Economic Impact of Resource Optimisation in Cloud Environment Using Different Virtual Machine Allocation Policies

Bilal Ahmad[1](✉), Zaib Maroof[2], Sally McClean[1], Darryl Charles[1], and Gerard Parr[3]

[1] School of Computing, Ulster University, Coleraine, UK
ahmad-b@ulster.ac.uk
[2] National Defence University, Islamabad, Pakistan
[3] University of East Anglia, Norwich, UK

Abstract. Exceptional level of research work has been carried in the field of cloud and distributed systems for understanding their performance and reliability. Simulators are becoming popular for designing and testing different types of quality of service (QoS) matrices e.g. energy, virtualisation, and networking. A large amount of resource is wasted when servers are sitting idle which puts a negative impact on the financial aspects of companies. A popular approach used to overcome this problem is turning them ON/OFF. However, it takes time when they are turned ON affecting different matrices of QoS like energy consumption, latency, consumption and cost. In this paper, we present different energy models and their comparison with each other based on workloads for efficient server management. We introduce a different type of energy saving techniques (DVFs, IQRMC) which help toward an improvement in service. Different energy models are used with the same configuration and possible solutions are proposed for big data centres that are placed globally by large companies like Amazon, Giaki, Onlive, and Google.

Keywords: Cloud computing · Energy optimisation · Resource optimisation · Economic impact · Service quality · Green computing · Virtualisation

1 Introduction

Cloud Computing is growing day by day with the development of IT services. The reason for this development is improving cost effectiveness and quality of experience from a user's perspective. The IT industry is becoming adaptable to cloud computing technologies for the achievement of improved service intelligence and good user experience. The cloud generally has three types of services SaaS (software as a service), PaaS (platform as a service) and IaaS (infrastructure as a service). Along through provisioning of enhanced quality of service cloud providers can move towards more

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

M. H. Miraz et al. (Eds.): iCETiC 2019, LNICST 285, pp. 46–58, 2019.
https://doi.org/10.1007/978-3-030-23943-5_4

profits by saving resources e.g. energy, bandwidth consumption, cost effectiveness etc. In a cloud environment, servers have a significant part in the design of cloud infrastructure in addition of resource allocation. The quality of cloud service primarily depends upon economic resource allocation and scheduling which servers perform during their operation. If servers have advanced resource allocation and scheduling algorithms services could be improved automatically [1].

All types of resource allocation and scheduling is related to a server's physical design and resource allocation policies. A system designer's major task is to determine trade-offs between quality of service factor and energy consumption [6]. Idle servers can be turned off for power saving purpose and expense to profit ratio can be improved. However, 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 suggestions and ideas have been proposed for energy consumption for jobs arriving in cloud servers. The quality of cloud service depends upon how much stable resource allocation is provided to the user requesting the service. For the achievement of this goal virtualisation is carried out by the service providers. Large scale data centres consist of thousands of hosts and nodes resulting in the consumption of a large amount of energy. As a result, cloud servers are being designed in such a way that they become automatically adaptable to the service requested by the users [2]. Dynamically scaling up or down is carried out by the servers and virtual machines are created and destroyed depending upon the load servers are receiving from users across the globe. This dynamic approach helps to maintain a quality of service while managing the resources efficiently from both perspectives i.e. user and service providers. Techniques require to be developed and deployed on the cloud servers that support elastic management of tasks that are being run on the server with different workloads (gaming, big data and internet of things, web hosting, social networking, etc.). These applications are quite challenging depending upon the user's location, service requests, time, weather and interaction patterns. Hence, to attain a good quality of service and experience dynamic, provisioning techniques are required to be designed and implemented by researchers that are compatible across the globe. However, so far resource allocation is still a challenge in terms of video streaming, gaming in which data is streamed online globally [3].

Recent advances that are being made around the world have turned the idea of cloud gaming into reality. The use of elastic resource utilisation and globally placed servers has made it possible for users to enjoy service on a pay as you go basis. Issues related to a bulk amount of data streaming to cloud servers are being addressed resulting in improvement of user experience. User satisfaction has been mainly improved because of dedicated servers that are placed globally for solving latencies and data offloading issues. High definition 3D issues related to gaming have been addressed over the cloud environment which makes it a reality for gamers to enjoy single and multi-player games over the cloud environment. The basic architecture design of cloud gaming consists of a game that is hosted on the cloud server which is located globally. The player whether in single or multi player mode streams the game scenes in the form of video by using the internet as a communication media. The player sends the commands over the cloud environment and these commands are processed by the graphical processing unit and are sent back to the user through a thin client. All these actions are required to be executed in an order of milliseconds therefore, a service provider has a small margin of error [4].

Several factors are required to be managed for hosting of gaming application over the cloud environment. In a virtual cloud environment factors like quality of service, energy consumption and cost need to be managed. An efficient energy solution is required for power saving in big data centres. Certain types of techniques e.g. dynamic voltage frequency scaling, virtual machine migration and load balancing are required to be designed and implemented by researchers. By implementing these techniques not only improved value of facility be provided, it can also decrease of carbon dioxide emission from the servers [5]. The rest of the paper is organised as follows: Sect. 2 (Cloud Computing Background and Platforms), Sect. 3 (Related Work), Sect. 4 (Experimentation), Sect. 5 (Experimental Framework), and Sect. 6 (Results and Discussion) with a Conclusion at the end.

2 Cloud Computing Background and Platforms

There are number of advantages of using services over the cloud environment from which a normal user can benefit e.g. facilities and designer tools. People carrying out the research work can benefit from these tools without any device limitations. These devices could be in any form from a small tablet to large computer servers that could be placed in a commercial environment for development purpose. The emerging field of Cloud Computing where servers are placed globally provides its user with immense advantages as compared to old technologies. This include concept of working anywhere any time without limitation of devices, storage, cost and virtualization concept.

In the era of development and progress in this century still many people around the globe are unable to enjoy these services for number of reasons e.g., (a) Restricted capability of electronic devices (performance, swiftness, visuals) (b) System limitations (bandwidth, topographical location) (c) Delay of service from core computers (weak networks) [6]. Consequently, to accommodate these problems simulation platforms (Cloud Sim, iFog Sim, Green Cloud, iCaroCloud, Cisco, Cloud Analyst, Network Cloud Sim, iCanCloud etc.) have been developed which provide users with a means to overcome the problems they are facing. The cloud architecture is in the form of layers. Each layer has a defined functionality and is interconnected as shown in Fig. 1. The service provision of IaaS and PaaS is performed using middleware whereas, SaaS services are provided by lower layer services using physical resources in a cloud environment. Third-Party service providers develop and provide SaaS/PaaS services in the cloud environment. Different cloud layers and their functionality [7] are discussed below (Fig. 1).

CloudSim is a rich platform that provides it users with the ability to simulate and calculate energy consumption of large servers by using the Dynamic Voltage and Frequency Scaling (DVFS) technique. The input parameters are used as host for the cloud environment and energy calculations are provided as an output results. In this way, a researcher that is carrying an investigation towards green computing can measure amount of energy which will be required for calculation of designed tests. This platform is very flexible and provides its user with the leverage of dynamic experimentation [2].

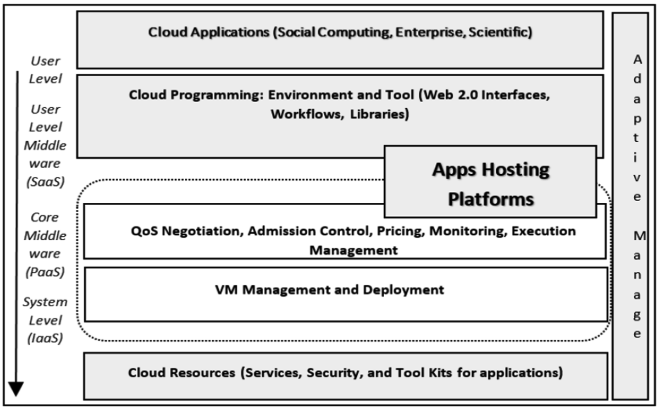
User Level
User Level Middle ware (SaaS)
Core Middle ware (PaaS)
System Level (IaaS)
Cloud Applications (Social Computing, Enterprise, Scientific)
Cloud Programming: Environment and Tool (Web 2.0 Interfaces, Workflows, Libraries)
Apps Hosting Platforms
QoS Negotiation, Admission Control, Pricing, Monitoring, Execution Management
VM Management and Deployment
Cloud Resources (Services, Security, and Tool Kits for applications)
Adaptive Manage

Fig. 1. Cloud computing layered architecture.

3 Related Work

The IT industry is evolving day by day from the domain of grid, parallel and distributed computing. With the ongoing development of the industry different simulation software tools have been developed for cloud-based environments 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 the quality of service and quality of experience [3]. Present research shows that distributed computing is more advantageous towards economical resource provisioning. It can provide centralised access, tolerance and coordination capability. It also allows cloud developers to manage and provision resources as desired. Quality of service is maintained by the broker which is present between user and server and is located globally. In this model, resources are provisioned to the users depending upon their requirements. e.g. Amazon EC2 server [7]. Ahmad, B., et al., uses the concept of static load and allocates resources based on under or overloaded hosts. Energy consumption comparison is performed by using different virtual machine allocation policies based on host workload. The results are analysed and compared with the DVFS technique. Dynamic allocation of resources helps in saving more resources as compared to static allocation. The experimentation is performed in the CloudSim environment and configured using Eclipse Luna and Java IDE [13]. The authors focus on the basic of resource optimisation and proposes an algorithm that manages resources for efficient performance. Resource optimisation requires service providers to manage resources in such a way that no user should be denied resources when requested. The work focuses on one of the non-preemptive resource allocation technique i.e. a game-based approach. Different trade-offs are also required to be carried that may hamper the quality of service at the user end. A game based resource allocation model has been suggested but lacks the experimentation results whether it is feasible for cloud gaming or not [8].

Saving of energy in the cloud environment has been one of the critical factors that have been addressed by researchers across the globe. Several algorithms and techniques have been proposed for this purpose. Recently, a rack architecture design has been suggested by Hamilton which uses AMD Athlon processors consisting of low power devices. Along with this, the Hadoop platform has also been used by Atiewi, S. and S. Yussof. for testing of energy saving techniques [10]. In [11] two of the major platforms are discussed, which are available for experimentation involving energy relating scenarios by researchers across the world. Green Cloud is limited to energy experimentation whereas CloudSim is also capable of dealing with other factors of quality of service. Green Cloud does provide an attractive user interface as compared to CloudSim [11]. Other researchers also focus on saving of energy and this involves the use of virtual machines. In this work, Nguyen, B.M., suggest that data images should be maintained on servers and when a user requests a certain amount of data it can be provided. This can save an amount of energy which will be wasted in switching the servers ON and OFF again and again [10]. The authors focus on energy consumption issues in a cloud data centre that are located globally and are providing services to users 24/7. The suggestion is limited to small data workload applications based upon hardware requirements (server, memory and network). A further study emphases that many techniques that are implemented in cloud servers work towards reliable availability of services when requested by users [4].

Ahmad et al., uses the concept of changing energy and time scaling technique. Based on this energy consumption is analyzed for observation of quality of service. A gaming workload is used, and optimum energy solution is suggested. Results show that DVFS behaves better as compared to non-power aware techniques for the same test scenario [9]. Many techniques are being used in the mobile gaming world for reduction of power. Song J. et al., propose that if a platform embedded with a GPU (graphical processing unit) is used with dynamic voltage frequency scaling algorithms, energy consumption can be reduced. It defines the frame complexity model that recognises the importance of GPU and its working efficiency towards the energy cost saving problem. However, the approach lacks the practical capability for testing the ideas on real gaming servers like Onlive and Gaikai [12]. In an approach based on the switching of machine states is used i.e. active and inactive states. This algorithm works based on the calculations that are performed about waiting time of jobs which are present on the servers and uses queue theory for energy resource optimisation. This work suggests working based on two states of servers i.e. ON/OFF. The algorithm is efficient for small jobs but can be overloaded when jobs are long. Back draw is that it results in high energy consumption and latency. Servers need a lot of power to come active gain when requested by users [10].

4 Experimentation

The experimentation has been carried in CloudSim and measures one of the service parameters i.e. energy for data centers. The economic impact of energy consumption has been analysed and tested by using different techniques, e.g. interquartile range, changing power and time scrambling and non-power aware for the same workload. The tests scenarios that are designed will be implemented and tested using the CloudSim platform. The platform is combination of software and platforms and is built using Java IDE along with Eclipse Luna. In this simulation platform number of methods have been implemented for calculation of power consumption. The workload that has been used for testing purpose consist of popular game World of War Craft. The workload consists of data traces from servers which are located over three continents and has runtime of 1107 days, 660723 sessions, 91065 avatars. The approach compares the used power levels, service provisioning and excellence of facility by using a gaming workload for testing the behaviour of the proposed technique [9].

4.1 Dynamic Voltage Frequency Scaling

DVFS one of the methods that helps big servers to save energy. It uses the frequency scaling technique. In DVFS the CPU power consumption is directly proportional to the workload which is provided to it. Therefore, if the CPU has more load, it will consume more power and vice versa. This allows it to consume less power in the idle state. This does not affect additional features of the server (I/O devices, random access memory, bandwidth etc.) as they are dependent upon CPU frequency. DVFS provides the user with four states that are available for the user. These states allow the device to select its operation depending upon the workload e.g. G0 (power ON), G1 (partial sleeping), G2 (partially OFF but still being powered by the power supply), G3 (power off state). These states provision users to have their algorithms designed according to the workload [9].

4.2 Inter Quartile Range

Inter Quartile Range is a statistical dispersion metric that calculates the different the third *Q3*. It dynamically calculates the threshold level of CPU utilizations by the following equation:

$$IQR = Q3 - Q1 \tag{1}$$

$$T(n) = 1 - sf \times IQR \tag{2}$$

Whereas, '*sf*' is the protection factor. It defines the maximum safe limits for the user in a cloud environment. Its minor values signify maximum acceptance level of fluctuations in central processing unit [14].

4.2.1 Maximum Correlation VM Selection Policy

The concept behind maximum correlation relates to how resources are being used on the servers. If an application running on servers have a higher correlation with resource usage then the chance of servers overloading will be increased. Therefore, VM performance is analysed and VMs having high correlation with CPU utilization are migrated to other VMs in the system. Thus, multiple correlation is used for this purpose which evaluates the quality of independent constants [14].

4.2.2 Minimum Migration Time Policy

After overloaded system detection, virtual machines migration is carried out for resource optimization (power usage) and to refrain from the low service quality standards. In this algorithm only those VMs are migrated that need minimum time for migration from the system. This is done on the basis of bandwidth consumption of every virtual machine that is allocated for each individual user [15]. Therefore, VMs that are required to be moved across the network can be calculated by the following equation,

$$\left(\frac{RAM_u(v)}{NETj}\right) \leq \left(\frac{RAM_u(a)}{NETj}\right), \quad v \in Vj|\forall a \in Vj, \tag{3}$$

Whereas, Vj shows total number of VMs with host j, $RAM_u(a)$ is the amount of RAM that is used by the virtual machine (a), $NETj$ equals available bandwidth from host 'j' [13].

4.2.3 Minimum Utilisation Selection Policy

Due to the resource consumption, a virtual machine can be migrated in overloaded hosts. This technique will migrate VMs from underutilized or over loaded hosts to reduce the overhead caused in CPU utilization. This migration of virtual machines allows the systems to meet SLA violation and helps in saving of unwanted energy consumption [13].

5 Experimental Framework

There are two main techniques that have been used in this experimentational framework one of which is called nonpower aware and second one is called dynamic voltage and frequency scaling technique. These two methods differ from each other based on resource allocation phenomena they use. Dynamic voltage and frequency scaling technique adjusts the amount of power required for each host based on the level of the workload which is present. In the power adjustment and optimisation other elements of the system including random access memory, storage, bandwidth allocation remains same throughout the testing [9]. If changing power level are used within the central processing unit energy usage for all parts of the system can be reduced. Central processing unit within any computer system has limitations therefore its states are limited for current and frequency. Therefore, results that are achieved after experimentation are better than other simpler methods used in general. As a result when system will be implanted using dynamic power management lot of energy will be wasted as compared to dynamic frequency scaling approach [13]. In this experimental setup, broker plays a key role which is responsible for resource allocation and virtualisation. This consist of 800 hosts that are physically present on the system along with 1000 virtual machines. These virtual machines are assigned to the hosts dynamically depending upon the load present. The system constitutes of HP ProLiant model having two Xeon 3040 and Xeon 3075 standards.

Both these systems are dual core and have a processing speed of 1860 MHz and 2660 MHz [13]. Detailed parameters are given in Table 1. All the users that are present in this system will be adjusted based upon usage of central processing unit. Random access memory used is four giga bytes with a bandwidth rate of one giga bits per second for each system.

Table 1. System specification.

System (HP ProLiant)	Host MIPS	Host RAM	Host Bw	Host PE (s)	Hard Disk
ML110G4 (Xeon3040)	1860	4096 MBs	1 Gbit/s	02	1 GB
ML110G5 (Xeon3075)	2660	4096 MBs	1 Gbit/s	02	1 GB

The system is tested using a workload that consists of modern multiplayer game called world of war craft. This dataset has been collected over 1107 days and consist of different features and values. The data set specification consists of traces of avatars, sessions and data size. The data provides information of game location, time when it was played and for how long it was played, game positions information, level of graphics it used etc. for analyzing quality of services [16]. The time which is required by each to execute is provided based on energy consumption levels. The time-shared policy is being used to fulfill this scenario. In this experimentation, all the instructions are executed at one speed and total speed is sum of all the instructions that are executed. Therefore, to have good user experience better quality of service is required to provided [15].

All the resource allocation and virtualization is done on priority basis i.e. a virtual machine which requires immediate service it will be treated first and all resources will be allocated to it. If this condition is ignored quality of service and quality of experience is hampered and this leads to bad user experience. It provides service provider with higher level of violations and drops the standard of service. Therefore, a good mechanism is required that provides switching of resource when required urgently by the hosts with immediate service requirement. performance level causing SLA violations. Therefore, service quality can be met if two matrices are kept in mind i.e. number of violations and each hosts performance per unit time [16].

6 Results and Discussion

The system has been evaluated based on the performance of the designed model which is verified with diverse virtual machine placement and selection policies. The performance of the systems has been analysed based on the comparison with different energy saving techniques e.g. DVFS, Non-Power Aware, Inter Quartile Range (IQR) technique. Further, interquartile range uses three different virtualisation policies i.e. maximum correlation selection policy (MC), minimum migration time selection policy (MMT). IQR performs CPU utilization based upon these algorithms and does virtual machine allocation deallocation based upon the available workload. The virtual machine allocation has been performed using MC and MMT selection policy. Based upon these allocations and selection policies, simulation has been performed in CloudSim for the designed data centre. This facilitates analysis and assessment of different parameters such as energy efficiency, service provision level and its violations, how many violations occur every second and time it takes to shut down the server after processing.

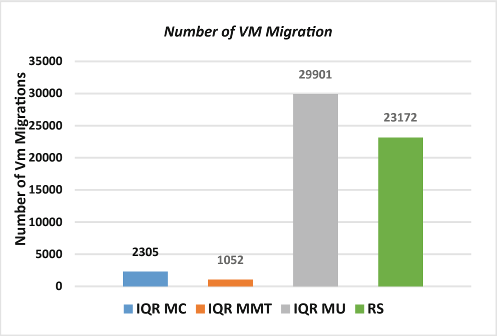

Fig. 2. VM migration.

System performance has been observed by using different virtual machine allocation and placement algorithm. From the results (Fig. 2) it becomes obvious that IQR MMT performs minimal virtual machine migration which leads to minimum downtime probability. However, when it comes to SLA violations MMT performs minimum violations leading to the best quality of service (Fig. 3). System performance has been observed by using different virtual machine allocation and placement algorithm. From the results (Fig. 2) it becomes obvious that IQR MMT performs minimal virtual machine migration which leads to minimum downtime probability. However, when it comes to SLA violations MMT performs minimum violations leading to the best quality of service (Fig. 3).

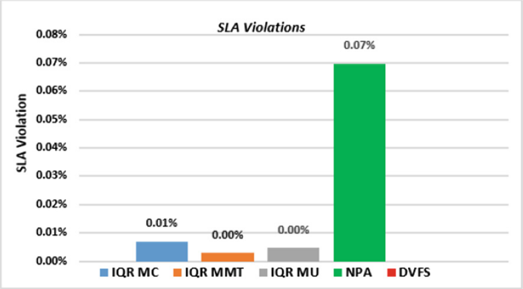

Fig. 3. Number of SLA violations

In (Fig. 4), performance degradation of virtual machines can be analysed, and it becomes clear from the results that by using MU approach a minimum number of degradations is performed at the virtual machine level. On the other hand, maximum correlation selection policy has the highest value of service level agreement violations. Minimum Utilisation involves a smaller number of virtual machines leading to less SLA violations for the system.

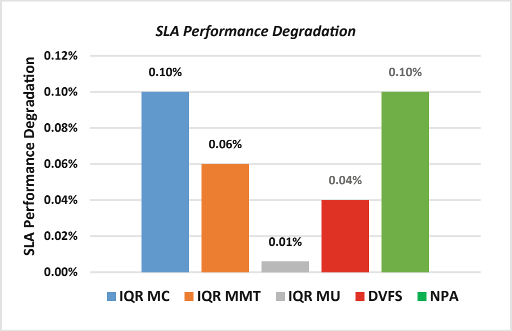

Fig. 4. SLA performance degradation.

MMT has better service quality and results and less violations are performed for every second of execution (Fig. 5). After performing the desired tasks, the hosts that are created are closed. The highest number of host shutdowns is performed by MU thus leading to better reliability of the system (Fig. 6).

Energy consumption has also been analysed in all these algorithms. It has been observed that MC has minimum energy utilisation when it comes to interquartile range algorithm (Fig. 7). Thus, by using our approach, energy could be saved and quality of service can be improved by adjusting other service parameters, i.e. service level agreements, service level agreement violations, number of hosts created and shutdown etc.

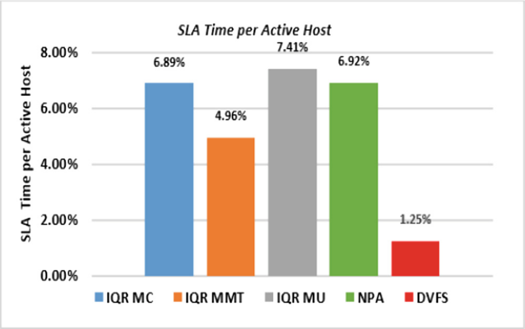

Fig. 5. SLA time for active host.

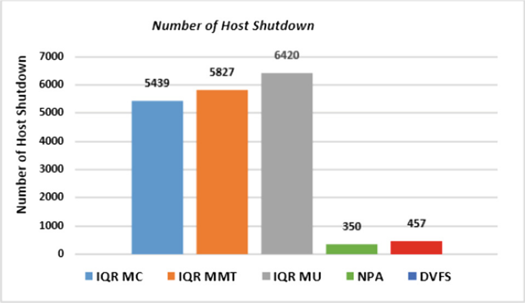

Fig. 6. Total sum of host shutdown.

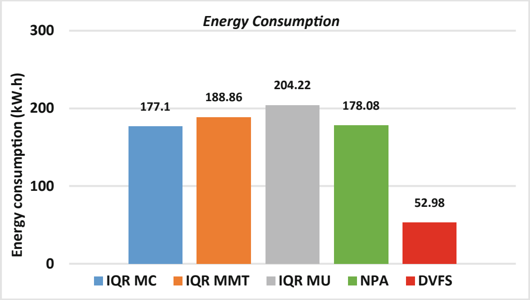

Fig. 7. Energy consumption comparison.

From the above results, it can be deduced that minimum energy consumption is dependent upon two factors, minimum virtual machine migration and maximum number of host shutdowns. It can also be noted that energy consumption by the physical host and service level agreement violations are indirectly proportional to each other. When the hosts use more energy, they have a smaller number of service level agreement violations and vice versa. Therefore, better quality of service and resource optimisation could be performed if these factors are controlled in the real time computing world. Dynamic frequency scaling techniques provides better results as we can demonstrate using the interquartile virtualisation technique.

7 Conclusion

In this paper, tests have been carried out in relation to different factors that affect resource optimisation in cloud computing. Virtual machine migration is done based on under or overutilization of resources, service level agreement that are achieved and which are violated. The economic impact of servers in terms of the energy efficiency factor has also been considered for resource optimisation. Implementation of interquartile range algorithm shows that MMT has a minimum number of virtual machine migration and service level agreement violations which lead to better resource optimisation of quality of service. Therefore, the suggested scheme can be expanded and implemented for virtualised cloud servers, leading to better efficiency, cost saving and better quality of service. Therefore, the dynamic frequency scaling technique can save more energy when used on a bigger scale as demonstrated using interquartile virtualisation.

References

1. Chen, K.T., Huang, C.Y., Hsu, C.H.: Cloud gaming onward: research opportunities and outlook. In: IEEE International Conference on Multimedia and Expo Workshops (ICMEW) (2014)
2. Long, S., Zhao, Y.: A toolkit for modeling and simulating cloud data storage: an extension to CloudSim. In: International Conference on Control Engineering and Communication Technology (2012)
3. Calheiros, R.N., et al.: CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Softw. Pract. Exper. **41**(1), 23–50 (2011)
4. Shuja, J., et al.: Survey of techniques and architectures for designing energy-efficient data centers. IEEE Syst. J. **10**(2), 507–519 (2016)
5. Yannuzzi, M., et al.: A new era for cities with fog computing. IEEE Internet Comput. **21**(2), 54–67 (2017)
6. Alsaffar, A.A., et al.: An architecture of IoT service delegation and resource allocation based on collaboration between fog and cloud computing. Mobile Information Systems **2016**, 15 (2016)
7. Rawat, P.S., et al.: Power consumption analysis across heterogeneous data center using CloudSim. In: 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (2016)
8. Godhrawala, H., Sridaran, R.: A survey of game based strategies of resource allocation in cloud computing. In: 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (2016)
9. 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)

10. Nguyen, B.M., Tran, D., Nguyen, Q.: A strategy for server management to improve cloud service QoS. In: IEEE/ACM 19th International Symposium on Distributed Simulation and Real Time Applications (DS-RT) (2015)
11. Atiewi, S., Yussof, S.: Comparison between Cloud Sim and green cloud in measuring energy consumption in a cloud environment. In: 3rd International Conference on Advanced Computer Science Applications and Technologies (2014)
12. Song, J., et al.: FCM: Towards fine-grained GPU power management for closed source mobile games. In: International Great Lakes Symposium on VLSI (GLSVLSI), pp. 353–356 (2016)
13. Ahmad, B., et al.: Energy optimisation in cloud servers using a static threshold VM consolidation technique (STVMC). In: Proceedings of the 13th International FLINS Conference on Data Science and Knowledge Engineering for Sensing Decision Support (FLINS2018) (2018)
14. Abdelsamea, A., et al.: Virtual machine consolidation enhancement using hybrid regression algorithms. Egypt. Inform. J. **18**, 161–170 (2017)
15. Theja, P.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**(2), 97–112 (2016)
16. Lee, Y.-T., et al.: World of warcraft avatar history dataset. In: Proceedings of the Second Annual ACM Conference on Multimedia systems, pp. 123–128. ACM, San Jose (2011)