Chapter 5
Green Cloud Computing with Efficient Resource Allocation Approach

Fei Cao
University of Central Missouri, USA

Michelle M. Zhu
Southern Illinois University – Carbondale, USA

Chase Q. Wu
New Jersey Institute of Technology, USA

ABSTRACT

Due to the increasing deployment of data centers around the globe escalated by the higher electricity price, the energy cost on running the computing, communication and cooling together with the amount of CO₂ emissions have skyrocketed. In order to maintain sustainable Cloud computing facing with ever-increasing problem complexity and big data size in the next decades, this chapter presents vision and challenges for energy-aware management of Cloud computing environments. We design and develop energy-aware scientific workflow scheduling algorithm to minimize energy consumption and CO₂ emission while still satisfying certain Quality of Service (QoS). Furthermore, we also apply Dynamic Voltage and Frequency Scaling (DVFS) and DNS scheme to further reduce energy consumption within acceptable performance bounds. The effectiveness of our algorithm is evaluated under various performance metrics and experimental scenarios using software adapted from open source CloudSim simulator.

INTRODUCTION

The flexible utility-oriented pay-as-you-go Cloud computing model has demonstrated tremendous potential for both commercial and scientific users to access and deploy their applications anytime from anywhere at reasonable prices depending on their QoS specifications. Gartner estimated that the market opportunity for Cloud computing will be worth around $150 billion by 2014 (Gartner Newsroom,
Green Cloud Computing with Efficient Resource Allocation Approach

2007). The computing power of the Cloud environment is supplied by a collection of data centers that are typically installed with hundreds to thousands of servers which are built on virtualized compute and storage technologies. Meanwhile equally massive cooling systems are required to keep the servers within normal operating temperatures. Servers and cooling systems make up about 80% of all the electricity used within a data center (Data Center Energy Characterization Study Site Report, 2007). However, these infrastructures consume tremendous amounts of energy. For example, a typical data center with 1000 racks consumes about 10 Megawatt of power during normal operation (Garg, Yeo, Anandasivam, & Buyya, 2009). The average data center consumes as much energy as 25,000 households (Kaplan, Forrest, & Kindler, 2008). Over the past decade, the cost of servers running and cooling systems has increased by 400%, and such cost is expected to continue to rise (Filani et al., 2008). Following the current usage and efficiency trends, energy consumption by data centers could nearly double in another five years to more than 100 billion kWh (U.S. Environmental Protection Agency ENERGY STAR Program, 2007). A simple metric to gauge the energy efficiency of a data center is the computer power consumption index, which is the fraction of the server power consumption to total power consumption including servers, cooling systems, lighting and space, etc. A realistic maximum power consumption index lies between 0.8 to 0.9, depending on the climate condition with common values from 0.30 to 0.75 (Greenberg, Mills, Tschudi, Rumsey, & Myatt, 2006). Besides the energy cost, these data centers also produce considerable amount of CO₂ emissions which significantly contribute to the growing environmental issue of Global Warming. Gartner estimated that the Information and Communication Technologies (ICT) industry generates about 2% of the total global CO₂ emissions in 2007 (Gartner Newsroom, 2007). As the governments start to impose CO₂ emissions limits on various industries such as automobile industry (EUbusiness, 2007), Cloud providers should also ensure that their data centers are CO₂ emission regulation compliant to meet the future permissible restrictions (Brill, 2009; Dunn, 2010). Reducing energy consumption for modern data centers has been recognized as an ever increasingly important technique for operation cost, environment footprint and system reliability. Furthermore, less energy consumption means less heat generated to maintain the system in a relatively cool temperature to reduce the hardware related failures for longer Mean Time Between Failures (MTBF). Hence, energy-efficient Cloud computing technologies are highly desirable for future sustainable ICT (Berl et al., 2010) for cost effectiveness, environmental friendliness as well as stable system operation.

We want to study the problem of supporting scientific applications over the Cloud. There are many scientific Cloud projects such as Nimbus (Nimbus, 2014), Eucalyptus (Nurmi et al., 2009) and OpenNebula (Nebula, 2014) as well as some Cloud systems such as DOE Magellan (Magellan: Cloud Computing for Science, 2014) to discover the potential and challenges for scientific Cloud computing. One distinguished feature is that many scientific applications are modeled as workflows, which can be as simple as a single task or as complex as a Directed Acyclic Graph (DAG). The dependency and parallelism embedded in a workflow requires that the tasks be dispatched to a group of distributed VMs in order to maximize the execution efficiency. Also the Cloud resource availability map is time-dependent as many Cloud providers support both on-demand and reservation VM allocations. Our Cloud model accommodates these issues to establish a realistic and useful testbed for energy-efficient scientific computing research.

Data centers make use of virtualization which allows a single server to run multiple operating instances simultaneously (Weiss, 2007) to achieve efficient computing resource sharing. The startup, idle and shutdown time compose the overhead of resource utilization. The average resource utilization in most data centers can be as low as 20% and the energy consumption on idle resources can be as high as 60% (Valentini, Khan, & Bouvry, 2012). The overhead could result in resource under-provisioning, which