Chapter 23
Resource Provisioning in the Cloud: An Exploration of Challenges and Research Trends

Ming Mao
University of Virginia, USA

Marty Humphrey
University of Virginia, USA

ABSTRACT

It is a challenge to provision and allocate resources in the Cloud so as to meet both the performance and cost goals of Cloud users. For a Cloud consumer, the ability to acquire and release resources dynamically and trivially in the Cloud, while being a powerful and useful aspect, complicates the resource provisioning and allocation task in the Cloud. While on the one hand, resource under-provisioning may hurt application performance and deteriorate service quality; on the other hand, resource over-provisioning could cost users more and offset Cloud advantages. Although resource management and job scheduling have been studied extensively in the Grid environments and the Cloud shares many common features with the Grid, the mapping from user objectives to resource provisioning and allocation in the Cloud has many challenges due to the seemingly unlimited resource pools, virtualization, and isolation features provided by the Cloud. This chapter focuses on surveying the research trends in resource provisioning in the Cloud based on several factors such as the type of the workload, the VM heterogeneity, data transfer requirements, solution methods, and optimization goals and constraints, and attempts to provide guidelines for future research.

DOI: 10.4018/978-1-4666-6178-3.ch023

INTRODUCTION

The Cloud has become a significant computing platform. It has attracted many businesses and individual users by offering on-demand computing power and storage capacity. The economies of scale and *pay-as-you-go billing model* could save users large up-front capital investments and long term operation costs (Armbrust et al., 2010). A key feature of the Cloud is the *elasticity*, the ability to dynamically acquire and release computing resources in response to demand (Mell & Grance 2011). For successful resource management in the Cloud, one needs to first determine and acquire the required amount and type of resources that may be needed to satisfy a computing job, which is the act of *provisioning*; and then place computing activities onto each of the resources in a dynamic and efficient manner, which is the act of *allocation*. The provisioning and allocation of Cloud resources is a challenging problem because the mapping from user objectives to the resource provisioning and allocation plans is not trivial (Mell & Grance, 2011; Buyya et al., 2009).

Resource provisioning and allocation in the Cloud needs to consider the following factors:

- A performance goal may be achieved through different types of resources with different costs.
- A fixed budget may be used to rent a wide variety of resource configurations for varying durations.
- Task precedence constraints may need to be preserved for a job.
- The workload may experience unexpected peaks.
- The performance requirements and cost constraints may change dynamically.

One of the main advantages of the Cloud is the rapid elasticity or dynamic scalability (Armbrust et al., 2010), which enables the dynamic acquisition and release of Cloud resources in response to demand. It is a key enabler of Cloud adoption. This elasticity saves the Cloud users large up-front capital investments and allows the computing resources to grow according to business demand. The elasticity has become one of the main forces to drive application migration to the Cloud.

Another important factor when considering Cloud adoption is the cost aspect (Armbrust et al., 2010). Maximizing the return and minimizing the cost of Cloud investment are the two main goals for Cloud users. The cost savings result from the economies of scale and dynamic scalability. For the same business scenarios, if it is more expensive to develop and maintain the applications in the Cloud, the Cloud loses its advantages. Therefore, cost savings is a significant concern for Cloud users. In addition to being a goal, sometimes cost could become a constraint for some Cloud applications. For example, Cloud consumers may have a budget limit that they are allowed to spend on Cloud purchasing which restricts the running cost of the acquired resources from exceeding a certain amount. In such cases, cost essentially determines the maximum size of the acquired resource pool.

In summary, the key benefit of the Cloud is to be able to acquire resources in response to demand dynamically and only pay for the resources used. This benefit can only be realized when the Cloud users can determine the right size of the resource pool and allocate the resources in a cost-effective way. While resource over-provisioning can cost users more than necessary, which essentially offsets the Cloud advantages; resource under-provisioning hurts the application performance and could violate the service level agreements that service providers on the Cloud have with their customers, causing customers to turn away. Essentially, the Cloud adopters should understand what resources should be acquired or released