

Distributed Meta-Scheduling for Grids

Janko Heilgeist, Thomas Soddemann, Harald Richter

Abstract

Grid computing, a special form of distributed computing, stands for the effort undertaken mainly by computing centers to open up and combine their resources for an enhanced availability. With the increasing size of these compute infrastructures, there is a growing demand for an automatic balance of inter-infrastructure resource requests. Existing middleware such as UNICORE and Globus Tool Kit is ill-suited to this job since it requires the user to provide the location of suitable resources and only facilitates the migration process. Other projects like Gridway or LSF Multiclustor suffer (at least currently) from missing interoperability.

We describe a distributed meta-scheduling architecture that allows the automatic exchange of jobs between resource providers, aiming at improved resource utilization, automatic load-balancing, as well as reduced turn-around times. Additionally, the architecture is resilient to link and site failures due to its distributed nature. Finally, the system tries to achieve grid-wide improvements while still preserving the autonomy of resource providers. This is accomplished by making all decisions locally.

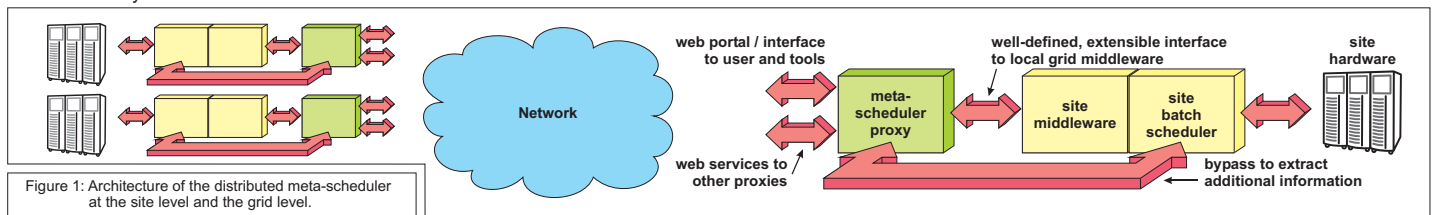


Figure 1: Architecture of the distributed meta-scheduler at the site level and the grid level.

Resource Discovery

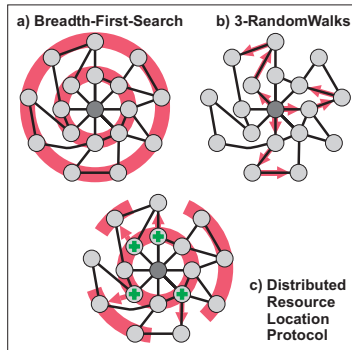


Figure 2: Three forwarding algorithms and their unique message distribution schemas.

The first step towards automatic request migration is the discovery of compatible resources, that is, to find resources fulfilling the requirements of a compute job with regard to software, hardware, and administrative conditions. This discovery process is achieved through communication between the meta-scheduler proxies utilizing forwarding-based algorithms from the area of peer-to-peer (P2P) networks. Usually, forwarding algorithms suffer from the fact that high-quality results lead to a considerable cost to the network in terms of propagated messages and, the other way around, a reduction of the number of distributed messages leads to insufficient results.

Forwarding algorithms are therefore inadequate when used on their own. 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 a search is issued in. The selection of the most appropriate algorithm ensures that an algorithm can play to its advantages. Thus, results can be achieved through the combination of multiple algorithms, that would not be possible with a sole algorithm.

Decision Making

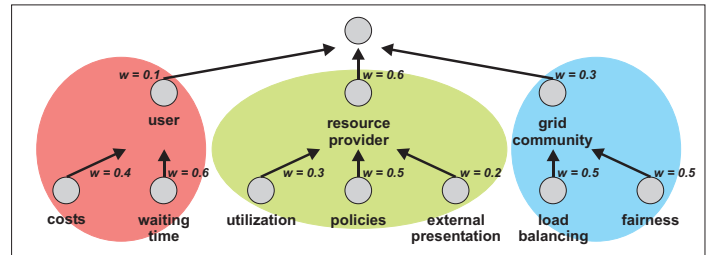


Figure 3: Sample hierarchy of criteria relevant to decision making in a grid. The different parties user (red), resource provider (green) and grid (blue) are placed in separate sub-trees.

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 goals into a final decision, such that over the long run all parties are maximally 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. Additionally, a simple method to compute the weights supports even inexperienced users, that would like to influence the decision.

Conclusions

We have described a distributed meta-scheduling architecture, that employs P2P search algorithms and a multi-criteria decision-making algorithm to achieve automatic load-balancing. It is planned to implement and deploy the described approach in the Distributed European Infrastructure for Supercomputing Applications (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 *Int. Conf. on Adv. Eng. Comp. and Appl. in Sciences (ADVCOMP'07)*, 2007. And references therein.
2. T. Saaty. *Math. Methods of Operations Research*. Dover Publications Inc., 2004. And references therein.