Volume 3, Issue 7, July 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Review of Different Energy Saving Techniques in Cloud Computing

Keffy Goyal,
Research fellow, CSE Deptt.
Sri Guru Granth Sahib World University
Fatehgarh Sahib, Punjab(INDIA)

Supriya Kinger
Assistant Professor, CSE Deptt.
Sri Guru Granth Sahib World University
Fatehgarh Sahib, Punjab(INDIA)

Abstract: Cloud Computing is a widely accepted Computing Technology that uses the internet and center remote servers to maintain the data and applications. So it is necessary to develop the large scale data centers. In addition to traditional IT infrastructure designers of such data centers increasingly needs to deal with issues of power consumption, heat dissipation and cooling provisioning. Thus, Green Cloud Computing takes the initiative to build up the "Green Data Centers". Green Cloud Computing provides different Power and Temperature based energy saving techniques to decrease the energy consumption. Our paper addresses different power and Temperature based techniques for power saving.

Keywords- Cloud Computing, Green Cloud Computing, Power Management, Temperature Management, Virtualization.

I. Interoduction

1.1 Cloud Computing

Cloud Computing is a model of enabling ubiquitous, convenient and on demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and service) that can be rapidly provisioned and released with minimal management effort or service interaction according to NIST (National Institute of Standards and Technology)[1]. Cloud computing is a product from mixing traditional computer techniques and network technologies, such as grid computing, distributed computing, parallel computing, utility computing, network storage, virtualization, load balancing etc .

Fig. 1 show the architecture of the Cloud computing. Cloud computing system scales applications by maximizing concurrency and using computing resources more efficiently, One must optimize locking duration, statelessness, sharing pooled resources such as task threads and network connections bus, cache reference data and partition large databases for scaling services to a large number of users. IT companies with innovative ideas for new application services are no longer required to make large capital outlays in the hardware and software infrastructures. By using Clouds as the application hosting platform, IT companies are freed from the trivial task of setting up basic hardware and software infrastructures. Thus they can focus more on innovation and creation of business values for their application services. Some of the traditional and emerging Cloud-based application services include social networking, web hosting, content delivery, and real time instrumented data processing. Each of these applications types has different composition, configuration and development requirements [2]. Cloud computing also describes applications that are extended to be accessible through the Internet. These cloud applications use large data canter and powerful servers that host Web applications and Web services. Anyone with a suitable Internet connection and a standard browser can access a cloud application.

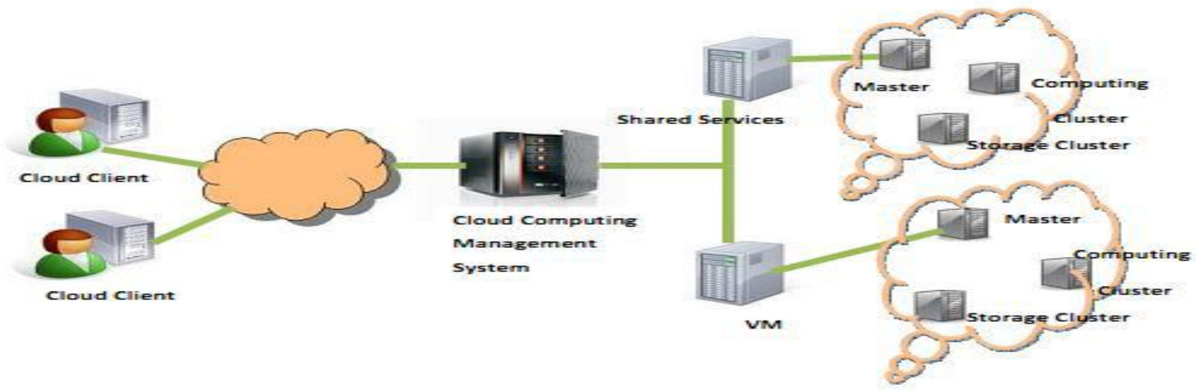


Fig. 1 Cloud Computing Architecture

1.2 Green Cloud Computing

In recent years, different IT service providers[3]—like IBM, Microsoft, Google and other organizations—have deployed data centers for the provision of Cloud computing. Usually, such data centers have thousands of servers and switches. The increasing growth of such data centers, and expansion of the existing ones, has prompted many in-depth studies related to their energy consumption. In 2007, USA directed the Environmental Protection Agency (EPA) to conduct an analysis on the data centers' electricity demands. The EPA study estimated that the data centers consumed 61 TWh of electricity in 2006 which was projected to grow to 107 TWh in 2011. The electricity consumption by these data centers was 1.5 % of the total electricity sales in the USA for the year 2006. About 80% of this electricity usage growth is attributable to servers, while network devices and storage equipments account for 10 %, each in individual capacity. This is due to the fact that the servers are the most used commodity in a data center. It has been estimated, in the said EPA report, that the annual electricity consumption growth of data centers will be 76 % in 2010. Lowering the energy usage of data centers [4] is a challenging and complex issue. Computing applications are growing so quickly that increasingly larger servers and disks are needed to process them fast enough within the required time period.

Green Cloud computing is envisioned to achieve not only efficient processing and utilization of computing infrastructure, but also minimizes energy consumption. This is necessary for ensuring that the future growth of Cloud computing is sustainable. Otherwise, Cloud computing with increasingly pervasive front-end client devices interacting with back-end data centers will cause an enormous escalation of energy usage. To address this problem, data center resources need to be managed in an energy-efficient manner to drive Green Cloud computing. In particular, Cloud resources need to be allocated not only to satisfy QOS requirements specified by users via Service Level Agreements (SLA), but also to reduce energy usage. Energy consumption can be reduced with planned power management, two main power management technologies are: (a) static power management (SPM) systems that utilize low-power components to save the energy, and (b) dynamic power management (DPM) systems that utilize software and power-scalable components to optimize the energy consumption. We can apply other energy efficiency techniques such as device reduction, dynamic voltage/frequency scaling (DVFS), network port reduction, and improved server and storage efficiency will lead to a saving of 36 % in the current electricity bill. Electricity can also be saved from reduced device demands, uninterruptible power supply (UPS), transformer and cooling efficiency.

2. Energy Efficient Techniques in Green Cloud Computing

Green Cloud Computing can be achieved by using two techniques Power management and Thermal Management as shown in Fig. 2.

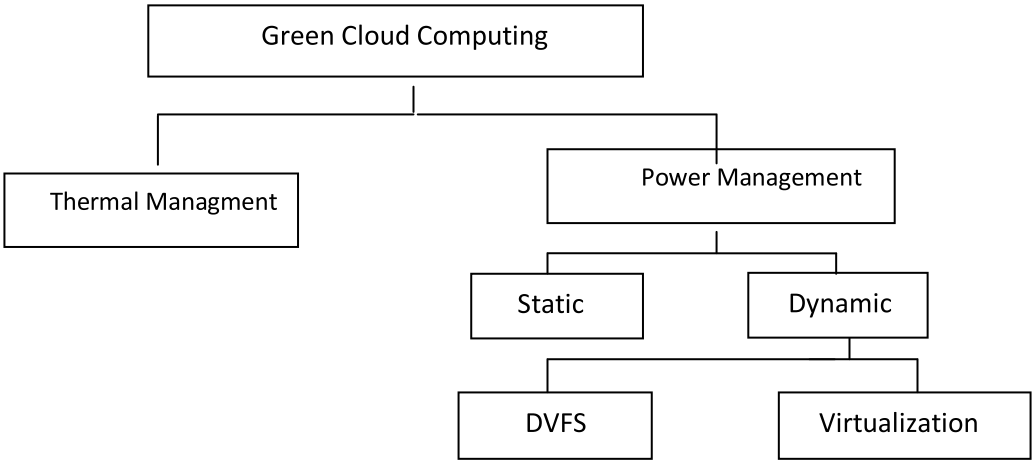


Fig. 2 Thermal and Energy management in Cloud Computing

2.1 Power management techniques

A large volume of research has been done in the area of power and energy efficient resource management in computing systems. As power and energy management techniques are closely connected, from this point we will refer to them as power management. There are two main categories of the power management: (a) Static Power Management (SPM), and (b) Dynamic Power Management (DPM). SPM technologies use energy-efficient hardware equipment (e.g. CPUs and power supplies) to reduce energy usage and peak power consumption. DPM techniques are based on the knowledge of current resource utilization and application workloads to reduce energy usage.

2.1.1 Static Power Management (SPM)

SPM contains all the optimization methods that are applied at the design time. Power optimization requires careful design at several levels of the system architecture. These levels are CPU level and System level. CPU is the main target of power consumption analysis. Many power-aware models were presented and integrated into already-in-use performance simulators in order to investigate power consumption of the CPU, on a unit basis or for the processor as whole. These investigations were mainly held on two abstraction levels: the register-transfer level (or cycle-level) processor energy consumption is generally estimated by its (RTL's) simulators. The instruction-level this technique estimates the energy consumed by a program by summing the energy consumed by the execution of each instruction. Instruction-by-instruction energy costs are precharacterized once for all for each target processor. System Level: There is little benefit in optimizing only the CPU core if other elements participate or sometimes even dominate the energy consumption. To effectively optimize system energy, it is necessary to consider all of the significant components. Different researchers investigate the power consumption on different system levels, targeting both hardware and software on different levels of abstraction [5].

State-level models and measurements are used to account the energy consumption of the whole system based on the state each device is in or transiting from or to. Measurements are used to find the system power consumption and help targeting the hotspots in applications and operating system procedures. This approach tries to reduce energy consumption by acting on the application- and OS-level. One another approach is complete system level simulation tool which models the CPU, memory hierarchy and a low power disk subsystem and quantifies the power behavior of both the application and operating system.

2.1.2 Dynamic Power Management (DPM)

The concept of DPM was first proposed by Benini et al.[6] Dynamic Power management (DPM) techniques include methods and strategies for run-time adaptation of a system's behavior according to current resource requirements or any other dynamic characteristic of the system's state. The major assumption enabling DPM is that systems experience variable workloads during their operation allowing the dynamic adjustment of power states according to current performance requirements. The second assumption is that the workload can be predicted to a certain degree. DPM techniques can be distinguished by the level at which they are applied: hardware or software. Hardware DPM varies for different hardware components, but usually can be classified as dynamic performance scaling (DPS), such as DVFS, and partial or complete dynamic component deactivation (DCD) during periods of inactivity. In contrast, software DPM techniques utilize interface to the system's power management and according to their policies apply hardware DPM. The introduction of the Advanced Power Management (APM) and its successor, the Advanced Configuration and Power Interface (ACPI), has drastically simplified the software power management and resulted in broad research studies in this area. The problem of power-efficient resource management has been investigated in different contexts of device-specific management, OS-level management of virtualized and non-virtualized servers, followed by multiple-node system such as homogeneous and heterogeneous clusters, data centers, and Clouds.

(i)Dynamic voltage/frequency scaling (DVFS)

DVFS technique is based on the fact that processing chip's power consumption depends on voltage supplied to it and is described by the following equation,

$$P = V^2. F$$

Where P is the power consumed, V is the voltage and f is the corresponding frequency. Therefore, by reducing the voltage or the switching frequency, the power consumption can be reduced. The frequency reduction only applies to the CPU power since bus, memory, and disks do not depend on the CPU frequency. Moreover, the hardware support is necessary to implement the DVFS technique. Many manufactures nowadays adopted, the Advanced Configuration and Power Interface (ACPI) specification, which is an OS-independent power management and configuration standard. They defines four power states for a server first is G_0(S0) is a subset of S0, where monitor is off but background tasks are running. G1 is the Sleeping state and subdivides into the four states S1 through S4.S1 all processor caches are flushed, and the CPU(s) stops executing instructions. Power to the CPU(s) and RAM is maintained, devices that do not indicate they must remain on may be powered down.S2: CPU powered off. Dirty cache is flushed to RAM. S3: Commonly referred to as Standby, Sleep or Suspend to RAM. S4: Hibernation or Suspend to Disk. All content of main memory is saved to non-volatile memory such as a hard drive, and is powered down. G2 (S5) is almost the same as G3 Mechanical Off, except that the PSU still supplies power, at a minimum so the computer can "wake" on input from the keyboard, clock, modem, LAN, or USB device. G3, Mechanical Off, the computer's power has been totally removed via a mechanical switch. Most of the DVFS schemes depend on the ACPI standard implementation [3].

(ii)Virtualization

Virtualization was first developed in the 1960s by IBM as a way to logically partition mainframe computers into separate virtual machines. These partitions allowed mainframes to run multiple applications at the same time. Since mainframes were expensive at the time, they were designed for partitioning as a way to fully leverage the investment. The technology was effectively abandoned when the computing paradigm shifted from mainframes and terminals to inexpensive x86 servers and personal computers. Organizations typically run only one application per server to avoid the risk of vulnerabilities in one application affecting the availability of another application on the same server. Due to the massive power growth of the servers, this has gradually led to low infrastructure utilization. Using virtualization, we can convert servers into virtual machines and run them on fewer physical machines without losing the advantage of having the services running on distinct servers. This process is commonly called server consolidation (illustrated in Fig.3). It makes virtualization perfectly fit for the energy efficiency in data centers. Virtualization is the most adopted power management and resource allocation technique used by the data center operators. The use of live migration of a virtual machine is a recent concept. Live migration can be applied to Green computing in order to migrate away machines. VMs can be shifted from low load to medium load servers when needed. Low load servers are subsequently shutdown when all VMs have migrated away, thus conserving the energy required to run the low load idle servers. When using live migration, the user is completely unaware of change and there is only a 60 to 300ms delay, which is acceptable by most standards.

Beloglazov et al. [7] used live migration of VMs for power saving. In their experiments the jobs were concentrated on a few physical nodes so that the rest of the nodes can be put in a power saving mode. New VMs allocation is done by sorting all VMs in a Modified Best First Decreasing (MBFD) order with respect to the current utilization. A VM is then allocated to a host based on the least deterioration in the power consumption among the hosts. The current allocation of VMs is optimized by selecting the VMs to be migrated on the basis of heuristics related to utilization thresholds. If the current utilization of a host is below a threshold, then all the VMs from that host should be migrated and the host is put in the power saving mode.

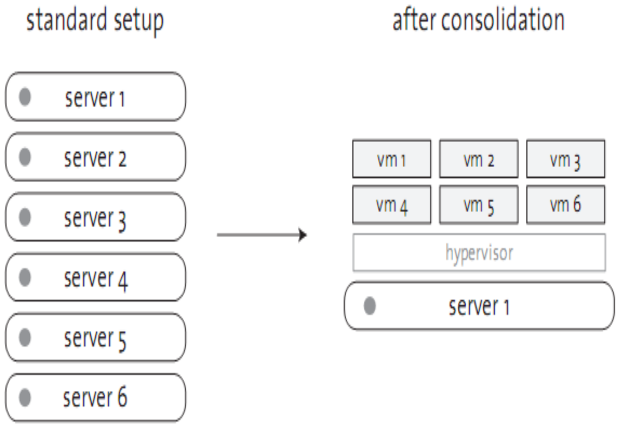


Fig.3: Six hardware servers consolidated into one server, which runs six virtual machines.

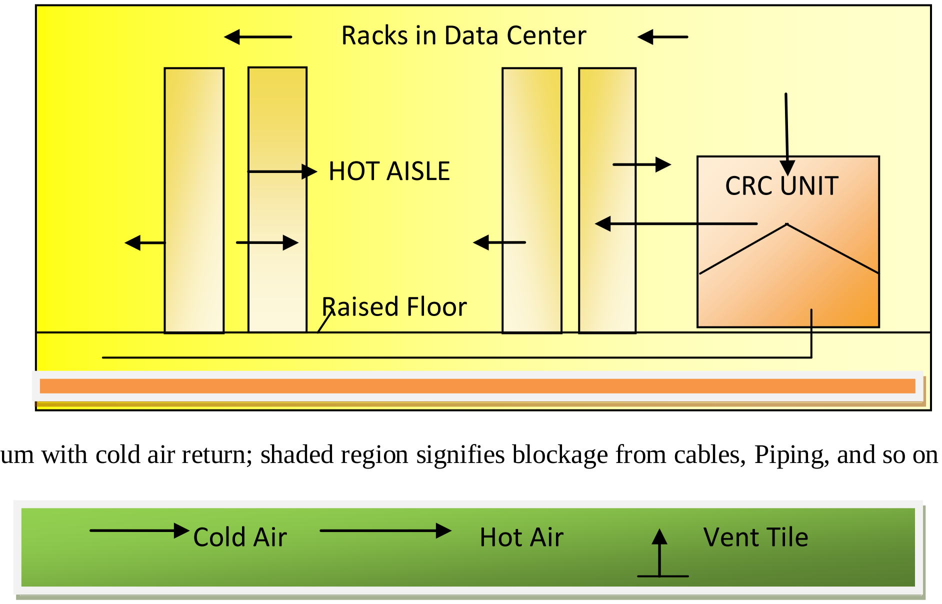
2.2 Thermal Management

Thermal management is an increasingly important architectural thought for high-performance computing. It presents challenges that can apply to chips, servers, racks, and data centers [8]. A thermal management policy considers facilities components, such as CRAC units and the physical layout of the data center, and temperature-aware IT components. A temperature-aware IT component can maintain an efficient thermal profile within the data center, resulting in reduced annual cooling costs. In the event of a thermal emergency, IT-level actions include scaling back on server CPU utilization, scaling CPU voltages, migrating or shifting workload , and performing a clean shutdown of selected servers. In thermal-aware scheduling jobs are scheduled in a manner that minimizes the overall data center temperature. The goal is not always to conserve the energy used to the servers, but instead to reduce the energy needed to operate the data center cooling systems. There are some thermal management benefits [9] that are following:

(i)Decrease cooling costs
(ii)Increase hardware reliability
(iii)Decrease response times to transients and emergencies
(iv) Increase compaction and improve operational efficiencies
(v) Enable holistic IT-facilities scheduling

2.3Thermal Management in data centers

As we know in Cloud, to store huge amount of data large data centers are required but it consumes a lot of power. So, it is necessary to balance the power of the data centers. K. Sharma et al. [8], they provide the dynamic thermal management for internet data centers. They describe the small PDC (Programmable Data Centre) architecture in which a CRAC unit is used and the racks of servers are used. The function of CRAC (Computer Room Air Conditioning) is to circulate the cool air between the racks and exhaust the heat from the servers. As shown in Fig.4 the cooled air enters the machine room through floor vent tiles in alternating aisles between the rows of racks.



Plenum with cold air return; shaded region signifies blockage from cables, Piping, and so on

Fig. 4 Data Center Cooling System

But CRAC fails at some extent because of unequal distribution of cool air or can say mixing of cool air with hot air. Therefore, objectives of thermal aware workload scheduling are to reduce both the maximum temperature for all compute nodes and the imbalance of the thermal distribution in a data center. For this, k.Sharma et al.suggest the migration of workloads among the servers in order to balance the thermal load distribution across the PDC.

II. Conclusion

With the increasing tendency of Cloud Computing the need for energy saving mechanisms also increases. The modern research focuses on the issues of workload energy efficient resource scheduling, virtualization and automatic power and thermal management etc. In this paper we discussed power aware and temperature aware techniques to maximize energy savings both for physical servers and cooling systems used in data centers. Large power consumption produces huge CO_2 emissions and significantly contributes to the growing environmental issues of global warming. Thus Green Computing will be one of the fundamental components of the next generation of cloud computing technologies.

References

[1]. Fang liu et al ,"NIST Cloud Computing Reference Architecture", National Institute of standard and Technology U.s Department of commerce, Special Publication 500-292.

[2]. Sandeep Tayal, "Tasks Scheduling optimization for the Cloud Computing Systems", (IJAEST) International Journal Of Advanced Engineering Sciences And Technologies, Vol No. 5, Issue No. 2, 111 – 115.

[3]. Junaid Shuja, Sajjad A. Madani **,"**Energy-efficient data centers**",** © Springer-Verlag 2012, Received: 3 January 2012 / Accepted: 10 August 2012 / Published online: 2 September 2012.

[4]. Rajkumar Buyya et al., "Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges".

[5]. Wissam Chedid and Chansu Yu, "Survey on Power Management Techniques for Energy Efficient Computer Systems", 2121 Euclid Avenue, SH 332, Cleveland, OH 44115

[6]. Anton Besloglazov, "A Taxonomy and Survey of Energy-Efficient Data Centers and Cloud Computing Systems", Advances In Computers, Vol. 82 47 Copyright © 2011 Elsevier Inc. ISSN: 0065-2458/DOI: 10.1016/B978-0-12-385512-1.00003-7

[7]. Anton Beloglazov et al., "Energy Efficient Allocation of Virtual Machines in Cloud Data Centers" 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing.

[8]. Ratnesh K. Sharma et al., "Balance of Power Dynamic Thermal Management for Internet Data Centers" Published by the IEEE Computer Society 1089-7801/05/$20.00 © 2005 IEEE.

[9]. Justin Moore et al. ,"Weatherman: Automated, Online, and Predictive Thermal Mapping and Management for Data Centers", 1-4244-0175-5/06/$20.00 02006 IEEE.

[10]. S. Lizhe Wang et al. ,"Thermal aware workload placement with task-temperature profiles in a data center", © Springer Science+Business Media, LLC 2011.