Future Generation Computer Systems 28 (2012) 1133–1144

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

High level QoS-driven model for Grid applications in a simulated environment

Reda Albodour *, Anne James, Norlaily Yaacob

Distributed Systems and Modeling Research Group, Faculty of Engineering and Computing, Coventry University, Coventry, CV1 5FB, UK

ARTICLE INFO

Article history:
Received 10 November 2010
Received in revised form
11 May 2011
Accepted 29 June 2011
Available online 31 October 2011

Keywords:
Grid
QoS
Cost constraints
Time constraints
Simulation

ABSTRACT

This paper presents a model for high-level Quality of Service (QoS) maintenance within business-context domains and associated simulation results achieved via an expansion of the GridSim toolkit. Grid Computing traditionally has been linked with scientific environments, where heterogeneous resources were networked and employed for carrying out compute-intensive and data-intensive scientific experiments or applications that may have not been possible before. The natural progression is that business-oriented applications will build on this success and utilise the large number of heterogeneous Grid resources, potentially available. The success of introducing these applications into the mainstream is directly related to whether Grid Resource Providers can deliver a suitable level of QoS to the Grid Resource Consumer (GRC) and the ability of the GRC to request high level QoS such as the numbers of CPUs required or the RAM required, on demand. Moreover, we present dynamically calculated metrics for measuring QoS such as reliability, using up-to-date information on resources. We introduce a novel model, Business Grid Quality of Service (BGQoS), for a new generation of commercial and business-oriented Grid applications which may wish to make use of Grid environments. BGQoS allows GRCs to specify varying types of high level QoS requirements which are delivered via querying up-to-date resource information, matchmaking and monitoring operations. In addition, testing is required and this has posed a problem where testing on physical Grid test-beds is either impractical or not viable economically. Simulation is therefore important.

1. Introduction

The main aim of Grid Computing is resource sharing and the utilisation of heterogeneous and geographically distributed resources. This approach has come from different scientific and research institutions and organisations who wanted to carry out compute-intensive and data-intensive applications that required a large number of resources while also requiring them to be completed within a realistic time frame within which the results would still be applicable and viable. However, after Grids established themselves within scientific domains and environments and have allowed some of the biggest experiments in human history to be carried out, such as the Large Hadron Collider (LHC) at CERN [1,2], it was inevitable that Grid Computing would evolve to be utilised within other domains. Initially, the concept was employed to make use of unutilised resources within the same community [3] or administrative domain, progressively evolving to the current state, where resources can be in different geographical locations and owned by different entities [4]. By enabling the dynamic and coordinated sharing of resources, businesses, enterprises and individual users referred to within this paper as *Grid Resource Consumers (GRCs)*, have been able to access resources that were not accessible before Grids. Moreover, Grid Computing concepts have been used to organise, coordinate and utilise resources within the same organisation. Either way, Grid Computing provided a solution for faster, more responsive, dynamic and larger business applications [5] and commercial solutions [6]. The success of this integration between multi-domain task execution and Grids is directly related to guarantees that can be given to the GRC that they are getting the services they have asked for from the resources they acquire. Moreover, in terms of resources that are made accessible by resource providers referred to within this paper as *Grid Resource Providers (GRPs)*, it is important that they get guarantees that the policies they set are adhered to. Mainstream applications have been emerging. Using the resources Grids could offer opened up doors for carrying out applications that would not have been possible before. These new applications [6] are wrapped within a business context [7] as opposed to the traditional scientific context associated with Grid applications.

The commercialisation of Grid Computing [6,8,7,9], the introduction of variable and heterogeneous resources to new domains, provided by different GRPs, as well as a new type of user, the GRCs, which cannot be expected to have knowledge of the underlying Grid infrastructure, requires a specific set of guarantees and

* Corresponding author.
E-mail addresses: r.albodour@coventry.ac.uk, Reda.albodour@gmail.com (R. Albodour), a.james@coventry.ac.uk (A. James), n.yaacob@coventry.ac.uk (N. Yaacob).

0167-739X/$ – see front matter © 2012 Published by Elsevier B.V.
doi:10.1016/j.future.2011.06.013

guidelines for how the relationships between the different parties should be governed. These are referred to as Quality of Service (QoS) requirements. More specifically, high level QoS must be supported if the integration between different domains and Grid Computing is to be successful. This is even more important because of the inherent dynamic, heterogeneous and complex nature of Grids. This paper presents a model that controls and coordinates the required relationships between the different parties involved in this process. It also presents a new model for QoS delivery within this environment that guarantees that each GRC's requirements are met [10,11].

The rest of this paper is structured as follows. Section 2 provides an overview of related work and sets the context of the research. Section 3 presents an explanation of the research problem and how our model, Business Grid Quality of Service (BGQoS), attempts to tackle the main issues. Section 4 introduces the high level components of BGQoS. Section 5 explains the model phases of execution and QoS requirement specification. Section 6 describes the simulation approach we have taken to test our theory and includes a description of our extension to GridSim. Section 7 gives our evaluation results and Section 8 offers some conclusions and directions for further research.

2. Related work

Recently, with Grid Computing expanding towards different commercial domains [9], the provision of QoS within Grids has become an active field of research. The majority of currently existing QoS efforts in Grids concentrates on local optimal QoS scheduling. Although these approaches do take user information and resource information into consideration when allocating resources to different tasks, their concentration is on local resources and they are not designed to meet global requirements of QoS scheduling. Our proposed framework provides a direct solution to this issue, while maintaining a local approach to scheduling.

Golconda and Ozguner compared different QoS based scheduling efforts in [12]. Al Ali et al., propose the Grid QoS Management framework (G-QoSM) [13]. Their framework uses service abstraction in the OGSA (Open Grid Services Architecture) and extends it for QoS properties. G-QoSM reserves quantitative resources, such as CPUs, then allocates and monitors these resources, independently. Another, reservation-based approach is presented by Venugopal et al. [14] where they use the alternate offer protocol to make advance reservations. In both these approaches, it is assumed the all resources involved understand the reservation and negotiation protocols.

A quorum based resource allocation and management scheme is introduced by Nam and Youn [15]. Each resource Quorum includes two entities, a middleware entity and a network entity. Both of the entities can satisfy a user's QoS requirements. A heuristic algorithm is proposed by them in order to optimise the performance and cost of every Quorum. Virtual Application Service (VAS) [16] is essentially an extended Grid Service with interfaces that deal specifically with the negotiation of Service Level Agreements (SLAs). The main objective of VAS is to ensure that time-sensitive applications are carried out within the time that they are allowed and before a specific deadline, hence the user only needs to provide the time constraint when submitting a request. The system contains application information and application modelling information that are used to determine the computation resources needed to carry out a task, and reserves them. The General purpose Architecture for Reservation and Allocation (GARA) [17] is a general purpose architecture proposed by Foster et al. GARA's simple and useful reservation capability has made it popular in the Grid community with its capability to *create, modify, bind* and *cancel* reservations. Moreover, it supports flow-specific end-to-end QoS specification and resource monitoring. Curescu and Tehrani [18] propose an approach where the bandwidth is assigned such that the utility of the whole process, over time is minimised. Ghosh et al. [19] propose QoS optimisation algorithms for allocating resources to tasks in multi-processor environments. Their algorithms pick a QoS reference point, identify the number of replicas required, create the replicas, place the replicas and finally identify the number of processors required in order to maximise overall system performance QoS. Dogan and Ozguner proposed a solution to allocating individual resources according to multiple QoS requirements [20]. In their model, the cost of resources is a main factor and is not a constant, but varies during the scheduling process.

A framework providing service selection mechanisms based on QoS is presented by Taher et al. [21]. A selection manager is used by Yu and Lin [22] as a solution for the service selection problem in complex Grid services with multi-QoS requirements. The Selection Manager can be implemented as a combinatorial model or a graph based model. An heuristic is proposed for the combinatorial model based on the algorithms used for solving the multi-option, multi-dimension knapsack problem, also used by Wieczorek et al. [23], who propose an approach for modelling scheduling problems as an extension to the knapsack problem solution. The graph model, on the other hand, is based on the algorithm proposed as a solution to the multi-constraint optimal path problem in [22]. The main objective is maximising the utility of the system. To achieve this, a utility function is proposed and the algorithms attempt to maximise this function, increasing user satisfaction. Their approach is specifically tailored for the user, without taking the resource provider into consideration. A generalised resource management model is presented by Czajkowski et al. [24] where the Service Negotiation and Acquisition Protocol (SNAP) is used to map resource interactions to platform independent SLAs.

3. QoS guarantees, a new approach to the problem

The QoS requirements of mainstream GRCs are an optional set of requirements that are chosen by authorised users to describe the services they require from the resources they request. While they are optional, as some GRCs may opt for a *Best Effort* approach or may not be authorised to make specific requirements and are restricted to best effort, QoS are an essential, necessary and a vital component in maintaining a sustainable process between Grids and the mainstream. There are a number of challenges that need to be addressed:

- GRCs must be able to specify their high-level QoS requirements, if they are authorised to do so.
- Appropriate resources must be identified, located and selected.
- The consumer must be guaranteed the level of QoS the Grid Resource Providers (GRPs) promise of their resources.
- Appropriate monitoring, feedback, failure detection and reallocation methods must be in place.

3.1. The environment

As a general definition, traditionally, the resource allocation process used within Grid Computing can be split into two distinct phases, resource discovery (also sometimes merged with resource selection) and resource allocation. However, when BGQoS was designed, we separated the resource discovery and resource selection processes from each other and chose to define each to be associated with a distinct operational process.

Much previous work and research has been conducted under the headline of resource discovery where specification languages