Chapter 8
Dynamic Rightsizing with Quality-Controlled Algorithms in Virtualization Environments

Ming-Jeng Yang
Mackay Medical College, Taiwan

Chin-Lin Kuo
National Taiwan Normal University, Taiwan

Yao-Ming Yeh
National Taiwan Normal University, Taiwan

ABSTRACT

Virtualization and partitioning are the means by which multiple application instances can share and run multiple virtual machines supported by a platform. In a Green Cloud environment, the goal is to consolidate multiple applications onto virtual machines associated by fewer servers, and reduce cost and complexity, increase agility, and lower power and cooling costs. To make Cloud center greener, it is beneficial to limit the amount of active servers to minimize energy consumption. This paper presents a precise model to formulate the right-sizing and energy-saving mechanism, which not only minimizes energy consumption of the server but also maintains a service quality through the $M/M/V_t$ strategy of queuing theory. The authors map the complicated formula of the energy-saving mechanism to an approximation equation and design the fast decidable algorithms for calculating the right size of virtual machines in constant time complexity for power management systems.

1. INTRODUCTION

A 2010 article "A View of Cloud Computing" by Armbrust et al. (2010) defined the cloud computing as refers to both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services. The cloud itself comprises many services running on a set of highly configurable and dynamically configurable hardware and software resources (Winkler, 2009). The cloud technology offers a wide range of services such as infrastructure-as-a-service (IaaS), software-as-a-service (SaaS), and platform as a service (PaaS)

DOI: 10.4018/978-1-4666-2065-0.ch008

(Prodan & Ostermann, 2009). In the characteristic view, the cloud technology has ultra-large-scale, virtualization, high reliability, versatility, high extendibility, on demand service, as well as non-expensive, etc. (Zhang, Zhang, Chen, & Huo, 2010).

Basically, Cloud Computing is a model in which IT infrastructure and software are offered as services to users over the Internet. Winkler (2009) mentioned that cloud computing holds great promise in energy reduce and greenhouse effect. Cloud computing, he said, was the "green computing option". The potential and the reasons are as follows (Winkler, 2009):

- Shared resources in cloud can "eliminate redundancies".
- "Dynamically-assigned resource pools" means that spare capacity isn't sitting around drawing power in as many places or as many configurations.
- "Location independence" could mean the ability to move services to physical facilities where power is cleaner or used more efficiently. The tyranny of the speed of light will certainly limit "follow the moon" - changing longitudes daily to take advantage of evening cooling. But it could work for some applications; more will be able to "follow the seasons"; shifting biannually by latitude to take advantage of winter cooling.
- "Properly instrumented", clouds will be able to inform consumers of their environmental (energy, carbon, etc.) impact, to enable them to be accountable for their choices.
- Clouds can provide an "elastic infrastructure" to enable retro-commissioning of more traditional infrastructure.
- Clouds hold the potential for rapid connection between disparate sources of people and data to "accelerate innovation" to address all sorts of environmental challenges.

The above potential means that if we put it on the action and take advantage of cloud computing, we could integrate environment resources to reduce energy consumption. However, as the cloud computing is available and popularized, the data center energy consumption in cloud is growing. Therefore it is imperative to reduce energy consumption of servers of cloud's data center. Since the server is considered as a vital supply resource, the work pattern of the servers is to keep open and turning regardless of if they actual perform of the services. But now we must face the green efficiency considerations (Blackburn, 2008) required to abandon the above concept.

Recently, Cloud services have driven the growth of server farms in Cloud centers. Most of these servers are vastly over-provisioned. The analysis of server usage patterns will reveal the potential for 'right-sizing'. Unneeded capacity can be turned off, but the server farm can still provide sufficient resiliency for agreed upon service levels. Modern Cloud centers use virtualization (Xen, VMware, and Hyper-V) to get better performance through resource consolidation and live migration (Beloglazov & Buyya, 2010). Consolidating multiple servers running in different virtual machines (VMs) on physical machines (PMs) increases the overall utilization and efficiency of the equipment across the whole deployment.

In this paper, we present the Mt/M/Vt strategy of queuing theory model to address how to efficiently manage energy in virtual servers. Also the formulations are calculated to show the best way that can control some of the virtual machines into the energy-saving states. In addition, Abdelsalam et. Al. (2010) have proposed schemes by calculating model in the data center through validating average distribution of workload to achieve overall minimum energy consumption for servers. In this paper, we propose an energy-saving strategy of cloud data center by rightsizing the virtual machine supported by virtual platform. Based on efficient consideration, we develop computational models to minimize energy consumption of server but