Condor on WAN

D. Bortolotti[1], *T. Ferrari*[2], *A. Ghiselli*[2], *P. Mazzanti*[1], *F. Prelz*[3], *F. Semeria*[1], *M. Sgaravatto*[4], *C. Vistoli*[2]

[1] INFN-Bologna, Italy
[2] INFN-Cnaf, Italy
[3] INFN-Milano, Italy
[4] INFN-Padova, Italy

Abstract

Condor is a software project to support High Throughput Computing in a large collection of computing resources, developed by the Computer Science Institute of the Wisconsin University. In collaboration with the Condor team, a Condor pool on Wide Area network has been deployed as general purpose computing resource for INFN. The characteristics of the INFN WAN Condor pool have been defined as result of a experimental phase. The phase objectives were to identify and address specific INFN requirements: suitability for INFN computing, policy and rules for execution machine access, network aware checkpointing.

The article describes how these requirements have been satisfied through a specific sub-pool and checkpoint domain design. In particular the network aware checkpointing mechanism allows to optimize checkpoint function for large jobs and to limit and control network traffic. A network monitoring system evaluates link bandwidth between checkpoint servers and execution machines and updates the bandwidth parameter on each execution machines. This mechanism allows a dynamic and reliable association between execution machine and checkpoint server according to the job network demand.

The present Condor WAN pool consists of about 180 machines, distributed in 20 sites, connected to the national research network Garr-B, and 6 checkpoint domains.

Keywords: HTC, WAN, Checkpoint domain, Sub-pool

1 Introduction

This document describes the configuration of a wide area Condor batch system at INFN to provide the INFN users with a global computing resource.

High Throughput Computing (HTC) on WAN can be deployed to satisfy the computing needs of INFN by accessing the huge CPU capacity distributed in all INFN sites: thank to the increasing CPU power of PCs and WSs and the decrease in cost the total computing capacity of the Institute has been increasing substantially.

Since the Condor philosophy aims at using only idle CPU cycles, Condor has been identified as the right candidate to satisfy the computing needs at INFN: Condor optimizes the usage of existing computing capacity. Local users still access local machines with higher priority. Resource sharing is controlled by local policies: for example subsets of machines (sub pools) can be defined. In each subset (e.g. all the machine belonging to some research group) specific jobs can have absolute priority on other jobs, that may be eventually vacated when high priority jobs need to execute.

The implementation of the Condor System on the INFN wide area network as described in this document is the result of a project developed in collaboration with the Condor-Team from the Computer Science Department of the Wisconsin-Madison University.

2 Test phase

During the first part of the project an experimental Condor pool has been set up in order to check the reliability and efficiency of the system on WAN, and its suitability to INFN computing needs. Tests of CPU intensive jobs gave good results in terms of very good workload, whereas jobs with an high frequency I/O were less efficient of CPU usage. In latter case performance improves if jobs run in a uniform file system: this means that with appropriate configuration even with I/O intensive jobs can run efficiently. Future mechanisms like caching or dedicated file systems will be investigated to improve the efficiency of I/O intensive programs. Other findings are the need of flexible policies in CPU usage of the machines and the importance of an adequate location of checkpoint servers.

3 Implementation phase

The choice to have only one "pool" (i.e. a pool with only one Central Manager) was made in order to optimize the CPU usage of all the INFN hosts available for Condor. The need of providing guarantees to local jobs in CPU usage can be satisfied by configuring sub-pools, while the overall efficiency of the system can still be achieved through a suitable setting of a set of checkpoint servers.

4 Sub-pool

A sub-pool is a collection of machines configured in order to give higher priorities to jobs belonging to local users or to research group users. A sub-pool can be local to one INFN site or distributed between different sites connected through WAN. Sub-pool policies must be defined by local management in agreement with the responsible for the research groups.

5 Checkpoint topology

The need of an appropriate checkpoint server topology stems from the decision to limit the impact of checkpoint file transfers especially when the number of machines will increase up to several hundreds. The optimal checkpointing policies are the following:

- Checkpointing a big size file should be accomplished in short time in order to let the owner access its machine without a visible delay.
- Sub-pools may make use of a dedicated checkpoint server.
- The definition of the best checkpoint server should be network adaptive

Obviously checkpoint should not limit the overall computing throughput. The solution adopted will be implemented in two steps:
1) Configuration of checkpoint domains
2) Implementing a distributed (dynamic) checkpointing as a new feature of the Condor System.

The most important characteristic of the solution adopted is that the network has been defined as resource of Condor : network bandwidth between checkpoints server and execution machines is a machine ClassAds attribute, dynamically updated, and used by checkpointing for the better choice between execution machines and checkpoint servers.

Initially each execution machine will have a fixed checkpoint server associated with it, then the association between execution machine and checkpoint server will be dynamically decided according to the network load.

6 INFN Condor topology

Most of the INFN sites have several machines in the WAN Condor pool and they are connected to the research network GARR-B with access speeds ranging from 2Mbps up to 8Mbps. The logical topology is described below.

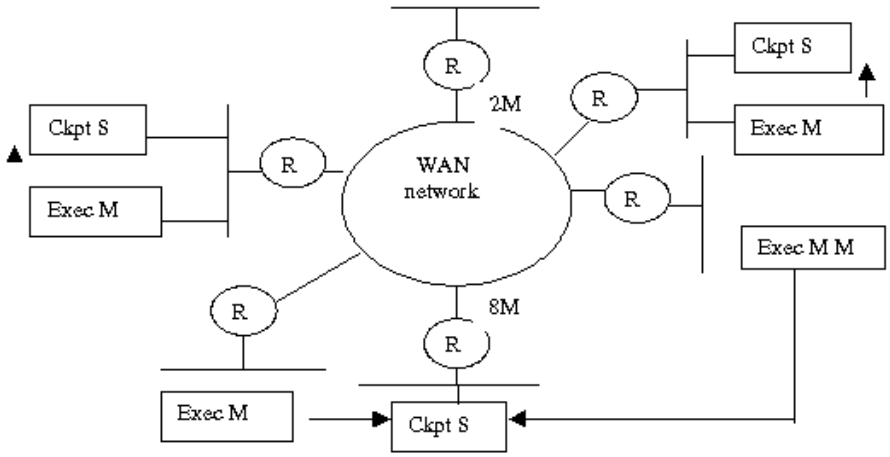

Figure 1: INFN Condor topology

The checkpoint server domain is defined according to the following guidelines:

- Presence of a sufficiently large CPU capacity.
- Presence of a set of machines with an efficient network connectivity.
- Set of site policies (eg. jobs have to run only locally).

The initial topology will have at least 10 checkpoint servers and the idea is to increase the number of the machines in order to have one checkpoint server in each site.

7 Management

7.1 Central management

The Admin Group will act as central management group and has to provide:

- Configuration, tuning and overall maintenance of the INFN Condor Wan pool
- management tools
- activity reports
- Condor resource usage statistics (CPU, Network, Ckpt-server)
- Which Condor release has to be installed
- Help desk for users and local administrators.
- Interface to condor support in Madison.

7.2 Local management

Local management has to provide:

- release installation in agreement with the central management

- local condor usage policies (e.g. sub-pools)

7.3 Steering Committee

The Steering committee should:

- consider the status of the condor system and suggest when upgrade the software
- interact with the Condor Team and suggest possible modifications of the system
- define the general policy of the condor pool organize meeting for condor administrators (and users)

8 Condor pool usage

The present number of machines in the pool is about 180, most of which are Alpha and pc-Linux machines.

Below is the table of the total allocation time as an example of the computing power provided by the INFN Condor pool during the last year. As can be seen the pool has provided more then 36 years of total allocation time.

Table I: Allocation Time of the last year

Month	**Hours**
Feb 1999	32877.5
Mar 1999	39471.1
Apr 1999	30427.6
May 1999	9418.8
Jun 1999	23027.5
Jul 1999	25845.1
Aug 1999	24797.5
Sep 1999	34185.3
Oct 1999	17834.9
Nov 1999	35247.4
Dec 1999	35432.3
Jan 2000	13360.4
Total	321925.4 (36.7 years)

Project documentation is available in http://www.infn.it/condor.

9 Conclusions

The Condor WAN pool test layout, where machines are distributed over 20 INFN sites and connected through the national research network GARR, gave the possibility to prove the reliability and robustness of the system and to study the most suitable “checkpoint domain” topology in order to optimise checkpoint operations and to limit geographic network traffic as much as possible. This goal can be achieved by considering the “network” as Condor resource. A “Network Manager” for Condor has been developed in collaboration with the authors of Condor. Furthermore in each site machines belonging to the pool can be configured in order to give absolute priority to the local research group jobs as if they were in a dedicated “sub-pool”. The choice to have only one “pool” does not represent a single point of failure because the last releases of Condor make the definition of central manager backups possible for higher stability of the WAN pool.