Chapter VII
Quality of Service of Grid Computing

Ming Wu
Illinois Institute of Technology, USA

Xian-He Sun
Illinois Institute of Technology, USA

ABSTRACT

Rapid advancement of communication technology has changed the landscape of computing. New models of computing, such as business-on-demand, Web services, peer-to-peer networks, and Grid computing have emerged to harness distributed computing and network resources to provide powerful services. The non-deterministic characteristic of the resource availability in these new computing platforms raises an outstanding challenge: how to support Quality of Service (QoS) to meet a user’s demand? In this chapter, we conduct a thorough study of QoS of distributed computing, especially on Grid computing where the requirement of distributed sharing and coordination goes to the extreme. We start at QoS policies, and then focus on technical issues of the enforcement of the policies and performance optimization under each policy. This chapter provides a classification of QoS metrics and policies, a systematic understanding of QoS, and a framework for QoS of Grid computing.

INTRODUCTION

With the advance of network technology, many new distributed computing models are being constructed to harness geographically distributed computing and communication resources, such as business-on-demand, Web services, peer-to-peer networks, and Grid computing. Typical examples of these systems include WebSphere, Gnutella, Skype, Seti@home, Condor, PPLive (a P2P television network), and Globus (Wu, 2006). The system size of these systems scales...
From hundreds of nodes to tens of thousands of nodes, and even more. In these systems, resources are shared and collaborated to provide services/functionalities such as online shopping, online telephony and television, teleimmersion, online control of scientific instrumentation, and resource pooling. Much effort is being made in the standardization of protocols and interface for service orchestration and resource collaboration in these environments (Foster and Kesselman, 2004). With the maturity of these systems, when more and more users to use them as day-to-day computing infrastructure, Quality of Service (QoS) of these newly-emerged computing platforms is becoming more and more important.

QoS study was focused on QoS control and delivery in a dedicated environment where resources are controlled and managed in a centralized mechanism. In a shared network environment like a Grid, where resources are shared among different applications and managed within different organizations and domains, there are several new issues related to QoS support that do not arise in a single computer system. The first issue is the variation of resource availability, the accessibility of a system resource to an application. This variation may be due to resource contention, dynamic system configuration, software or hardware failures, and other factors beyond the control of a user. The uncertainty of resource availability presents a big challenge on guaranteed application quality delivery. The second issue is parallel processing. The total workload of a large scale application is often partitioned into smaller pieces, called subtasks. These subtasks are then allocated to resources in a distributed system to be processed concurrently. The challenge of parallel processing in a shared network environment lies on that the computing resources may be heterogeneous and have individual availability patterns. The third issue is non-centralized control. In a general Grid environment, the computing resources are autonomous. Local schedulers schedule local jobs and the Grid scheduler does not have the control of the local jobs.

Because of these difficulties, a suitable and broadly applicable QoS solution has been elusive. This is especially true for Grid computing, where the requirement of distributed sharing and coordination goes to the extreme. QoS is a known technical hurdle preventing a broader adoption of Grid computing for which there has been no well-conducted QoS study to balance the need of Grid tasks and local jobs. Some efforts have been made to address the issues of sharing. Distributed systems, such as Condor, NetSolve, Nimrod, and Globus (Foster and Kesselman, 2004), support Grid computing and facilitate resource sharing and collaboration. These systems adopt different QoS policies, usually implicitly, and try to provide a satisfactory QoS under their adopted policies. These policies often support QoS for one side and sacrifice that of the other – they perform well for certain applications but do not provide a satisfactory solution for general Grid computing.

Without a better understanding of the impact of resource reservation on QoS an appropriate decision cannot be made. Recently, a prototype of QoS system, Grid Harvest Service (GHS) has been developed at Scalable Software System lab in Illinois Institute of Technology (Wu and Sun, 2006). GHS is based on a fundamental understanding of QoS of Grid computing in two stages: policymaking and optimization mechanisms. Policymaking decides the QoS policy of resource sharing among Grid tasks and local jobs. Optimization mechanisms obtain an optimum QoS under each QoS policy. They are integrated solutions of advanced performance modeling, resources management, and scheduling algorithms. These QoS optimization mechanisms provide a comprehensive investigation of the impact of system characteristics, such as resource sharing, non-centralized control, heterogeneity, and dynamics; and application characteristics, such as parallel processing, computation or communication, hard guarantee or soft guarantee, on the application QoS delivery in Grid computing (Wu et al, 2006).