Efforts Made Towards Achieving Energy Efficiency in Cloud Computing - A Survey

[1]Sasikala.P, [2]Dr. Suresh.S
[1] P.G Scholar, [2]Associate Professor,
Department of Computer Science and Engineering,
Adhiyamaan College of Engineering, Hosur,
Tamilnadu, India

ABSTRACT-- **Cloud computing has become an attractive technology to business persons due to its features such as virtualization and "pay-as-you-go" model. Now-a-days, balancing performance and energy consumption is a major issue in large data centers. Energy has become a leading design constraint for developers and they try to explore new techniques to reduce energy consumption. This work is to address the need for achieving energy efficiency, and to survey various recent techniques used to reduce energy consumption and compare those techniques based on their characteristics. This paper aims to investigate this issue and review the current state of energy consumption in large datacenters and identify the common pitfalls among surveyed results.**

Index Terms-- **Cloud Computing, Energy efficiency, DVFS, Sleep states**

---

1. INTRODUCTION

Cloud computing is a natural evolution of the widespread adoption of multiple technical advances in distributed computing including virtualization, grid computing, autonomic computing, utility computing and software-as-a-service [1]. It has become an efficient technology to offer computational capabilities as services on a "pay-as-you-go" model.

As this technology gets smarter it has to address many challenges like automated service provisioning, virtual machine migration, server consolidation, energy management, traffic management and analysis, data security, software frameworks, storage technologies and data management and novel cloud architectures as shown in figure 1[2].

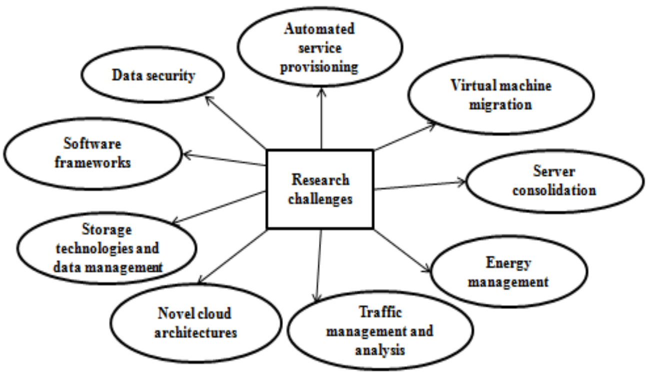

Fig 1. Challenges in cloud computing

2. THE NEED FOR ENERGY CONSERVATION

Improving energy efficiency is considered as the major issue in cloud computing. It has been estimated that the cost of powering and cooling accounts for 53% of the total operational expenditure of data centers. In 2006, data centers in the US consumed more than 1.5% of the total energy generated in that year, and the percentage is projected to grow 18% annually. Hence infrastructure providers are under enormous pressure to reduce energy consumption. The goal is not only to cut down energy cost in data centers, but also to meet government regulations and environmental standards [2].

Energy conservation is a major concern today. It has become a leading design constraint for various developers. They try to explore new techniques to reduce energy consumption in large datacenters [3]. Energy saving is important not only for economical and environmental reasons, but also for sustainable growth of cloud computing. On the other hand there should not be any compromise in achieving the desired level of Quality of Service (QoS) between cloud providers and users.

The large power consumption and high concentration of nodes in data centers leads to increased node failures. It has been observed that a 15 degree Celsius rise increases the failure rates in hard-disks by a factor of two. Hence, maintaining the computer systems at proper temperature is important for ensuring maximum reliability, longevity, and large return on investment. Finally, large power consumption also has adverse environmental impact, e.g. large carbon emission. In this paper we highlight the need of achieving energy efficiency in data centers and survey various techniques used to achieve energy efficiency.

3. ENERGY EFFICIENT TECHNIQUES

In the recent years researchers have proposed many algorithms and techniques for managing power consumption in data centers. We are going to focus on techniques which are based on Dynamic Voltage Frequency Scaling (DVFS) and switching the idle nodes to sleep mode. Based on these parameters / characteristics the techniques can be further classified. In the table 1 the various methodologies used are tabulated.

TABLE 1

YEAR	TITLE	METHODOLOGY USED
2015	A Stochastic Approach to Analysis of Energy-Aware DVS-Enabled Cloud Datacenters	Dynamic Voltage Frequency Scaling
2010	Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers	Switching idle nodes to sleep mode
2014	An Algorithm-Centric Energy-Aware Design Methodology	Energy evaluation method
2015	An Energy-Saving Task Scheduling Strategy Based on Vacation Queuing Theory in Cloud Computing	Energy-saving task scheduling algorithm based on vacation queuing theory
2013	Bee-MMT: A Load Balancing Method for Power Consumption Management in Cloud Computing	Bee-MMT (artificial bee colony algorithm- Minimal migration time), Migrating the virtual machines and Switching idle nodes to sleep state
2010	Capacity Planning and Power Management to Exploit Sustainable Energy	Power supply based on time-varying workloads hosted in data centers
2013	Energy-Aware Scheduling Scheme Using Workload-Aware Consolidation Technique in Cloud Data Centers	Energy-aware Scheduling algorithm using Workload-aware Consolidation Technique (ESWCT) and the Energy aware Live Migration algorithm using Workload-aware Consolidation Technique (ELMWCT), Switching idle nodes to sleep mode
2015	Energy-Efficient Resource Allocation and Provisioning Framework for Cloud Data Centers	Prediction based resource estimation, Switching idle nodes to sleep state
2011	How Data Center Size Impacts the Effectiveness of Dynamic Power Management	Analyze the effectiveness of Dynamic power management system
2013	Integrated Approach to Data Center Power Management	Power proportional approach and Green data center approach
2005	Managing Server Energy and Operational Costs in Hosting Centers	A hybrid mechanism from steady state queuing analysis and feedback control theory
2012	Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers	Live migration and Switching idle nodes to sleep mode
2009	Optimal Power Allocation in Server Farms	Predicting the optimal power to be allocated
2009	Power-aware Provisioning of Cloud Resources for Real-time Services	Dynamic Voltage Frequency Scaling
2013	Simple and Effective Dynamic Provisioning for Power-Proportional Data Centers	Predicting the future workload and Switching idle nodes to sleep mode

DVFS Based Techniques

Xia and Zhou [5] proposed a novel stochastic framework for energy efficiency and performance analysis of DVS enabled cloud. This framework used virtual machine request arrival rate, failure rate, repair rate, and service rate of datacenter servers as model inputs and DVFS as a key technique in exploiting the hardware characteristics of cloud datacenters to save energy by lowering the supply voltage and operating frequency.

Gmach et al [10] describes an approach for designing a power management plan that matches the supply of power with the demand for power in data centers. The demand for power is mainly determined by the time-varying workloads hosted in the data center and the power management policies implemented by the data center.

Kim et al [18] investigated power-aware provisioning of virtual machines for real-time services. Their approach is (i) to model a real-time service as a real-time virtual machine request; and (ii) to provision virtual machines of datacenters using DVFS (Dynamic Voltage Frequency Scaling) schemes.

State transitioning based techniques

Gandhi et al [4] investigated the regime of sleep states that would be advantageous in data centers. They considered the benefits of sleep states across three orthogonal dimensions: (i) the variability in the workload trace, (ii) the type of dynamic power management policy employed, and (iii) the size of the data center.

Ghafari et al [9] developed a new power aware load balancing, named Bee-MMT (artificial bee colony algorithm-Minimal migration time), to decline power consumption in cloud computing; as a result of this decline, CO2 production and operational cost were decreased. According to this purpose, an algorithm based on artificial bee colony algorithm (ABC) has been proposed to detect over utilized hosts and then migrate one or more VMs from them to reduce their utilization; following that they detected underutilized hosts and, if it was possible, migrated all VMs which have been allocated to these hosts and then switched them to the sleep mode.

Hongyou et al [11] proposed two algorithms called the Energy-aware Scheduling algorithm using Workload-aware Consolidation Technique (ESWCT) and the Energy aware Live Migration algorithm using Workload-aware Consolidation Technique (ELMWCT). Both algorithms investigated the problem of consolidating heterogeneous workloads. They try to execute all Virtual Machines (VMs) with the minimum amount of Physical Machines (PMs), and then powered off unused physical servers to reduce power consumption.

Dabbagh et al [12] proposed an integrated energy-aware resource provisioning framework for cloud data centers. The proposed framework: i) predicts the number of virtual machine (VM) requests, to be arriving at cloud data centers in the near future, along with the amount of CPU and memory resources associated with each of these requests, ii) provides accurate estimations of the number of physical machines (PMs) that cloud data centers need in order to serve their clients, and iii) reduces energy consumption of cloud data centers by putting to sleep unneeded PMs. This framework is evaluated using real Google traces collected over a 29-day period from a Google cluster containing over 12,500 PMs.

Ganesh et al [14] proposed an integrated approach combining two distinct approaches: the power-proportional approach focuses on reducing disk and server power consumption, while the green data center approach focuses on reducing power consumed by support infrastructure like cooling equipment, power distribution units, and power backup equipment. Their solution enforces power-proportionality at the granularity of a rack or even an entire containerized data center; thus, they powered down not only idle IT equipment, but also their associated support-infrastructure.

Gandhi and Balter [17] found that the optimal power allocation varies for different scenarios. In particular, it is not always optimal to run servers at their maximum power levels. There are scenarios where it might be optimal to run servers at their lowest power levels or at some intermediate power levels. Their analysis shows that the optimal power allocation is non-obvious and depends on many factors such as the power-to-frequency relationship in the processors, the arrival rate of jobs, the maximum server frequency, the lowest attainable server frequency and the server farm configuration.

Lu et al [19] explored how much gain knowing future workload information can bring to dynamic provisioning. In particular, they developed online dynamic provisioning solutions with and without future workload information available. First, revealed an elegant structure of the offline dynamic provisioning problem, which allowed them to characterize the optimal solution in a “divide and conquer” manner. A fundamental observation is that future workload information beyond the full-size look-ahead window will not improve dynamic provisioning performance.

Other techniques

Buyya et al [6] proposed a novel technique for dynamic consolidation of VMs based on adaptive utilization thresholds, which ensures a high level of meeting the Service Level Agreements (SLA). They validated the high efficiency of the proposed technique across different kinds of workloads using workload traces from more than a thousand PlanetLab servers.

Hajj et al [7] idea revolves around identifying and measuring components of code with high energy consumption. There are two major contributions of this brief: 1) a method for identifying components with high energy consumption in compute-intensive applications. To this end, they target operations called kernels, which are frequently used operations in the algorithm; 2) a method for estimating software energy for the identified software components, in particular for kernels and load/store operations. The energy evaluation method involves isolated code with assembly injection. Furthermore, to ensure reliable results, they use physical energy measurements conducted on specially instrumented circuit boards to provide actual and not just simulated measurements.

Cheng et al [8] proposed an energy-saving task scheduling algorithm based on the vacation queuing model for cloud computing systems. First, they use the vacation queuing model with exhaustive service to model the task schedule of a heterogeneous cloud computing system. Next, based on the busy period and busy cycle under steady state, they analyze the expectations of task sojourn time and energy consumption of compute nodes in the heterogeneous cloud computing system.

Gandhi and Balter [13] analyzed the effectiveness of dynamic power management in data centers. Dynamic power management aims to reduce power wastage in data centers by turning servers off when they are not needed. However, turning a server back on requires a setup time, which can adversely affect system performance. Thus, it is not obvious whether dynamic power management should be employed in a data center. They analyzed the effectiveness of dynamic power management in data centers under an M/M/k model via Matrix analytic methods.

Chen et al [15] analyzed the techniques for power management that include shutting down these servers and/or modulating their operational speed, can impact the ability of the hosting center to meet SLAs. In addition, repeated on-off cycles can increase the wear-and-tear of server components, incurring costs for their procurement and replacement. They presented a formalism to this problem, and proposed three new online solution strategies based on steady state queuing analysis, feedback control theory, and a hybrid mechanism borrowing ideas from these two.

Beloglazov and Buyya [16] analyzed that dynamic consolidation of virtual machines (VMs) using live migration and switching idle nodes to the sleep mode allow Cloud providers to optimize resource usage and reduce energy consumption. However, the obligation of providing high quality of service to customers leads to the necessity in dealing with the energy-performance trade-off, as aggressive consolidation may lead to performance degradation. Due to the variability of workloads experienced by modern applications, the VM placement should be optimized continuously in an online manner. To understand the implications of the online nature of the problem, they conducted competitive analysis and prove competitive ratios of optimal online deterministic algorithms for the single VM migration and dynamic VM consolidation problems. Furthermore, They proposed a novel adaptive heuristics for dynamic consolidation of VMs based on an analysis of historical data from the resource usage by VMs.

REFERENCES

[1] Juhnyoung Lee, “A View Of Cloud Computing”, International Journal of Networked and Distributed Computing, Vol. 1, No. 1 (January 2013), 2-8
[2] Q. Zhang, L. Cheng, and R. Boutaba, “Cloud computing: state-of-the-art and research challenges,” Journal of Internet Services and Applications, vol.1, no.1, pp.7-18, 2010.
[3] Susanne Albers, “Energy-Efficient Algorithms”, Communications of the ACM, May 2010, Vol. 53, no. 5.
[4]A. Gandhi, M. Harchol-Balter, and M. Kozuch, “Are sleep states effective in data centers?”, in International Green Computing Conference, 2012.
[5] Yunni Xia, Mengchu Zhou, Xin Luo, Shanchen Pang, And Qingsheng Zhu, “A Stochastic Approach to Analysis of Energy-Aware DVS-Enabled Cloud Datacenters”, IEEE Transactions on Systems, Man, And Cybernetics: Systems, Vol. 45, No. 1, January 2015
[6] Anton Beloglazov and Rajkumar Buyya, “Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers”, Communications of ACM, December 2010
[7]H. Hajj, W. El-Hajj, M. Dabbagh, and T.R. Arabi, “An algorithm-centric energy-aware design methodology,” IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2013.
[8] Chunling Cheng, Jun Li, and Ying Wang, “An Energy-Saving Task Scheduling Strategy Based on Vacation Queuing Theory in Cloud Computing”, Tsinghua Science and Technology, February 2015, 20(1): 28-39
[9] Seyed Mohssen Ghafari, Mahdi Fazeli, Ahmad Patooghy, Leila Rikhtechi, “Bee-MMT: A Load Balancing Method for Power Consumption Management in Cloud Computing”, IEEE, 2013
[10] Daniel Gmach, Jerry Rolia, Cullen Bash, Yuan Chen, Tom Christian, Amip Shah, Ratnesh Sharma, Zhikui Wang, “Capacity Planning and Power Management to Exploit Sustainable Energy”, IEEE, 2010
[11] LI Hongyou, WANG Jiangyong, PENG Jian, WANG Junfeng, LIU Tang, “Energy-Aware Scheduling Scheme Using Workload-Aware Consolidation Technique in Cloud Data Centres”, China Communications, December 2013
[12] Mehiar dabbagh, Behir hamdaoui, “Energy-Efficient Resource Allocation and Provisioning Framework for Cloud Data Centers”, IEEE Transactions on Network and Science Management, 2015
[13] Anshul Gandhi and Mor Harchol-Balter, “How Data Center Size Impacts the Effectiveness of Dynamic Power Management”
[14] Lakshmi Ganesh, Hakim Weatherspoon, Tudor Marian, and Ken Birman, “Integrated Approach to Data Center Power Management, IEEE Transactions on Computers, VOL. 62, NO. 6, JUNE 2013

[15] Yiyu Chen, Amitayu Das, Wubi Qin, Anand Sivasubramaniam, Qian Wang, Natarajan Gautam, "Managing Server Energy and Operational Costs in Hosting Centers"
[16] Beloglazov and R. Buyya, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers," Concurrency and Computation: Practice and Experience, vol.24, no.13,pp.1397-1420, 2012
[17] Anshul Gandhi, Mor Harchol-Balter, Rajarshi Das, Charles Lefurgy, "Optimal Power Allocation in Server Farms", Communications of ACM, 2009
[18] Kyong Hoon Kim, Anton Beloglazov, Rajkumar Buyya, Power-aware Provisioning of Cloud Resources for Real-time Services, Communications of ACM, 2009
[19] Tan Lu, Minghua Chen, Lachlan and L.H. Andrew, "Simple and Effective Dynamic Provisioning for Power-Proportional Data Centers", IEEE Transactions on Parallel and Distributed Systems, VOL. 24, NO. 6, JUNE 2013