A Survey on Green Cloud Computing

Bethala Pravallika & Bethala Shirisha
[1]Assistant Professor, MTech, Department of IT, IARE, JNTUH
[2]Assistant Professor, MTech, Department of CSE, CMRIT, JNTUH

Abstract

Cloud computing is offering utility oriented IT services to users world wide. It enables hosting of applications from consumer, scientific and business domains. However data centres hosting cloud computing applications consume huge amounts of energy, contributing to high operational costs and carbon footprints to the environment. With energy shortages and global climate change leading our concerns these days, the power consumption of data centers has become a key issue. Therefore, we need green cloud computing solutions that can not only save energy, but also reduce operational costs. The vision for energy efficient management of cloud computing environments is presented here. A green scheduling algorithm which works by powering down servers when they are not in use is also presented.

Keywords: Cloud computing, Green computing, DVFS, Resource Allocator

1. Introduction

In 1969, Leonard Kleinrock , one of the chief scientists of the original Advanced Research Projects Agency Network (ARPANET) which seeded the Internet, said: *"As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of „computer utilities" which, like present electric and telephone utilities, will service individual homes and offices across the country."[1]* This vision of computing utilities based on a service provisioning model anticipated the massive transformation of the entire computing industry in the 21st century whereby computing services will be readily available on demand, like other utility services available in today's society. Similarly, users (consumers) need to pay providers only when they access the computing services. In addition, consumers no longer need to invest heavily or encounter difficulties in building and maintaining complex IT infrastructure.

In such a model, users access services based on their requirements without regard to where the services are hosted. This model has been referred to as *utility computing*, or recently as *Cloud computing* . The latter term denotes the infrastructure as a “Cloud” from which businesses and users can access applications as services from anywhere in the world on demand. Hence, Cloud computing can be classified as a new paradigm for the dynamic provisioning of computing services supported by state-of-the-art data centers that usually employ Virtual Machine (VM) technologies for consolidation and environment isolation purposes .[2] Many computing service providers including Google, Microsoft, Yahoo, and IBM are rapidly deploying data centers in various locations around the world to deliver Cloud computing services.

Cloud computing delivers infrastructure, platform, and software (applications) as services, which are made available to consumers as subscription-based services under the pay-as-you-go model. In industry these services are referred to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) respectively.[3] A recent Berkeley report stated “Cloud Computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service”.

Clouds aim to drive the design of the next generation data centers by architecting them as networks of virtual services (hardware, database, user-interface, application logic) so that users can access and deploy applications from anywhere in the world on demand at competitive costs depending on their QoS (Quality of Service) requirements .[4]

2. Need of Cloud Computing

The need of cloud computing can be explained with the help of an example. The following graph shows the number of users who log on to the Australian Open web page.

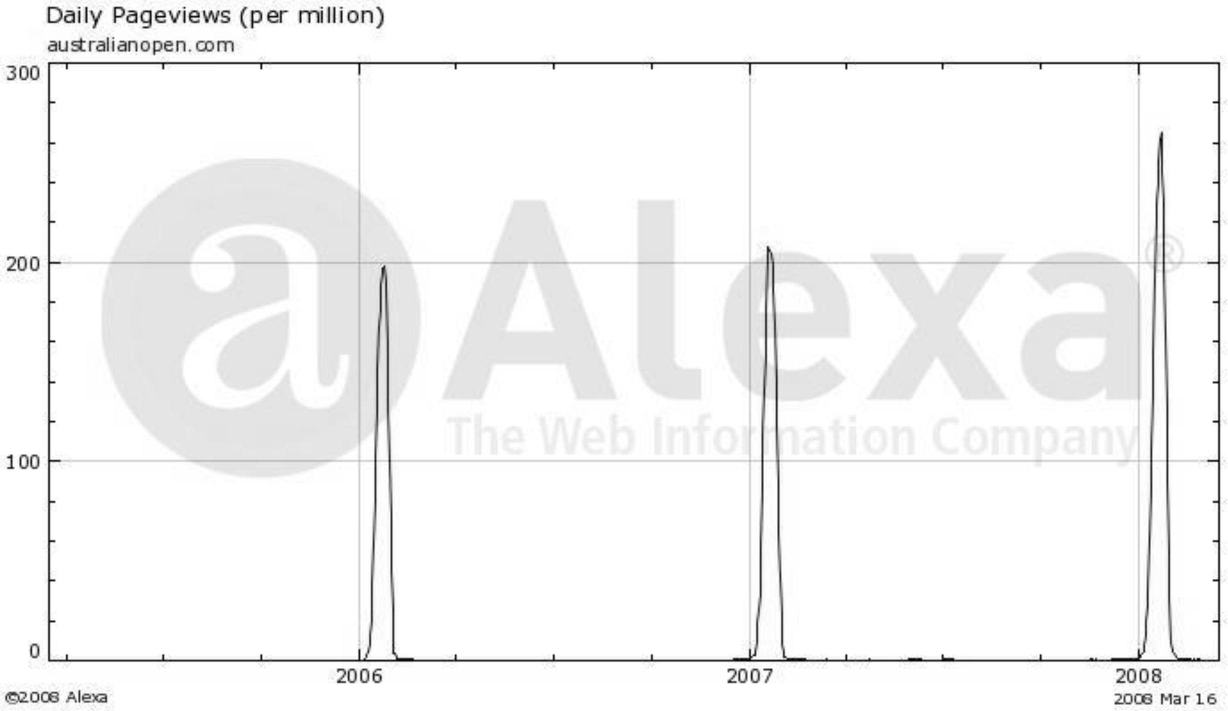


fig 1:monthly page views of australian open official website

The spikes correspond to the month of January during which the tournament is going on. The site remains almost dormant during the rest of the year. It would be wasteful to have servers which can cater to the maximum need,as they wont be needed during the rest of the year.[5] The concept of cloud computing comes to the rescue at this time. During the peak period, cloud providers such as Google,Yahoo,Microsoft etc.can be approached to provide the necessary server capacity.

In this case, Infrastructure is provided as a service(IaaS) through cloud computing. Likewise,cloud providers can be approached for obtaining software or platform as a service. Developers with innovative ideas for new Internet services no longer require large capital outlays in hardware to deploy their service or human expense to operate it .[6] Cloud computing offers significant benefits to IT companies by freeing them from the low-level task of setting up basic hardware and software infrastructures and thus enabling focus on innovation and creating business value for their services.

3. Green Computing

Green computing is defined as the atudy and practice of designing , manufacturing, using, and disposing of computers, servers, and associated subsystems—such as monitors, printers, storage devices, and networking and communications systems—efficiently and effectively with minimal or no impact on the environment." The goals of green computing are similar to green chemistry; reduce the use of hazardous materials, maximize energy efficiency during the product's lifetime, and promote the recyclability or biodegradability of defunct products and factory waste. Research continues into key areas such as making the use of computers as energy-efficient as possible, and designing algorithms and systems for efficiency-related computer technologies.[7]

There are several approaches to green computing,namely

- Product longetivity
- Algorithmic efficeincy
- Resource allocation
- Virtualisation
- Power management etc.

4. Need of green computing in clouds

Modern data centers, operating under the Cloud computing model are hosting a variety of applications ranging from those that run for a few seconds (e.g. serving requests of web applications such as e-commerce and social networks portals with transient workloads) to those that run for longer periods of time (e.g. simulations or large data set processing) on shared hardware platforms. The need to manage multiple applications in a data center creates the challenge of on-demand resource provisioning and allocation in response to time-varying workloads. Normally, data center resources are statically allocated to applications, based on peak load characteristics, in order to maintain isolation and provide performance guarantees. Until recently, high performance has been the sole concern in data center deployments and this demand has been fulfilled without paying much attention to energy consumption. The average data center consumes as much energy as 25,000 households [8]. As energy costs are increasing while availability dwindles, there is a need to shift focus from optimising data center resource management for pure performance to optimising for energy efficiency while maintaining high service level performance. According to certain reports,the total estimated energy bill for data centers in 2010 is $11.5 billion and energy costs in a typical data center double every five years.

Data centers are not only expensive to maintain, but also unfriendly to the environment. Data centers now drive more in carbon emissions than both Argentina and the Netherlands . High energy costs and huge carbon footprints are incurred due to massive amounts of electricity needed to power and cool numerous servers hosted in these data centers.[9] Cloud service providers need to adopt measures to ensure that their profit margin is not dramatically reduced due to high energy costs. For instance, Google, Microsoft, and Yahoo are building large data centers in barren desert land surrounding the Columbia River, USA to exploit cheap and reliable hydroelectric power . There is also increasing pressure from Governments worldwide to reduce carbon footprints, which have a significant impact on climate change. For example, the Japanese government has established the Japan Data Center Council to address the soaring energy consumption of data centers . Leading computing service providers have also recently formed a global consortium known as The Green Grid to promote energy efficiency for data centers and minimize their environmental impact.

Lowering the energy usage of data centers is a challenging and complex issue because computing applications and data 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 minimise energy consumption.[10] This is essential 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.

Architecture of a green cloud computing platform

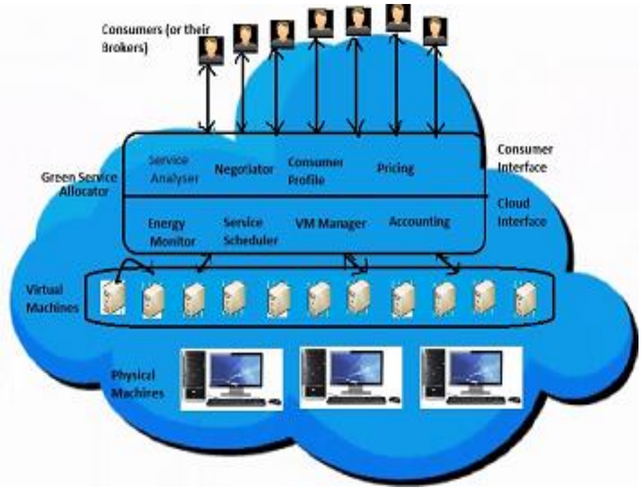


Figure 2: architecture of a green cloud computing environment

Figure 2 shows the high-level architecture for supporting energy-efficient service allocation in Green Cloud computing infrastructure.

There are basically four main entities involved:

a) Consumers/Brokers**:** Cloud consumers or their brokers submit service requests from anywhere in the world to the Cloud. It is important to notice that there can be a difference between Cloud consumers and users of deployed services. For instance, a consumer can be a company deploying a Web application, which presents varying workload according to the number of users accesing it.

b) *Green Resource Allocator*: Acts as the interface between the Cloud infrastructure and consumers. It requires the interaction of the following components to support energy-efficient resource management:

1. Green Negotiator: Negotiates with the consumers/brokers to finalize the SLA with specified prices and penalties (for violations of SLA) between the Cloud provider and consumer depending on the consumer's QoS requirements and energy saving schemes. In case of Web applications, for instance, QoS metric can be 95% of requests being served in less than 3 seconds.[11]

2. Service Analyser: Interprets and analyses the service requirements of a submitted request before deciding whether to accept or reject it. Hence, it needs the latest load and energy information from VM Manager and Energy Monitor respectively.

3. Consumer Profiler: Gathers specific characteristics of consumers so that important consumers can be granted special privileges and prioritised over other consumers.

4. Pricing: Decides how service requests are charged to manage the supply and demand of computing resources and facilitate in prioritising service allocations effectively.

5. *Energy Monitor*: Observes and determines which physical machines to power on/off.

6. *Service Scheduler*: Assigns requests to VMs and determines resource entitlements for allocated VMs.

It also decides when VMs are to be added or removed to meet demand.

□ *VM Manager*: Keeps track of the availability of VMs and their resource entitlements. It is also in charge of migrating VMs across physical machines.

□ *Accounting*: Maintains the actual usage of resources by requests to compute usage costs. Historical usage information can also be used to improve service allocation decisions.

c) ***VMs*:** Multiple VMs can be dynamically started and stopped on a single physical machine to meet accepted requests, hence providing maximum flexibility to configure various partitions of resources on the same physical machine to different specific requirements of service requests. Multiple VMs can also concurrently run applications based on different operating system environments on a single physical machine. In addition, by dynamically migrating VMs across physical machines, workloads can be consolidated and unused resources can be put on a low-power state, turned off or configured to operate at low-performance levels (e.g., using DVFS) in order to save energy.

d) *Physical Machines*: The underlying physical computing servers provide hardware infrastructure for creating virtualised resources to meet service demands. [12]

5.Making cloud computing more green

Mainly three approaches have been tried out to make cloud computing environments more environmental friendly. These approaches have been tried out in the data centres under experimental conditions. The practical application of these methods are still under study. The methods are:

- ***Dynamic Voltage frequency scaling technique(DVFS)***:- Every electronic circutory will have an operating clock associated with it. The operatin frequency of this clock is adjusted so that the supply voltage is regulated. Thus, this method heavily depends on the hardware and is not controllabale according to the varying needs. The power savings are also low compared to other approaches. The power savings to cost incurred ratio is also low.

- ***Resource allocation or virtual machine migration techniques***:- In a cloud computing environment,every physical machine hosts a number of virtual machines upon which the applications are run. These virtual machines can be transfered across the hosts according to the varying needs and avaialble resources.The VM migration method focusses on transferring VMs in such a way that the power increase is least. The most power efficient nodes are selected and the VMs are transfered across to them. This method is dealt in detail later.
- ***Algorithmic approaches***:- It has been experimently determined that an ideal server consumes about 70% of the power utilised by a fully utilised server. (See figure 3).

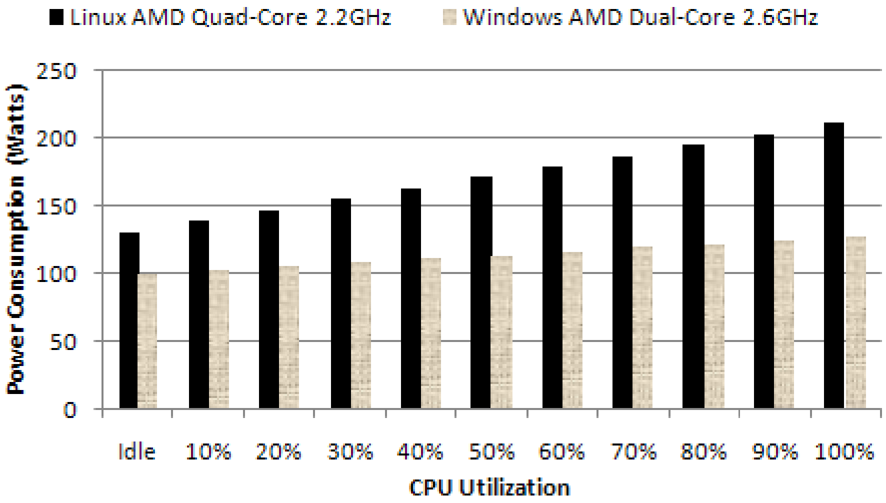


Fig 3: Power consumption under different work loads.

Using a neural network predictor,the green scheduling algorithms first estimates required dynamic workload on the servers. Then unnecessary servers are turned off in order to minimize the number of running servers, thus minimizing the energy use at the points of consumption to provide benefits to all other levels. Also,several servers are added to help assure service-level agreement. The bottom line is to protect the environment and to reduce the total cost of ownership while ensuring quality of service.

VM Migration

The problem of VM allocation can be divided in two: the first part is admission of new requests for VM provisioning and placing the VMs on hosts, whereas the second part is optimization of current allocation of VMs. [13]

Optimization of current allocation of VMs is carried out in two steps: at the first step we select VMs that need to be migrated, at the second step chosen VMs are placed on hosts using MBFD algorithm. We propose four heuristics for choosing VMs to migrate. The first heuristic, **Single Threshold (ST)**, is based on the idea of setting upper utilization threshold for hosts and placing VMs while keeping the total utilization of CPU below this threshold. The aim is to preserve free resources to prevent SLA violation due to consolidation in cases when utilization by VMs increases. At each time frame all VMs are reallocated using MBFD algorithm with additional condition of keeping the upper utilization threshold not violated. The new placement is achieved by live migration of VMs .

The other three heuristics are based on the idea of setting upper and lower utilization thresholds for hosts and keeping total utilization of CPU by all VMs between these thresholds. If the utilization of CPU for a host goes below the lower threshold, all VMs have to be migrated from this host and the host has to be switched off in order to eliminate the idle power consumption. If the utilization goes over the upper threshold, some VMs have to be migrated from the host to reduce utilization in order to prevent potential SLA violation. We propose three policies for choosing VMs that have to be migrated from the host. [14]

□ Minimization of Migrations (MM) – migrating the least number of VMs to minimise migration overhead.

□ Highest Potential Growth (HPG) – migrating VMs that have the lowest usage of CPU relatively to the requested in order to minimise total potential increase of the utilization and SLA violation.

□ Random Choice (RC) – choosing the necessary number of VMs by picking them according to a uniformly distributed random variable.

6. *Experimental Setup*

As the targeted system is a generic Cloud computing environment, it is essential to evaluate it on a large-scale virtualised data center infrastructure. However, it is difficult to conduct large-scale experiments on a real infrastructure, especially when it is necessary to repeat the experiment with the same conditions (e.g. when comparing different algorithms). Therefore, simulations have been chosen as a way to evaluate the proposed heuristics. The CloudSim toolkit has been chosen as a simulation platform as it is a modern simulation framework aimed at Cloud computing environments. In contrast to alternative simulation toolkits (e.g. SimGrid, GandSim), it supports modeling of on-demand virtualization enabled resource and application management. It has been extended in order to enable power-aware simulations as the core framework does not provide this capability. Apart from the power consumption modeling and accounting, the ability to simulate service applications with variable over time workload has been incorporated.[15]

There are a few assumptions that have been made to simplify the model of the system and enable simulation-driven evaluation. The first assumption is that the overhead of VM migration is considered as negligible. Modeling the cost of migration of VMs is another research problem and is being currently investigated . However, it has been shown that application of live migration of VMs can provide reasonable performance overhead. Moreover, with advancements of virtualization technologies, the efficiency of VM migration is going to be improved. Another assumption is that due to unknown types of applications running on VMs, it is not possible to build the exact model of such a mixed workload . Therefore, rather than simulating particular applications, the utilization of CPU by a VM is generated as a uniformly distributed random variable. In the simulations we have defined that SLA violation occurs when a VM cannot get amount of MIPS that are requested. This can happen in cases when VMs sharing the same host require higher CPU performance that cannot be provided due to consolidation. To compare efficiency of the algorithms we use a characteristic called SLA violation percentage, or simply SLA violation, which is defined as a percentage of SLA violation events relatively to the total number of measurements.

A data center that comprises 100 heterogeneous physical nodes was simulated. Each node is modeled to have one CPU core with performance equivalent to 1000, 2000 or

3000 Million Instructions Per Second (MIPS), 8 Gb of RAM and 1 TB of storage. According to this model, a host consumes from 175 W with 0% CPU utilization and up to 250 W with 100% CPU utilization. Each VM requires one CPU core with 250, 500, 750 or 1000 MIPS, 128 MB of RAM and 1 GB of storage. The users submit requests for provisioning of 290 heterogeneous VMs that fills the full capacity of the simulated data center. Each VM runs a web-application or any kind of application with variable workload, which is modeled to create the utilization of CPU according to a uniformly distributed random variable. The application runs for 150,000 MIPS that equals to 10 minutes of execution on 250 MIPS CPU with 100% utilization. Initially, VMs are allocated according to the requested characteristics assuming 100% utilization. Each experiment has been run 10 times and the presented results are built upon the mean values.

Simulation Results

For the benchmark experimental results we have used a Non Power Aware (NPA) policy. This policy does not apply any power aware optimizations and implies that all hosts run at 100% CPU utilization and consume maximum power. The second policy applies DVFS, but does not perform any adaptation of allocation of VMs in run-time. For the simulation setup described above, using the NPA policy leads to the total energy consumption of 9.15 KWh, whereas DVFS allows decreasing this value to 4.4 KWh.[16]

The simulation resuts of various policies are explained in the next sections.

Energy Consumption and SLA vioaltion of ST policy

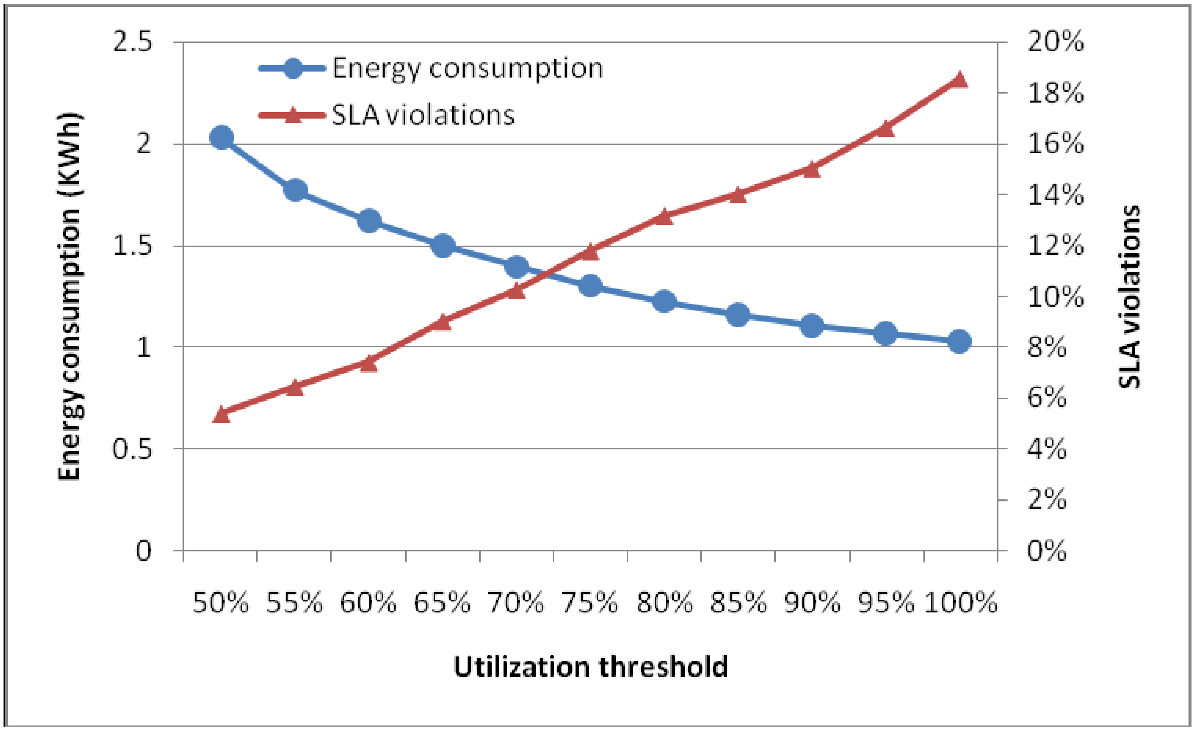


Fig 4:simulation results of ST policy

To evaluate ST policy we conducted several experiments with different values of the utilization threshold. The simulation results are presented in Figure 4. The results show that energy consumption can be significantly reduced relatively to NPA and DVFS policies – by 77% and 53% respectively with 5.4% of SLA violations. They show that with the growth of the utilization threshold energy consumption decreases, whereas percentage of SLA violations increases. This is due to the fact that higher utilization threshold allows more aggressive consolidation of VMs, however, by the cost of the increased risk of SLA violations.

Energy consumption and SLA violations of other policies

We have compared MM policy with HPG and RC policies varying exact values of the thresholds but preserving 40% interval between them. The results (Figures 5 & 6) show that these policies allow the achievement of approximately the same values of energy consumption and SLA violations. Whereas the number of VM migrations produced by MM policy is reduced in comparison to HPG policy by maximum of 57% and 40% on average and in comparison to RC policy by maximum of 49% and 27% on average.

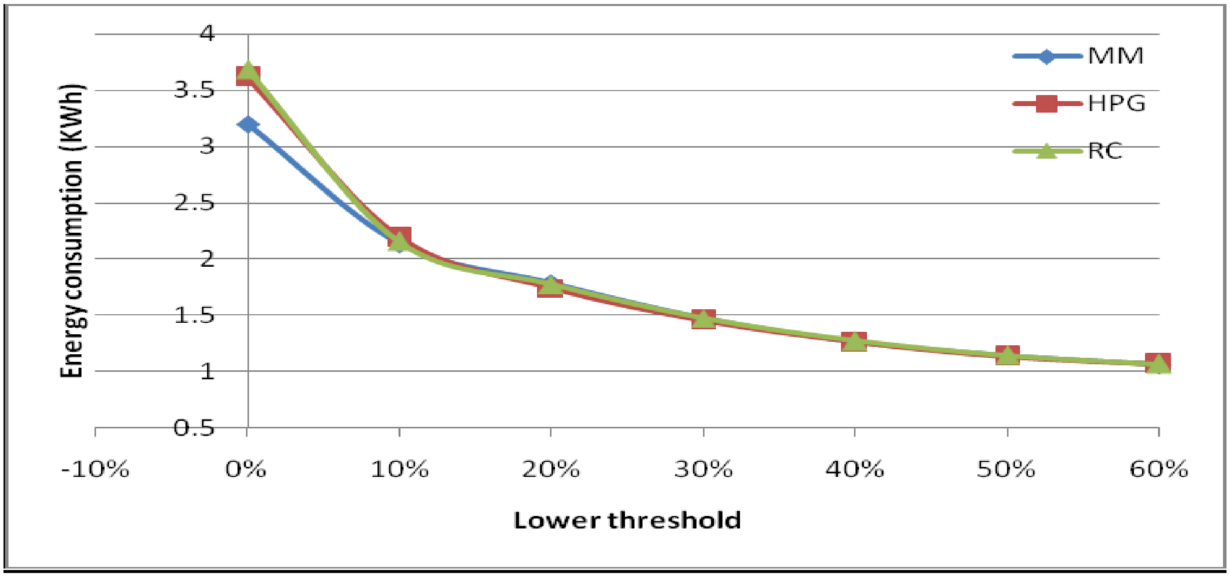


Fig 5:energy consumption of different policies

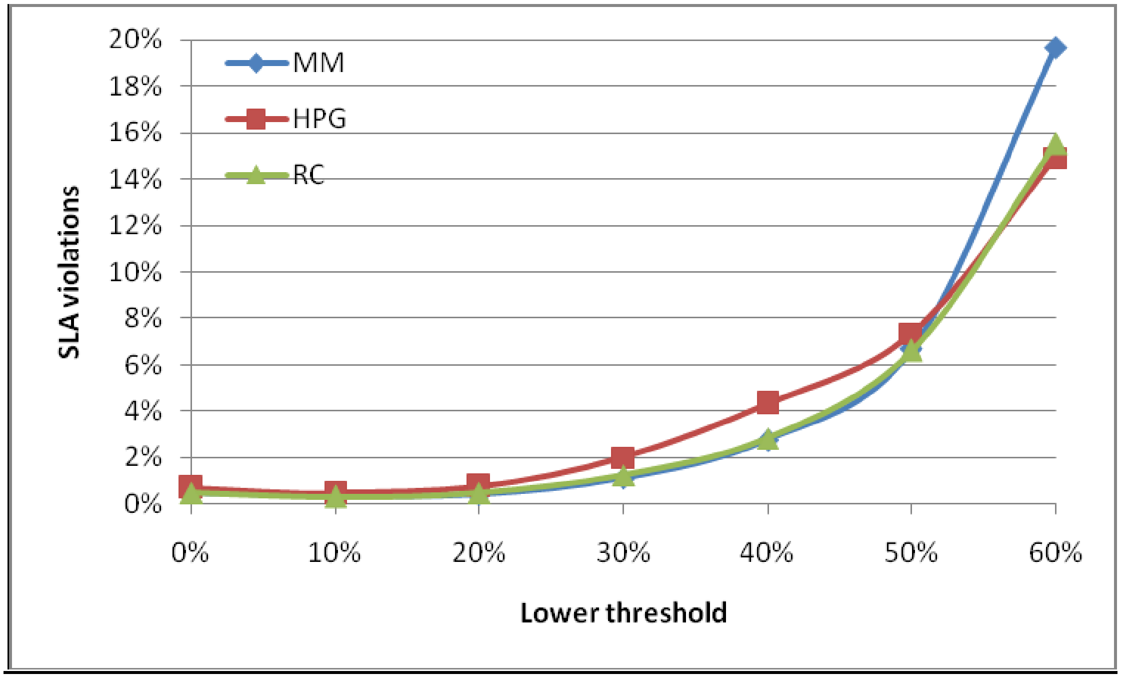
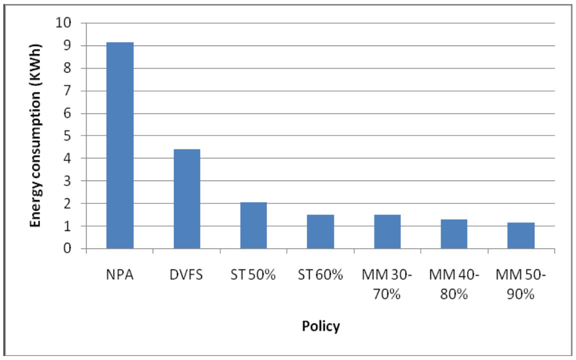


Fig6:SLA violations of different policies under different thresholds
Comparison with respect to CPU utilisation



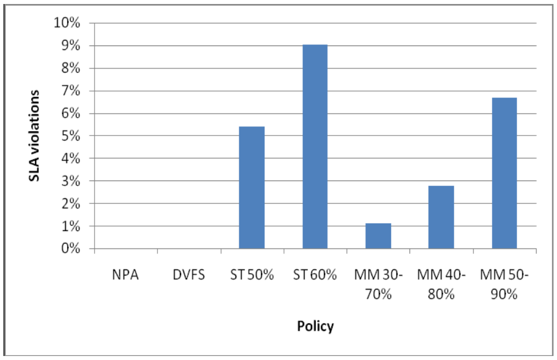


Fig 7:comparison of diff policies under different workloads with respect to cpu utilsation

As can be seen,the Non Power Aware (NPA) policy which is currently being followed results in higher power loss,even though it doesnt result in SLA violations. The optimal case occurs when MM policy under a workload of 30-70%is used.[17]

7. Conclusion

Applying green technologies is highly essential for the sustainable development of cloud computing. Of the various green methodologies enquired, the DVFS technology is a highly hardware oriented approach and hennce less flexible. The reuslt of various VM migration simulations show that MM policy leads to the best energy savings: by 83%, 66% and 23% less energy consumption relatively to NPA, DVFS and ST policies respectively with thresholds 30-70% and ensuring percentage of SLA violations of 1.1%; and by 87%, 74% and 43% with thresholds 50-90% and 6.7% of SLA violations. MM policy leads to more than 10 times less VM migrations than ST policy. The results show flexibility of the algorithm, as the thresholds can be adjusted according to SLA requirements. Strict SLA (1.11%) allow the achievement of the energy consumption of 1.48 KWh. However, if SLA are relaxed (6.69%), the energy consumption is further reduced to 1.14 KWh. Single threshold policies can save power upto 20%,but they also cause a large number of SLA violations. Green scheduling algorithms based on neural predictors can lead to a 70% power savings. These policies also enable us to cut down data centre energy costs, thus leading to a strong,competitive cloud computing industry. End users will also benefit from the decreased energy bills.

8. References

[1] D. Cavdar and F. Alagoz, (Eds.), "A Survey of Research on Greening Data Centers", Proceedings of the IEEE Global Communications Conference (GLOBECOM), (2012) December 3-7; Anaheim, CA.

[2] A. Jain, M. Mishra, S. Kumar Peddoju and N. Jain, (Eds.), "Energy Efficient Computing-Green Cloud Computing", Proceedings of the International Conference of the Energy Efficient Technologies for Sustainability (ICEETS), (2013) April 10-122; Nagercoil.

[3] T. Vinh T. Duy, Y. Sato and Y. Inoguchi, (Eds.), "Performance Evaluation of a Green Scheduling Algorithm for Energy Savings in Cloud Computing", Proceedings of the IEEE International Symposium of the Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), (2010) April 19-23; Atlanta, GA.

[4] F. Satoh, H. Yanagisawa, H. Takahashi and T. Kushida, (Eds.), "Total Energy Management system for Cloud Computing", Proceedings of the IEEE International Conference of the Cloud Engineering (IC2E), (2013), March 25-27; Redwood City, CA.

[5] C. Belady, (Ed.), "How to Minimize Data Centre Utility Bills", US (2006).

[6] R. Beik, (Ed.), "Green Cloud Computing: An Energy-Aware Layer in Software Architecture", Proceedings of the Spring Congress of the Engineering and Technology (S-CET), (2012), May 27-30; Xian.

[7] "Green Grid Metrics—Describing Data Centres Power Efficiency", Technical Committee White Paper by the Green Grid Industry Consortium, (2007) February.

[8] S. Greenberg, E. Mills, B. Tschudi, P. Rumsey and B. Myatt, (Eds.), "Best Practices for Data Centres: Results from Benchmarking 22 Data Centres", Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings, (2006) April, pp. 3-76, -3-87.

[9] T. Kgil, D. Roberts and T. Mudge, "Pico Server: Using 3D Stacking Technology to Build Energy Efficient Servers", vol. 4, no. 16, (2006).

[10] N. Rassmussen, (Ed.), "Electrical Efficiency Modelling of Data Centres", American Power Conversion (APC) White Paper #113, (2007) October, pp.1-18.

[11] B. Priya, E. S. Pilli and R. C. Joshi, (Eds.), "A Survey on Energy and Power Consumption Models for Greener Cloud", Proceeding of the IEEE 3rd International Advance Computing Conference (IACC), (2013), February 22-23; Ghaziabad.

[12] D. Kliazovich and P. Bouvry, (Eds.), "Green Cloud: A Packet-level Simulator of Energy-aware Cloud Computing Data Centers", Proceeding of the IEEE Global Telecommunications Conference (GLOBECOM), (2010), December 6-8; Miami, FL.

[13] S. K. Garg, C. S. Yeo and R. Buyya, (Eds.), "Green Cloud Framework for Improving Carbon Efficiency of Clouds", Proceedings of the 17th International European Conference on Parallel and Distributed Computing (2011), August-September, Bordeaux, France.

[14] A. Beloglazov and R. Buyya, (Eds.), "Energy Efficient Allocation of Virtual Machines in Cloud Data

Centres", Proceedings of the 10th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid), (2010) May 17-20; Melbourne, Australia.

[15] M. N. Hulkury and M. R. Doomun, (Eds.), "Integrated Green Cloud Computing Architecture", Proceedings of the International Conference on Advanced Computer Science Applications and Technologies (ACSAT), (2012), Washington DC, USA.

[16] S. K. Garg, C. S. Yeo and R. Buyya, (Eds.), "Green Cloud Framework for Improving Carbon Efficiency of Clouds", Proceedings of the 17th International European Conference on Parallel and Distributed Computing (EuroPar), (2011) August-September. Bordeaux, France.

[17] A. R. Nimje, V. T. Gaikwad and H. N. Datir, (Eds.), "Green Cloud Computing: A virtualized Security Framework for Green Cloud Computing", Proceeding of the International Journal of Advanced Research in Computer Science and Software Engineering.