

# 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

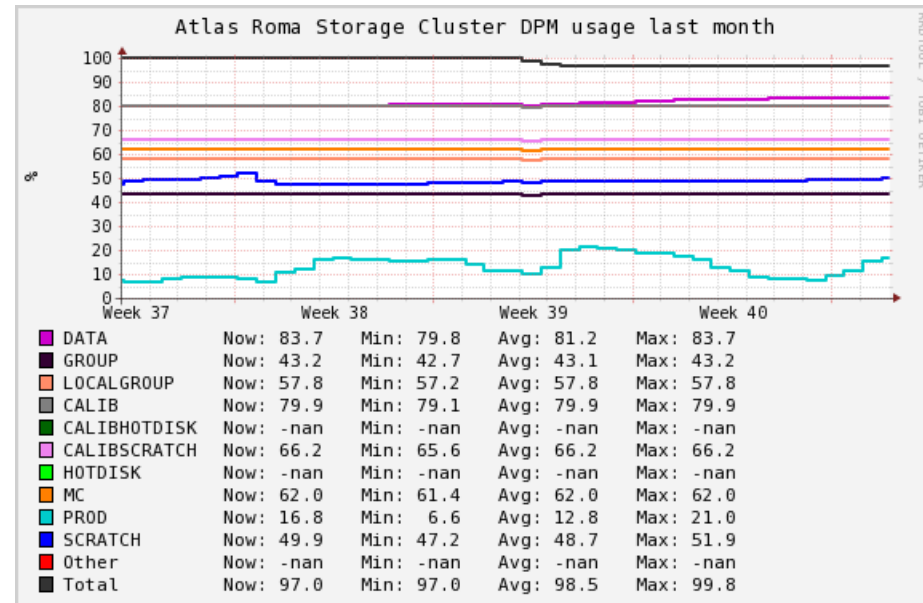
# Italy in DPM collaboration



- 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.

# DPM setup at INFN-Roma1 Tier2

- **1.3 PB of disk space**
  - 16 disk servers, attached to 6 SAN systems (used in direct attach mode) via 4/8/16 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 Foreman + Puppet
  - Currently using a custom **puppet module executing yaim**
  - Will move to full puppetized configuration when all the modules will be officially available from WLCG, 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
- **Monitoring**
  - Ganglia/nagios via custom plugins
  - Will add the DPM standard plugins



# DPM setup at INFN-Napoli Tier2

- ~ 2 PB of disk space (700 TB recently acquired by the resources of the RECAS project, see next slide)
  - 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 **1.8.8-4** on SL6 for head node, still SL5 on disk nodes .
  - MySQL DB (rel. 5.1.73) 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 than 12 h.
- **Monitoring**
  - Ganglia/nagios via custom plugins

# DPM at Napoli RECAS infrastructure



investiamo nel vostro futuro

- Funded in November 2011 by the Italian Ministry of Education, University and Research in the framework of the National Operational Program
- The purpose of the project was to empower and federate four Data Centers in the South of Italy to build a distributed infrastructure, based on both the Grid and Cloud Paradigms.
- Initially the main goal was to create the computing infrastructure for the SuperB experiment, to be built in Italy; as the SuperB project was cancelled in Oct 2012, the LHC experiments became the new major users, together with Belle2, KM3Net and other physics projects.
- The distributed infrastructure built with the project fundings consists of 120 racks, with a total availability of more than 10000 cores and 5.5 Pbyte of storage, distributed in the four Data Centers located in Bari, Catania, Napoli and Cosenza.
- 700 TB of storage were recently acquired by the RECAS project and added to INFN-NAPOLI-ATLAS storage system, to be used by Napoli ATLAS Tier2 .
- About 900TB more are being installed with DPM in the RECAS-NAPOLI site, to serve the other VOs supported by the project.

# DPM setup at INFN-FRASCATI Tier2

- **850 TB of available disk space**
  - 9 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
- **Setup:**
  - DPM release 1.8.8-4
  - 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 at INFN-Frascati

Installed at the end of summer 2013 for the first time, re-installed in Nov '13 from EPEL +EMI3 SL6 repos (DPM 1.8.7-3).

## Setup

- DPM head node: [atlasdisk1.lnf.infn.it](http://atlasdisk1.lnf.infn.it)
- The head node is installed on a VM (2 cores, RAM 4GB)
- the MySQL server is on the physical server
- Only one disk server with 500GB of disk space
- 1Gbps links for LAN and WAN connections

## Installation

- SL6.4 , DPM 1.8.9 just installed
- Puppet configuration with Foreman (in Roma1)
- XRootD, WebDAV/https enabled

- April '14 - new EPEL-testing release DPM 1.8.8 tested:
  - Upgrade from 1.8.7-3 to 1.8.8 with yaim reconfiguration
  - Parametrized Puppet module developed
    - Fully generic module, using Foreman as ENC, can be used to install and configure any site
    - For the moment just the ATLAS xrootd federation/setup is supported, but the rest can be easily added
    - New puppet installation from EPEL-testing
- May '14 – Roma production site upgraded as soon as 1.8.8 available in EPEL

# EPEL testbed at INFN-Roma1

Instantiated at the end of April 2014 for the first time, based on OpenStack. Still “private” testbed but fully functional and re-installable very quickly.

## Setup

- DPM head node (atlas-dpm-01.roma1.infn.it) + 1 Disk Pool (atlas-dpm-pool-01.roma1.infn.it) running on OpenStack (persistent nodes on cinder volumes)
- MySQL hosted in the headnode VM, but planning to use the existing Percona external DB cluster (8 nodes available)
- Currently 100G of cinder (gluster) volume via gfapi, to be extended to 1T soon
- 1Gbps external connectivity via OS neutron (VLAN)

## Installation









- CentOS6.5, but can easily test other distros
- DPM 1.8.9 just installed
- Puppet configuration and foreman integration
- XRootD, WebDAV/https enabled
- ACLs on the border router to be opened soon

- Same configuration (and same foreman/puppet) used as for the Frascati site



## Hosts

dpm x Q Search ▾ New Host

<input type="checkbox"/>	Name	Operating system	Environment	Model	Host group	Last report	
<input type="checkbox"/>	 atlasdisk1.lnf.infn.it	 Scientific...	production	KVM	ATLAS_LNF/DPM_T...HeadNode	2 minutes ago	Edit ▾
<input type="checkbox"/>	 atlas-dpm-01.roma1.infn.it	 CentOS 6.5	production	OpenStack Nova	ATLAS/Service/D...HeadNode	22 minutes ago	Edit ▾
<input type="checkbox"/>	 atlas-dpm-pool-01.roma1.inf...	 CentOS 6.5	production	OpenStack Nova	ATLAS/Service/DPM_T...Disk	23 minutes ago	Edit ▾
<input type="checkbox"/>	 atlaswn091.lnf.infn.it	 Scientific...	production	PowerEdge 1950	ATLAS_LNF/DPM_T...DiskNode	2 minutes ago	Edit ▾



# EPEL testing activity

- Oct '14 - EPEL testing release 1.8.9 ongoing tests:
  - Upgrade of both Frascati (Physical Hw) and Roma1 testbeds (OpenStack)
- No noticeable problem encountered in DPM upgrade, both for head and disk nodes
- Some parameter fixed in the parametrized puppet modules
- Few small troubles with dpm-xrootd upgrade:
  - Dependency error: dmlite-libs doesn't update, even if new release 0.7.0 is present in epel-testing
  - Inaccuracies in twiki pages notified to developers

# DPM in HA

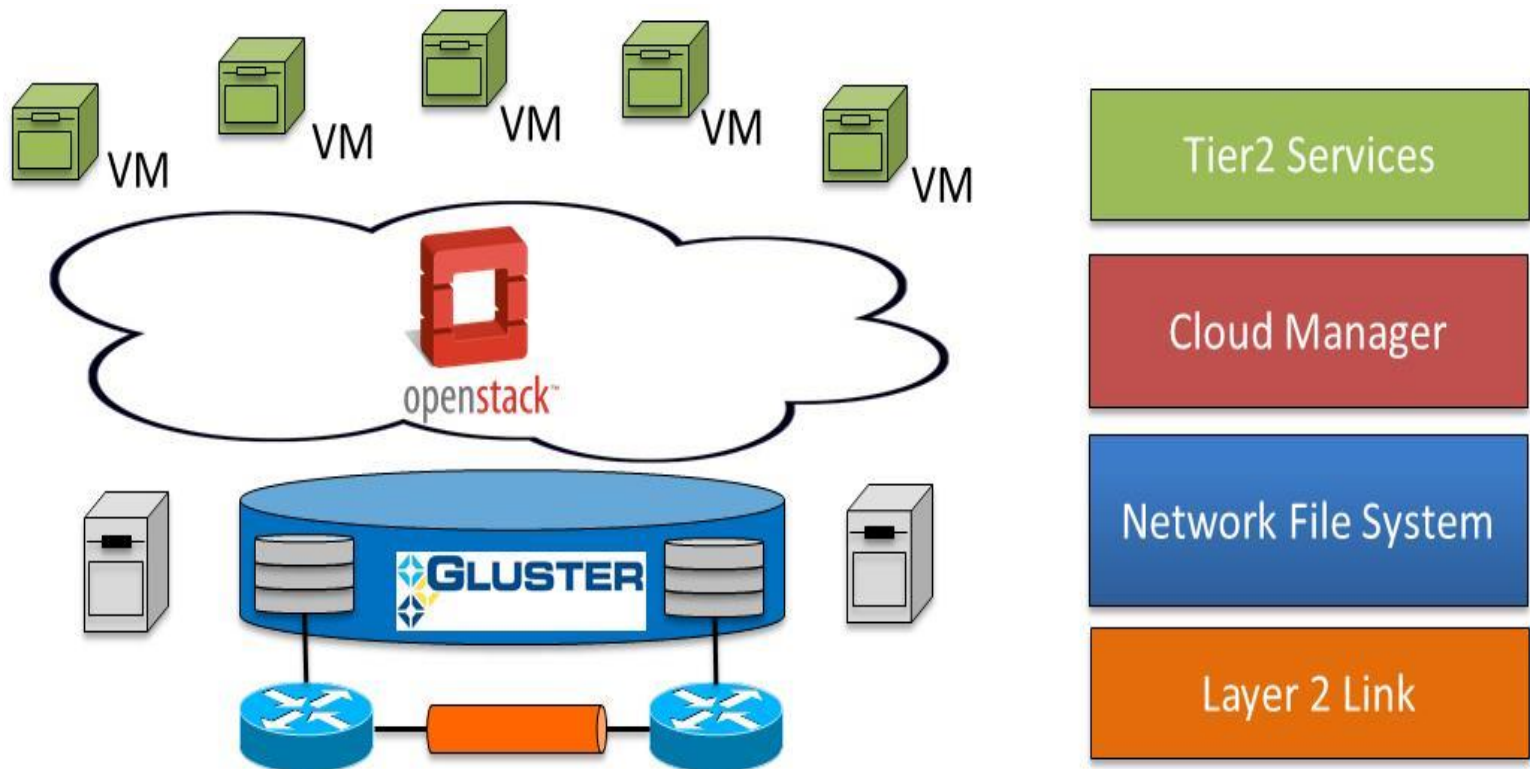
- The DPM head node can be virtualized
  - Need to evaluate the performance impact on this
- Can be done via OpenStack
  - Already using OS in Roma and soon in Napoli
  - Can be set up both in failover or load balancing, if using a distributed DB like Galera/Percona
  - Already some experience with load balancing in OS
  - Using GlusterFS with GFAPI or Ceph for performance reasons
  - Working DPM testbed in Roma, running on OpenStack
- Database in HA mode via Percona XtraDB cluster
  - Using (multiple instances of) HAproxy as load balancer, can be integrated with OpenStack
- Head node HA
  - Using OS Heat
- Pool nodes
  - May want to experience with GlusterFS or Ceph for the testbed

# http/WebDAV federation

- An http(s) redirector has been installed at CNAF.
- Deployed with Storm on a test endpoint
  - http(s) in Storm now supports WebDAV
  - ...but there are performance problems, due to logfile flooding
    - The developers are working to solve this issue
- With DPM the http(s) access has already been tested and it works without problems
- No LFC interface is anymore used in ATLAS, but still we need a Rucio interface for HTTPS
- No monitoring yet
- Tested with ROOT and direct file access
  - TDavixFile, plain HTTPS file, etc (Vincenzo Lavorini)
  - Rucio file renaming (all sites, including Storm)

# The Napoli-Roma distributed T2 prototype

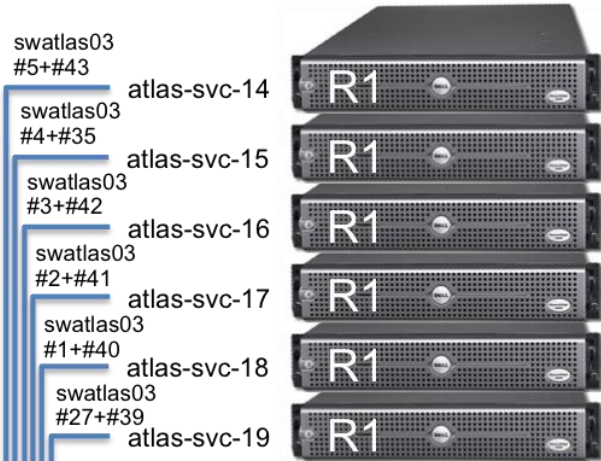
- **Dedicated L2 network link @ 1 Gbps between NA-RM**
  - Used to build up a distributed T2
  - Backbone infrastructure based on (synchronous) Gluster replicated storage for VM instances
  - OpenStack Cloud infrastucture for services (currently only in RM, to be expanded to NA)
  - Possibility to extend the link bandwidth and technology (using MPLS) to join more sites



# ATLAS RM1 Cloud Setup



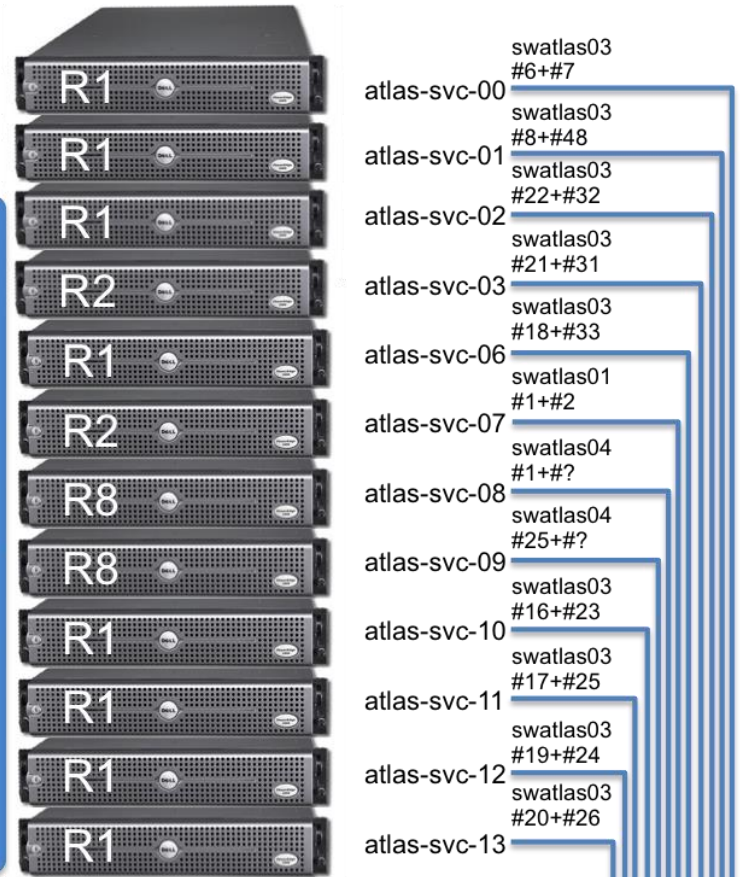
## Cloud Compute Nodes



## Cloud Controllers



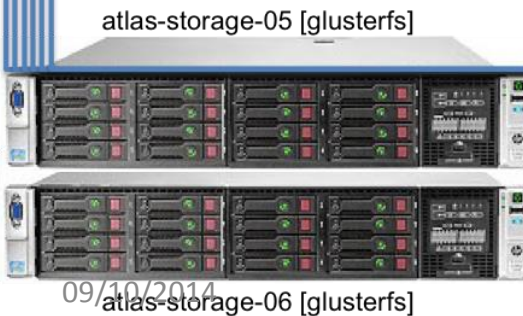
Pacemaker / Corosync / Percona XtraDB



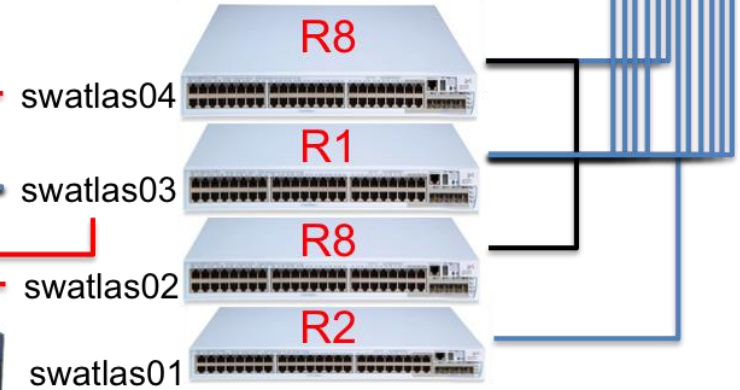
[atlas-foreman.roma1.infn.it](http://atlas-foreman.roma1.infn.it) → [atlas-svc-07.roma1.infn.it](http://atlas-svc-07.roma1.infn.it)



- Copper 1 Gbps
- Fibre 10 Gbps
- CX4 10 Gbps



Ex-4500 #?  
Ex-4500 #?



# DPM prototype on the distributed T2

- Our goal is to start cloudifying a distributed DPM
- Headnode
  - Cloudification of the headnode and HA/load balancing with heat
  - Distributed/replicated DB using Galera/Percona
- Pool nodes
  - Same as the headnodes, but locality with the storage is needed here
- Storage
  - Too expensive to add redundancy to the full disk space
  - We'll stick to the scenario where the disks will NOT be replicated, but only the services will be
    - It means that in case of failure we'll have some of the files not accessible, but not all the files