Value of information based scheduling of cloud computing resources

Ladislau Bölöni, Damla Turgut

PII: S0167-739X(16)30447-2
DOI: http://dx.doi.org/10.1016/j.future.2016.10.024
Reference: FUTURE 3196

To appear in: *Future Generation Computer Systems*

Received date: 8 May 2016
Revised date: 21 September 2016
Accepted date: 23 October 2016

Value of information based scheduling of cloud computing resources

Ladislau Bölöni and Damla Turgut
Invited paper
Dept. of Computer Science, University of Central Florida

Abstract

Traditionally, heavy computational tasks were performed on a dedicated infrastructure requiring a heavy initial investment, such as a supercomputer or a data center. Grid computing relaxed the assumptions of the fixed infrastructure, allowing the sharing of remote computational resources. Cloud computing brought these ideas into the commercial realm and allows users to request on demand an essentially unlimited amount of computing power. However, in contrast to previous assumptions, this computing power is metered and billed on an hour-by-hour basis.

In this paper, we are considering applications where the output quality increases with the deployed computational power, a large class including applications ranging from weather prediction to financial modeling. We are proposing a computation scheduling that considers both the financial cost of the computation and the predicted financial benefit of the output, that is, its *value of information* (VoI). We model the proposed approach for an example of analyzing real-estate investment opportunities in a competitive environment. We show that by using the VoI-based scheduling algorithm, we can outperform minimalistic computing approaches, large but fixedly allocated data centers and cloud computing approaches that do not consider the VoI.

Keywords:

1. Introduction

We often forget that computation and networking costs money. We do not see the dollars counted down when we talk on the cellphone, have the hardware depreciation cost displayed on our laptops or the power bill showing up on the screen when we are playing on a gaming PC with a 700 Watt power source (of course, in a room illuminated by a 7W LED lightbulb). Even for high performance computing, where the computation and networking costs can be substantial, financial considerations used to come into picture only at the time of the investment decisions. Once a new supercomputer, data center or Beowulf cluster had been purchased, users were encouraged to utilize them to their maximum capacity. The grid computing model emerging in the late 1990s [1, 2] introduced the ability of requesting computational power on demand, on the analogy of the power grid. Nevertheless, all grid systems had been financed by national research foundations and thus, the objective of maximum utilization remained in place.

Cloud computing introduced a significant change. In some ways, cloud computing is fulfilling the promise of grid computing of providing on demand, on a very short notice, an amount of computational, storage and networking capacity, normally in the form of virtualized resources. For instance, computational capacity can be offered in form of virtual machines or containers. For instance, virtual machine on demand services provide the user with a remotely allocated virtual machine, in which the client can run its own operating system. These services can be *public clouds*, fully managed services offered by a third party vendor, such as in the case of Amazon's EC2 service [3, 4], Rackspace's OpenStack on Demand service or VMWare's vCloud Air service. In these cases, the cloud provider charges in actual dollars for the provided, metered computing capacity. The client makes a request for a computing unit, and in less then a minute, he can log in and start the computation. While computing, the client will pay an hourly fee. Once the computation is terminated, the client discards the virtual machine. Alternatively, companies can create *private clouds* in which they are offering similar services for their internal divisions. This allows companies to more efficiently share their existing computational resources. Large companies have developed their internal software architectures to provide computing on demand (this being, for instance, the casse of Google's Borg system [5]). Other companies can use available open source software to provide computing-on-demand solutions, for instance using OpenStack Nova system [6] for virtual machines on demand, or Apache Mesos or Google's Kubernetes for containers on demand. In Borg, company subdivisions need to purchase computing quotas with actual money, thus the principles of operation are similar to the one in public clouds.

In such systems, the amount of computation that a client can obtain is limited, in principle, only by the client's ability to pay. For instance, a client can pay $168, and run a parallel computation on 10,000 computer cores, a computing power that was out of reach to anyone without a million dollar investment. For variable and unpredictable loads, such as the sudden popularity and just as sudden fading of a game or app, cloud computing might be the only approach that can trace the variations of the computing demand. Of course, if one would use the 10,000 computer