A Study on the Effect of Application and Resource Characteristics on the QoS in Service Provisioning Environments

Theodora Varvarigou, National Technical University of Athens, Greece
Konstantinos Tserpes, National Technical University of Athens, Greece
Dimosthenis Kyriazis, National Technical University of Athens, Greece
Fabrizio Silvestri, Italian National Research Council, Italy
Nikolaos Psimogiannos, University of the Aegean, Greece

ABSTRACT

This article deals with the problem of quality provisioning in business service-oriented environments, examining the resource selection process as an initial matching of the provided to the demanded QoS. It investigates how the application and resource characteristics affect the provided level of QoS, a relationship that intuitively exists but has not yet being mapped. To do so, it focuses on identifying the application and resource parameters that affect the customer-defined QoS parameters. The article realistically centres upon modeling a data mining application and simple PC nodes in order to study how they affect response times. It moves on, by proving the existence of these specific relations and maps them using simple artificial neural networks so as to be able to wrap them in a single mechanism for resource selection based on customer QoS requirements and real time provider QoS capabilities.

Keywords: Application Modeling, QoS, Resource Modeling, Service Provisioning

INTRODUCTION

The matching of available resources to application instances is considered a fundamental problem in distributed computing, even in its current cornerstone, that is cloud computing. Cloud computing enables the virtualization of infrastructure capabilities that are delivered as a service to the customer. Even though this virtualization layer usually hides other abstraction...
layers, set up for serving interoperability and/or dynamic service provisioning such as Grid middlewares, distributed operating systems and distributed applications, they all end up in an actual resource infrastructure with several interconnected computing, storage, network and perhaps other types of resources. The conclusion is that the infrastructure (or service) provider will always want to manage the resources’ usage so as to support the application requirements in an optimal way.

This optimization aims at improving the rate of the revenue for service provision/offered quality of the service. However, quality of service (QoS) and specifically the provided level of quality, is affected by various parameters which are related to the application (i.e. security, accessibility, availability and reliability), the resources (i.e. availability, reliability, throughput and utilization) as well as to the customer-defined requirements (i.e. cost and time). It is usually possible for the service provider to “play” with the application and resource parameters in order to satisfy the customer-defined ones. So, apart from the matching of applications to resources, the service provider has to also make sure that this matching will satisfy the customer’s QoS requirements.

Service providers will always apply a scheme for load balancing however mapping the actual service provisioning level to the quality requirements of the customer is an issue that still needs to be investigated. The reason is that the problem requires to be examined by both the resource capabilities and the application requirements point of view. A supercomputing node may not be adequate to run the simplest algorithm efficiently if the data generated and consumed are huge, if the network support is inadequate or if the current load simply does not permit it.

What is needed, is for the service provider to identify the current (real time) resource characteristics and map them to the application characteristics so as to satisfy QoS requirements. For example, in a 3D rendering application service, a service provider would have to ensure that the processor that will undertake the task to render 100 frames will be able to do so in 3 seconds because the customer demands imply that (implicitly or explicitly). Nonetheless, using such a mapping mechanism would make sense only before applying the provider’s business models because this would require the incorporation of market parameters together with the QoS parameters and market parameters are extremely variable. A first mapping of quality demand to quality provision would work as a pre-selection phase for the resources that are capable to fulfil the SLA conditions.

In this article, we investigate the correlation of application and resource characteristics and how it affects the provided quality. Intuitively, we understand that there are sets of parameters that represent these characteristics and we try to examine what is their relation. Thus, we create a simple modeling scheme for describing in a unique way the resource and the application characteristics with regards to the QoS. We put our assumption to the test, by focusing on a data-mining application deployed on a Grid middleware which poses computational, storage and network challenges. Our tests lead to the development of a model that provides solid indications to a service provider of how this mapping can take place in the service provisioning environment using neural networks. The purpose of this model is not to constitute a ready exploitable mechanism but to evaluate the assumptions made in an already-deployed environment. Having stated that, our model is designed in a way to reduce assumptions, to simplify the understanding of the results and to deliver evidence that such a mapping can be implemented by a more sophisticated mechanism that will still be based on the proposed model.

The document is structured as such: The following section (State of the Art) gives the state of the art in application and resource modeling, providing examples of indicative works using neural networks; it also explains how our work progress the state of the art. The Section entitled Application Modeling, explains what application modeling means and how we can apply it specifically in a data mining
Related Content

Service-Oriented Development of Workflow-Based Semantic Reasoning Applications

Resource Co-Allocation in Grid Computing Environments
[www.irma-international.org/chapter/resource-allocation-grid-computing-environments/40814/](www.irma-international.org/chapter/resource-allocation-grid-computing-environments/40814/)

Load Balancing to Increase the Consistency of Replicas in Data Grids
[www.irma-international.org/article/load-balancing-increase-consistency-replicas/47426/](www.irma-international.org/article/load-balancing-increase-consistency-replicas/47426/)

A Dynamic Load Balancing Strategy with Adaptive Thresholds (DLBAT) for Parallel Computing System

Face Recognition: A Tutorial on Computational Aspects
[www.irma-international.org/chapter/face-recognition/139855/](www.irma-international.org/chapter/face-recognition/139855/)