Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 194 (2009) 856–872
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

O.R. Applications

Maximizing business value by optimal assignment of jobs to resources in grid computing

Subodha Kumar [a,*], Kaushik Dutta [b,1], Vijay Mookerjee [c,2]

[a] Information Systems and Operations Management, Michael G. Foster School of Business, University of Washington, Box 353200, Seattle, WA 98195, USA
[b] Decision Science and Information Systems, Florida International University, 11200 SW 8th Street, University Park Miami, FL 33199, USA
[c] Information Systems and Operations Management, School of Management, University of Texas, Dallas, Richardson, TX 75083-0688, USA

Received 24 December 2006; accepted 17 December 2007
Available online 31 December 2007

Abstract

An important problem that arises in the area of grid computing is one of optimally assigning jobs to resources to achieve a business objective. In the grid computing area, however, such scheduling has mostly been done from the perspective of maximizing the utilization of resources. As this form of computing proliferates, the business aspects will become crucial for the overall success of the technology. Hence, we discuss the grid scheduling problem from a business perspective. We show that this problem is not only strongly NP-hard, but it is also non-approximable. Therefore, we propose heuristics for different variants of the problem and show that these heuristics provide near-optimal solution for a wide variety of problem instances. We show that the execution times of proposed heuristics are very low, and hence, they are suitable for solving problems in real-time. We also present several managerial implications and compare the performance of two widely used models in the real-time scheduling of grid computing.

Keywords: Evolutionary computations; Parallel computing; Scheduling; Heuristic; Integer programming

1. Introduction

Recent developments in processing power, data storage, and networking have made possible the migration of grid technology from the educational and research fields into the business realm (Berman et al., 2003; Glasgow, 2003). A grid is essentially a shared pool of networked resources that are modular, cost-effective, flexible, balanced, scalable, distributed, and standards-based (Castro-Leon and Munter, 2005; Intel, 2004a). In the grid computing, storage, data, and CPUs from multiple systems are connected into a managed and flexible computing environment (Intel, 2004b). Grid computing enables users to access idle or unutilized IT resources (De Roure et al., 2003). A grid becomes useful and meaningful when it both encompasses a large set of resources and serves a sizable community (Castro-Leon and Munter, 2005).

Grid computing has been put into use and achieved great success in the research arena. The most well-known example is the SETI@home project initiated by the University of California at Berkeley. However, the grid is not only of interest to the researchers, but its deployments are encompassing a broad swath of industry verticals that will take the grid well beyond its roots (Castro-Leon and Munter, 2005). Some of the successful examples of grids include Legion (Legion, 2007), NetSolve (Seymour et al., 2005), Nimrod/G (Buyya et al., 2000), and DISCWorld (Coddington, 2002). For a detailed tutorial on grid computing and its architecture and protocols, the reader is directed to Foster et al. (2002) and Castro-Leon and Munter (2005).

There are three sets of entities involved in the grid computing – job owners, resource owners, and the scheduler.

* Corresponding author. Tel.: +1 206 543 4777; fax: +1 206 543 3968.
E-mail addresses: subodha@u.washington.edu (S. Kumar), kaushik.dutta@fiu.edu (K. Dutta), vijaym@utdallas.edu (V. Mookerjee).
[1] Tel.: +1 305 348 3302; fax: +1 305 348 3497.
[2] Tel.: +1 972 883 4414; fax: +1 972 883 2089.

0377-2217/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2007.12.024

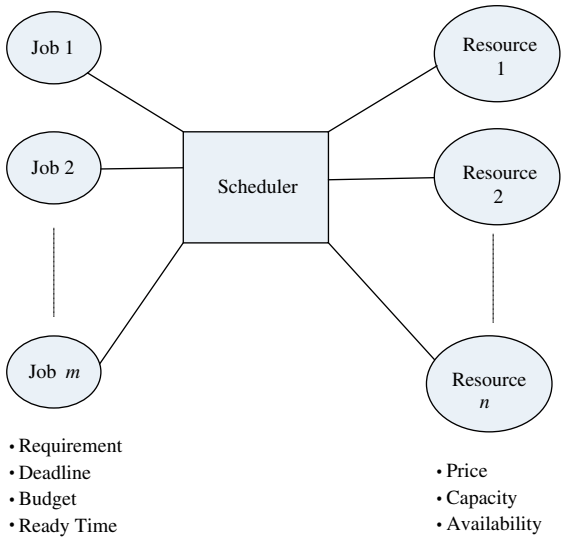

Fig. 1. Entities involved in grid computing and their parameters.

As illustrated in Fig. 1, both job owners and resource owners interact with the scheduler. Job owners specify their job requirements, the time by which the job must be finished (i.e., deadline), the budget available for their job, and the time at which the job will be ready to be processed (i.e., the ready time of the job) (Wolski et al., 2003). On the other hand, the resource owners set the price for using their resources, and specify their capacities and availabilities. In different settings of grid networks, the resource owners may decide the price using different mechanisms as explained by Wolski et al. (2003). Examples of resource pricing can be found at Sun Grid (Sun Grid, 2006; Utility Computing, 2006) and the cluster of Tsunamic Technologies Inc. (2006).

Many important research issues have arisen with the advent of grid computing. Some of these are in the area of security, compatibility, and functionality. However, the value of a grid system will be realized in enterprise systems only when the grid is used in an optimal way to serve a business objective (Cheliotis et al., 2005). An important issue in a grid is the optimal scheduling of jobs, i.e., given a set of resources and computing jobs, what is the best assignment of jobs to resources. There are two formal approaches to the allocation problem in grid: performance based, and economics based. The first is based on the assumption that the system response can be accurately modelled by various system parameters such as network usage, disk usage and the usage of CPU cycles. However due to the inherent nature of a grid, being composed of heterogeneous platforms distributed geographically across the world, developing such model remains elusive, making it difficult to define formally tractable performance based mechanisms (Wolski et al., 2003).

On the other hand, economics based approaches are attractive for several reasons. Firstly, the concept of *efficiency* or cost effectiveness is well defined; although it is different from the notion of efficiency typically understood in a computer performance evaluation setting. Secondly, economic systems and the assumptions upon which they are based seem familiar, making common intuition a more valuable research asset. Economics based approaches have garnered quite a bit of attention as a way to allocate grid resources (e.g., in the area of computational economy). As more grid systems will be deployed for commercial purposes, an economic approach to grid scheduling will become more appropriate (Cheliotis et al., 2005). Currently most economics based approaches to grid scheduling are studied using an auctions perspective (Schnizler et al., 2008). However, auctions based scheduling may not always be suitable for enterprise wide grid deployments.

For example, Merrill Lynch (ML) is working towards developing an enterprise wide grid that can be used by its various divisions to do complex financial computations (Private Communications). ML, one of the world's leading financial management and advisory companies, with offices in 36 countries and total client assets of approximately $1.6 trillion, requires complex, CPU intensive financial computations to allocate and manage their assets. In the proposed grid, various branches will be required to submit computational jobs to the centralized grid with a specification of the deadline, opportunity cost of not executing the job, budget, and estimated numbers of instructions required to finish the job. These jobs will then be allocated to appropriate resources meeting the constraints specified by the job owner. Following the spirit of a decentralized organizational structure, each of these branches will be required to pay to use the grid. In such scenarios, it is appealing to consider a schedule that optimizes an overall business objective of cost, rather than use auctions that essentially lead to competition between multiple divisions of the organization (Private Communications). Therefore, we study the problem of resource scheduling in grid computing networks with the goal of optimizing an economic objective, such as cost or value.

We use the real-world examples to illustrate the problems considered in this research. Let us first discuss the example of World Wide Grid (WWG) (World Wide Grid, 2007), which is represented in Fig. 2. World wide grid consists of computers in five continents and has been used for drug design (BioGrid, 2001). The economic model in WWG considers deadline and budget of jobs and dollar cost of resources to decide the assignment of jobs to resources. The few resources used in WWG are as given in Table 1 with their respective parameters. The details of existing resources in WWG are available at http://gridbus.cs.mu.oz.au/sc2003/list.html. Currently, WWG has 218 resources. The values of job parameters related to drug design case in WWG are shown in Table 2.

Recently, Buyya et al. (2005) proposed an economic approach for scheduling jobs in a grid. They considered budget, deadline, and processing time of jobs and the cost of cpu time of resources to develop Nimrod/G scheduler as