Survey on Energy Efficient Resource Allocation Methods in Cloud Environment

Vinisha Sasidharan
S3 Software Engineering
Dept. of Computer Science
TKM Institute of Technology, Kollam

P. Mohamed Shameem
Associate Professor
Dept. of Computer Science
TKM Institute of Technology, Kollam

ABSTRACT
Cloud computing is emerging as a new paradigm of large-scale distributed computing. It is a framework for enabling convenient, on demand network access to a shared pool of computing resources. Cloud computing environments provide scalability for applications by providing virtualized resources dynamically. It offers utility-oriented IT services to users worldwide. Based on a pay-as-you-go model, it enables hosting of pervasive applications from consumer, scientific, and business domains. However, data centers hosting Cloud applications consume huge amounts of electrical energy, contributing to high operational costs and carbon footprints to the environment. Therefore, Green Cloud computing solutions that can not only minimize operational costs but also reduce the environmental impact is essential. This paper discusses the various methods used to reduce energy consumption and scheduling algorithms in cloud computing.

General Terms
Cloud Computing, Energy Efficiency, Green Computing.

Keywords
Scheduling, Cloud Computing, Energy Efficiency, Virtualization.

1. INTRODUCTION
Cloud computing [1] is emerging as a new paradigm of large scale distributed computing. It has moved computing and data away from desktop and portable PCs, into large data centers. It provides the scalable IT resources such as applications and services, as well as the infrastructure on which they operate, over the Internet, on pay-per-use basis to adjust the capacity quickly and easily. It helps to accommodate changes in demand and helps any organization in avoiding the capital costs of software and hardware. Thus, cloud computing is a framework for enabling a suitable, on demand network access to a shared pool of computing resources (e.g. networks, servers, storage, applications, and services). These resources can be provisioned and deprovisioned quickly with minimal management effort or service provider interaction. This further helps in promoting availability. Due to the exponential growth of cloud computing, it has been widely adopted by the industry and there is a rapid expansion in data-centers. This expansion has caused the dramatic increase in energy use and its impact on the environment in terms of carbon footprints. The link between energy consumption and carbon emission has given rise to an energy management issue which is to improve energy-efficiency in cloud computing to achieve Green computing [2].

Energy Efficiency is one of the critical issues in cloud computing [3]. To achieve energy efficiency in cloud environment, tasks need to be scheduled efficiently. In cloud computing, the underlying large-scale computing infrastructure is often heterogeneous, not only because it's not economic and reliable to procure all the servers, network devices and power supply devices in one size and one time, but because different application requires different computer hardware, e.g. workflow extensive computing might need standard and cheap hardware; scientific computing might need specific hardware other than CPU like GPU or ASIC. There are kinds of resources in the large-scale computing infrastructure need to be managed, CPU load, network bandwidth, disk quota, and even type of operating systems. To provide better quality of service, resources are provisioned to the users or applications, via load balancing mechanism, high availability mechanism and security and authority mechanism. To maximize cloud utilization, the capacity of application requirements shall be calculated so that minimal cloud computing infrastructure devices shall be procured and maintained. Given access to the cloud computing infrastructure, applications shall allocate proper resources to perform the computation with time cost and infrastructure cost minimized. Proper resources shall be selected for specific applications.

According to Amazon.com's [4] estimates, at its data centres (as illustrated in figure 1), expenses related to the cost and operation of the servers account for 53% of the total budget (based on a 3-year amortization schedule), while energy-related costs amount to 42% of the total, and include both direct power consumption (~19%) and the cooling infrastructure (23%) amortized over a 15-year period.

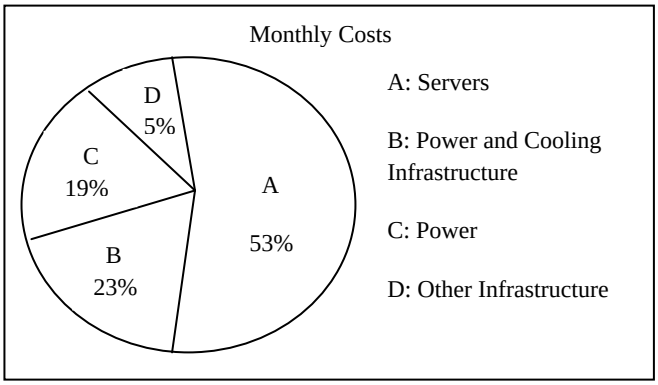

Figure 1: Energy distribution in the data centre.

This paper presents a survey of the energy efficient resource allocation algorithms in cloud computing for reducing energy consumption.

The remainder of this paper is organized as follows. Section 2 describes the need for energy management in cloud. Section 3 describes the existing techniques for achieving energy efficiency in cloud. Section 4 describes conclusion.

2. NEED FOR ENERGY MANAGEMENT IN CLOUD

Nowadays, many people are devoted to a widespread adoption of Information and Communications Technologies (ICTs). The increasing accumulation of greenhouse gases is changing the world's climate, creating serious problems such as droughts, floods and higher temperatures. In order to stop the accumulation of these gases in the atmosphere, it is necessary to stop the global growth of emissions, in which the generation of electricity plays a major role not only because of the carbon dioxide which results from the coal and oil used in this process, but also because it releases sulphurs and other pollutants into the atmosphere. Green IT emerges as a new perspective for designing, developing and managing computing infrastructure aiming for more efficient processes and mechanisms to avoid waste of resources and considering the environmental implications of its use and disposal.

Energy efficiency is increasingly important due to the increasing energy costs and the need to reduce greenhouse gas emissions and also to decrease the overall energy consumption, storage and communications.

Servers are unfriendly to environment [5] and IT industry contributes to 2% of worlds total CO2 emissions (eg: 2.8 tons from US power plants). A typical datacentre consumes as much as energy as 25000 households. Servers consume 0.5% of the world's total electricity usage. More than 15% [6] of the servers are running without being used actively. So we need energy efficient resource management methods that minimize energy consumption at the same time meet the job deadlines.

The energy consumption for computing could be divided according its use in two edges [7], the first regarding to the energy consumed by the clients conformed by PCs, peripherals and all types of mobile devices and the second refers to the energy consumed by servers, networks and cooling systems in data centers. Due to the need to maintain the quality of service that customers expect and the continuous expansion of the industry, energy consumption in the "*data center edge*" is increasing along with their performance increase. With the aim of minimizing the negative environmental impact of ICTs, emerges a different perspective to perform and use computing infrastructure named "*Green Computing*" or "*Green IT*". Different approaches for energy efficiency are

- Energy Efficient Hardware [4]
- Virtualization [8]
- Dynamic Voltage and Frequency Scaling [9]
- Energy-aware job Scheduling
- Request Batching [10]
- Multi-speed Disks [11]
- Server Consolidation [12]

3. EXISTING TECHNIQUES FOR ENERGY EFFICIENCY

3.1 Adaptive Spin Down for Mobile Computers [13]

Adaptive energy conservation techniques include Powering down when devices are not used and changing speed of CPU dynamically to save power. This paper explains an algorithm which decides when to the disk to be powered down. Disk subsystem on computers consume large amount of energy say about 30% or more.

Disk spin down in computers consumes some energy, so in order to maximize energy savings, algorithm deciding when to spin down disk is necessary. This paper uses a new algorithm called share algorithm which decides when to spin down the disk. Share algorithm calculates delay or timeout which decides how long an idle disk needs to be powered on before spinning down. This algorithm spins down disk if it remains idle more than the computed timeout.

Energy used by time_out = idle time,
if idle time<= time_out
= time_out+spindowncost
if idle time> time_out
Energy used by optimal = idle time,
if idle time<= spindowncost
= spindowncost
if idle time> spindowncost
Excess energy = Energy used by time_out - Energy used by optimal
Loss = Excess energy/spindowncost

The performance of algorithm is measured in 'seconds of energy used' (difference in energy consumed between a spinning disk and a spun down disk). Input to algorithm is other algorithms which predicts and it keeps one weight per predict and predict with weighted average of expert's prediction. The goal of share algorithm is to combine the predictions of all predicting algorithm thereby minimizing the total error. After each trial weights of each prediction algorithms are changed. That means if an expert is misleading, its weight is reduced drastically. Two parameters are used, they are $n>1$ and $0<\alpha<1$, how rapidly weights of misleading experts are reduced and how rapidly weight of a poor predicting expert recovers when it begins predicting well.

3.2 SLA-Based Scheduling Of Bag-Of-Tasks Applications on Power-Aware Cluster Systems [14]

Dynamic Voltage Scaling (DVS) is an efficient way to manage energy dissipation during task computation. The DVS scheme reduces dynamic power consumption of processor by adjusting the supply voltage in an appropriate manner. This paper explains a power-aware scheduling algorithm for bag-of-tasks applications with deadline constraints on DVS enabled cluster systems in order to minimize power consumption as well as to meet the deadlines specified by application users. A cluster system is composed of multiple Processing Elements (PEs) and a central resource controller. Each Processing Element executes its submitted jobs. Users submit their task to cluster system. Upon receiving the job by cluster system, the resource controller plays an important role for admission control based on information from PEs in the system.

A cluster system is defined as (N, Q), where N is the number of PEs and Q is the processing performance of each PE in terms of MIPS. The main power consumption is composed of dynamic and static power.

Edynamic = k1V²ddNcycl ---------- (1)
where Vdd is the supply voltage and Ncycl is the number of clock cycles of the task.

Estatic = k2Edynamic ---------- (2)

A user's job is defined as (p, {l1, l2, · · · , lp}, d), where p is the number of sub-tasks, li is the number of instructions of the i-th task in Million Instructions (MIs), and d is the deadline. When a job arrives, the job admission and execution process pass through 4 steps:

Job submission, Schedulability test & Energy estimation, Acknowledgement of schedulability and energy amount and Selection of PEs.

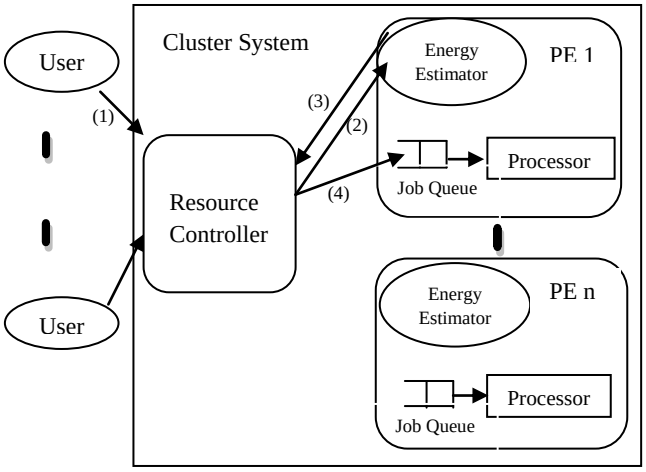

Figure 2: Resource Allocation Framework

Upon submission, resource controller asks schedulability test and energy estimation time to each PE. PEs replies with energy amount and depending upon the replies, best PE is selected by resource controller and allocates jobs to PEs. Each PE has its own processor scaling capabilities which scales up or down the voltage depending upon the current processor utilization.

3.3 An Energy-Efficient Scheduling Approach Based on Private Clouds [15]

This paper proposed a hybrid energy efficient scheduling approach that can save more time for users, conserve more energy and achieve higher level of load balancing. Most of the energy efficiency algorithms use Virtual machine migration and scheduling. There are two challenging problems in VM scheduling:

a) How to reduce the coming request's response time, and

b) How to balance the workloads, when the data center is running on low-power mode.

These two problems happen because:

a) Powering up a sleeping node via remote access is kind of time-consuming;

b) Traditional energy-efficient algorithms always fall short to sharing workloads across hosts.

To solve these two problems, a hybrid energy-efficient scheduling approach is proposed, which is comprised of two algorithms namely pre-power technique and least-load-first algorithm.

First algorithm, pre-power technique is used to reduce Response time. First an administrator specified threshold (left capacity i.e. available capacity of awake hosts) is set. Therefore powering up or down a node is not passively controlled by the idle threshold, which is difficult to set, but fully controlled by cloud scheduler.

There is always a trade-off problem between energy conservation and load balancing. One of the traditional ways to balance workloads across compute nodes is Round-robin (RR) [16], which assigns workloads to each host in equal portions and in circular order, handling all nodes without priority. It is simple and easy to apply to application, but not energy-efficient. Another traditional way to conserve energy is Greedy [17], which schedule workloads on the first node and sticks to it until it is saturated. It can save much energy by turning the idle host down when it passed the given threshold; however, the sharing of workloads across nodes is neglected.

In this paper, through the desired spectrum of left capacity and the least-load-first algorithm, a hybrid energy-efficient scheduling approach is proposed. Jobs are arranged in the least load first order, and then load is managed based on the left capacity.

3.4 Energy-Aware Resource Allocation Heuristics for Efficient Management of Data Centers for Cloud Computing [18]

This paper proposed an architectural framework and principles for energy-efficient Cloud computing.

Figure below shows the high-level architecture for supporting energy-efficient service allocation in a Green Cloud computing infrastructure [19].

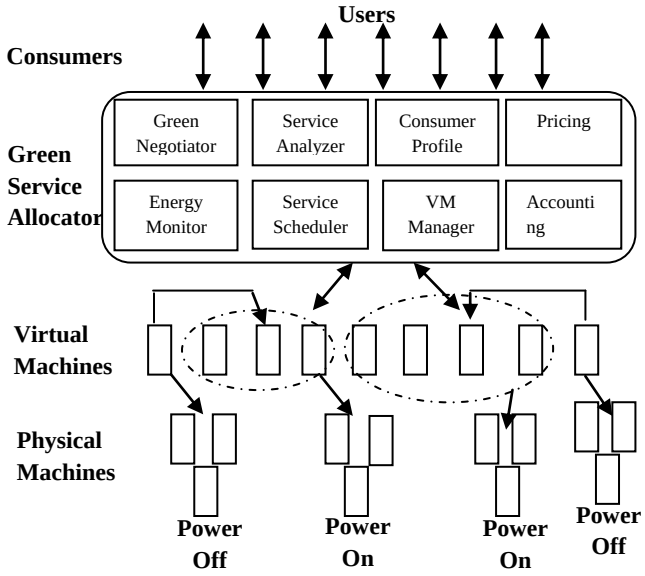

Figure 3: Green Cloud Infrastructure

Virtual Machine migration and placement is a technique for minimizing energy consumption. There are two problems of VM placement

1) Placing the VMs on hosts,

2) Optimization of the current VM allocation.

First problem can be solved by Modified Best Fit Decreasing. That is, sort all VMs in decreasing order of their current CPU utilizations, and allocate each VM to a host that provides the least increase of power consumption due to this allocation. Second problem is carried out in two steps

1) Select VMs that need to be migrated

2) Chosen VMs are placed on the hosts using the MBFD algorithm.

To determine when and which VMs should be migrated, three double-threshold VM selection policies

1) The minimization of migrations policy

2) The highest potential growth policy

3) The random choice policy

The Minimization of Migrations (MM) policy selects the minimum number of VMs needed to migrate from a host to lower the CPU utilization below the upper utilization threshold if the upper threshold is violated. The algorithm sorts the list of VMs in the decreasing order of the CPU

utilization. Then, it repeatedly looks through the list of VMs and finds a VM that is the best to migrate from the host.

The Highest Potential Growth (HPG) policy migrates VMs that have the lowest usage of the CPU relatively to the CPU capacity.

The Random Choice (RC) policy relies on a random selection of a number of VMs needed to decrease the CPU utilization by a host below the upper utilization threshold.

3.5 Load Balancing and Unbalancing for Power and Performance in Cluster-Based Systems [20]

This paper focuses on power and energy conservation for clusters of workstations or PCs. Approach used for conserving power and energy is to develop systems that can leverage the widespread replication of resources in clusters. In particular, systems are developed that can dynamically reconfigure the cluster nodes. That is putting on cluster when current cluster nodes cannot handle task load and putting off when some of the nodes are underutilized.

Load balancing and unbalancing is a technique for developing more energy efficient cluster. When performing load balancing, the goal is to evenly spread the work load over the available cluster nodes in such a way over utilized nodes task can be split across underutilized nodes and performance can be promoted. The inverse of the load balancing operation concentrates work in fewer nodes; underutilized nodes can be turned off. This load balancing or unbalancing operation saves the power consumed by the powered-down nodes, but can degrade the performance of the remaining nodes and potentially increase their power consumption. In more detail, the algorithm periodically considers whether nodes should be added to or removed from the cluster.

After an addition or removal decision is made, additional load should be re-distributed. If the decision is to add one or more nodes, the algorithm must determine what portion of current task to be distributed across added nodes. Obviously, the load to be migrated should come from over utilized nodes. If the decision is to remove one or more nodes, the algorithm must determine which nodes should be removed. Underutilized nodes should be shut down and additional load should be moved to remaining existing nodes.

Table 1: Energy Efficient Resource Allocation Methods in Cloud

Technique	Environment	Description	Attained	Not Attained
1) Adaptive Spin Down For Mobile Computers	Mobile Computers	1) Explains an algorithm for deciding when to power down disks. 2) Uses share algorithm to decide when to spin down disk 3) Algorithm calculates delay or timeout which shows how long idle disk is kept powered on before spinning down. 4) Spin down if remains idle more than the computed timeout.	1) Extends the battery life	1) Single Component level energy reduction
2) SLA-Based Scheduling Of Bag-Of-Tasks Applications On Power-Aware Cluster Systems	Cluster Computing	1) Proposes DVS (Dynamic Voltage Scaling) scheme which reduces dynamic power consumption by adjusting the supply voltage in an appropriate manner. 2) Steps are Job submission, Schedulability test & Energy estimation for each task of the job by resource controller to all PEs, Acknowledgement of schedulability and energy amount from PEs, Selection of PEs by resource controller 3) Each PE has its own job queue and scales voltages according to the tasks.	1) Minimizes energy consumption 2) For applications with deadline	1) Cannot be applied to real time applications
3) An Energy-Efficient Scheduling Approach Based On Private Clouds	Private Cloud Computing	1) Two challenging problems in VM scheduling: reducing response time and balancing the workloads. 2) To reduce Response time, pre-power technique is used. 3) Proposed a hybrid energy-efficient scheduling approach, which is comprised of pre-power technique and least-load-first algorithm. 4) Jobs are arranged in the least load first order, and then load is managed based on the left capacity.	1) Energy and time conservation for private clouds	1) Response time too long is beyond consideration 2) migration should be cautiously applied
4) Energy-Aware Resource Allocation Heuristics For	Cloud Computing	1) Proposed a method that solves problems of VM placement: Placing the VMs on hosts and Optimization of the current VM allocation. 2) First problem can be solved by Modified Best Fit	1) Reduction in energy consumption in Cloud	1) Virtual Server Sprawl

Efficient Management Of Data Centers For Cloud Computing		Decreasing. It sorts all VMs in decreasing order of their current CPU utilizations, and allocates each VM to a host that provides the least increase of power consumption due to this allocation. 3) Second problem is carried out in two steps • Select VMs that need to be migrated • Chosen VMs are placed on the hosts using the MBFD algorithm. 4) To determine when and which VMs should be migrated, three double-threshold VM selection policies • The minimization of migrations policy • The highest potential growth policy • The random choice policy	datacenters	
5) Load Balancing And Unbalancing For Power And Performance In Cluster-Based Systems	Cluster Computing	1) During Load Balancing algorithm evenly spread the work over the available cluster resources. 2) During load unbalancing operation concentrates work in fewer nodes, idling other nodes that can be turned off. 3) Algorithm periodically considers whether nodes should be added to or removed from the cluster, based on the expected performance and power consumption that would result, and decides how the existing load should be re-distributed in case of a configuration change.	1) Reduction in Power Consumption	1) Can degrade the performance of the remaining nodes and potentially increase their power consumption

4. CONCLUSION

Cloud computing is emerging as a new paradigm of large scale distributed computing. It has moved computing and data away from desktop and portable PCs, into large data centers. Energy Efficiency is one of the critical issues in cloud computing. The issue of energy consumption in information technology equipment has been receiving increasing attention in recent years and there is growing recognition of the need to manage energy consumption across the entire information and communications technology (ICT) sector. It is estimated that data centers accounted for approximately 1.2% of total United States electricity consumption in 2005.

In this paper, several methods for energy efficiency are surveyed on. Energy efficient method proposed earlier focused on single hardware level like: processor (DVS), disk (multi-speed disks) etc. But to achieve energy efficiency in cloud environment, tasks need to be scheduled efficiently. These methods include Virtualization, Server Consolidation, VM migration, dynamic resource allocation etc. All these techniques mainly focused on reducing energy consumption. Various scheduling and consolidation techniques can be applied for reducing the CPU utilization and bring about drastic changes in energy efficiency.

5. REFERENCES

[1] Alexa Huth and James Cebula, The Basics of Cloud Computing, US-CERT 2011.

[2] Jayant Baliga, Robert W. A. Ayre, Kerry Hinton, and Rodney S. Tucker, Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport, IEEE 2010.

[3] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic, Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, ELSEVIER 2009.

[4] Andreas Berl, Erol Gelenbe, Marco di Girolamo, Giovanni Giuliani, Hermann de Meer, Minh Quan Dang and Kostas Pentikousis, Energy-Efficient Cloud Computing, Computer Journal 2010.

[5] Luiz André Barroso and Urs Hölzle, The Case for Energy Proportional Computing, IEEE 2007.

[6] Server Energy and Efficiency Report. Technical report, 1E, 2009.

[7] Ismael Solis Moreno, Jie Xu, Energy-Efficiency in Cloud Computing Environments: Towards Energy Savings without Performance Degradation.

[8] F. Tusa, M. Paone, M. Villari and A. Puliafito, CLEVER: A CLoud-Enabled Virtual EnviRonment IEEE 2010.

[9] G. von Laszewski, L. Wang, A. Younge, X. He, Power-aware scheduling of virtual machines in DVFS-enabled clusters, IEEE 2009.

[10] Enrique V. Carrera, Eduardo Pinheiro, and Ricardo Bianchini, Conserving disk energy in network servers, ACM.

[11] Sudhanva Gurumurthi, Anand Sivasubramaniam, Mahmut Kandemir, and Hubertus Franke, Drpm: Dynamic speed control for power management in server class disks, in: Computer Architecture, International Symposium 2003.

[12] Ching-Hsien Hsu, Shih-Chang Chen2, Chih-Chun Lee, Hsi-Ya Chang, Kuan-Chou Lai, Kuan-Ching Li and Chunming Rong, Energy-Aware Task Consolidation Technique for Cloud Computing, IEEE 2011.

[13] David P. Helmbold, Darrell D.E. Long, Tracey L. Sconyers and Bruce Sherrodm, Adaptive Spin Down For Mobile Computers, Journal on Mobile Network and Applications 2000.

[14] Kyong Hoon Kim, Rajkumar Buyya, Jong Kim, SLA-Based Scheduling Of Bag-Of-Tasks Applications On Power-Aware Cluster Systems, Ieice Trans. Inf. & Syst. 2010.

[15] Jiandun Li , Junjie Peng , Zhou Lei , Wu Zhang , An Energy-Efficient Scheduling Approach Based On Private Clouds, Journal of Information & Computational Science 2011.

[16] Round-robin (RR) on Wikipedia: http://en.wikipedia.org/wiki/Round-robin_scheduling.

[17] Greedy on Wikipedia: http://en.wikipedia.org/wiki/Greedy algorithm.

[18] Anton Beloglazov, Jemal Abawajy, Rajkumar Buyya, Energy-Aware Resource Allocation Heuristics For Efficient Management Of Data Centers For Cloud Computing, ELSEVIER 2012.

[19] R. Buyya, A. Beloglazov, J. Abawajy, Energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges, in: International Conf on Parallel and Distributed Processing Techniques and Applications 2010.

[20] Eduardo Pinheiro, Ricardo Bianchini, Enrique V. Carrera, and Taliver Heath, Load Balancing and Unbalancing for Power and Performance in Cluster-Based Systems.