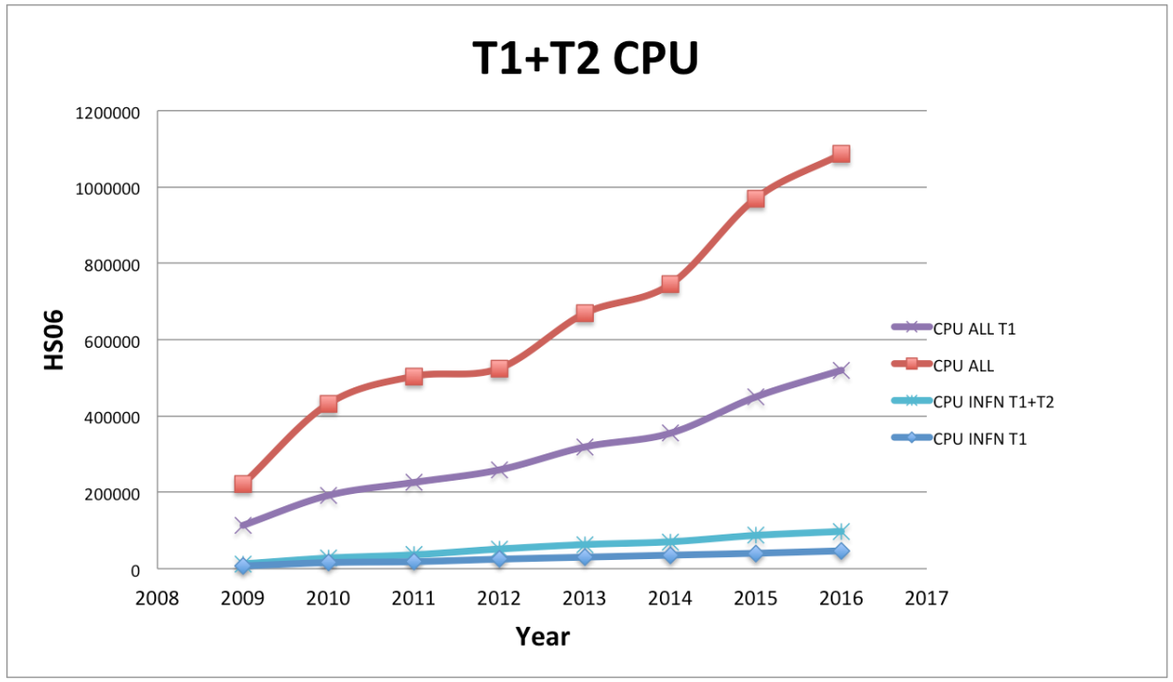
IT Cloud site perspectives

Alessandro De Salvo

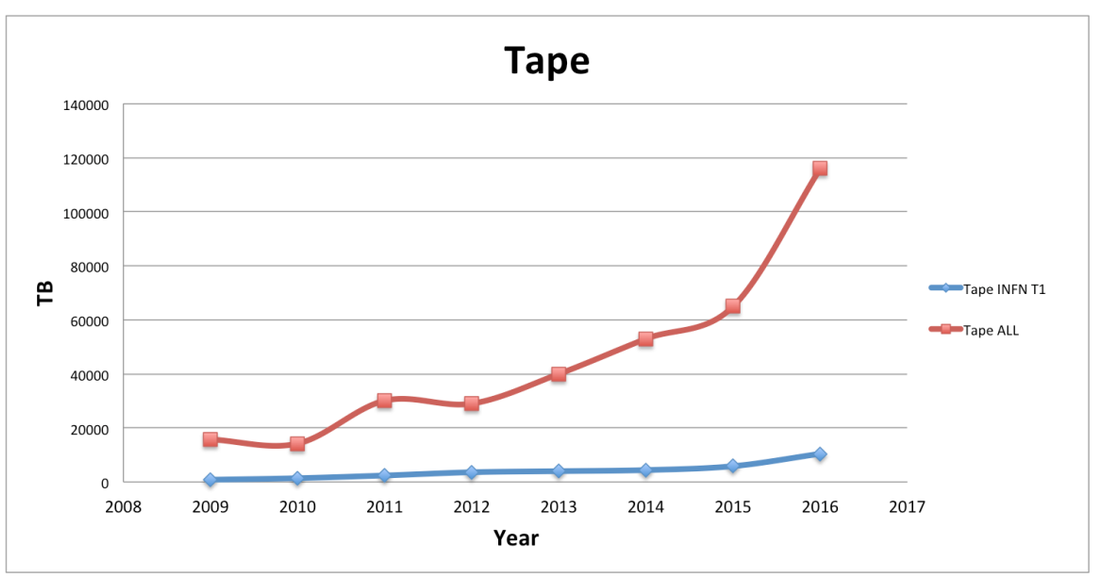
ATLAS Sites Jamboree, 27-1-2016

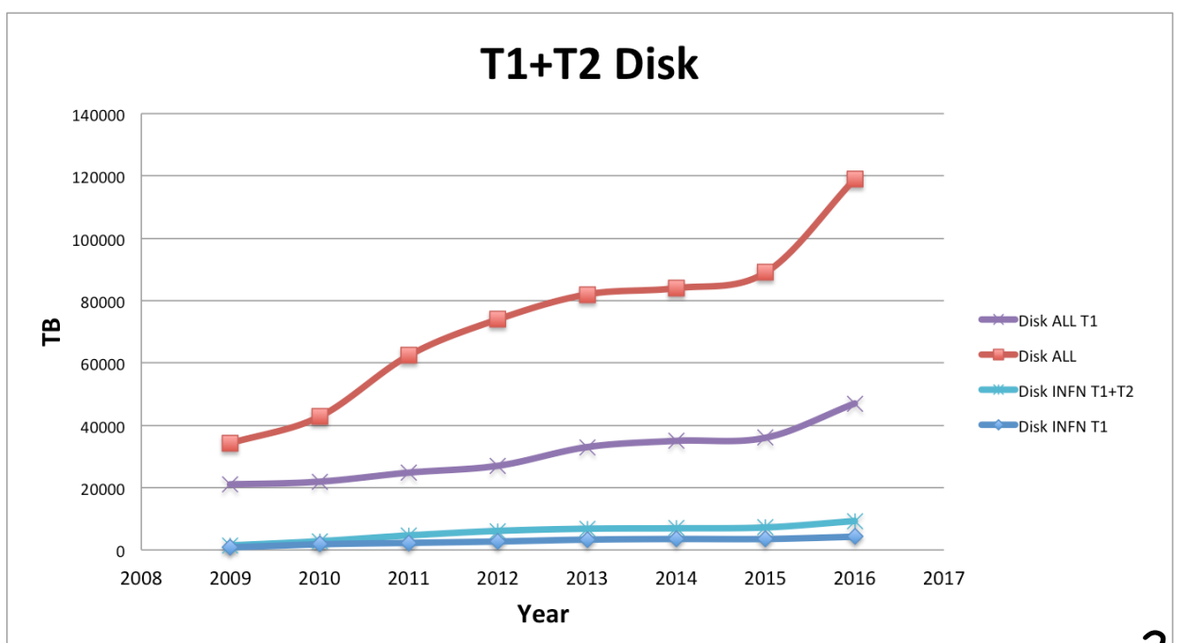
IT Cloud sites & resources

- **11+4 sites**
  - 1 T1: CNAF
  - 4 T2s: Milano, Frascati, Roma, Napoli
  - 6+4 T3s:
    - Genova, Bologna, Roma2, Roma3, Lecce, Cosenza
    - South Africa: WITS, UJ
    - Greece: Auth, Kavala
  - 1 T3 retired at the end of 2015 (Pavia)



- **Pledged resources in 2016**
  - 46.8 (T1) + 50.9 (T2) = 97.7 kHS06
  - 4.2 (T1) + 5.0 (T2) = 9.2 PB Disk
  - 10.4 PB Tape





IT Cloud funding

- **Budget discussed and defined every year**
  - In the past years we had some help in the funding from external projects (e.g. the Recas project in Napoli and Cosenza), but now they are over
  - Not clear what the future will be, trying to keep up with the model but it's not completely guaranteed we will be able to
- **We try to keep up with the flat budget**
  - But over pledge CPU resources may not be totally available, due to a lack of prompt replacement in the sites
  - Partial replacement of the old CPUs in the Italian T2s since 2014
    - Partially replacing in2015 the CPUs acquired in 2011 (3 years of maintenance), the rest will be replaced in 2016
    - No CPU replacement in 2016
    - But since 2014 we are acquiring CPU resources with a cycle of 4 years of maintenance
  - No changes in the disk space policy
    - All pledged disk fully under maintenance

IT Cloud sites organization

- **Cloud organization**
  - **Two-level support**
    - **High level site and user support**
    - **Site level support**
  - **Cloud squad both proactively monitoring the sites'performance and reacting to users or site admins requests**
  - **Well organized ecosystem, experienced people and share of know-how**
  - **Current support model well suited for the sites' operations**
  - **Periodic meetings of the italian sites**

- **In the next slides we'll focus on the activities of the Italian sites only**

Roadmap: CPU and Job Processing

- **New PRIN (Research Project of National Interest) project submitted in January, focused on the R&D on the access to Computing and Storage resources for BigData analysis**
  - **Main goals are:**
    - **The transition to Cloud infrastructures**
    - **High availability of the sites and transaprent use of remote, federated resources**
    - **Porting of the software to low-power architectures, to enhance the cost effectiveness of the whole infrastructure**
    - **Exporing the introduction of hardware accelerators, GPUs and FPGA in the scientific software area**
- **Natural evolution of the PRIN successfully ending in February 2016**
- **All the Italian LHC Tier1/Tier2 sites participate in the project, common effort among the different parties**

Roadmap: Storage [DPM]

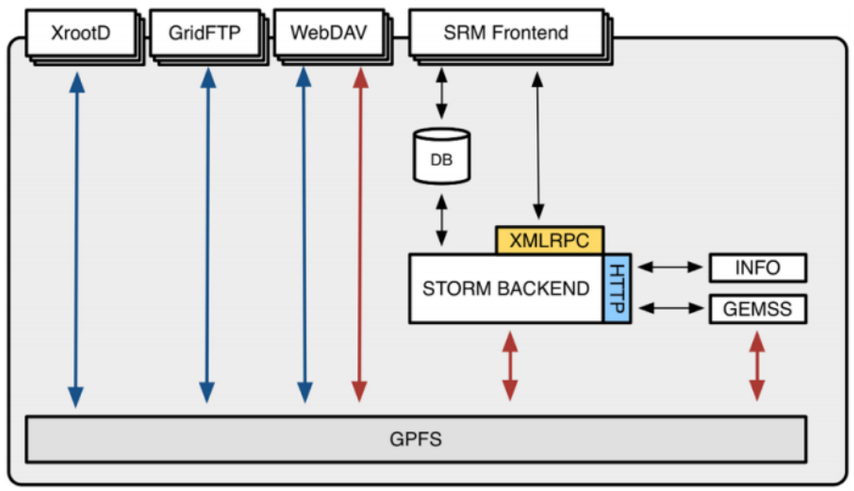
- ¾ of the Italian T2s use and will keep using DPM
- The Italian Cloud participates to the DPM development team too
  - Test of the new DPM releases in pre-production
  - Test of the deployment procedures and fine tuning of the automatic configurations
  - Test of pure grid features (SRM) and storage federation
    - Willing to test also the remote pools, to use single endpoints and pools distributed geographically, both for manageability and high availability
  - DPM can be a good candidate for the WLCG proposal of next generation of storage model in the T2 sites (caches)
  - The DPM collaboration is a very active area and made of well motivated people, the Italian Community will continue following this item as it's strategically important

Roadmap: Storage [StoRM]

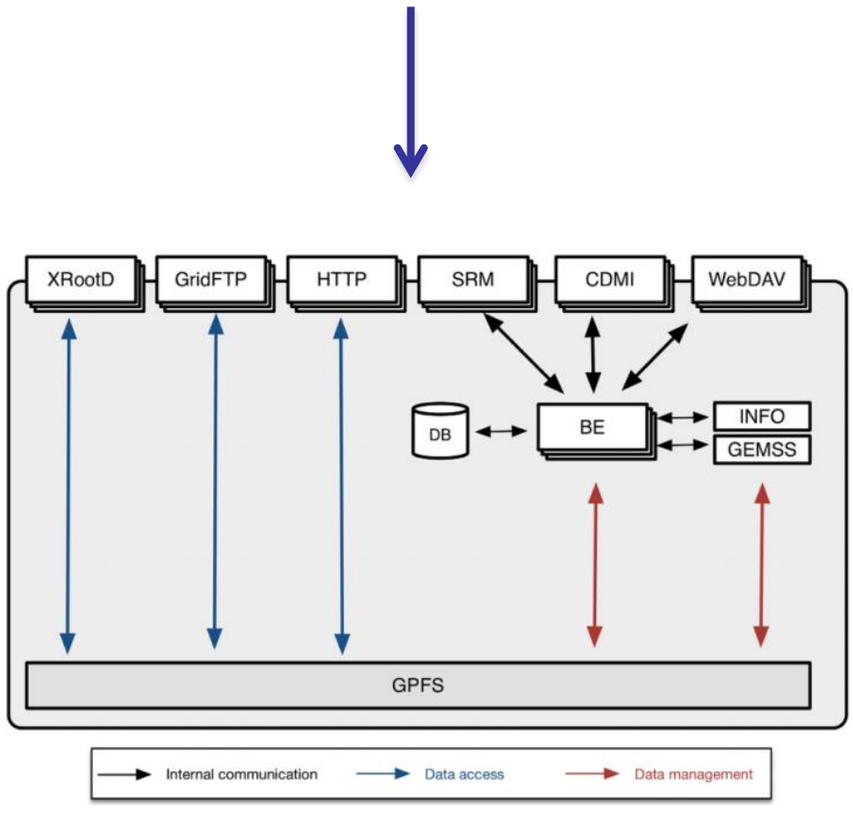
- **Mini workshop on the StoRM evolution in November 2015**
  - General discussion on the future of StoRM and the WLCG roadmap on the Storage

- **Current status**
  - All storm components are packaged for CentOS7
  - Planning to release Storm 1.11.10 very soon

- **Short/medium term plans**
  - Switch from YAIM to Puppet
  - Several improvements foreseen in space reporting
    - POSIX-based, for group quotas
    - Break srm monopoly for overall space by using WebDAV
    - Extension to nearline reporting under consideration
  - Extending Argus callouts to GridFTP level
  - Token-based authentication for HTTP access



- **Medium/long term plans**
  - Enhancement toward a better factorization of the storage manager and the specific interfaces (i.e. srm, WebDAV CDMI, ...)
  - Horizontal scalability for all StoRM services
  - Reduce and simplify evolution costs (mainly due to srm features), operation and deployment



Batch systems

- **Batch systems in the sites**
  - **No big issues, besides the well-known ones, but sites with PBS are suffering for the rigidity of the system (same for LSF, although mitigated by external agents that are/can be put in place)**
  - **Still work in progress for the migration to Condor of most the sites, no defined time scale**
    - **Needs an accurate documentation on the ATLAS side for the migrating sites**
  - **1 site (Milano) already using Condor**

- **Batch configuration in the sites**
  - **Generally limiting the ATLAS jobs only the WallTime of the jobs, demanding the limits on memory and disk space to the pilot**
  - **No cgroups enabled, but some T1/T2 sites may be able to enable it in the future**
    - **CNAF (LSF9, ready)**
    - **Milano (Condor, ready)**
    - **Roma (LSF7, needs to upgrade to LSF9 first)**

- **All CEs in the sites are Cream, no plan to change for now**

WLCG MW readiness activities

- **Cream CE MW readiness:**
  - **Dedicated ATLAS queue in Napoli to test Cream CE updates, in collaboration with WLCG MW readiness group.**

- **gfal utils for I/O of production jobs**
  - **The same queue as above is used to test gfal-copy (instead of lcg-cp) as copy tool for prod jobs**
  - **Copytool parameter changed in AGIS, for the test queue**
  - **gfal2 release in cvmfs tested with success**
  - **Updated release (gfal2-util-1.3.1) to be tested. Work in progress.**

GLUE values for accounting

- **GLUE parameters for accounting**
  - **We realized that, even among the Italian sites, there wasn't a uniform way to publish values for shares, LogicalCPUs, Benchmarks (GlueCECapability, GlueSubClusterLogicalCPUs, etc….)**
  - **Those values, used by ATLAS for the accounting, were wrong in some cases**
  - **We had some brainstorming about the content of the values**
    - **Logical CPUS are cores, job slots… ?**
    - **Can the intra-VO shares (per role) be published?**
    - **How to calculate the benchmarks?**
- **These troubles have been reported to the WLCG Information System Task Force**
- **They are preparing the definitions forthe GLUE2 values and they asked for some input to provide a clearer definitions for sites**

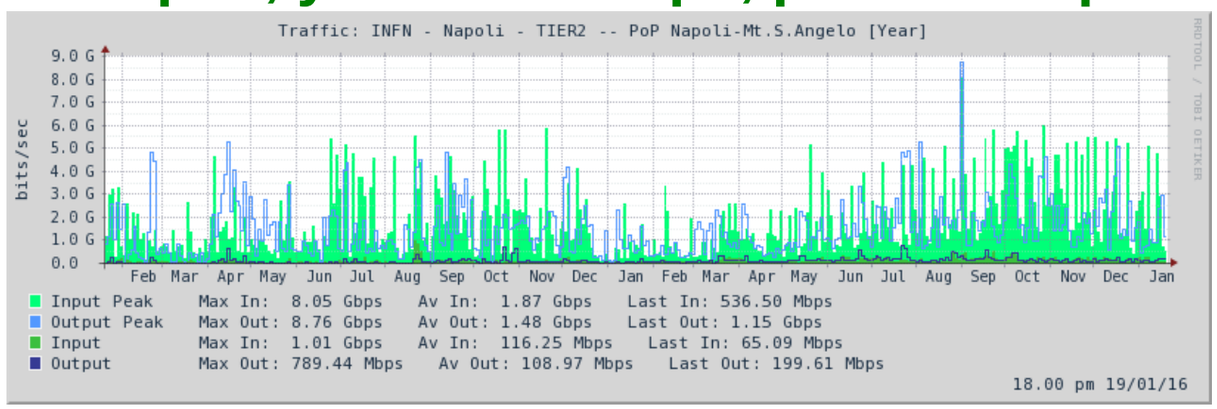
Sites' evolutions in the coming years

- **Same procurements model as before**
  - Same amount of memory
  - May still create "high memory" blobs by aggregating smaller/multiple slots
  - Probable migration of the WNs local connectivity to 10 Gbps, in response to the increase of the number of cores per logical unit

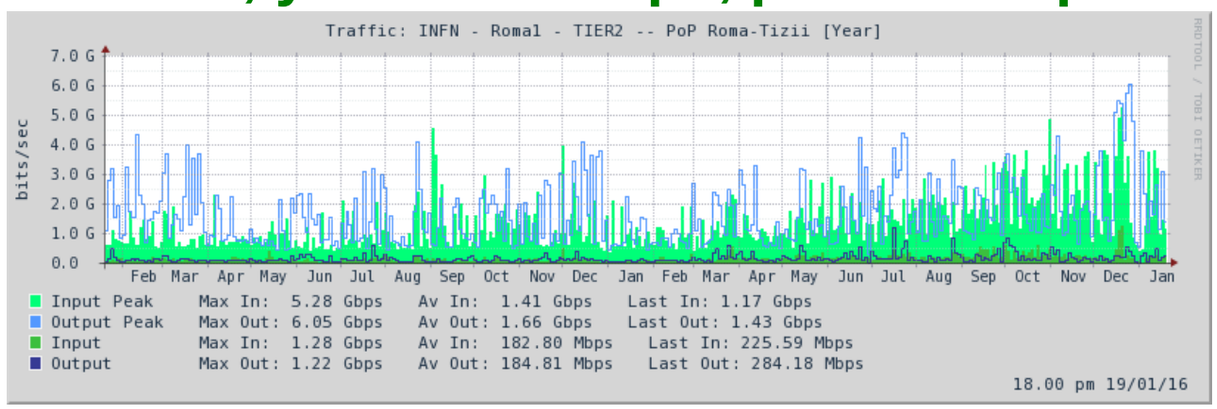
- **Low power CPUs can be attractive, but increasing the complexity of the system and with an higher initial cost**
  - For the moment we are considering them just as an R&D

- **Network bandwidth, currently 10 Gbps in the T2 and 40 in the T1, should at least double in the coming 3 5 years in most of the T2s, or even reach 100 Gbps in some case (e.g. Napoli and CNAF)**
  - The network will generally be able to cope well with the amount of CPU and Storage in the sites

Napoli, year av. 2 Gbps, peak. 9 Gbps



Roma, year av. 2 Gbps, peak. 6 Gbps



Issues & desiderata

- **General**
  - **More integration among the various ATLAS tools (e.g. AGIS and VOMS) and more/easier automation, to decrease the load on the sysadmins**
    - **The dark data automatic cleanup is a good example of what we should achieve**

- **Documentation**
  - **Sometimes not very clear, especially for sites starting to work in ATLAS (even accessing the documentation pages can be an issue here)**
  - **Obsolete documentation/links to be cleaned up**
  - **Clear instructions on the people to contact for the known issues, to be updated every time a new issue is identified**