Distributed Operating System for Resource Discovery and Allocation in Federated Clusters

Emmanuel Jeanvoine, EDF R&D – IRISA, INRIA Paris team

Nowadays, numerical simulation has an important place in several scientific fields where real experimentation is expensive or impossible. Important computing resources must be used to realize these experimentations and workstations are not suitable to satisfy these requirements. After having aggregated computers in clusters or in federated clusters (or grid), a challenge for laboratories or large companies is to efficiently and easily use this large amount of resources. Several projects like PBS [Bode00] or Kerrighed [Morin04] offer solutions for cluster resources management. Other projects like Condor [Litzkow88], Globus [Foster97] or NetSolve [Agrawal03] focus on federated clusters or grid computing.

We present the design of a distributed operating system for resource discovery and allocation in federated clusters.

To design the system, we have had several goals. The first one is to offer a single system image vision for resources use through the federation. This would allow for instance to launch an application execution from anywhere without taking care about real execution location.
Our second goal is to support heterogeneous clusters, in terms of hardware or software architecture. So users must be able to specify their application needs when they submit tasks to the federation in order to ensure the execution with the suitable resources.
The last goal is to deal with large scale. Typically, several thousands of clusters may be federated with our operating system. We assume that clusters are distributed through different geographical areas, like different sites of a large company. We give a lot of attention to this point in the design of the system by avoiding any type of centralization and by supporting dynamic behavior and tolerating failures.

The operating system for federated clusters is materialized by a daemon executed on every clusters, typically on the front node of each cluster. We briefly present the components of the daemon.
First of all, a communication module based on a structured peer-to-peer overlay network ([Rowstron01]) is used to allow communications between clusters. Thus, it permits the remote execution of applications. Peer-to-peer overlays perfectly fit for large scale and dynamic behavior constraints.
In order to find suitable resources for the execution of a particular application, our system has a resource discovery module based this time on an unstructured peer-to-peer overlay ([Ganesh01]). Currently, only flooding and random walk policies have been implemented.
When the resource discovery step is fulfilled, the resource allocation module is requested to find the most suitable resource for execution of the application according several policies. For instance, a policy could be: “minimization of files fetching through the network”. Currently, only a basic policy has been implemented. That consists in choosing the resource with the lower load.
As far as it is impossible to plan the real cost of an execution at the resource allocation step, load unbalances might appear. To correct these unbalances, we have designed a module, based on process migration, that performs load balancing within a federation.
Finally, we have developed an application manager module to supervise the execution of applications and to handle user queries during execution.

Contributions of our system rely on its fully distributed architecture and in the transparency of use it provides to users. Furthermore, we have implemented a modular prototype that will allow us to create new resource discovery and allocation policies. We will also be able to experiment it with the Grid'5000 testbed (http://www.grid5000.org).

[Agrawal03] Agrawal S., Dongarra J., Seymour K., Vadhiyar S., NetSolve: Past, Present, and Future – A Look at a Grid Enabled Server, Making the Global Infrastructure a Reality, A. eds. Wiley Publishing, 2003.
[Bode00] Bode B., Halstead D., Kendall R., Lei Z., Jackson D., The Portable Batch Scheduler and the Maui Scheduler on Linux Clusters,4th Annual Linux Showcase & Conference, Atlanta, 2000.
[Foster97] Foster I., Kesselman C., Globus : A Meta-computing Infrastructure Toolkit, The International Journal of Supercomputer Applications and High Performance Computing, vol 11, pp 115-128, 1997.
[Ganesh01] Ganesh A., Kermarrec A., Massoulie L., Scamp: Peer-to-peer lightweight membership service for large-scale group communication, Third International Workshop on Networked Group Communications, London, UK, 2001.
[Litzkow88] Litzkow M., Livny M., Mutka M. Condor - A Hunter of Idle Workstations, Proc of the 8th International Conference on Distributed Computing Systems, IEEE Computer Society, pp 104-111, 1988.
[Morin04] Morin C., Gallard P., Lottiaux R., Vallée G., Towards an Efficient Single System Image Cluster Operating System, Future Generation Computer Systems, Elsevier Science, 20(2), 2004.
[Rowstron01] Rowstron A., Druschel P., Pastry : scalable, decentralized object location and routing for large-scale peer-to-peer systems, Proc of the 18th IFIP/ACM International Conference on Distributed Systems Platforms Middleware, Phoenix, Arizona, 2001.

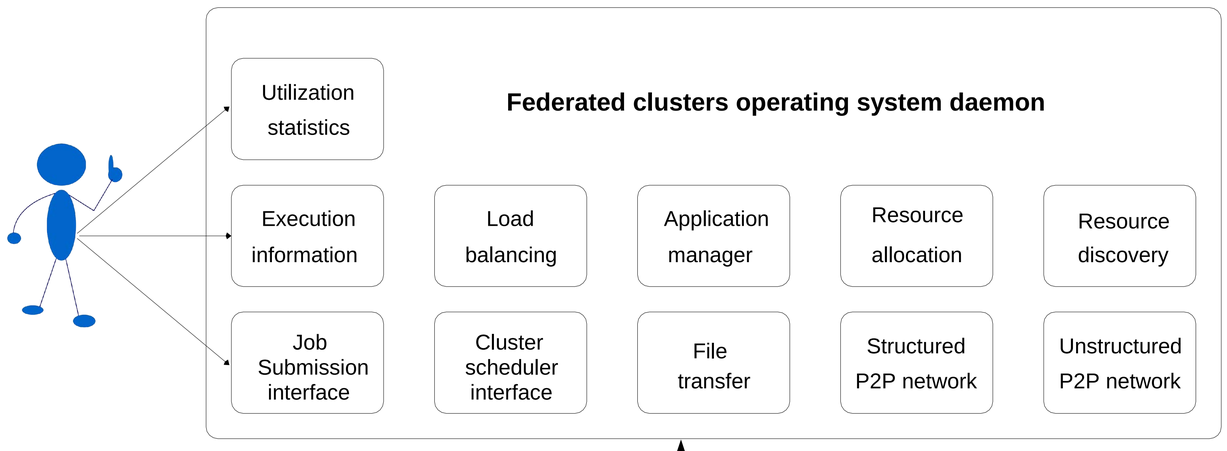
Distributed Operating System for Resource Discovery and Allocation in Federated Clusters

Emmanuel Jeanvoine
EDF R&D – IRISA, INRIA Paris project team

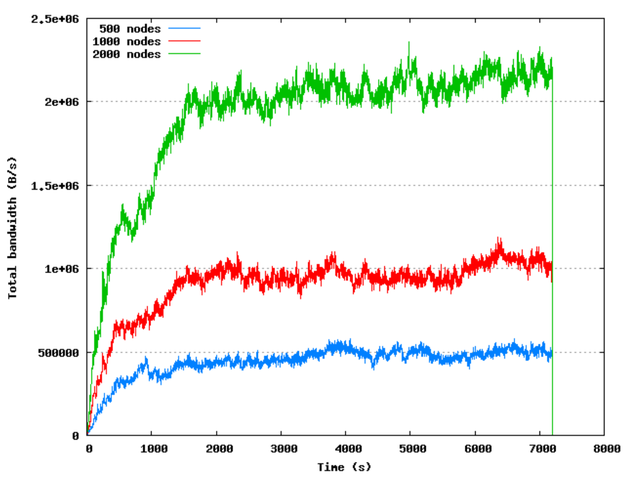
Objectives

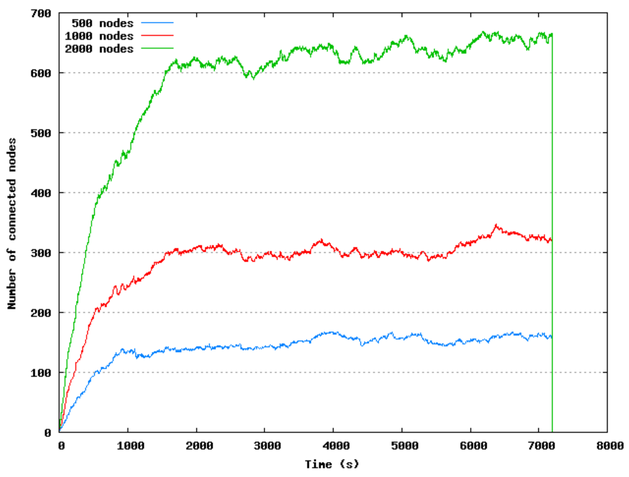
- Transparency of use (Single System Image)
- Large scale (several thousand clusters)
- Dynamic behavior support
- Heterogeneous cluster support
- Fully distributed architecture
- Support for multiple resource discovery and scheduling policies

Architecture

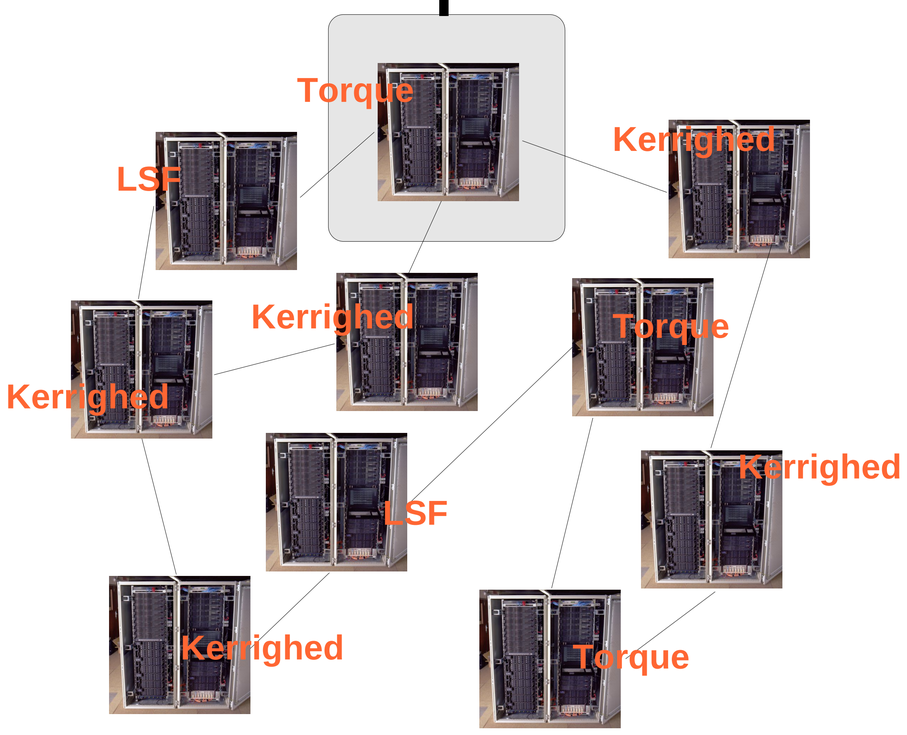


Early experimentation





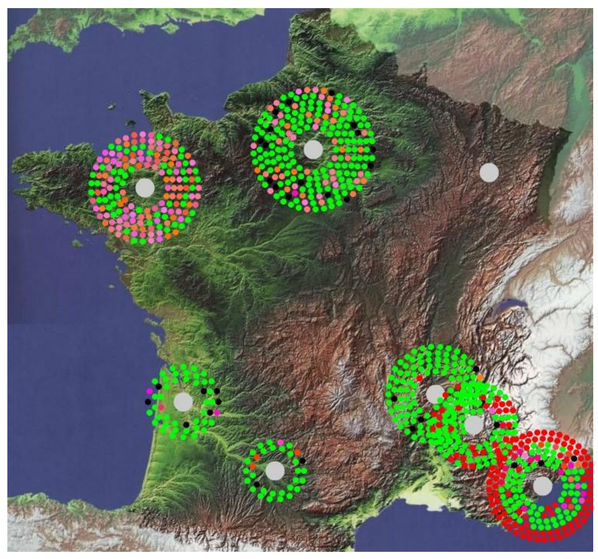
- Infrastructure scalability with 500, 1000 and 2000 nodes, each representing a cluster
- Dynamic behavior (arrival and departure of clusters) simulated by a Poisson law



Early results

Whatever the number of clusters and with a relative dynamic behavior (30 minutes up and 5 minutes down in average), bandwidth consumption per cluster is constant

Grid'5000 Testbed



Contact information

Mail: Emmanuel.Jeanvoine@irisa.fr
Web: http://www.irisa.fr/paris/pages-perso/Emmanuel-Jeanvoine/
Postal: IRISA - INRIA
Campus de Beaulieu
35042 – Rennes Cedex
France
Tel: +33 (0)2 99 84 25 56

EDF IRISA