
| *Performance* | ★★★★★ | ★★★ |
| *Community* | ★★★★ | ★★★★ |
| *Scalability* | ★★★★ | ★★★ |
| *Metadata performance* | ★★★ | ★★★★ |
| *Quota* | ★★★★ | ★★★★ |

Features

# Resume

Bandwidth doesn't represent a bottleneck but 10Gbit/s between nodes is needed.

For 48 Core Nodes we expect a throughput of 1.5GB per Node

File system choice is a crucial aspect to guaranteed the scalability

GlusterFS is a very interesting product that show promising feature, in term of reliability and Performance.

HDFS as general purpose FS doesn't show exiting performance .

# To Do

Test GlusterFS performance with other SuperB code and with other setup.

Compare Our setup With standard ATLAS like setup by implementing a Storage Element on 10Gbit/S Net

Test GlusterFS and HDSF for data movement

Test the HDSF and GlusterFS on geographical scale through the Naples, Bari and Catania sites.

# Interesting Issue

- Already available grid data movement solutions
  - Dirac
  - Phedex
  - PD2P

- This evaluation will be interesting in order to understand which one could be useful for SuperB computing model