Istituto Nazionale di Fisica Nucleare
COMMISSIONE DI CALCOLO E RETI

Workshop della
Commissione Calcolo e Reti dell'I.N.F.N.
La Biodola, Isola d'Elba
3 - 7 giugno 2019

AMiCO
Apparato Milanese per il Calcolo Opportunistico

Authors: Leonardo Carminati[1], Franco Leveraro[1], Francesca Milanini[1], Laura Perini[1], Francesco Prelz[2], David Rebatto[2], Paolo Salvestrini[3], Miguel Villaplana[2]
[1] Università degli Studi di Milano - Dipartimento di Fisica - [2] INFN - Sezione di Milano - [3] CNR - Istituto di Fotonica e Nanotecnologie

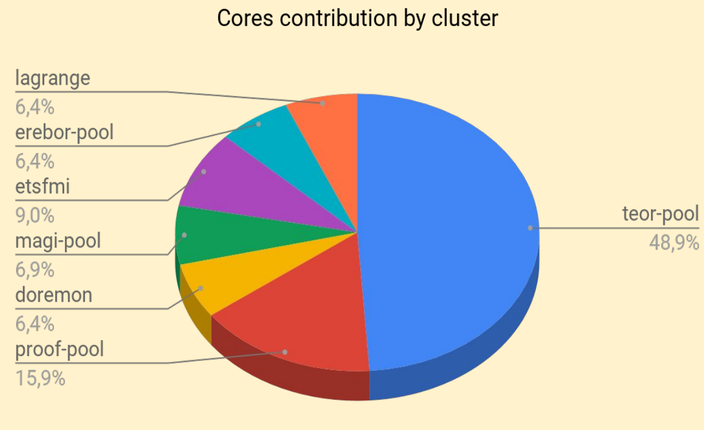
INFN Istituto Nazionale di Fisica Nucleare SEZIONE DI MILANO
UNIVERSITÀ DEGLI STUDI DI MILANO

OVERVIEW

- **AMiCO** is a project aiming to federate heterogeneous computing clusters at **Physics Department of Università degli Studi** and **INFN Milano**.
- Many research groups in the Department proceeded in sparse order with the purchase of computing resources:
  - usually one rackful of worker nodes with Infiniband for the execution of MPI jobs,
  - usually sized for spike usage, underused for a significant fraction of time.
- To achieve an *a posteriori* rationalisation, we enabled clusters to:
  - spill out their excess jobs to unused resources when they are overloaded,
  - preserving the possibility for cluster owners to preempt alien jobs.
- We found a recipe on the **HTCondor** Wiki fitting this common scenario. On top of that, we added support for:
  - dynamic slots,
  - parallel scheduling,
  - **Docker** jobs.
- For jobs needing **data access** while running outside their home cluster we provide:
  - A **CEPH** readable/writeable storage, accessible via S3 on RADOS gateway,
  - **CVMFS**, mounted on worker nodes and inside Docker containers.

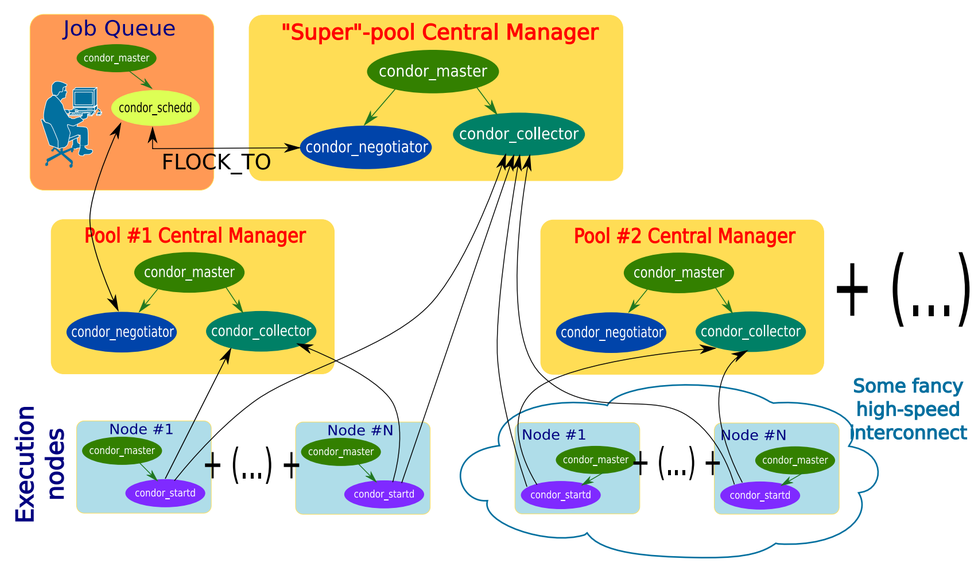
Some numbers

- 14 research groups
- ~400 active users
- 1870 cores (4500 when T2 comes in)
- 7 clusters (3 with **Infiniband** network)
- ~500 Terabytes of local storage
- ~450 Terabytes of **CEPH** storage



WORKLOAD MANAGEMENT

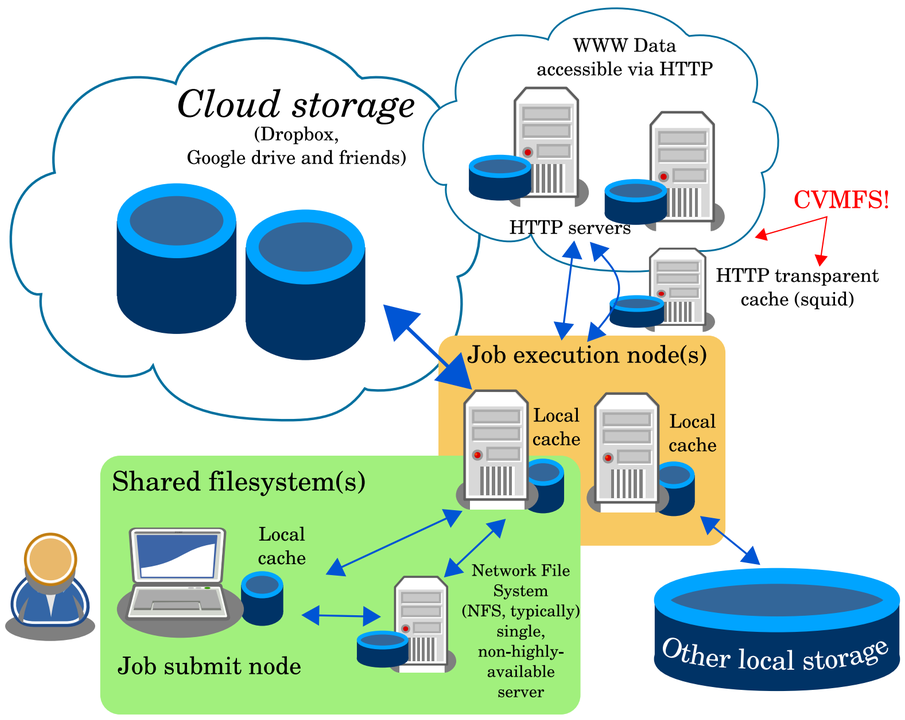
- Each cluster is an independent Condor pool
- The nodes are registered both in the local central manager *and* in a "superpool" central manager
- When a job can't find a match in the local negotiator, it flocks to the superpool
- The pools have additional condor policies to let local jobs or interactive load suspend and evict the external jobs on a machine.



STORAGE MANAGEMENT

Execution nodes have access to:

- a central CEPH storage, either as object storage or as a block device (RBD) mounted locally;
- a shared file system on the master node;
- (possibly) an area on SAN dedicated to the pool
- CVMFS for Cern experiments.



PADDOCK: parallel Docker for MPI

- Experimental set of BASH scripts meant to replace the real 'docker' command in the Condor config DOCKER config variable.
- Docker containers with all the required dependencies to *mpirun* our local MPI applications.
- Preliminary characterization of the MPI applications shows that the magnitude of the wall-time **is comparable to single-node execution**.