DPM Italian sites and EPEL testbed in Italy

Alessandro De Salvo (INFN, Roma1), **Alessandra Doria** (INFN, Napoli),
Elisabetta Vilucchi (INFN, Laboratori Nazionali di Frascati)

Outline of the talk:

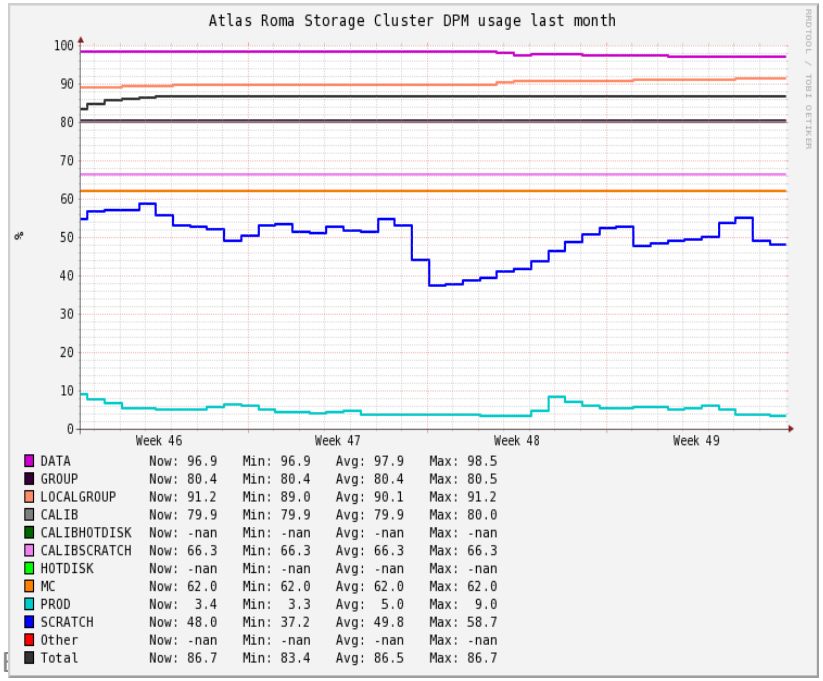
- DPM in Italy
- Setup at the DPM TIER2 sites
- Setup of the EPEL testbed
- EPEL testing activity
- Other activities related to storage access in Italy

Italy in the DPM collaboration

- Storage systems at ATLAS italian sites:
  - DPM at Frascati, Napoli, and Roma Tier2.
  - STORM at CNAF Tier1 and at Milano Tier2.
- The 3 people involved in the DPM collaboration are the Site Managers of the 3 DPM Tier2s.
- Alessandro D.S. is the ATLAS Italian Computing Coordinator.
- Some smaller sites also use DPM.

DPM setup at INFN-Roma1

- > 1 PB of disk space
  - 18 disk servers, attached to 7 SAN systems (used in direct attach mode) via 4/8 Gbps Fibre Channels
  - All servers equipped with **10 Gbps** Ethernet connection to a central switch
  - WAN connection via **LHCONE at 10 Gbps**
- Installation
  - Supervised by Puppet (some machines already using Foreman + Puppet)
  - Currently using a custom **puppet module executing yaim**
  - Will move to full puppetized configuration soon, the infrastructure is ready
- Backup Policy
  - MySQL main replica in the DPM head node
  - **Slave replica** on a different node
  - Backup, daily snapshots from the slave node
  - Currently evaluating the possibility to move the database to our Percona XtraDB cluster (6 DB machines on-line), already used for the ATLAS global Installation System, hosted in Roma.
- Monitoring
  - Ganglia/nagios via custom plugins
  - Will add the DPM standard plugins soon



DPM setup at INFN-Napoli

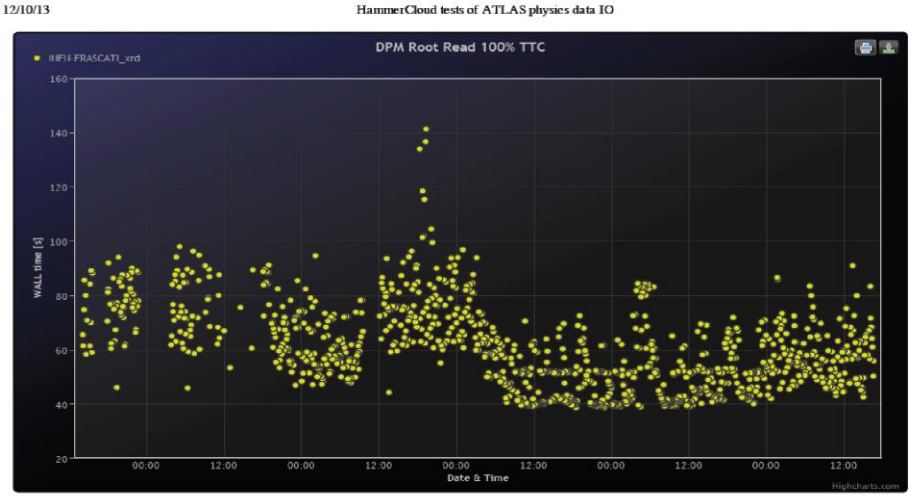
- 1.2 PB of disk space
  - 22 disk servers, attached to 9 SAN systems (used in direct attach mode) via 4/8 Gbps Fibre Channels.
  - All servers connected to a central switch with **10 Gbps** Ethernet FC .
  - WAN connection via **LHCONE at 10 Gbps**.
- Setup
  - DPM release **1.8.7-3** on SL5, **yaim** configuration
  - MySQL DB (rel. 5.0.95) is on the same server as the DPM head node.
- Backup policy
  - MySQL DB daily backups, with Percona xtrabackup, are saved for 7 days.
  - Full backup of the the whole head node (DB included) is done by rsync twice a day on a secondary disk of another server. In case of hw failure, the other server can boot from this disk, starting an exact copy of DPM head node, not older then 12 h.
- Monitoring
  - Ganglia/nagios via custom plugins

DPM setup at INFN-FRASCATI (production site)

- 700 TB of disk space
  - 7 disk servers, attached to 7 SAN systems (used in direct attach mode) via 4/8 Gbps Fibre Channels
  - All servers equipped with **10 Gbps** Ethernet connection to a central switch
  - WAN connection via LHCONE at 10 Gbps soon (by the end of Feb '14)
- Setup:
  - DPM release 1.8.6. Upgrade to 1.8.7 by the end of the year
  - DPM DB on a separated server: MySQL 5.0.95
- Backup
  - DPM DB replication with MySQL master/slave configuration;
  - If MySQL server crashes it's enough to change the DPM_HOST variable in the DPM head node with the MySQL slave hostname and run yaim configuration;
  - Slave DB daily backup with mysqldump;
- Monitoring
  - Ganglia/Nagios custom plugins

EPEL testbed setup at INFN-Frascati

- Setup
  - DPM head node: atlasdisk1.lnf.infn.it
  - The head node is installed on a VM (2 cores, RAM 4GB)
  - the MySQL server is on the physical server (8 cores, RAM 16GB) .
  - Only one disk server with 500GB of disk space.
  - 1Gbps links for LAN and WAN connections.
- Installation
  - SL6.4 , DPM 1.8.7-3
  - yaim configuration, to be moved to puppet a.s.a.p.
  - XRootD, WebDAV/https enabled
- Added to Wahid **Hammercloud test** page .



EPEL testing activity

- Installed at end of summer 2013 for the first time, too late for the first EPEL release validation !
- Re-installed again in November ’13 from EPEL +EMI3 SL6 repos:
  - EPEL-testing repo enabled, no autoupdate.
  - Installation guide followed step by step https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/Admin/Install to check its consistency.
  - Very easy, no problem found: only a reference to XRootD setup could be useful.
- Ready to test upgrades or reinstallation of new releases from scratch , according to the collaboration needs.
- The HW is quite old and Ethernet connection is 1Gbps:
  - testbed tailrored for functionality tests
  - HW improvements for performance tests could be planned if needed.
- We are planning to test a High Availability setup.

DPM in HA

- The DPM head node is virtualized
  - Need to evaluate the performance impact on this
- Two different ways of virtualization possible
  - Standard virtualization (KVM) over a shared filesystem (GlusterFS or Ceph)
  - OpenStack-based virtualization
- Both solutions are possible, more experience in Italy with the first option
  - Will try to use either GlusterFS with GFAPI or Ceph for performance reasons, need very recent versions of QEMU and libvirt
  - We also have an OpenStack Havana testbed in Roma, we will experience on that (it is also GlusterFS GFAPI enabled)
- Database in HA mode via Percona XtraDB cluster
  - Using (multiple instances of) HAproxy as load balancer
- Head node HA
  - Pacemaker/Corosync if standard virtualization
  - Still have to be discussed for the Cloud option
- Pool nodes
  - May want to experience with GlusterFS or Ceph for the testbed
- Configuration operated via Foreman/Puppet

ATLAS data access at DPM sites

- **Different access modes** can be set for the queues in Panda, the ATLAS workload management system for production and distributed analysis.
- For DPM sites the usual configuration is:
  - Analysis jobs access **directly local input files via XRootD**
  - MC production at DPM sites copies input files to WNs.

- The 3 DPM Italian Tier2s and the Tier1 are part of **FAX**
  - **Federated ATLAS storage systems using XRootD.**
  - FAX uses redirection technology to access data in the federated systems.

- The use of FAX can be enabled in Panda, with ad-hoc configuration of the Panda queues.
- FAX redirection allows the failover to storage systems in different sites in case a required file is not found locally.
- The FAX redirection in Panda is not yet enabled for the Italian sites.
- Some tests with FAX will be showed .

PROOF on Demand (PoD) and tests with FAX in the ATLAS IT cloud

"PROOF-based analysis on the ATLAS Grid facilities: first experience with the PoD/PanDa plugin", E. Vilucchi, et al.

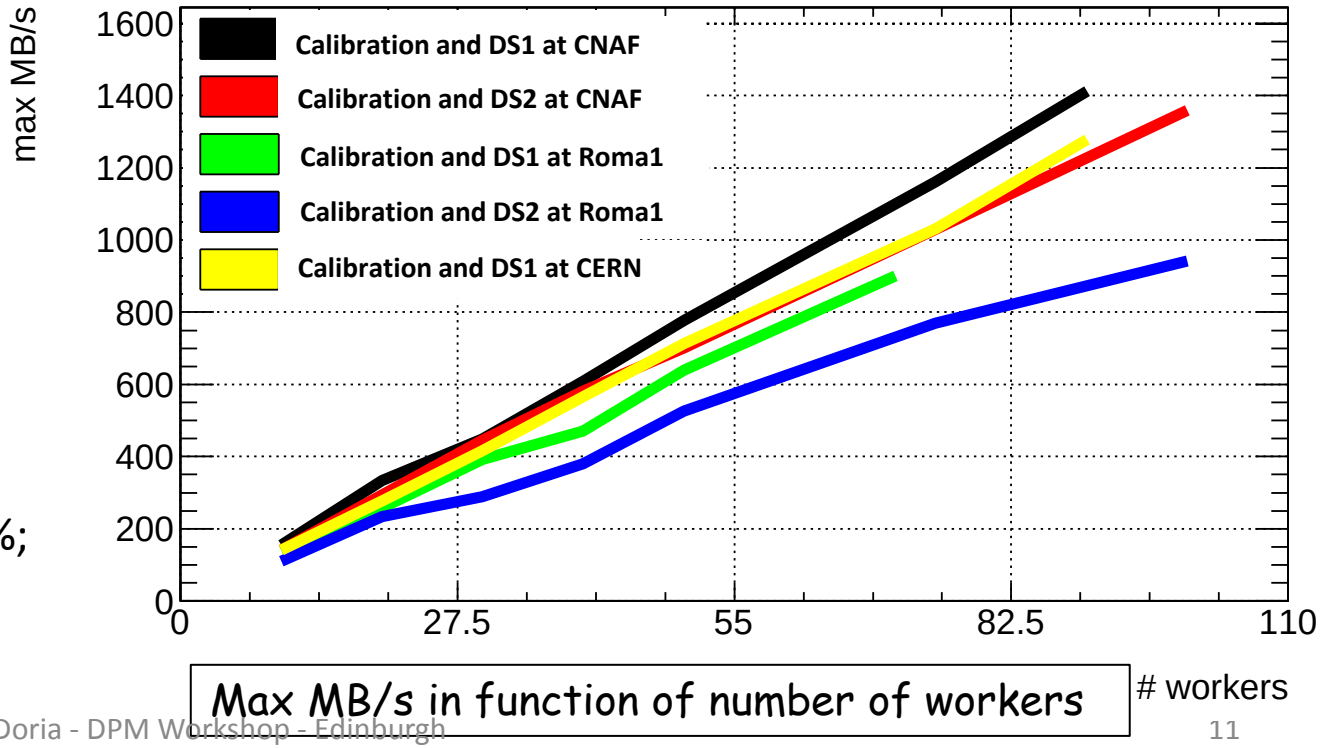
We presented this work at **CHEP2013** about the use of **PoD**, a tool enabling users to allocate a PROOF cluster on Grid resources to run their analysis.

- We tested the new **Panda plugin for PoD,** using real analysis examples, in the ATLAS Italian cloud and at CERN.
- We also run **performance tests** over the DPM, StoRM/GPFS and EOS storage systems and with different access protocols:
  - XRootD both on LAN and WAN;
  - file protocol over StoRM/GPFS;
  - XRootD on Storm/GPFS and EOS over WAN;
  - FAX infrastructure was tested for XRootD access;
  - Few tests of *root* access with HTTP for DPM.

Results of local data access tests

- Calibration tool performs only data access (no computation load):
  - It's a PROOF job reading the whole content of each entry of the TTrees from the files.
  - To check the site efficiency before to run analysis with PoD;
  - To obtain an upper limit on the performance in terms of MB/s as a function of number of used worker nodes.
- Sample datasets of real analysis used as input with PoD
  - *DS1 (100 files), DS2 (300 files):* standard *D3PD* dataset (~100KB per event).

- STORM/GPFS and EOS have comparable performance.
- *EOS and GPFS file* access perform better, but the difference with DPM XRootD is only about 25%;



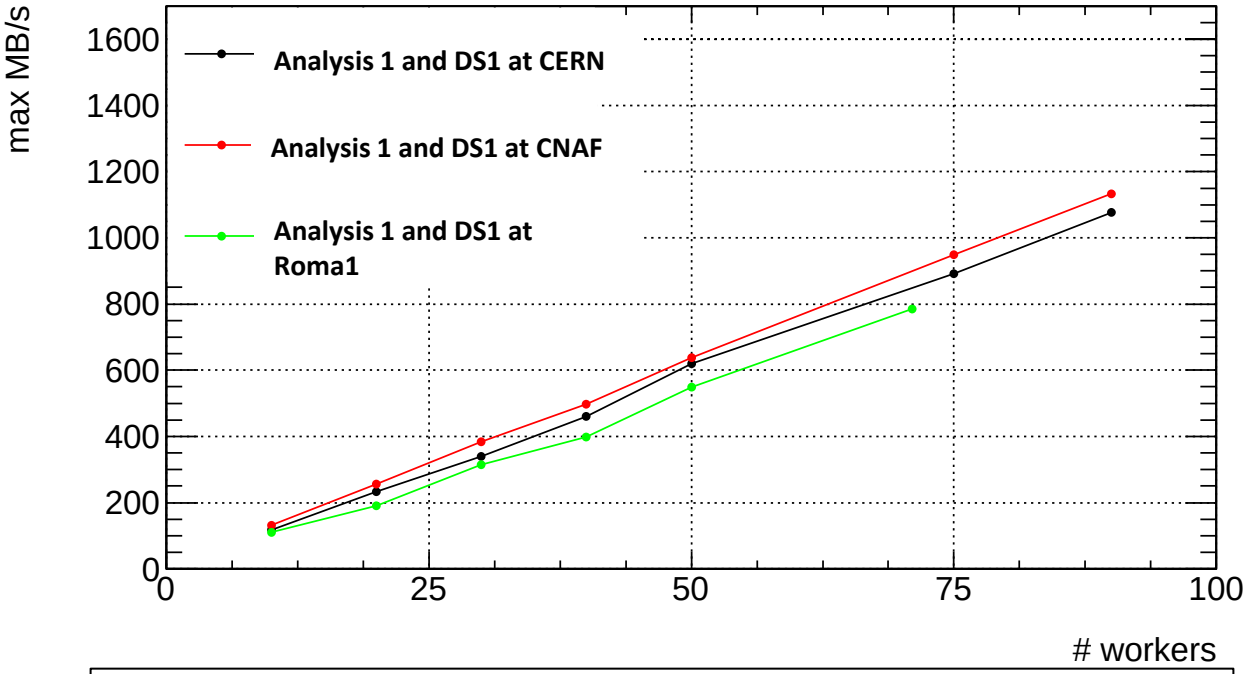
Max MB/s in function of number of workers

Analysis tests

- Analysis tests with three real analyses,
  - on different input datasets (D3PD, user-defined ntuples)
  - with different event processing load.
- I/O performance scalability at CERN, CNAF, Roma1
- Compared with the results of the calibration, in this analysis the weight of the additional event processing is about 16%.

- The performance scales linearly in a cluster up to 100 WNs, no bottleneck found due to data access.

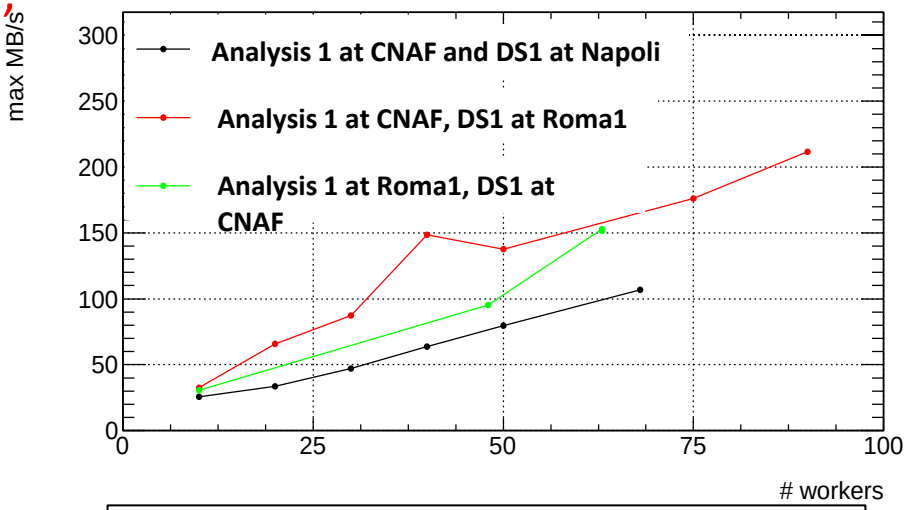
- Some preliminary tests with http protocol gave slightly lower performance, but comparable with XRootD.



Max MB/s versus number of workers running analysis 1

Remote XRootD access over WAN and FAX infrastructure

- We run analyses with PROOF cluster and input datasets in different sites.
  - Direct XRootD access (without *FAX*);
  - *FAX* redirection.
- *FAX* works smoothly. Accessing, through a central *FAX* redirector, to an input container distributed across different sites, the request is forwarded to the appropriate storage elements, selecting the local Storage when possible
- the mechanism is transparent for the user.
- Performance depends on network connections.



Max MB/s in function of # of WNs for analysis 1 with DS1 over WAN

WebDAV/http Federation

- An http(s) redirector has been installed at CNAF.
- Deployed with Storm on a test endpoint
  - http(s) in Storm now supports WebDAV
- With DPM the http(s) access has already been tested and it works without problems.
- The LFC interface has been enabled for the http redirector
  - Which LFC is addressed? CERN or CNAF?
- When ATLAS will switch from LFC to Rucio, a plugin interface for Rucio will be needed .
- Shortly the redirector will use the production endpoint at CNAF.
  - DPM sites will be ready to join the http(s) federation.
- No monitoring yet
- To be tested with ROOT
  - the ROOT support is limited until now for https, TDavixFile needed .

Conclusions

- We are happy with our DPM systems and with the latest DPM developments.
- A ot of interesting work can be done in testing new solutions with DPM.
- Thanks to the DPM collaboration!