I Tier-3 di CMS-Italia: stato e prospettive

Hassen Riahi
Claudio Grandi

Workshop CCR GRID 2011

Outline

- INFN Perugia Tier-3
  - Computing centre: activities, storage and batch system
  - CMS services: bottlenecks and workarounds
  - R&D
- INFN Bologna Tier-3
  - Computing centre: configuration, activities and resources
  - Functionalities
- Conclusions

INFN Perugia Tier-3

INFN Perugia computing centre activities

- Computing centre was setup since 2004
- Main experiments making use of the computing farm are CMS, Babar, SuperB, NA48/62, Theo, Erna, Glast, Ise, Virgo and Grid.
- ~ 50 active users for the last 6 months

Storage

- dCache based disk-only storage managing ~ 35 TB
  - 14 pools of ~ 2TB: equal size pools for better load balancing
  - 1 pool of 5 G dedicated for storage certification jobs
- 2.5 TB NFS for POSIX-like storage

Batch system

- Scheduler accepting local and Grid jobs
- Support of the local submission using CRAB

CHOSEN PARAMETERS

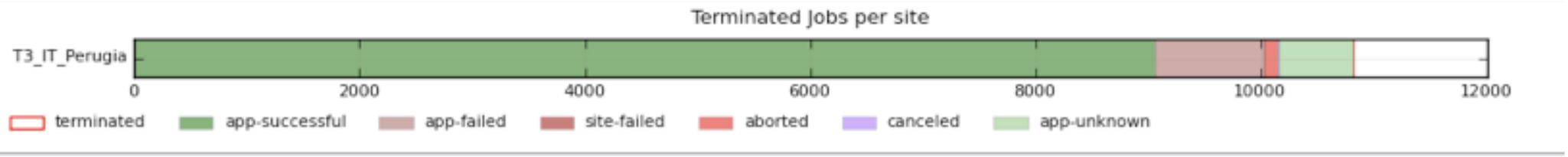
Sites:
T3_IT_Perugia

Activities:
all

Time Range:
From: 2010-10-01
To: 2011-04-30

Granularity:
Daily

Sort by:
Sites



➡ ~ 80 % of CMS analysis jobs are submitted locally using CRAB

- Torque/Maui scheduling:
  - 3 internal queues for each VO (VO-short, VO-medium and VO-long)
  - 1 Grid queue for each VO
  - 1 queue for certification jobs
- Maui fair share policy of ~ 260 cores in a private network (25% CMS and 75% others)

CMS services: bottlenecks and workarounds

Services

- Xen-based virtualization of services:
  - SE-dCache admin node,
  - LCG-CE (CREAM-CE on-going),
  - UI and SLC5 nodes
  - Phedex
    - ➡ more than 30 TB of data transferred since 2007
    - ➡ dCache is used as SE since 2007 and has allowed to use the entire bandwidth available
- Squid server to access conditions data,
- NoSQL database (couchDB) instance for CRAB development and tests.

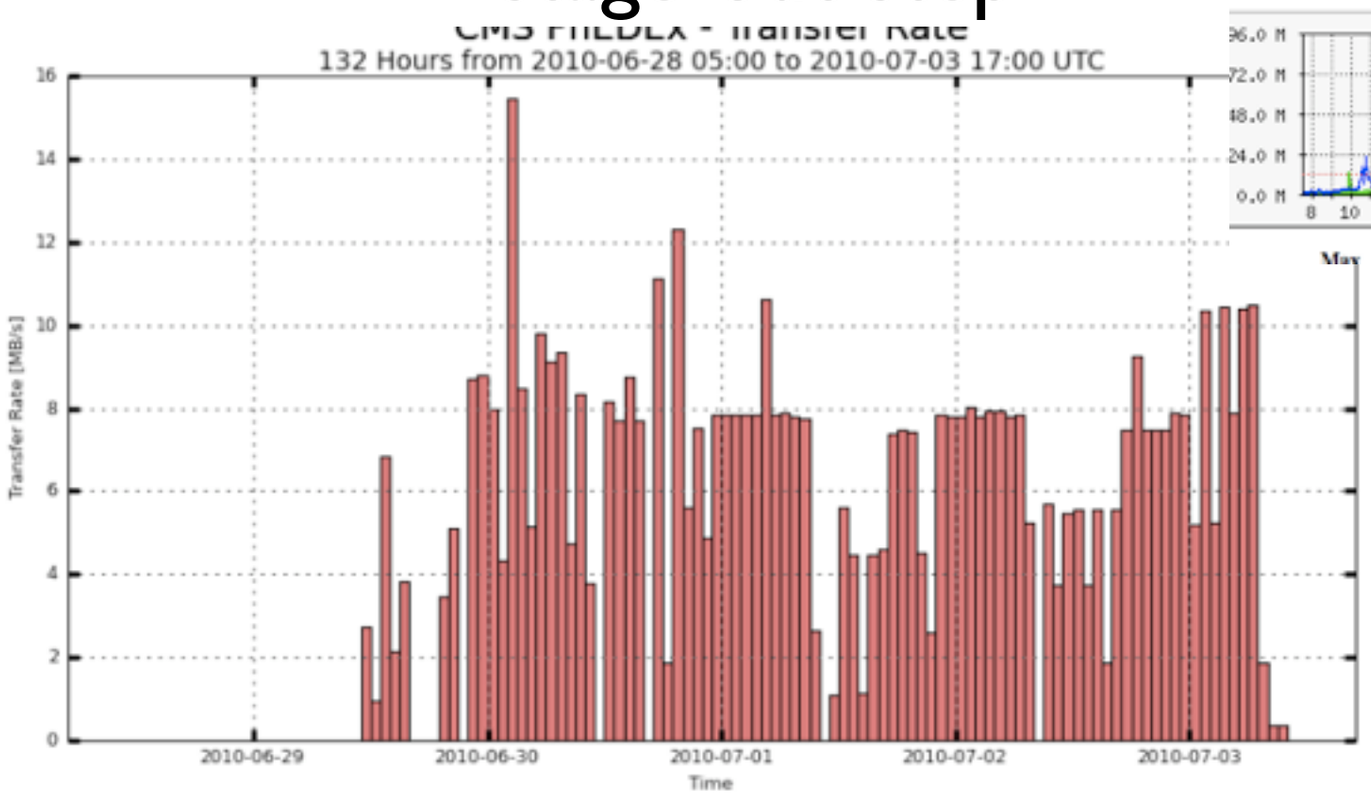
Bottlenecks

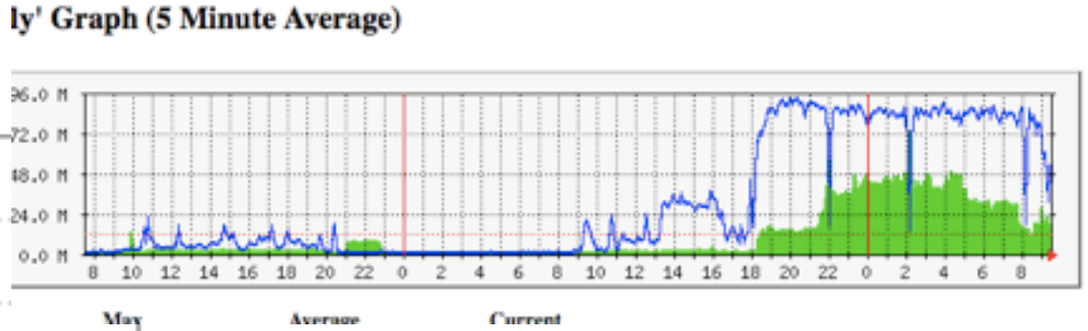
- Very small external bandwidth (100Mb)

➡ Low data transfer rate using Phedex

➡ Users are penalized by the use of the entire bandwidth

➡ Large number of CRAB jobs failing during the stage-out step





Workarounds

- **Phedex**

➡ Limit the farm traffic to 40 - 50 Mb during the working hours to avoid penalizing other users ➡ 2 Phedex configurations

- **CRAB**

➡ Limiting the GridFTP queue in dCache has reduced the failure rate by ~ 20%

➡ CRAB 3 will reduce much more the failure rate (stage-out asincrono)

➡ Tests the stage-out of users outputs in T2_IT_Bari and access them using **xrootd** from there (see Giacinto's talks):

➡ Promising results:

➡ From users point of view, the performances in accessing data residing in Bari are the same as accessing data residing in Perugia

➡ No known problems

R&D in Tier-3 Perugia

Storage optimization

- Looking for a storage system which can provide (compared to dCache):
  - better read performance
  - higher fault tolerance

In collaboration with CMS Tier-2 Bari (Giacinto Donvito)

➡ Data replica in Lustre is not easy to configure (see Giacinto's talk in CCR workshop 2010)

➡ Hadoop can provide needed performance and scalability by means of commodity HW

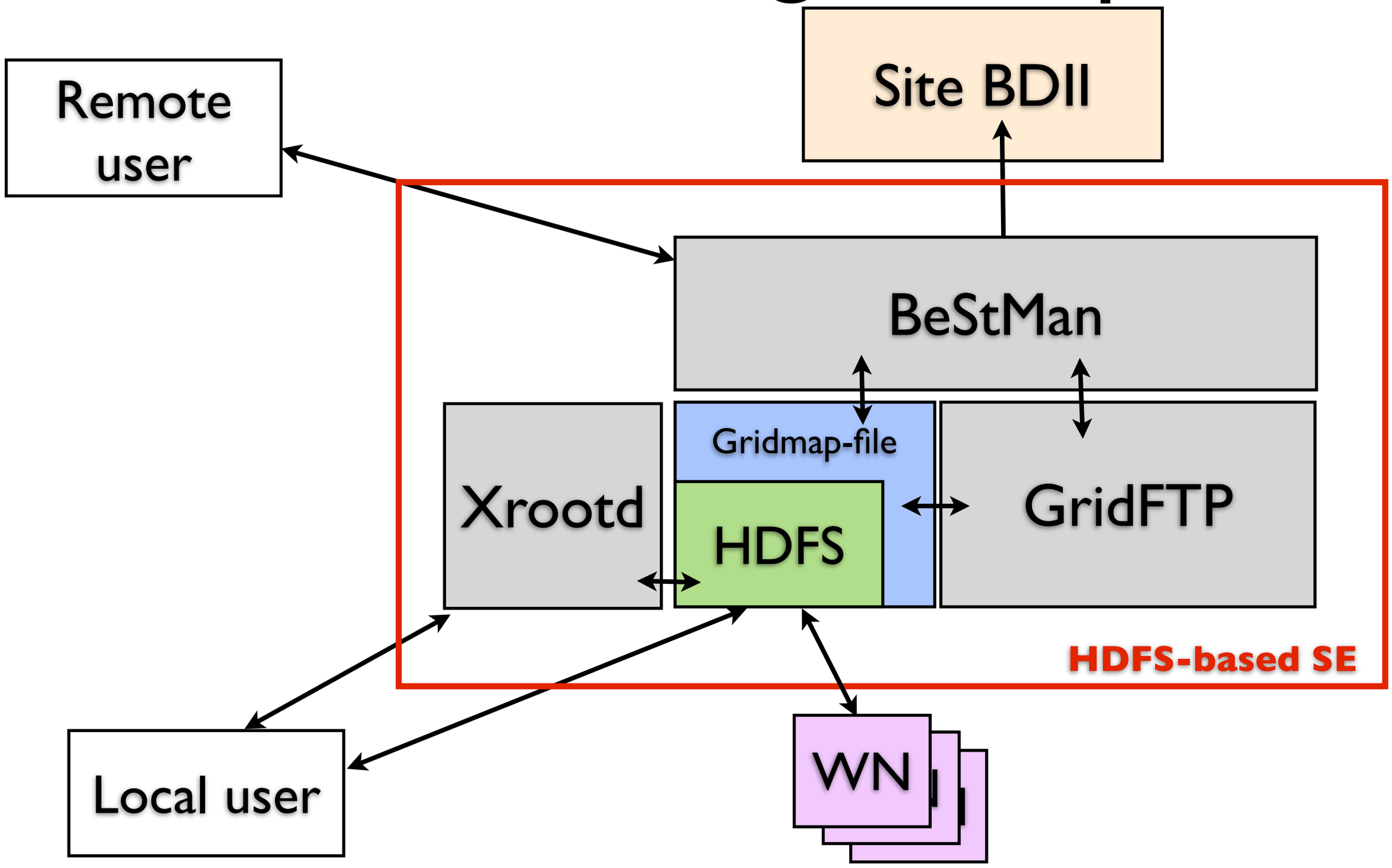
➡ Storm can work only with filesystems which support ACL and it is not the case of HDFS

➡ BeStMan can be configured to work with HDFS

Why Hadoop?

- Small/medium sites use commodity hardware more than large sites → exposed to hardware failure. HDFS provides:
  - ➡ Native data replication (block replication)
  - ➡ DataNode failure ~ transparent
  - ➡ Rack awareness
- Splitting files in different DataNode when storing them in HDFS → can improve the performance when reading them back.

HDFS-based SE schema - Tier-3 Perugia setup



Setup

1. GridFtp:
   - Compile HDFS plugin for GridFTP used in OSG sites: `http://t2.unl.edu:8094/browser/gridftp_hdfs/src/`
   - Start gLite GridFTP server to read/write from/to HDFS: `globus-gridftp-server -p 2811 -dsi hdfs -debug`
2. SRM v2.2:
   - Load fuse module in order to be able to mount HDFS (hadoop-fuse-0.19.1-15.el5)
   - Install BeStMan (2.2.1.3) and configure it using `--with-gridmap-path-local` and `--with-tokens-list` (to manage experiments storage area in HDFS) options
3. Install xrootd/hdfs: `rpm -ivh xrootd-1.3.2-6.el5.x86_64.rpm --nodeps` (rpms downloaded from here http://vdt.cs.wisc.edu/hadoop/testing/1.0/rhel5.5//x86_64/)
4. Provider script: publishes dynamic information by calling srm-ping command.

Future directions

- System deployment in INFN-Grid production:
  - Develop scripts for system configuration using INFN-Grid profile
  - Maintain the system
  - Monitor the system measuring its performance in the production environment
- System tuning:
  - Tune the GridFTP plugin
  - Tune the system configuration

Other R&D

- WN on demand: dynamic virtual machine manager for heterogeneous execution environment
  - Optimize the use of computing resources for specific purposes applications
- Performance optimization of Maui cluster scheduler
  - Development of a tool to predict the system behavior when a new set of maui configuration parameters is set

INFN Bologna Tier-3

INFN-Bologna Tier-3

- Joint project of INFN and Università di Bologna
  - Supports local research groups including LHC experiments
- Integrated with the INFN-CNAF Tier-1 infrastructure
  - ~ 50 dual quad-core boxes accessible through CNAF-LSF batch system and managed through the WNoDeS system
  - 150 TB on CNAF-GPFS storage system
    - home directories and software areas mounted through CNFS
  - The boundaries of the Tier-1 and Tier-3 will be virtual
    - Each site can expand into the other according to policies in LSF
    - The profile of the virtual machine defines to which site it belongs
  - Computing and Storage Elements are independent
- Differently from the Tier-1 also offers interactive services to local users
- The site is certified and currently is being commissioned for the experiments

Functionalities

- Standard Grid site for the supported experiments and for the local researchers of the Università di Bologna
  - E.g. for CMS: PhEDEx service, standard software installations, CMS SAM/Nagios tests, ...

Sitename	Service Type	Service Name	mc	sft-job	analysis	prod	basic	frontier	squid	swinst	cr-basic	cr-squid	cr-analysis	cr-sft-job	cr-swinst	cr-prod	cr-mc	cr-frontier	lcg-cp	user	get-pfn-from-tfc
T3_IT_Bologna	CE	cebo-t3-02.cr.cnaf.infn.it	ok	ok	ok	ok	ok	ok	ok	ok											
	CREAMCE	cebo-t3-01.cr.cnaf.infn.it									ok	ok	ok	ok	ok	ok	ok	ok			
	SRMv2	sebo-t3-01.cr.cnaf.infn.it																	ok	warn	ok

  - Local users support:
    - Direct access to a dedicated LSF queue
    - Read/Write access to a portion of the GPFS storage
    - Read access to the SE area (Write through SRM)
    - No access to the Tier-1 queues and storage!
    - Interactive access to a few nodes
    - R&D: Virtual Interactive Pools (VIP) UI on demand managed by the WNoDeS system

Conclusions

- We have described 2 different configurations of Tier-3 sites, each one has its own issues and development directions:

1. **INFN Perugia Tier-3:**

➡ Computing centre was setup since 2004

➡ Support of CMS services:

➡ CRAB development

➡ Local support of CRAB

➡ Data transfer using Phedex

➡ Small external bandwidth

➡ CRAB jobs failing during the stage-out step

➡ Tuning dCache configuration has reduced the failure rate

➡ CRAB 3 will reduce much more the failure rate

➡ First tests to access outputs stored at Tier-2 Bari using xrootd seem promising

➡ The functionalities tests of HDFS based SE in INFN-Grid environment are done with success.

Conclusions

2. **INFN Bologna Tier-3:**

- There is a clear benefit in term of hardware provisioning in being attached to a big site
  - High quality hardware and infrastructure
  - All basic services managed at the "Tier-1" level by CNAF people: GPFS, LSF, Squids, monitoring, ...
- The complexity increases and a good communication is needed
  - Strong division of responsibilities: often an intervention requires people both from CNAF and Bologna
  - Unforeseen interactions between components: need to be careful not to impact the Tier-1 operations
- Perfect playground for R&D activities
  - High quality site without explicit duties to the collaboration

Backup

CMS jobs submission

For the last 6 months

CHOSEN PARAMETERS

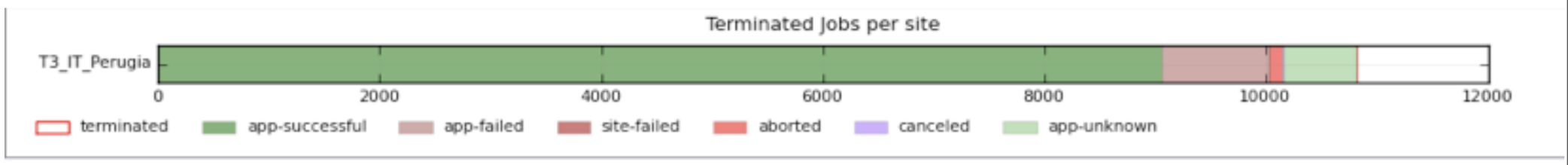
Sites:
T3_IT_Perugia

Activities:
all

Time Range:
From: 2010-10-01
To: 2011-04-30

Granularity:
Daily

Sort by:
Sites



- Local jobs submission using PBS plugin of CRAB and Grid jobs submission are both supported

➡ ~ 80 % of analysis jobs are submitted locally using PBS plugin of CRAB

➡ Only ~ 50 % of CMS jobs submitted locally uses PBS plugin of CRAB