Cluster Computing
https://doi.org/10.1007/s10586-019-02946-w

Economic impact of energy saving techniques in cloud server

Bilal Ahmad[1] · Zaib Maroof[2] · Sally McClean[1] · Darryl Charles[1] · Gerard Parr[3]

Received: 11 September 2018 / Revised: 7 April 2019 / Accepted: 27 May 2019

Abstract

In recent years, lot of research has been carried in the field of cloud computing and distributed systems to investigate and understand their performance. Economic impact of energy consumption is of major concern for major companies. Cloud Computing companies (Google, Yahoo, Gaikai, ONLIVE, Amazon and eBay) use large data centers which are comprised of virtual computers that are placed globally and require a lot of power cost to maintain. Demand for energy consumption is increasing day by day in IT firms. Therefore, Cloud Computing companies face challenges towards the economic impact in terms of power costs. Energy consumption is dependent upon several factors, e.g., service level agreement, virtual machine selection techniques, optimization policies, workload types etc. We address a solution for the energy saving problem by enabling dynamic voltage and frequency scaling technique for gaming data centers. The dynamic voltage and frequency scaling technique is compared against non-power aware and static threshold detection techniques. This helps service providers to meet the quality of service and quality of experience constraints by meeting service level agreements. The CloudSim platform is used for implementation of the scenario in which game traces are used as a workload for testing the technique. Selection of better techniques can help gaming servers to save energy cost and maintain a better quality of service for users placed globally. The novelty of the work provides an opportunity to investigate which technique behaves better, i.e., dynamic, static or non-power aware. The results demonstrate that less energy is consumed by implementing a dynamic voltage and frequency approach in comparison with static threshold consolidation or non-power aware technique. Therefore, more economical quality of services could be provided to the end users.

Keywords Energy saving technique · Economic impact · Dynamic frequency scaling · Static threshold and non-power aware technique · Service level agreement · Quality of service

1 Introduction

Cloud Computing is growing day by day with the development of IT services. The reason for this development is cost effectiveness and quality of experience from user's perspective. IT industry is becoming adaptable to cloud computing technologies for achievement of quality of service and quality of experience matrices. Along with provisioning of better quality of service cloud providers can scramble towards more profits by saving resources e.g. energy, bandwidth consumption etc. In cloud environment, it can be administered that servers play an important role in the design of cloud infrastructure and resource allocation. With the era of globalization, computing is also being transformed into a model where service is provided based on user requirements instead of hosting them permanently [1]. This provides the industry with the liberty to reach the users doorstep for the provision of services [2]. Cloud

✉ Bilal Ahmad
ahmad-b@ulster.ac.uk

Zaib Maroof
zaib.maroof@fui.edu.pk

Sally McClean
si.mcclean@ulster.ac.uk

Darryl Charles
dk.charles@ulster.ac.uk

Gerard Parr
g.parr@uea.ac.uk

1 School of Computing, Ulster University, Coleraine, UK

2 School of Management, Foundation University, Islamabad, Pakistan

3 School of Computing, University of East Anglia, Norwich, UK

Published online: 07 June 2019

Computing provides users with multiple advantages, e.g., services, resources, and developer tools. It facilitates researchers to develop, tests and implement their ideas. It also provisions them to use the latest services on different devices (tablets, phones, home appliances etc.). Cloud Computing has unmatchable advantages to its predecessors because of technological advancements, e.g., virtualization, storage, processing, memory, performance, low cost, ease of excess, mobility, high expansibility, reliability, and fast bandwidth etc. These advancements and innovations in the field of cloud technology provisions the industries to have unlimited computational power while maintaining good quality of service (QoS). Cloud industries must maintain several service level agreements (SLAs) to meet high quality of service requirements from the user and service provider perspective. The service provider is also responsible for the availability of the resources whenever and wherever they are required by the user. This also presents challenge how energy consumption can be reduced while having minimum service level agreement violations (SLAVs) (Figs. 1, 2, 3, 4).

All type of resource allocation and scheduling is related to server's physical design and resource allocation policies. System designer major task is management of trade-offs between quality of service factor and energy consumption. Idle servers can be turned off for power saving purpose and expense to profit ratio can be improved. But this can also hamper the quality of service factor i.e. latency when they must be turned on as requested by the users. To date, many suggestion and ideas have been proposed for energy consumption for jobs arriving in cloud servers. The quality of cloud service depends upon the fact that how much stable resource allocation is provided to the user requesting

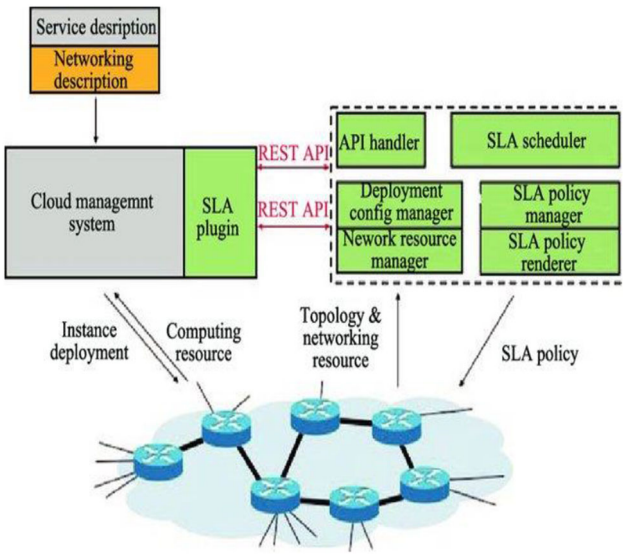

Fig. 1 Overview of system architecture

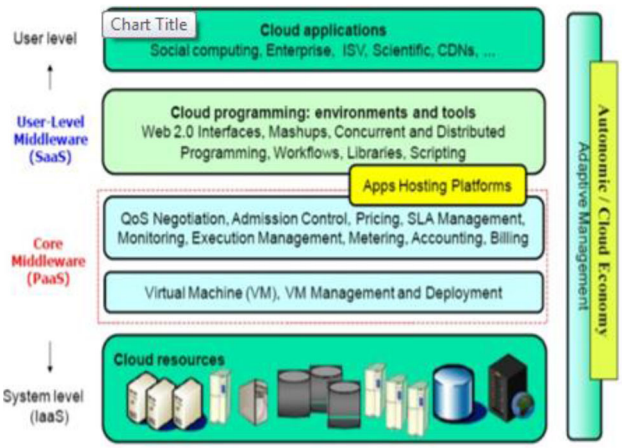

Fig. 2 Layered CloudSim architecture overview

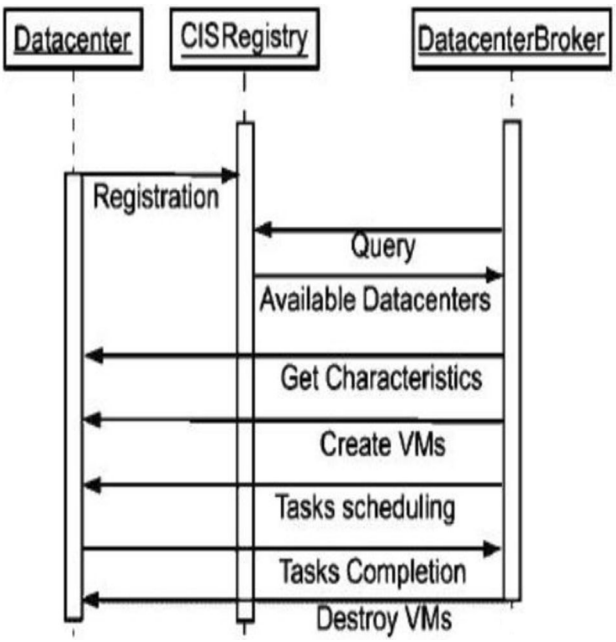

Fig. 3 Simulation data flow in CloudSim

the service. For the achievement of this goal virtualisation is carried out by the service providers. Large scale data centre consists of thousands of hosts and nodes resulting in consumption of large amount of energy. As a result, cloud servers are being designed in such a way that they become automatically adaptable to the requested service by the users [2].

The corresponding large amount of data management and streaming leads to an increase in energy consumption. All kind of services (gaming, internet of things, Big Data etc.) that are hosted over the cloud environment are maintained using large data centers that are placed globally. When observed closely, it can be seen that these servers are not running at their full performance, i.e., 100%

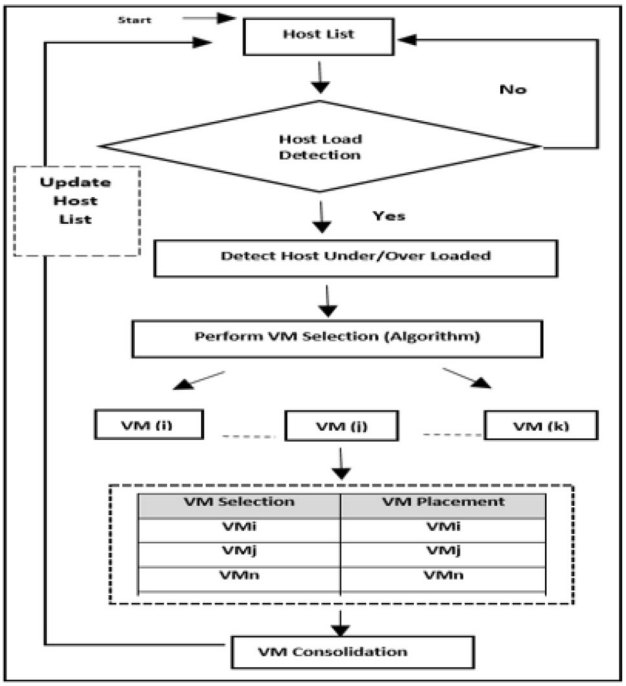

Fig. 4 Flow chart for the VM consolidation

utilization while remaining idle at other times. Therefore, an ample amount of energy is wasted to keep these servers running 24/7. This causes a major rise in cost and threat to the environment as large amount of carbon dioxide (CO_2) is produced by these data servers [3]. Consequently, data centers are becoming unmaintainable. Therefore, a lot of work is being carried out, researchers are investigating different kinds of algorithms and techniques. There are different procedures in which this workload can be handled ranging from dynamic to static threshold and non-power aware technique. If virtualization is used in these big gaming server's energy consumptions can be reduced, and better quality of service could be provided. The hosts that are under or overloaded can be relocated, and energy could be saved in this aspect. Services provided by Cloud provisioners varies with time and have different workloads that require dynamic or static allocation of resources especially for Big Data Applications and Multiplayer Games. The migration of virtual machine can help in saving of energy, but it can also degrade the quality of service on the other hand. Tradeoff is required to be managed between user experience and quality of service. Therefore, such techniques are required to be implemented in gaming with awareness of dynamic and static workloads. This can help in the reduction of energy consumption while maintaining a quality of service and quality of experience [4].

For testing of new algorithms in IT industry researcher needs to have a secure platform. The selected platform should be fail-safe and must avoid risk to customers data privacy and data impairment [5]. Most cloud computing platforms are software based as it is very difficult and expensive to set a cloud server for test and trials purposes for each researcher. For example, it is practically difficult for a researcher to use a data server consisting of 200 physical machines because of maintenance costs, (e.g., energy, space, expense, power, and cooling requirements) [6]. There is also no specific platform due to the following reasons: the relocation of the virtual machine, confidentiality and data integrity, a need for energy management, and cost modelling [7]. The main purpose of carrying this research is, therefore, to find how resource optimization can be performed in the gaming data centres. In our work, we consider the following aspects of service quality: energy consumption and service level agreements, by using online gaming data in our experiments. In this paper, DVFS, Non-Power Aware and Static Threshold virtual machine consolidation technique will be tested and implemented for the improvement of energy consumption and SLAs. Better results are expected to be achieved using dynamic voltage and frequency technique as compared to a static threshold or non-power aware technique; this hypothesis will be verified using real-time gaming workload.

The rest of this paper is organized as follows, Section 2 describes the related work; Section 3 presents the basics about platform and techniques; Section 4 addresses the simulation environment; Section 5 discusses performance analysis and provides a discussion of our approach while, conclusions and future work close the article.

2 Related work

The concept of dynamic voltage and frequency scaling has been used by Ahmad et al. Tests were performed using gaming data. The results show that dynamic voltage and frequency scaling technique saves more energy as compared to non-power aware technique [1]. Work has been carried in the field of cloud computing particularly relating to the cluster servers and virtualized servers. Here, the authors use a single system by implementing and comparing three different energy saving concepts, i.e., the supply voltage of underloaded servers is reduced, idle servers are left in sleep mode and thirdly, the two techniques are combined for analysis. The author proposes that DNS and changing voltages together provide better results for energy saving. However, the paper lacks cost comparison for quality of service matrices [8]. A solution is provided to save cost and to earn more profit on a large data scale by managing the scheduling of heterogeneous machines with multiple users. This work is limited to just

one quality of service metric, i.e., cost from the service provider perspective [9]. Another algorithm was designed to optimize energy by using the concept of multi objective workflow and dynamic voltage scaling. However, the user was given the ability to choose between the cost or energy criterion [10]. In the field of computing, distributed computing provides the user with fault tolerance, organization, and support for resources. Typically, resources are allocated to the users based on load balancing technique. In this, all resources are allocated to the broker that is wholly responsible for the provisioning of resources when required [11].

IT industry is evolving day by day from the domain of grid, parallel and distributed computing. With the ongoing development of the industry many simulation software for cloud-based environment have evolved e.g. GridSim, CloudSim, Green Cloud, iCan Cloud etc. This software help researcher around the world to design and test their algorithms and techniques for improvement of quality of service and quality of experience [12]. The author addresses the issue related to quality of service and service level agreement by using the energy constraint as a core parameter. Virtualization concept has been implemented in graphics card and central processing unit. The test helps in determining how latency factor can be improved by exploiting the game frames. The results predict that quality of service could be enhanced by exploiting a trade-off between different factors, e.g., data buffering, scalability, redundancy and game latency [13]. The virtualization concept has been used by the author for maintenance of quality of service. The idle virtual machines are migrated from servers for maintenance of load balancing. The technique suggests that energy can be saved in small online cloud servers. However, a live migration technique was used and can cause bottleneck in large and busy network [14]. In [15] the authors propose quality of service algorithms using scheduling policies. However, the work was related to virtualization mechanism only for large scale global data centers. Further work was carried out relating the energy saving mechanism to different kinds of workflow on the Green Cloud Platform using bi-objective scheduling to meet the quality of service matrices for energy consumption [16].

By looking at the related work it can be concluded that main research area involves single servers and unique tasks. However, these days' cloud computing platforms like Gaikai, OnLive, and Amazon EC2 have servers that are using multipurpose applications that are dispersed geographically. However, there is a research gap in the field of gaming especially for multiplayer games with users placed far apart from each other. On the other hand, some work about energy saving has been carried using Big Data with single purpose applications [5]. The concept of virtualization has been implemented by the author using local regression robust migration algorithm. Work suggests that latency and service quality can be achieved in Big Data servers by using this virtualization technique. However, a tradeoff is required between quality of service and quality of experience [17].

3 Basics about platform and techniques

CloudSim is one of the platforms which provides QoS parameters such as: energy, cost model, latency, virtual machine characteristics, federation policy, and analyzing the network communication model. Based on this platform, several popular models have also been designed, namely iFogSim, Cloud Analyst, Network CloudSim and iCaroCloud. Therefore, it provides enough leverage for researchers to use it to perform tests and develop new models as required. CloudSim has a layered architecture which provides user with the ability to design and implement applications. It supports core functions, such as handling of events, creation of cloud servers, hosts, brokers, and virtual machines [18]. The CloudSim simulation layer supports creation of hosts under virtual machines, application execution and application monitoring. A researcher who wants to implement an application relating energy, hosts, VM and data centers will be doing at this level. This layer supports the SaaS platform and provides users with defined quality of service levels with complex load reporting and application performance reports [19]. The topmost layer in the CloudSim architecture is where a user writes a code and it allows the user to define several virtual machines, hosts, data centers, brokers, tasks etc [20].

Therefore, it allows researchers to extend this layer and perform different tasks such as: generation of workload for monitoring designed experiments, designing of different cloud scenarios for robust testing and implementation of conventional applications in the cloud environment [18]. IaaS services can be simulated by extending different entities present in cloud environments such as data centers. Such data centers consist of many hosts which are assigned to more than one virtual machines depending upon the rules defined by the service provider [21]. The data center can also manage more than one host (physical components representing the computing server) which further manages virtual machines. Host provisioning supports single and multiple core nodes. Similarly, virtual machine allocation creates virtual machine scenarios on hosts for storage and memory related tasks [22]. After modelling and designing of the application, it is allocated to a running virtual machine through a specific defined procedure. The virtual machines required to host multiple applications are provided on a First Come First Serve basis depending upon

different hardware factors (storage, memory, cores etc.). Therefore, simulation test scenarios relating to CPU cores are dependent upon factors such as time usage, space sharing policy or allocating virtual machines as and when required [23].

It can analyse the system and its components properties, e.g., the number of virtual machines, data centers, resource provisioning policies and hosts [24]. It has the capability to support single and multi-cloud environments. The platform has a wide implementation in computing industry for testing of energy management systems and resource allocation scenarios (HP Labs in USA). It provides support for simulation of virtualized data centers in the cloud environment (memory, storage, bandwidth, and virtual machines). Cloud Sim has number of compelling features that provide support and speed up the development process of the applications [25]. These features include fast processing, flexible approach, support for modelling and simulation, self-contained platform, network support, federation policy, availability of virtualisation engine, ease of allocation energy and cost model, service area data type, availability programming language and graphical user interface [26].

3.1 DVFS technique

Dynamic voltage and frequency scaling (DVFS) is a technique that works by dynamically controlling the data in the hosts. It reduces the use of underutilized resources by dynamically controlling the frequency parameter and uses different strategies to reduce energy consumption by shifting load to the underutilized servers dynamically. Therefore, for the implementation of DVFS one needs to understand different factors like frequency and static power consumption.

CloudSim can calculate the power consumption of data centers using the DVFS technique. It uses the current metric for cloud host input and returns calculated power as an output. Energy consumptions model has been designed and the total power consumed can be calculated during the designed experiment. CloudSim also provides developers with the capability for experimentation of dynamic scenarios, i.e., a different number of data centers or hosts can be created and deleted for testing unpredictable events in which users can join and leave the cloud application [27]. The DVFS scheme is limited to CPU optimization and adjusts the CPU power according to the workload that is being run on it. However, other components of the system, i.e., memory, storage, RAM, bandwidth and network interfaces, keep running on the same original frequency, and no scaling is applied to them. The use of Dynamic Power Management (DPM) can turn down the power consumption for all the components of the system. The CPU has number of states for frequency and voltage which suggests that it provide better power performance as compared to basic approach [28]. Thus, the powering up of a system will require a large amount of energy using DPM as compared to DVFS technique [18].

3.2 Static threshold VM consolidation technique

In Static Threshold technique, upper and lower limits are set for the workload, and virtual machine allocation and relocation is done based on the defined threshold. In this virtual machine are selected depending on factors, e.g., minimum migration time, maximum correlation and minimum utilization. The amount of power that is being used in the data center can be managed by exploiting the trade-offs between service quality and service level agreement.

In this type of technique upper and lower threshold limits are defined for the CPU. The host under or overloading state is determined, and virtualization is performed. When this procedure is called it works by determining the current CPU utilization and differentiates it with the defined threshold level. On the basis of this differentiation, hosts are selected for relocation. The algorithm calculates the mean of 'n' latest CPU utilization and compares it with the defined threshold value. As a result, host over or underloaded state is determined. The resource provisioning is achieved by virtual machine relocation. The VMs that are allocated to the hosts initially are under or over utilized. Therefore, relocation helps in resource provisioning and helps in reduction of bottlenecks. The host node that is present on the systems is not turned off or sent in sleep mode. It remains active and helps in reduction of downtime and provides a better quality of service and can help in energy reduction [29]. The virtual machines that are present on the system can be selected for relocation using three different approaches defined below;

3.2.1 Minimum migration time policy

The selection of the virtual present on the host is performed on the basis of its migration time. The virtual machine that requires minimum migration time is selected. The time is calculated on the basis of RAM and bandwidth using the following equation.

$$\left(\frac{RAM_u(v)}{NETj}\right) \leq \left(\frac{RAM_u(a)}{NETj}\right), \quad v \in Vj | \forall a \in Vj \tag{1}$$

V_j represents total number of virtual machines that are associated with host *'j'*. Whereas, $RAM_u(a)$ is the RAM used by virtual machine *(a)* and NET_j shows total available bandwidth of host *'j'*.

3.2.2 Minimum utilization

The virtual machines that are required to be relocated in under or over utilized hosts are selected on their utilization criteria. The virtual machine that are having minimum utilization are selected for migration from one host to another when required.

3.2.3 Maximum correlation policy

In this type, virtual machine is selected based on maximum correlation. Virtual machine having higher value of resource utilization has higher probability of host overloading. Multiple correlation coefficient (MCC) is used for estimation of CPU utilization and intra virtual machine correlation. MCC coefficient has a squared correlation for dependent variable of real and predicted values [30].

4 Simulation

For the implementation and evaluation of the proposed experiments, CloudSim simulation platform is used to provide users with the ability to perform the desired tests. The experiments are carried out by using traces from a game as workload for the dynamic voltage frequency, non-power aware and static consolidation technique. The designed simulation consists of heterogeneous data centers consisting of 800 physical hosts and 1000 virtual machines which are dynamically allocated by the broker. Half of the hosts are HP ProLiant ML110G4 (Xeon3040) and the other half are HP ProLiant ML110G5 (Xeon3075) servers. The system's frequency characteristics are defined based on how many instructions can be executed in one second (MIPs). Therefore, HP ProLiant ML110G4 (Xeon3040) and ML110G5 (Xeon3075) have MIPs rating of 1860 MHz and 2660 MHz, both being dual-core servers [31]. The defined system specifications are suited to the hardware requirements for the experimental workloads and are shown in Table 1.

In DVFS and NPA, no dynamic allocation of virtual machines is performed, and host power adjustment is done based on their CPU utilization. Whereas, when tested with static threshold concept the virtual machine selection and consolidation is performed based on MTT, MU and MC policy. A fixed MIPs value is provided having a value of 1000 MIP per second for a virtual machine. The simulated model has a bandwidth rate of 1 Gbits per second and RAM 32 GB for each system. A fixed defined gaming workload is provided in this experiment that consists of traces from a popular multiplayer online game, namely World of Warcraft having a dataset size of 3.5 GB.

The data set consists of traces from real data of the popular massively multiplayer online game, World of Warcraft (runtime of 1107 days, 91065 avatars, 667032 sessions, users located globally in 3 continents with different time zones) collected to analyze the quality of service parameters and consisting of game time, race attributes, current position, profession info, game position information, game level etc. [28]. It provides execution time of each host and energy is calculated based on power consumed by individual host. It uses time shared policy and rating of the processing elements is calculated by having millions of instructions per second. The total MIPs, i.e., total execution time is the sum of all the MIPs from each processing element (PE). Here, it is assumed that all the processing elements have same rating in the used machine. The service level agreements are also required as it is necessary to maintain the quality of service matrices [32]. The detailed parameters are summarized in Table 2.

The reasoning behind the service level agreement violations (SLAV) time per active hosts is based on the observation that if there is an application that is managing the virtual machine migrations and it is busy with a host that has 100% utilization, it will not be able to address other hosts waiting for service provisioning. Therefore, virtual machines are deprived of the desired performance level causing SLA violations [33]. The mathematical definitions and formula are as follow,

$$SLAV(H) = \frac{1}{H(n)} \sum_{i=1}^{n} \frac{SLAH(t)i}{AH(t)i} \tag{2}$$

SLAV(H) is the violation of per unit time for active hosts, *H(n)* is number of hosts, *SLAH(t)i* represents the time duration that leads to service level agreement violations by reaching CPU utilization of 100% and *AH(t)i* is the total number of *hosts(i)* in the active state [34].

$$P(vm) = \frac{1}{VM(n)} \sum_{i=1}^{n} \frac{Pd(k)}{Cpu(k)} \tag{3}$$

P(vm) is the effect on the performance because of virtual machines migration, *VM(n)* represents the total number of virtual machines, *Pd(k)* represents the level of degradation in the service of a particular virtual machine when it is migrated, *Cpu(k)* represents the total utilization of CPU of

Table 1 Details of the system parameters

System (HP ProLiant)	MIPs rating (MHz)	Cores	RAM	Hard disk (GB)
ML110G4 (Xeon3040)	1860	Dual	32 GB	1
ML110G5 (Xeon3075)	2660	Dual	32 GB	1

Table 2 Detailed description of system parameters

Host MIPs	Host RAM (MBs)	Host PE(s)
1860	32768	02
2660	32768	02

a virtual machine [35]. Therefore, whenever a cloud server is considered for service level agreement violations it always depends on the above two factors independently described in Eqs. (2) and (3).

The SLA level is the product of two matrices, i.e., how many SLAV there are per unit time of active hosts and how much of the performance degradation is because of virtual machine migration, Eq. (4). Therefore, SLA is because of two factors: one is virtual machine migration and the other is when a host is overloaded resulting in *SLAV* as follows [33],

$$SLA = SLAV(H) \times P(vm) \tag{4}$$

The overall performance of cloud servers can be analyzed by using the following equation,

$$Perf(DC) = Energy \times SLAV \tag{5}$$

The CPU time is calculated from the following formula,

$$CPU(t) = \frac{C(Le)}{Pe \times (1.0 - C(Lo))} \tag{6}$$

CPU(t) = CPU Time, *PE* = MIPs of one Processing Element, *C(Le)* = length of cloudlet, and *C(Lo)* = load of cloudlet. Here, MIPs represent how many instructions can be executed in one second, *PE(x)* the number of MIPs of one processing element, *PE(y)* represents MIPs of *N* number of hosts,

$$Total\,MIPs = PE(x) + PE(y)N(host) \tag{7}$$

Cost per million instructions related to a resource can be calculated using Eq. (9). In this, *Cost(s)* = cost per second and *PE(MIPs)* = calculating MIPs of one processing element.

$$MI = \frac{Cost(s)}{PE(MIPs)} \tag{8}$$

The required execution time can be calculated using the following equation. Whereas, *Sys(t)* is current time in millisecond, *Exe(t)* is system execution time and *1000* is the defined MIPs rating.

$$Time = \frac{Sys(t) - Exe(t)}{1000} \tag{9}$$

Thus, energy consumed by each host, performance measure, CPU utilization, total execution time, and SLA violations count can be calculated by using the above equations [33]. Experimentation results are shown in Section V.

5 Performance analysis and discussion

This test calculates the energy performance across the data center in the given simulation environment. All the tests are carried out in the simulation environment, i.e., the CloudSim package which is configured using Eclipse Luna and Java IDE. DVFS, NPA and STVM techniques have been applied to analyze the gaming workload of the World of Warcraft multiplayer online game. The workload consists of data traces from servers which are collected over time of 1107 days. The above consolidation techniques are implemented for load management. Under or overloaded virtual machines are selected for relocation based on minimum migration time, maximum correlation and minimum utilization. A typical game workload has been provided for testing the behavior of the proposed techniques. The DVFS, NPA and STVM simulation models with the same specifications are used for power and service level agreement analyzation of same gaming workload. The main difference between the NPA and DVFS models lies in how resources are allocated to the hosts. All the parameters (RAM, bandwidth, storage, I/O file size etc.) are defined however, for DVFS, resources are allocated based on dynamic voltages and frequency fluctuations of the central processing unit for the active hosts.

In the NPA model hosts consume the maximum amount of power, thus increasing the cost of services and causing loss of profit for service providers. Figure 5 shows power consumption in the cloud environment with a fixed number of hosts and MIPs using DVFS and NPA. For DVFS, the data show a linear trend for CPU power consumption as compared to NPA technique. The results are by way of a reality check and verify the theoretical concept that in DVFS, the CPU adjusts frequency according to the workload to minimize the power consumption and thus provides a linear trend. The hosts using DVFS technique for the same gaming data consume less energy as compared to the NPA technique. In NPA technique hosts are loaded to maximum values and consume more energy resulting in greater values of CO_2 emissions.

Figure 6, shows different execution time by virtual machines using all three different techniques with the same workload and experimentation setup. It could be seen from the results that selection of the virtual in under or overloaded host takes minimum mean time. Whereas, migration of virtual machine from one host to another requires more time. Therefore, downtime in the network can be reduced if appropriate virtual machine relocation technique is selected.

The difference in the amount of energy consumption, service level agreement and quality of service degradation can be seen through the results which are estimated based

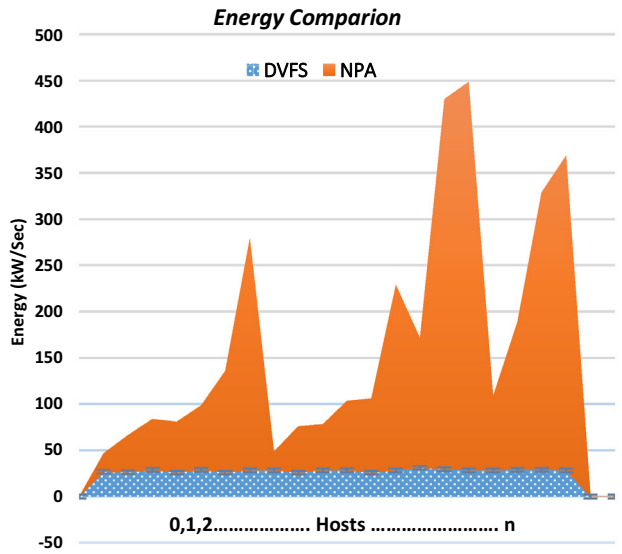

Fig. 5 Consumption in a data center

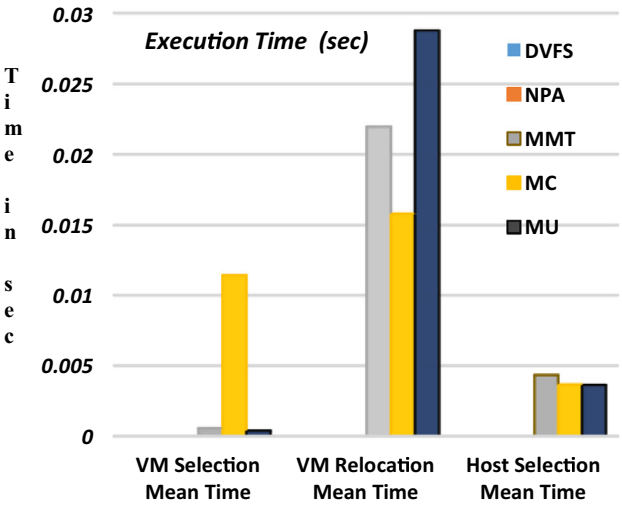

Fig. 6 VM execution time for each host

on CPU utilisation, static threshold and non-power aware technique (Fig. 7).

Comparison of three different approaches is carried out based on a service level agreement. Results show that minimum service level agreement degradation (SLAV) is achieved by using DVFS technique. Therefore, by using DVFS technique overall SLA violation can be reduced. The reduction in SLA performance degradation suggests that quality of service and quality of experience can be enhanced by using DVFS technique (Fig. 8).

In STVM, virtual machines that have minimum utilisation have higher rate of selection as compared to maximum correlation or minimum migration time. Therefore, this shows that more energy is saved in threshold techniques

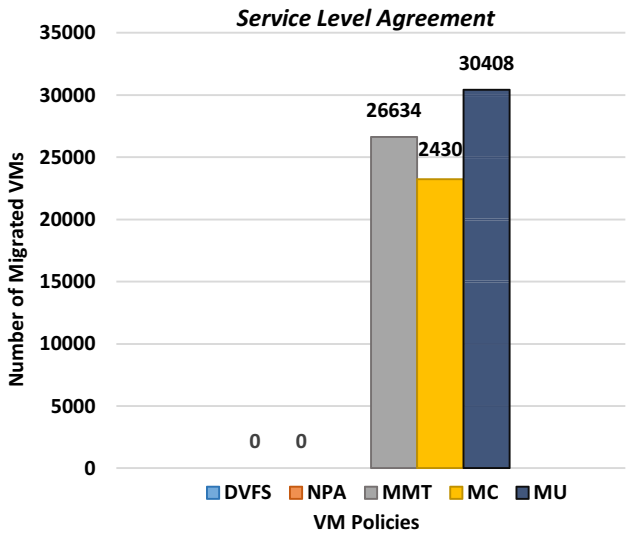

Fig. 7 Number of VM migration

when virtual machines are selected on the base of utilisation in underutilized hosts, as shown in Fig. 9.

The results also prove that quality of service is directly proportional to service level agreements, i.e., if QoS is not observed for a certain amount of time then we have SLA violation. Thus, by using DVFS, performance can be improved, and energy consumption can be minimized resulting in a lot of cost saving for Big Data from commercial point of view (Fig. 9). Whereas, STVM behaves better when workload is not of big size and is not changing dynamically.

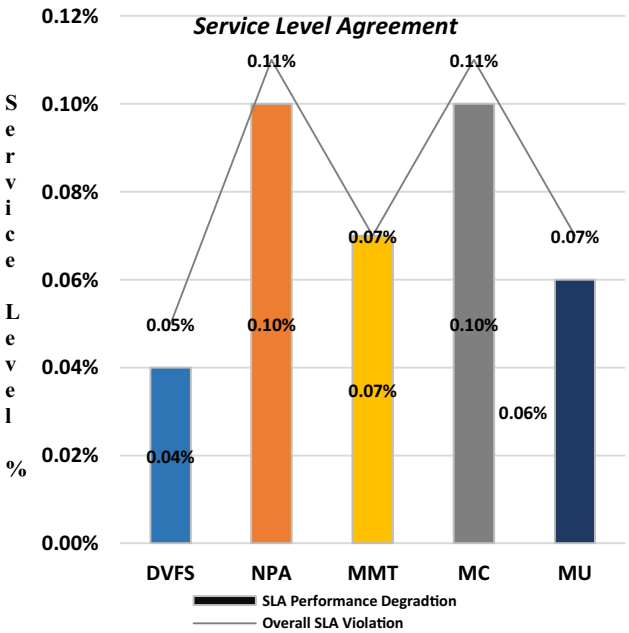

Fig. 8 Service level agreement violation (SLAV)

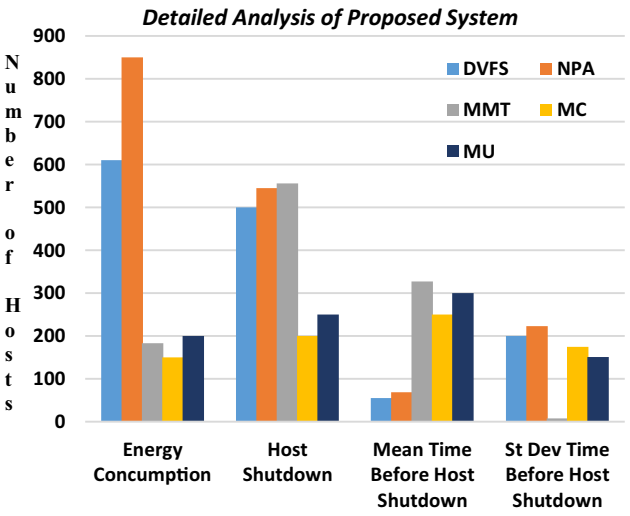

Fig. 9 Detailed analysis of proposed system

From the results, if the DVFS technique is used, the best results for energy utilisation are achieved and 14% of energy could be saved in comparison to the NPA technique using the same gaming workload (Fig. 10).

Whereas, the static threshold gives minimum energy consumption when used with maximum correlation policy. The reason static threshold performs minimum energy utilization is that the upper and lower threshold limits are defined in the system. Whereas, this approach will not be suitable with the dynamic workload environment. DVFS provides better trade-off for exploitation of SLAs per host for maintenance of quality of service and quality of experience. During the whole experiment DVFS uses fewer resources in the host when analyzed. Less energy

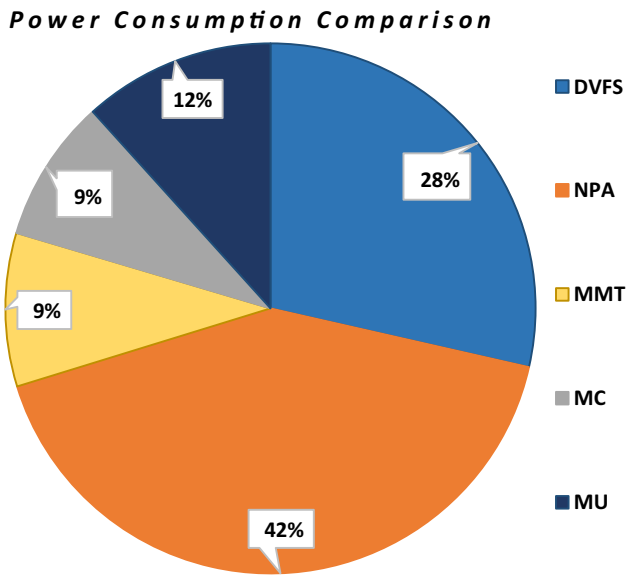

Fig. 10 Analysis of energy consumption the proposed system

consumption mean time and number of host shutdown are performed during the experimentation. These results show that overall the best quality of service can be achieved by implementing DVFS in gaming servers placed globally.

6 Conclusion and future work

The simulation tests that have been designed using CloudSim platform and are based on three different consumption approaches, i.e., dynamic voltage and frequency scaling, non-power aware and static threshold virtual machine consolidation technique. The same workload (game data) and data center specifications are set for testing which technique performs better for power saving and meet service level agreements. The workload provided demonstrates that dynamic voltage frequency scaling saves more energy as compared to general non-power aware or static virtual machine consolidation approach for dynamic workloads. It has less SLA violations which is important for maintaining QoS and QoE. Static virtual machine consolidation technique has a better ratio of service level agreement violation when used with maximum correlation virtual machine policy for workloads allocated statically.

CloudSim provides the ability to test the same workload scenario on two different approaches, i.e., static and dynamic. When compared to dynamic voltage and frequency technique, static virtual machine consolidation provides better results for small workloads under static allocation. In real-world for large cloud gaming servers, it is difficult to maintain upper and lower workload limits. Therefore, the effectiveness of this approach becomes impractical in dynamic environments. By using this simulation environment, a researcher can experiment and determine the amount of resources required, (e.g., the number of cloudlets, bandwidth, RAM, cost etc.) for maintaining the quality of service. Therefore, from the simulation results, it can be verified that cloud gaming data centers with the proposed DVFS technique can yield less energy consumption and providing more economical solutions while fulfilling service level agreements for maintaining a good quality of service leading to better quality of experience (QoE) for users placed globally.

In the future, this work will be enhanced, and better ways and techniques to save energy will be explored for Big Data, Internet of Things and Gaming data centers. Other extensions include that an analysis between number of users and submitted jobs can be carried out. This can help in energy improvement and optimization by carrying out failure-analysis in cloud environment. Along with this, an effort will be carried out to merge this workload in current CloudSim framework and make it public for research societies around the world.

Acknowledgements The authors would like to acknowledge partial support from the BT-Ireland Innovation Centre (BTIIC) and Ulster University.

References

1. Ahmad, B., et al.: Analysis of energy saving technique in CloudSim using gaming workload, In: Proceedings of the Ninth International Conference on Cloud Computing, GRIDS, and Virtualization, IARIA. (2018)
2. Sidana, S., et al.: NBST algorithm: A load balancing algorithm in cloud computing. In: Proceedings of 2016 International Conference on Computing, Communication and Automation (ICCCA). (2016)
3. Luo, H., et al.: The dynamic migration model for cloud service resource balancing energy consumption and QoS. In Proceedings of the 27th Chinese Control and Decision Conference (2015 CCDC). (2015)
4. Arroba, P., et al.: DVFS-aware consolidation for energy-efficient clouds. In: Proceedings of 2015 International Conference on Parallel Architecture and Compilation (PACT). (2015)
5. Tian, W., et al.: Open-source simulators for Cloud computing: Comparative study and challenging issues. Simul. Model. Pract. Theory **58**, 239–254 (2015)
6. Prazeres, C., Serrano, M.: SOFT-IoT: self-organizing FOG of things. In: Proceedings of 2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA). (2016)
7. Kliazovich, D., Bouvry, P., Khan, S.U.:. DENS: Data center energy-efficient network-aware scheduling. In: Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom). (2010)
8. Burge, J., Ranganathan, P., Wiener, J.L.: Cost-aware scheduling for heterogeneous enterprise machines (CASH'EM). In: Proceedings of 2007 IEEE International Conference on Cluster Computing. (2007)
9. Cao, F., Zhu, M.M., Wu, C.Q.: Energy-efficient resource management for scientific workflows in clouds. In Proceedings of 2014 IEEE World Congress on Services. (2014)
10. Buyya, R., Yeo, C.S., Venugopal, S.: Market-oriented cloud computing: vision, hype, and reality for delivering IT services as computing utilities. In: Proceedings of 2008 10th IEEE International Conference on High Performance Computing and Communications. (2008)
11. Beloglazov, A., Buyya, R.: Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers. Concurr. Comput. **24**(13), 1397–1420 (2012)
12. Rawat, P.S., et al.: Power consumption analysis across heterogeneous data center using CloudSim. In: Proceedings of 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). (2016)
13. Zhao, Z., Hwang, K., Villeta, J.: GamePipe: a virtualized cloud platform design and performance evaluation, pp. 1–8. (2012)
14. Shea, R., et al.: Cloud gaming: architecture and performance. IEEE Netw **27**(4), 16–21 (2013)
15. Varasteh, A., Goudarzi, M.: Server consolidation techniques in virtualized data centers: a survey. IEEE Syst. J. **11**(99), 1–12 (2015)
16. Yannuzzi, M., et al.: A new era for cities with fog computing. IEEE Internet Comput. **21**(2), 54–67 (2017)
17. Oikonomou, E., Panagiotou, D., Rouskas, A.: Energy-aware management of virtual machines in cloud data centers. In: Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS), pp. 1–6. ACM: Rhodes, Island, Greece (2015)
18. Calheiros, R.N., et al.: CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Softw. Pract. Exp. **41**(1), 23–50 (2011)
19. Tso, F.P., et al.: The Glasgow raspberry pi cloud: a scale model for cloud computing infrastructures. In: Proceedings of 2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops. (2013)
20. Wadhwa, B,. Verma, A.: Energy saving approaches for green cloud computing: a review, In: Proceedings of 2014 Recent Advances in Engineering and Computational Sciences (RAECS), pp. 1–6. (2014)
21. Keller, G., et al. DCSim: A data centre simulation tool. In Proceedings of 2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013). (2013)
22. Varasteh, A., Goudarzi, M.: Server consolidation techniques in virtualized data centers: a survey. IEEE Syst. J. **11**(2), 772–783 (2017)
23. Horvath, T., et al.: Dynamic voltage scaling in multitier web servers with end-to-end delay control. IEEE Trans. Comput. **56**(4), 444–458 (2007)
24. Atiewi, S., Yussof, S.: Comparison between cloud sim and green cloud in measuring energy consumption in a cloud environment. In: Proceedings of 2014 3rd International Conference on Advanced Computer Science Applications and Technologies. (2014)
25. Garg, S.K., Buyya, R.: NetworkCloudSim: modelling parallel applications in cloud simulations. In: Proceedings of 2011 Fourth IEEE International Conference on Utility and Cloud Computing. (2011)
26. Adhikary, T., et al.: Energy-efficient scheduling algorithms for data center resources in cloud computing. In: Proceedings of 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), pp. 1715–1720. (2013)
27. Long, S., Zhao, Y.: A toolkit for modeling and simulating cloud data storage: an extension to CloudSim. In: Proceedings of 2012 International Conference on Control Engineering and Communication Technology. (2012)
28. Lee, Y.-T., Chen, K.-T., Cheng, Y.-M., Lei, C.-L.: World of warcraft avatar history dataset. In: Proceedings of ACM Multimedia Systems 2011. http://mmnet.iis.sinica.edu.tw/dl/wowah/ (2011)
29. Shuja, J., et al.: Survey of techniques and architectures for designing energy-efficient data centers. IEEE Syst. J. **10**(2), 507–519 (2016)
30. Theja Perla, R., Babu, S.K.K.: Evolutionary computing based on QoS oriented energy efficient VM consolidation scheme for large scale cloud data centers. Cybern. Inf. Technol. **16**, 97 (2016)
31. Song, J., et al.: FCM: towards fine-grained GPU power management for closed source mobile games. In: Proceedings of International Great Lakes Symposium on VLSI (GLSVLSI), pp. 353–356. (2016)
32. Ahmed, A., Sabyasachi, A.S.: Cloud computing simulators: a detailed survey and future direction. In: Proceedings of 2014 IEEE International Advance Computing Conference (IACC). (2014)
33. Wang, J.V., et al.: A Stable Matching-based virtual machine allocation mechanism for cloud data centers. In: Proceedings of 2016 IEEE World Congress on Services (SERVICES). (2016)

34. Arroba, P., et al.: Dynamic voltage and frequency scaling-aware dynamic consolidation of virtual machines for energy efficient cloud data centers. Concurr. Comput. **29**(10), e4067 (2017)
35. Wickremasinghe, B., Calheiros, R.N., Buyya, R.: CloudAnalyst: a CloudSim-based visual modeller for analysing cloud computing environments and applications. In: Proceedings of 2010 24th IEEE International Conference on Advanced Information Networking and Applications. (2010)

Bilal Ahmad is a Ph.D. Researcher at Ulster University, Northern Ireland, UK. He is working in the field of Cloud Computing. His research interests include resource optimisation, quality of service, quality of experience, energy optimisation, multi-objective optimization, and game theory. Bilal, has been working in the industry after completion of his M.Sc. in Electrical and Electronic Engineering from University of Greenwich, UK since 2011. Mainly, the industry included Mobile Communication, Electronic Manufacturing and Embedded Electronics. Bilal, holds Associate Fellowship for Higher Education Academy (UKPSF) and has served as a reviewer and guest editor for several international journals and conferences.

Zaib Maroof is working as a Lecturer in Foundation University, Islamabad, Pakistan. She holds a Doctorate Degree in the field of Leadership and Management with Expertise in Finance and Economics. She has number of publications in the field of Economics and Finance and holds 03 Gold Medals in her educational career.

Sally Mcclean is working as Professor at Ulster University, Northern Ireland, UK. She is a Mathematician and works in the field of Cloud Computing. Her research interests include resource optimisation, quality of service, quality of experience, energy optimisation, multi-objective optimization, and game theory. She has been working in the academics for past 30 years. Currently, she is part of BTIIC (British Telecom Ireland Innovation Centre). Sally, has served as a reviewer and guest editor for several international journals and conferences.

Darryl Charles joined Ulster University in 2001 and is a member of the Computer Science Research Institute. Darryl is a Fellow of the Higher Education Academy, and teaches at both undergraduate and postgraduate level. His research background is in computational intelligence. He has applied his expertise in this area to serious games and apps as well as a number of research contexts such as intelligent interactive storytelling, machine learning in games, user modelling, cloud computing, and connected health. Recent research has been mainly focused on health contexts and creating software with natural user interfaces to help motivate people to engage with learning, exercise and rehabilitation. Since 1998, he has had over 80 papers published in peer reviewed journals and conferences as well as an authored book entitled 'Biologically Inspired Artificial Intelligence for Computer Games'. Darryl graduated from Queens University Belfast with a degree in Electrical and Electronic Engineering (Hons.) in 1988. After qualifying as a teacher at Stranmillis College, Belfast, he taught Technology and Design to A-Level at Portadown College until 1995 and during this spell completed an MSc in Microelectronics and Microcomputer Applications at the University of Ulster. After a spell as Head of IT at Cox Green School in Maidenhead he returned to full-time education to study for a Ph.D. on Unsupervised Artificial Neural Networks at the University of Paisley. While completing his PhD he was appointed as a lecturer and subsequently as a senior lecturer in Computing at Paisley before returning to Northern Ireland in 2001.

Gerard Parr is a Member of Advisory Committee for Northern Ireland at Ofcom since March 2004. He holds the full Chair in Telecommunications Engineering at the University of Ulster, Coleraine, and is Co-ordinator of the Internet Technologies Research Group in the Faculty of Engineering. He is also Head of Computing Department at University of East Anglia, Uk. Currently, he is part of BTIIC (British Telecom Ireland Innovation Centre).