Distributed Meta-Scheduling for Grids

Janko Heilgeist[†] jkh@rzg.mpg.de
Thomas Soddemann[†] tks@rzg.mpg.de
Harald Richter[‡] hri@tu-clausthal.de

[†]Computing Center of the Max-Planck-Society
Boltzmannstr. 2, 85748 Garching, Germany
[‡]Department of Computer Science
Technical University of Clausthal
Arnold-Sommerfeld-Str. 1, 38678 Clausthal-Zellerfeld, Germany

Grid computing stands for the effort undertaken mainly by computing centers to open up and combine their resources for an enhanced availability. There is a growing demand for an automatic balance of inter-infrastructure resource requests that existing middleware such as UNICORE and Globus Tool Kit is ill-suited to satisfy, as it it requires the user to provide the location of suitable resources and only facilitates the migration process. Other projects like Gridway or LSF Multicluster suffer (at least currently) from missing interoperability. We describe a distributed, failure-resilient meta-scheduling architecture that allows the automatic exchange of job requests between resource providers, aiming at improved resource utilization, automatic load-balancing, as well as reduced turn-around times. Additionally, the system tries to achieve grid-wide improvements while still preserving the autonomy of resource providers. This is accomplished by making all decisions locally.

Resilience to failures is achieved by a distributed design in which all sites are equally important to the meta-scheduler as a whole. A participating site deploys a software instance, that we call a *meta-scheduler proxy*, for each major resource that it wants to make available to the grid community. These proxies communicate with each other to discover free resources and decide if and where to migrate job requests to in order to realize load-balancing. Our approach relies on services provided by existing middleware like advanced reservation, data staging, scheduling information, etc to actually carry out its job. The proxies converse utilizing web services and may at a later stage be extended to adhere to the Open Grid Services Architecture (OGSA). Input is accepted in standard formats such as the Job Submission Description Language (JSDL) and well-defined interfaces (APIs) towards the grid middleware will allow the system to be adapted to different software solutions.

The first step towards automatic request migration is the discovery of resources, that fulfill the requirements of a compute job with regard to software, hardware, and administrative constraints. This discovery process is accomplished through communication between the proxies applying forwarding-based peer-to-peer algorithms. Usually, forwarding algorithms suffer from the fact that high-quality results lead to a considerable cost to the network in terms of propagated messages. To solve this problem, we propose to use a selection of different algorithms, each with distinct characteristics, and apply them dynamically depending on the situation. The dynamic selection of the most appropriate algorithm ensures that the selected algorithm can play out its advantages and thus results are achieved, that would not be possible with a single algorithm alone.

Each request migration decision depends on multiple criteria, that affect the interests of different groups like grid community, resource providers, or users. Multi-criteria optimization algorithms are required to incorporate these often conflicting criteria into a final decision, such that over the long run all parties are satisfied. Currently, we investigate the Analytic Hierarchy Process (AHP) [2], a decision-making algorithm employed e.g. in economics. It allows a tree-based representation of the decision criteria and describes a way to combine them into a final weighting of the alternatives. The special form of representation and the independent weights at each tree-node help to obey different interests and policies, meanwhile preserving the autonomy of a computing center.

It is planned to implement and deploy the described approach in the Distributed European Infrastructure for Supercomputing Applications (DEISA/DEISA2), a consortium of eleven leading European supercomputing centers, until 2009.

References

1. J. Heilgeist, T. Soddemann, and H. Richter. Algorithms for job and resource discovery for the meta-scheduler of the DEISA grid. In *International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP'07)*, 2007. And references therein.
2. T. Saaty. *Math. Methods of Operations Research.* Dover Publications Inc., 2004. And references therein.

MG '08, January 29-February 3 2008, Baton Rouge, LA, USA