Expert Systems with Applications 41 (2014) 5804–5816

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Hybrid shuffled frog leaping algorithm for energy-efficient dynamic consolidation of virtual machines in cloud data centers

Jian-ping Luo *, Xia Li, Min-rong Chen

*College of Information Engineering, Shenzhen University, Shenzhen 518060, PR China*

ARTICLE INFO

*Keywords:*
Cloud computing
Data center
Resource management
Intelligent algorithm

ABSTRACT

Cloud computing aims to provide dynamic leasing of server capabilities as scalable virtualized services to end users. However, data centers hosting cloud applications consume vast amounts of electrical energy, thereby contributing to high operational costs and carbon footprints. Green cloud computing solutions that can not only minimize the operational costs but also reduce the environmental impact are necessary. This study focuses on the Infrastructure as a Service model, where custom virtual machines (VMs) are launched in appropriate servers available in a data center. A complete data center resource management scheme is presented in this paper. The scheme can not only ensure user quality of service (through service level agreements) but can also achieve maximum energy saving and green computing goals. Considering that the data center host is usually tens of thousands in size and that using an exact algorithm to solve the resource allocation problem is difficult, the modified shuffled frog leaping algorithm and improved extremal optimization are employed in this study to solve the dynamic allocation problem of VMs. Experimental results demonstrate that the proposed resource management scheme exhibits excellent performance in green cloud computing.



## 1. Introduction

Cloud computing is a cost-effective service delivery model that makes IT management and maintenance easy; it can rapidly adapt to changing business needs. Cloud computing can be roughly divided into three categories in accordance with the service type, namely, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) (Gandhi, Harchol Balter, Das, & Lefurgy, 2009). Users can easily avail of these services on a pay-as-you-use basis without any geographical restrictions. Cloud computing is a convenient, necessary, and shared network computing resource with a dynamic configuration (including network, service, storage, and application); it can be managed easily and effectively by the resource manager at minimal management costs (Gandhi et al., 2009). Cloud service providers need to consider the convenience of cloud services without placing too much emphasis on the hardware facilities to extend the computing potential of a computing center. Thus, different cloud users can enjoy cloud computing capabilities easily and efficiently.

The core of the cloud computing environment is the cloud data center, which often consists of a number of highly configured hardware facilities. The computing capability of the data center is the main indicator considered by cloud service providers. With the appearance of increasingly large data centers, the energy consumption of data centers increases as well. This increase in energy consumption affects the environment. For instance, carbon dioxide emissions cause the greenhouse effect and have a serious impact on climate and the environment. These effects will eventually affect the operational benefits of a cloud service provider. Statistics show that the energy consumption of a common data center is approximately equal to that of 25,000 ordinary households (Kaplan, Forest, & Kindler, 2008). The energy consumption of data centers in the United States from 2008 to 2010 was supplied by 10 nuclear power stations. Clearly, more data centers cause higher energy consumption. However, the energy consumption of data centers has rarely elicited attention. Therefore, the computing capacity performance of data centers and the problem of huge energy consumption by these centers must be considered in addressing the resource management optimization issue of data centers. Green computing is the only method to enhance the operational efficiency of data centers and reduce damage to the environment. The designers of these centers should consider the efficient use of computing resources on the premise that the system maintains excellent computing service capacity. The hosts of data centers in idle or low utilization status also consume vast amounts of energy. For instance, the energy consumption of a

* Corresponding author. Tel.: +86 15889669910.
*E-mail address:* camelrock@126.com (J.-p. Luo).

http://dx.doi.org/10.1016/j.eswa.2014.03.039
0957-4174/

non-operative server (not turned off) accounts for approximately 70% of the full load energy consumption (Kusic, Kephart, Hanson, Kandasamy, & Jiang, 2009). Similar conclusions can also be found in the literature (Fernández-Montes, Gonzalez-Abril, Ortega, & Lefèvre, 2012).

In reality, hardware facilities in cloud data centers do not remain static for a long time. The state of most hardware facilities often change. First, adding a new server to the system causes the reconfiguration, restoration, or replacement of the existing server. Second, the resource pool often changes the operating status to satisfy the requirement of intermittent resource changes. Third, live migration causes virtual machines (VMs) to achieve rapid reconfiguration and consolidation in different physical nodes to achieve goals, such as load balancing. Fourth, some servers need to transfer VMs to the appropriate servers; shutdown, maintenance, and restart are operated after the live migration, thereby causing some servers to become unavailable. Similarly, some servers need to be temporarily open to address unpredictable access peak outbursts. Access to a server involves a certain degree of uncertainty. In addition, changes could occur within each server, such as changes in processing elements (PEs), memory size, hard disk storage, and bandwidth. Current servers generally support dynamic voltage frequency scaling (DVFS). In DVFS, the server can adjust the operating frequency based on the current load to achieve energy saving. Adopting a dynamic regulatory mechanism is necessary for the dynamics and uncertainty of the resource. Determining the transition state of the source, placing the new applied VMs reasonably, and optimizing the allocation of VMs that violate the service level agreements (SLAs) or VMs in the server with very low CPU utilization are also necessary. These measures are essential to ensure that the entire configuration of the VMs is as optimal as possible. A dynamic regulatory mechanism can be utilized to achieve automatic and dynamic management without managerial or staff intervention. In a cloud environment, the data center is usually presented in an over-provisioned state to meet the uncertain resource application peak, thereby resulting in large amounts of energy waste.

A complete data center resource management scheme is proposed for the Infrastructure as a Service (IaaS) cloud environment in this paper. The proposed scheme can not only guarantee user quality of service (QoS) specified by SLAs but can also achieve maximum energy saving and green computing goals. The main contribution of this study is the proposal of the complete dynamic resource management scheme. Consolidation of resources is achieved by VM migrations technology and low-utilized or idle hosts switched to power saving mode to achieve energy saving while ensuring that SLAs are adhered to. The intelligent method of modified shuffled frog leaping algorithm (SFLA) based on improved extremal optimization (EO) is applied to rapidly and efficiently complete the dynamic allocation of VMs.

This paper is organized as follows. Section 2 provides a brief review of related research on energy conservation and dynamic VM allocation technologies. Section 3 discusses the intelligent resource management scheduling framework. An intelligent hybrid algorithm for the dynamic VM consolidation problem is proposed in Section 4. Section 5 presents the experimental evaluations and result discussions. Section 6 provides the conclusion and suggestions for future work.

## 2. Related work

DVFS is a means of achieving hardware facilities energy conservation (Chase, Anderson, Thakar, Vahdat, & Doyle, 2001). DVFS, which is based on the different needs of the computing capacity by application program, dynamically adjusts the running frequency and voltage of the chip (for the same chip, the higher the frequency, the higher the voltage) to achieve energy saving. A low frequency can reduce power. However, energy cannot be saved by simply reducing the frequency because for a given task, energy consumption can only be reduced when the voltage and frequency are lowered simultaneously. The implementation of DVFS depends on the successful prediction of the number and execution time of processing tasks in the server. Clock frequency and voltage do not have a linear relation in a real-time system. Much uncertainty exists among task execution time, energy consumption, and processor voltage. Inappropriate frequent voltage adjustment will degrade processor performance. Predicting the number of tasks is difficult in most cloud environments.

DVFS generally requires power management through BIOS. Circuits designed by different manufacturers differ significantly. Intel, Microsoft, Toshiba, and other companies jointly developed the advanced configuration power interface (ACPI) specification to establish a common power management interface between the operating system and the hardware facilities (Venkatachalam & Franz, 2005). ACPI improves the original APM through BIOS and provides a relatively good power management mode and interface specification in configuration management. ACPI sets a maximum of six power states. Different states correspond to the different power consumptions of the processor, memory, and hard disk. Most processors at present support the states running, idling, sleeping, and closed.

Rusu, Ferreira, Scordino, and Watson (2006) proposed an energy consumption management strategy based on QoS in connection with the server cluster system. The strategy is divided into back-end management and local management. Local management supports DVFS. When the back-end manager detects that the system needs to close or open a server, the local manager controls power by the DVFS module and switches the server into the corresponding state. The system does not utilize live VM migrations technology; it involves the off-line calculation of the back-end server to decide whether to shut down or open the server. Such decision is limited for energy saving.

Lee and Zomaya (2010) proposed an efficient energy management strategy in a distributed cloud computing system. The researchers defined the optimized objective function as a relative superiority (RS) expression according to the relation between task processing time and energy consumption. The RS value for the assigned task is calculated first. The task is then allocated to the server with maximum RS value. This algorithm assumes that all servers are active and in good running condition and ignores the heterogeneity of the system. Only the allocation of the newly added VMs was considered in this study. In the actual process, the adjustment of VMs with SLA violation should be considered as well.

Kusic et al. (2009) defined the problem of energy management in virtual heterogeneous environments as a scheduling optimization problem and applied limited look-ahead control (LLC) during processing. LLC aims to maximize the profit of the resource service providers and minimize energy consumption and SLA violation. The system applies the Kalman filter to estimate the number of future client requests and predict the state of the system to achieve necessary resource integration. However, the system cannot be implemented in the IaaS cloud environment. In addition, the model is extremely complicated. Adjusting a system with 15 nodes requires 30 min. Thus, the model is difficult to apply to large-scale and real-time cloud data centers.

Verma, Ahuja, and Neogi (2008) modeled the energy-aware dynamic arrangements of VMs in the virtual heterogeneous environment and turned them into a continuous optimization problem; the placement of VMs was regarded as the minimum energy consumption and the maximum performance optimization issue in each time frame. The researchers utilized a heuristic algorithm to solve the packing problem and re-allocated the VMs of each time frame by live migration strategy. In their follow-up work