Chapter VIII
QoS in Grid Computing

Zhihui Du
Tsinghua University, China

Zhili Cheng
Tsinghua University, China

Xiaoying Wang
Tsinghua University, China

Chuang Lin
Tsinghua University, China

ABSTRACT

This chapter first summarizes popular terms of QoS related concepts and technologies in grid computing, including SLA, End-to-End QoS Provision and Virtualization. Then a three layered general grid QoS provision model based on MetaServices is proposed. Operating mechanisms are discussed in detail, and the model can maintain grid QoS by defining QoS requirements in different levels and solve the QoS problems hierarchically. A prototype named PMGrid is designed and implemented based on the QoS provision model. PMGrid is a grid system for astronomy data processing. The results show that the PMGrid can maintain the QoS requirements of astronomy data processing.

INTRODUCTION

Grid computing has been developing rapidly in recent years. Various Grid applications have involved the QoS (Quality of Service) provision problems. Since the Grid system is in essence a widely distributed and dynamically interactive system, QoS provision solutions for the Grid system are difficult and also very important.

Researches on Grid QoS started in the early 2000’s. There have been many research hotspots during these years in the Grid QoS provision area,
such as, resource reservation, QoS-based scheduling, Service Level Agreement, End-to-End QoS Provision, data transmission with QoS provision, and Virtualization technologies and so on. In this paper, we will try to clarify the basic concept of Grid QoS, establish a simple and unified Grid QoS provision model, and introduce some important Grid QoS technologies.

Grid QoS can be defined as a set of parameters on the collective behavior of one or more services in the Grid environment to prioritize different requirements. In fact, Grid QoS indicates the non-functional requirements for services provided by the Grid infrastructure. It provides criteria which can be used to measure the behavior and operations of a Grid system. Reliability, scalability and cost are typical examples of QoS requirements.

QoS requirements can be either descriptive (subjective) or accurate (objective). One QoS requirement can be the cumulative effect of multiple aspects of the behavior and operations in Grid services. Since the high-level users often use descriptive and comprehensive requirements, the Grid infrastructure should be capable of analyzing and parsing these requirements into some low-level and accurate requirements. Vice versa, the Grid infrastructure should also be able to integrate some low-level requirements into high-level and descriptive requirements.

From another point of view, Grid QoS requirements can be either elastic or inelastic. Elastic QoS requirements can take advantage of an arbitrary amount of services available in current Grid environment. On the contrary, inelastic QoS requirements often require the Grid services to reach a certain level in order to function - any more than required will be redundant and any less will cause the services non-functioning. These two different kinds of QoS requirements usually need different supporting or scheduling mechanisms.

In fact, Grid QoS provision is a must only when there is insufficient resource capacity and thus various services are competing in the Grid environment. Namely, the QoS provision is not necessary when there is enough capacity. Although the Grid infrastructure is designed in concept to provide powerful computing ability and massive storage, as the Grid users’ requirements increase with the fast deployment of Grid computing, Grid QoS provision becomes a very essential feature of the Grid infrastructure.

**BACKGROUND**

Before the descriptions of the Grid QoS provision model later in the part of “main focus”, technology hotspots of Grid QoS provision are discussed as the background in this section, including: Grid QoS negotiation protocols designed based on SLA, such as SNAP; End-to-End QoS Provision, concept borrowed from networks (Sander et al 2001); and Virtualization technologies which have been introduced into the area of Grid computing recently.

**Agreement Protocols**

In order to maintain a QoS-guaranteed communication path between services, the Service Level Agreement (SLA) is employed into Grid environment.

An SLA is a formal negotiation agreement between two parties. It is a contract that exists between customers and service providers. It records the QoS requirement of services, priorities, responsibilities, guarantee etc. (Wikipedia – Service Level Agreement).

In Grid computing, SLAs are instantiated via the Service negotiation and Acquisition Protocol (SNAP), which provides contract lifetime management (Czajkowski et al., 2002). SNAP proposes three different types of SLAs: Task service level agreements (TSLA), Resource service level agreements (RSLA), and Binding service level agreements (BSLA). It also defines an internal bundling model. Based on the standard form of SNAP, various Grid services keep consulting
Related Content

[www.igi-global.com/article/g2g-meta-grid-framework-convergence/45743?camid=4v1a](www.igi-global.com/article/g2g-meta-grid-framework-convergence/45743?camid=4v1a)

Algorithms for Maintaining Consistency of Cached Data for Mobile Clients in Distributed File System

Temporal Perspectives in Event Processing
Opher Etzion (2010). *Principles and Applications of Distributed Event-Based Systems* (pp. 75-89).
[www.igi-global.com/chapter/temporal-perspectives-event-processing/44396?camid=4v1a](www.igi-global.com/chapter/temporal-perspectives-event-processing/44396?camid=4v1a)

Towards a Mobile Augmented Reality System for Emergency Management: The Case of SAFE
[www.igi-global.com/article/towards-a-mobile-augmented-reality-system-for-emergency-management/171982?camid=4v1a](www.igi-global.com/article/towards-a-mobile-augmented-reality-system-for-emergency-management/171982?camid=4v1a)