Chapter 105
Novel Resource Allocation Algorithm for Energy-Efficient Cloud Computing in Heterogeneous Environment

Wei-Wei Lin
South China University of Technology, China

Liang Tan
Zhejiang University, China

James Z. Wang
Clemson University, USA

ABSTRACT

Energy efficiency is one of the most important design considerations for a cloud data center. Recent approaches to the energy-efficient resource management for data centers usually model the problem as a bin packing problem with the goal of minimizing the number of physical machines (PMs) employed. However, minimizing the number of PMs may not necessarily minimize the energy consumption in a heterogeneous cloud environment. To address the problem, this paper models the resource allocation problem in a heterogeneous cloud data center as a constraint satisfaction problem (CSP). By solving this constraint satisfaction problem, an optimal resource allocation scheme, which includes a virtual machine provision algorithm and a virtual machine packing algorithm, is designed to minimize the energy consumption in a virtualized heterogeneous cloud data center. Performance studies show that this proposed new scheme outperforms the existing bin-packing based approaches in terms of energy consumption in heterogeneous cloud data centers.

DOI: 10.4018/978-1-4666-6539-2.ch105

1. INTRODUCTION

With widely acceptance of cloud computing in industry (Buyya & Venugopal, 2008), reducing energy consumption has become important for cloud resource providers to decrease operating cost and maximize the profit (Srikantaiah & Kansal, 2009; Berl & Gelenbe, 2010). Hence, energy efficiency has emerged as one of the most important design considerations for cloud data centers (Beloglazov & Buyya, 2011). Many studies (Bichler, Setzer & Speitkamp, 2006; Khanna, Kar, & Kochut, 2006; Verma, Ahuja, & Neogi, 2008, Hermenier, Lorca, Menaud, Muller, & Lawall, 2009; Rietz, Macedo, Alves, & Carvalho, 2011) have been conducted recently to address the energy consumption issues in a cloud computing environment. Garg et al. (Garg, Yeo, Anandasivam & Buyya, 2011) proposed near-optimal scheduling policies which consider the dynamic change of numeric energy efficiency factors (such as energy cost, carbon emission rate, workload, and CPU power efficiency) across different data centers depending on their locations, architectural designs, and management systems. Beloglazov et al. (Beloglazov & Buyya, 2010) proposed an energy efficient resource management scheme for virtualized cloud data centers that reduces operational cost while providing the required Quality of Service (QoS). In this scheme, energy savings are achieved by continuous consolidation of VMs according to current resource utilization, virtual network topologies between VMs, and thermal states of computing nodes. Rodero et al. (Rodero, Jaramillo, Quiroz, Parashar & Guim, 2010) presented an energy-aware online provisioning approach for HPC applications on consolidated and virtualized computing platforms. Energy efficiency is achieved by using a workload-aware, just-right dynamic provisioning mechanism with an assumption of the ability to power off subsystems of a host system that are not required by the VMs mapped to it. Abdelsalam et al. (Abdelsalam, Maly, & Kaminsky, 2009) created a mathematical model for power management of a cloud computing environment that primarily serves clients with interactive applications such as web services. The mathematical model computes the optimal number of servers and the frequencies at which they should run. In the literature (Younge, Laszewski, & Wang, 2010) a new framework of a scalable cloud computing architecture was presented. In this framework, power-aware scheduling techniques, variable resource management, live migration, and a minimal virtual machine design are utilized to improve the overall system efficiency with minimal overhead. Chang et al. (Chang, Ren, & Viswanathan, 2010) studied the optimal resource allocation problem in clouds by formulating demand for computing power and other resources as a resource allocation with multiplicity. They proposed an algorithm with an approximation bound that can yield near optimal solutions in polynomial time.

However, all of these current approaches assumed the homogeneity of physical resources, i.e., all nodes have the same characteristics such as CPU speed, disk, memory, etc. In addition, most of these approaches considered the processor as the only resource that contributes to the energy consumption, leaving out other important system resources such as memory and disk storage. Sometimes, due to the massive aggregation of workloads in the same node, the present of bottle-necks in these other system resources may reduce the performance of the applications and increase the energy consumption as mentioned in the literature (Lee, & Zomaya, 2012).

The ability to reallocate VMs in run-time enables dynamic consolidation of the workload to reduce the energy consumption, as VMs can be moved to a minimal number of physical nodes and idle nodes can be switched to power saving modes. However, VM migration leads to time delays, performance overhead and extra power