Chapter 6.3
QoS-Based Job Scheduling and Resource Management Strategies for Grid Computing

Kuo-Chan Huang
National Taichung University, Taiwan

Po-Chi Shih
National Tsing Hua University, Taiwan

Yeh-Ching Chung
National Tsing Hua University, Taiwan

ABSTRACT

This chapter elaborates the quality of service (QoS) aspect of load sharing activities in a computational grid environment. Load sharing is achieved through appropriate job scheduling and resource allocation mechanisms. A computational grid usually consists of several geographically distant sites each with different amount of computing resources. Different types of grids might have different QoS requirements. In most academic or experimental grids the computing sites volunteer to join the grids and can freely decide to quit the grids at any time when they feel joining the grids bring them no benefits. Therefore, maintaining an appropriate QoS level becomes an important incentive to attract computing sites to join a grid and stay in it. This chapter explores the QoS issues in such type of academic and experimental grids. This chapter first defines QoS based performance metrics for evaluating job scheduling and resource allocation strategies. According to the QoS performance metrics appropriate grid-level load sharing strategies are developed. The developed strategies address both user-level and site-level QoS concerns. A series of simulation experiments were performed to evaluate the proposed strategies based on real and synthetic workloads.

DOI: 10.4018/978-1-4666-0879-5.ch6.3

1. INTRODUCTION

This article elaborates the *quality of service* (QoS) aspect of load sharing activities in a computational grid environment. Load sharing is achieved through appropriate job scheduling and resource allocation mechanisms. A computational grid usually consists of several geographically distant sites each with different amount of computing resources. Different types of grids might have different QoS requirements. In most academic or experimental grids the computing sites volunteer to join the grids and can freely decide to quit the grids at any time when they feel joining the grids bring them no benefits. Therefore, maintaining an appropriate QoS level becomes an important incentive to attract computing sites to join a grid and stay in it. This article explores the QoS issues in such type of academic and experimental grids. We first define QoS based performance metrics for evaluating job scheduling and resource allocation strategies. According to the QoS performance metrics appropriate grid-level load sharing strategies are developed. The developed strategies address both user-level and site-level QoS concerns. A series of simulation experiments were performed to evaluate the proposed strategies based on real and synthetic workloads.

2. BACKGROUND

Without grid computing users can only run jobs on their local site. A computational grid is an emerging platform for enabling resource sharing and coordinated computing work, which integrates resources across multiple geographically distant institutions. In most current academic and experimental grid systems, participating sites provide their resources for free with the expectation that they can benefit from the resource sharing in terms of improved job turnaround time. The improved job turnaround time is an example of QoS indicator which users care about. A grid system has to provide strong incentives concerning improved QoS for participating sites to join and stay in it. Job scheduling and resource allocation strategies fulfilling users' QoS requirements thus become a crucial research area. The QoS of a grid system can be explored at different levels. At the grid system level participating sites are concerned with the potential performance improvement of their local users' jobs once they join a grid. Therefore, fair resource sharing can be considered as the most important grid-level QoS requirement. We say the resource sharing is fair if it can bring performance improvement and the improvement is achieved in the sense that all participating sites benefit from the collaboration.

At the individual user level users concern mostly their jobs' turnaround times. Usually shorter turnaround time implies better QoS. Some research work on QoS based job scheduling associates each job with a hard completion deadline or a fixed budget which are strict QoS requirements. The deadline and budget are then taken into consideration when performing job scheduling. On the other hand, this article focuses on the academic and experimental computational grid environments where the resources are shared for free and usually the users just want their jobs to finish as soon as possible without strict deadlines associated with the jobs. Therefore, we do not discuss the issues related to the deadline and budget constraints. The improved job turnaround time is the sole concern of the user-level QoS requirement in the following studies.

Heterogeneity is another important issue in a computational grid. Many previous works (England & Weissman, 2005; Hamscher, Schwiegelshohn, Streit, & Yahyapour, 2000; Huang & Chang, 2006) have shown significant performance improvement for homogeneous grid environment. However, in the real world a computational grid usually consists of heterogeneous sites which differ at least in the computing speed. Heterogeneity puts a challenge on designing QoS-based scheduling methods. Methods developed for homogeneous grids have to be improved or even