A Performance Modeling Framework for Energy-based Scheduling in Cloud

Masnida Hussin, Raja Azlina Raja Mahmood, Nor Azura Husin, Noris Mohd. Norowi

Faculty of Computer Science & Information Technology
University Putra Malaysia, Selangor, Malaysia

{masnida, raja_azlina, n_azura, noris} @ upm.edu.my

***Abstract*: Cloud computing becomes a powerful trend in the development of ICT services. Cloud providers typically try to provide high performance to their users while minimizing energy consumption in their operation. However, there is lack of performance metric that analyzing trade-off between performance and energy consumption. Considering high volume of mixed users' requirements and diversity of services offered; an appropriate performance model for achieving better balance between Cloud performance and energy consumption is needed. One of resource management strategies for better performance is through effective scheduling approach. In this paper we investigate a relationship between performance metrics that chosen in existing scheduling approaches with energy consumption for energy efficiency. Through such relationship, we develop an energy-based performance modeling framework that provides a clear picture on parameter selection strategy for effective energy management. Our framework provides adaptive guideline for designing performance model explicitly for energy efficient scheduling. We believed that better understanding on how to model the scheduling performance will lead to green Cloud computing.**

***Keywords*:** energy efficient; scheduling; energy management; performance modeling.

1. Introduction

Cloud computing is a state-of-the-art technology where it provides computing services i.e., IaaS, PaaS and SaaS to users [1, 2]. Basically, Cloud providers e.g., Amazon, Microsoft and Google provide their users with resource sharing model where resources (e.g., processors, storage) can be added and released easily either they are needed or otherwise. They try their best to make resources available for large-scale number of users. Agreement between Cloud providers and users is defined through Service Level Agreement (SLA) that ensures users' requirements either processing or storage is fulfilled within the defined SLA. Such agreement approach practically eliminates hassles of procuring, managing and maintaining data centers at users' sides. It also able to save maintenance and operation costs.

Particularly, large computing and storage infrastructure like Cloud needs more energy to generate sufficient electricity and cooling systems thus more expenses need to be invested. Furthermore, computer systems not only consuming vast amount of power also emit excessive heat; this often results in system unreliability and performance degradation. It has been reported in previous studies (e.g., [3-5]) that system overheating causes system freeze and frequent system failures. According to the authors in [6] the highest energy cost of data centers are used to maintain the running servers. The servers need to be available and accessible throughout the year even though nobody is accessing them. Such situation incurs high cost in electric and cooling systems.

Noticeably, one of areas in Cloud that demands for green initiative is data center that becomes the heart of services. In order to sustain with good service reputation, the data centers needed to facilitate the processing and storage requirements. It is basically to fulfill the defined SLA between the Cloud providers and users. The issue of expanding and upgrading data centers to meet the defined SLA leads to excessive and ineffective energy consumption. Besides, the massive and rapid growth of Cloud ironically impacts the world negatively. The consequence of high resource availability in Cloud is not merely to financial expense but worse to environment through carbon footprint. There are more greenhouse gas (GHG) released in the atmosphere that leads to global warming, acid rain and smog [7]. The energy management in Cloud needs to be concerned not only to minimize the energy consumption; it also aimed for improving the system performance.

Although there have been many research efforts to reduce the energy consumption in computing operation, there is still lack of a decision support framework for metric selection in energy management. It is needed a clear guideline for assessing performance and energy efficacy of Cloud especially in its processing and communication activities, processes and protocols. Scheduling approach promotes better energy consumption by effectively scheduling users' requirements based on resource availability. The main issue in such area is how to balance between system performance and energy consumption for benefit both Cloud providers and users. It is a challenge to find the best trade-off between both metrics in dynamic computing environment like Cloud.

In this paper we focus on performance model for efficient energy consumption that influenced by parameter selection in the scheduling approaches. We develop energy-based performance modeling framework that aims to provide a guideline for analyzing the better trade-off between system performance and energy consumption in Cloud. It is wise to effectively manage energy consumption in Cloud computing; this would in turn be very beneficial cost of computation.

The reminder of this paper is organized as follows. A review of related work is presented in Section 2. In Section 3 we describe the parameter categories that chosen for efficient energy management. Section 4 details our performance model for energy efficiency. Finally, conclusions are made in Section 5.

2. Literature Review

The energy management has inspired many researchers [6, 8-13] to focus on green Cloud computing. The scope of energy assessment for modeling its performance should be stretched further incorporating parameter selection and energy model that being used. In [14], the authors emphasized resource utilization in their energy model. They utilized the virtualization as a core technology to contribute green IT because it able to share limited resources with varies workloads. The quantity of physical machines (PMs) can be reduced where processing is performed by virtual machines (VMs). VMs are very flexible that act as independent servers that maximized resources utilization while achieving energy efficiency. However, the selection of parameters in their performance model for energy efficiency is still an open issue. Some researchers focused on resource state in their energy models. According to the authors in [9], processors consumed approximately 32Watt when they are operated in idle mode that compared to storages merely used 6Watt. In peak processing mode the energy consumption of processors can be boosted up to more than 80Watt to 95Watt [15]. Energy efficient scheduling that proposed in [16] dynamically allocated users' tasks into processor to achieve better performance and minimize energy consumption. The system performance and energy consumption in their work is been measured throughout the task execution either during peak or idle state, then total energy consumption is recorded.

There are also some works that calculated the energy consumption through their modern scheduling approaches. Power-efficient scheduling in [17] assigned set of virtual machines (VMs) to physical machines (PMs) for data centers management. They used the consolidation fitness to determine the right VMs to replace PMs when existing PM is been switched off. However there is challenge to determine the most right VM due to unpredictable changes in the system workload. Power-aware mechanism in [18] is based on priority scheduling for efficient energy management. They adopted various heuristic for energy-aware scheduling algorithms that employed multi-objective function for diverse efficiency-performance tradeoffs. Their scheduling algorithms consist of three steps (i.e., job clustering, re-evaluating to give better scheduling alternatives and selecting the best schedule). They conducted the experiments in homogeneous environment that disguises Cloud computing environment. The existing researches that proposed energy efficient scheduling are able to reach appropriate balance between system performance and energy consumption. However, the dynamicity and heterogeneity on their computing environment are limited to some extent.

There are some researchers (e.g.,[7, 19, 20]) that highlighted the incorporation of low-energy computing nodes in heterogeneous distributed systems and able to achieve energy efficiency. Due to the scheduling approaches are subjected to system environment and scale, it required effective performance model for better evaluation on both performance and energy usage. However, there is lack of exclusive parameter selection strategy for energy management in order to balance between system performance and energy consumption. In Table I, we summarized the selection of performance metrics and system behaviors that used in those existing scheduling approaches to design energy model.

Table 1. Relationship between Scheduling Rule and Energy Model

No.	Scheduling Rule	Energy Model
1	Heuristic approach used for pre-scheduling processing. Such predictable rule satisfies energy saving where it consumed less processing time.	It considers busy and idle states of processing elements.
2	Threshold for processor, memory, disk and communication link in resource allocation. If the threshold chosen is too low it might reduces resource utilization while setting too high threshold leads to communication bottleneck.	It considers resource utilization and overhead in the system.
3	Migrating or remapping strategy in scheduling using a set of virtual machines (VMs).	Virtualization leads to energy efficiency.

3. Parameters in Scheduling Approaches for Energy Efficiency

In this paper we investigate the relationship of performance metrics that chosen in the scheduling approach with energy consumption to facilitate energy efficiency. The existing scheduling approaches (e.g.,[6, 12, 17, 21]) analyzed the performance of their energy management scheme through various scheduling algorithms. They have chosen several types of parameters such as time-based perspective, utilization and overhead.

3.1. Time-based Metrics

Time-based metric is one of the most popular parameters that chosen for focusing green Cloud. Such parameter used to measure the effectiveness of scheduling decision for uncertain, dynamic and large-scale environment. The time-based metrics, for example, execution time, waiting time and response time is designed to monitor and manage queuing problems in scheduling. The major problem in scheduling is to allocate various users' jobs to be mapped and executed by the right resources. Due to the scheduler needs to get information from users and resources, the scheduling decisions are most of the time consumed more processing energy compared than storage. The resources need to be available to perform the job execution at most of the time. Therefore, in order to improve the execution time for minimizing energy consumption we need to take into account the optimization technique in designing scheduling approach.

Furthermore, the issue on suitable queue length comes into a picture when the total waiting time of jobs leads for high energy consumption. Note that the determination of queue size is significantly related to the scale of computing

system. Specifically, the queue size contributes for better job waiting time that it relates to buffer management. The suitable size of queue needs to be identified to reduce data access time in the buffer. We do not want to power the entire memory module in lengthy time only for data accessing operation. Hence, the effective scheduling approach for dynamic environment required to define the (most) suitable queue length for reducing power consumption.

3.2 Utilization

There are scheduling approaches that calculate the energy consumption based on computing resources' busy and idle states. In particular, the resources that have high busy time means the system utilization is improved. Meanwhile the system is considered has low utilization when there are many resources that idle in a given observation time. However, from energy management perspective, high utilization may leads to large energy consumption. The resource that has high utilization certainly consumed high processing energy in order to complete the task execution. It is more crucial in the idle state of resources. There is huge percentage of power consumption (i.e., for electricity and cooling systems) to facilitate the running resources. The idle resources need to be available even though there is no processing happen in their space. Resource utilization only 53% of total energy consumed in data centers [22]. Therefore, the best solution is to effectively manage resource utilization for energy efficiency. There is a big role for the scheduler to monitor resource utilization in dynamic environment.

3.3. Overhead

The storage and processing resources in Cloud must be highly available that it reflects Quality of Service (QoS). It included the ability of Cloud to adapt with unexpected failures, e.g., storage overloaded, traffic congestion and performance fluctuation. Such scenarios needed extra time for the Cloud providers to solve and fix without notify the users. Some strategies used replication of objects and services, and using redundant processing and communication mechanisms to solve the unexpected failure. In order to implement these strategies they need more than one communication paths that used for disseminating the same information, and several processing elements for processing the same action. In scheduling approach, we need to design extra procedure or policy to manage such unexpected failure and implement reserve strategy. It explicitly incurred extra communication and processing overheads in the systems. For the sake of energy efficiency, overhead should be minimized while maintaining system performance. Hence, tunable parameters in experimental setting are significant to thoroughly identify the system behavior/action in order to achieving the target results (better trade-off).

4. Parameter Selection Strategy for Energy Efficiency

There are various performance models that adopted in scheduling approach for energy efficiency (e.g., [6, 12, 13, 17, 23]). Basically, in large-scale distributed system like Cloud where it needs to support large number of users, the performance degradation is unacceptable. Therefore, the priority in scheduling approach is mainly to maximize or at least sustain the system performance while minimizing the energy consumption. In this work, we specifically divide our energy-based performance model into three stages; (i) planning, (ii) formulating, and (iii) adjusting.

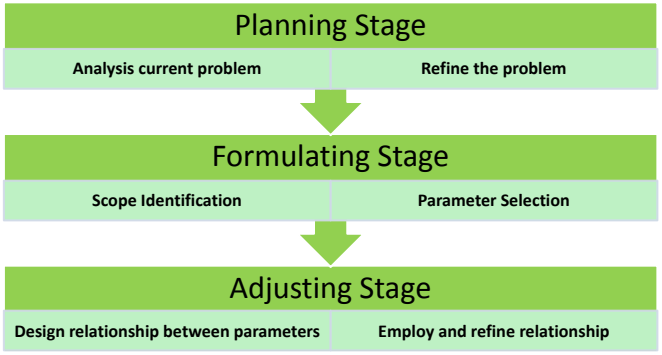

Figure 1. Energy-based Performance Model

4.1 Planning Stage

In planning phase, there is required to thoroughly understand the energy problem itself. There are several energy problems such as energy waste, inaccurate energy measurement etc. that identified by existing researchers (e.g., [7, 11, 24]). The investigation on energy consumption in large-scale data center can be commenced through its operational infrastructure. Such infrastructure can be classified based on power usage for physical equipment and processing condition. It leads to two different measurements of energy efficiency are; power usage effectiveness (PUE) and data center effectiveness (DCE) [19]. For PUE measurement, it concerns on total power used for IT equipment such as server, routers and cabling. Meanwhile DCE is calculated by the resource management strategy (i.e., scheduling, load balancing and security) that been applied in the server room or data center for complying the users' requirements. Both PUE and DCE are inter-related to each other for supporting computing operations.

The Cloud provider, in particular an administrator should prepare a regular report for analyzing the current energy consumed in the organization. Note that due to high demands from the users for processing and communication purposes, the energy consumption increased exponentially. For monitoring energy consumption from IT equipment perspectives, the energy distribution must be accurately measured. It is due to the usage of those equipment contributes to electricity bill that basically relates to operational and maintenance costs. For example, the server rooms or data centers needed of mechanical and electrical (M&E) infrastructure, also ventilating or cooling

infrastructure for supporting the operational in the room. Nowadays, the actual cost for managing the IT equipment either in server room or laboratory has become big issue in organization. Even though Cloud utilizes the virtualization technology, it still relies on the physical computing equipment at its end support.

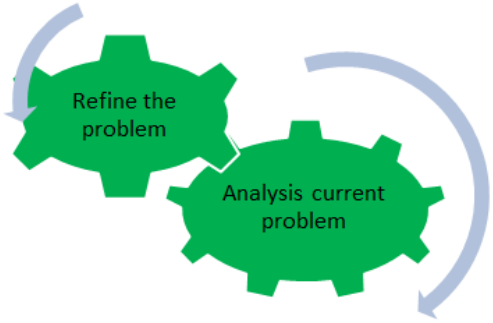

Figure 2. Planning Procedures

The performance model then should be focusing on how to utilize the usage of IT equipment hence the energy distribution can be optimally consumed. Several strategies such as scheduling, load balancing and authorization are needed to be highlighted in order to optimally manage and utilize the energy distribution. The administrator can analyzed the effectiveness of the processes by focusing which process leads to energy waste. For example, if FCFS policy is implemented in the system and it shows that network traffic is always in congestion for more than ten hours, and then the system should enhances with different scheduling policy. Another example is in the case of security, where the authorization procedure is too complex and steers to high complexity (it leads to excessive overhead). The administrator might be tailored some procedures to make the authorization quick and efficient. Such strategies will improve the energy consumption in many ways such as there is no needed to buy new IT equipment to meet the users' demands.

4.2 Formulating Stage

Cloud computing enables its services (i.e., IaaS, PaaS, SaaS) for various users at anywhere for anytime. It required to provide reliable in both computation and communication activities. In order to sustain the performance, thorough investigation of resource management is required. Note that the activities (i.e., computing and communication) in resource management are much influenced by the feature of system and characteristics of the users. Hence, we need to identify the system environment and its scale. Such criteria can be defined according to several levels that represent their complexity in communication and processing activities. In this work, we highlight three level of complexities; (i) homogenous/heterogeneous, (ii) static/dynamic and (iii) local/centralized/distributed. High in complexity means huge amount of energy consumption that consumed for communication and computing activities. For example, the system is considered consumed large amount of energy when there are needed a large number of processing elements (PE) to offer high availability in processing. In some cases, the small number of PE is yet consumed large percentage of energy. It happens when the PE is operated for 24/7 and it needed very cold room to control the heat releases. Note that excessive heat emitted by these PE causes notable power consumption for cooling them. In addition, the energy consumption is proportional to the users' requirements. The resources operated in extensive processing time if there is huge number of workload in the system. Such situation leads to increase the energy usage in the whole system operation.

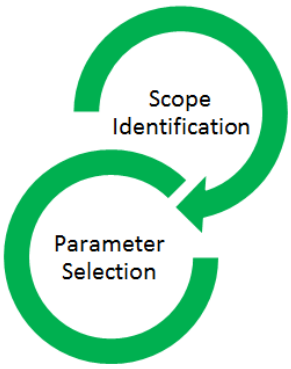

Figure 3. Formulation Procedures

In response to the system and user criteria, selection of performance metric to measure energy consumption is a huge challenge. The users in large-scale system normally demand for varies of processing and communication requirements. Hence, the system performance needed to fulfill the users' requirements in order to sustain the system's reputation. This is very important in Cloud because there are many Cloud providers that competing each other to provide better performance. Therefore, the performance metrics to evaluate energy consumption should be fragmented of a total/average of the system performance. The system performance might be reduced because its portion needs to embrace the energy consumed.

Table 2. Suggestion Metrics

No.	Evaluation Scope	Metrics
1	Hardware measurement	Real-time voltage and current, processing state/frequency.
2	Data Center	Time-based performance, utilization, processing overhead.
3	Mobile Cloud Computing	Communication overhead (transaction delay and traffic congestion).

In this work, we also express some metric suggestion for energy efficiency that can be chosen based on a given evaluation scope (Table 2). From Table 2, it is merely that item number 2 and 3 are related to scheduling approach where it can be formulated to maintain the system performance while monitoring the energy consumed. Due to the large scale constitution of Cloud environments, such parameter selection is considered to be unbounded and still an open issue.

ISBN: 978-1-941968-15-4 ©2015 SDIWC

4.3 Adjusting Stage

In response to the collection of performance metrics, we then raise issue on how to integrate each metric for energy efficiency. Basically, the metrics aim for better scheduling decisions that leads to improve the system performance. In response to energy efficiency, the design of scheduling approach should able to monitor both performance and energy consumption for a given time duration. It is wised to capture energy consumption for a specific time duration that can be based on incoming workload while calculating the average of the execution time. It is because; by frequently measuring the energy consumption it implicitly increased a percentage of power usage. The right proportional of the system performance e.g., total execution time to measure the energy consumption must be significantly concerned. For instances, the performance metric in Cloud data center is measured by processing overhead as follows.

$$\min_{\text{overhead}} = \left(\frac{\sum idle_{time} + \sum busy_{time}}{total\ number\ of\ task} \right)$$

The total energy consumption in the system can be calculated as given:

$$total\ energy\ consumption = \frac{\sum total\ overhead}{simulation\ time}$$

; assumed that the energy consumption is measured through simulation program. In such example, the total number of incoming task is recorded within the scheduling process that used to calculate the overhead. Meanwhile, the energy consumption is measured for entire simulation program. In some scheduling approaches (e.g.,[22, 23]), they used fix power consumption (in Watt) that identified at busy and idle states of processing element for calculating energy consumption.

Energy-based scheduling approach aims to schedule users' tasks to be run on the computer systems that consumed low energy consumption. It means that the computer system can tune the processing states while reducing the system energy consumption. In response to this, the computer system needs to collect the run-time information of applications, monitor the processing energy consumption states, notify about states changes and compute energy-based scheduling decisions. Through the monitoring process, the administrator is able to identify the processing activities that consumed large amount of energy. Then, it can perform modification in the scheduling strategy in order to balance between the system performance and energy consumption. Such strategy involves alteration techniques on how the systems and applications respond to the processing state changes and adjust the metrics that fit to the target goal.

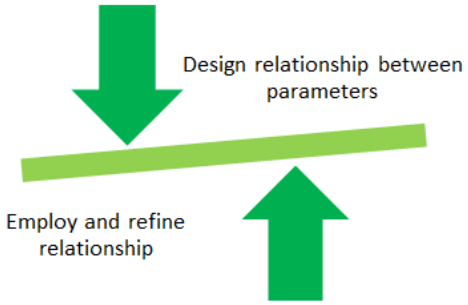

Figure 4. Alteration Strategy

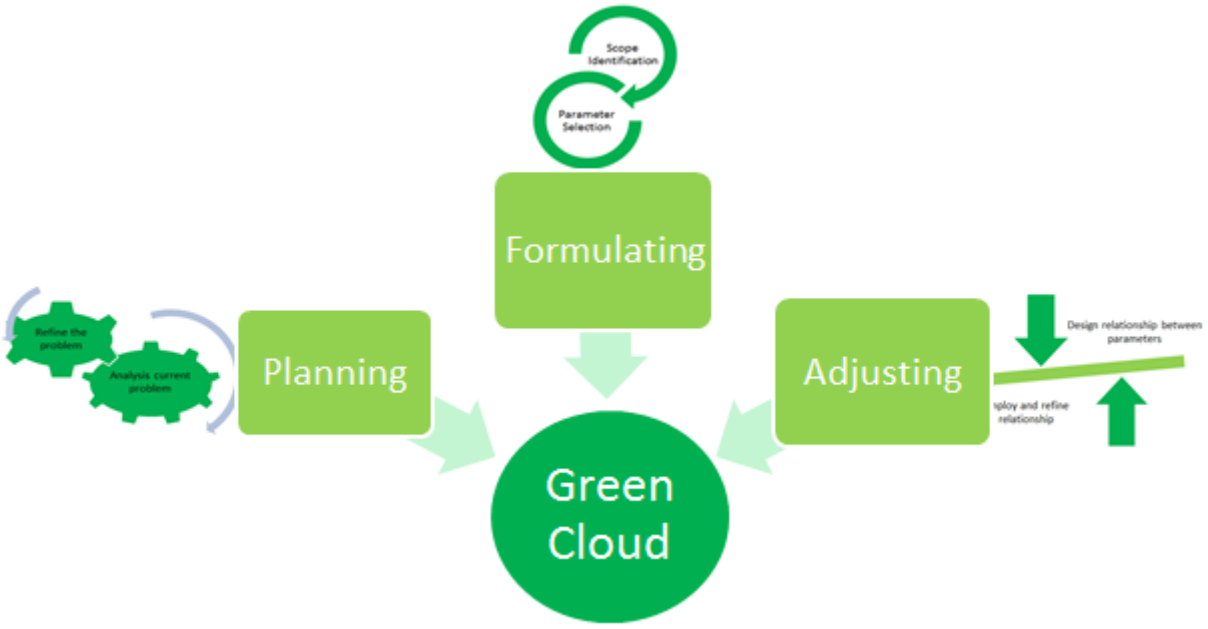

Figure 5. Decision Support Framework for Parameter Selection

5. Conclusion

The concept of green computing has begun to spread in the past few years and still gaining its popularity. It is due to its significant performance, environmental and economic implications. Basically, it is hard to obtain optimal scheduling decisions (in terms of energy consumption) with dynamic system environment like Cloud. In this paper, we analyze relationship between the parameter selection in scheduling approach and energy consumption that brings to energy efficiency. Our study aims to help other scheduling studies for designing better performance model in order to develop efficient energy management. Specifically, we develop the support decision framework for modeling the system performance where the energy efficiency becomes the next goal. Note that, the (near) optimal scheduling for energy efficiency is still an open issue. Hence, there is a lot of potential for more research on its performance model and design. Optimistically, Cloud able to achieve better trade-off between performance and energy consumption when there is clear guideline for designing the energy-based performance model.

References

1. Hwang, K., G.C. Fox, and J.J. Dongarra, *Distributed and Cloud Computing: From Parallel Processing to the Internet of Things.* 2012, MA USA: Elsevier.
2. Rountree, D. and I. Castrillo, *The Basics of Cloud Computing : Understanding the Fundamentals of Cloud Computing in Theory and Practise.* 2014, MA USA: Elsevier.
3. Hussin, M., A. Abdullah, and S.K. Subramaniam, *Adaptive Resource Management for Reliable Performance in Heterogeneous Distributed Systems*, in *Lecture Notes in Computer Science*, R. Aversa, et al., Editors. 2013, Springer International Publishing. p. 51-58.
4. Zunjare, P. and B. Sahoo, *Evaluating Robustness of Resource Allocation in Uniprocessor Real Time System.* International Journal of Computer Applications 2012. **40**(3): p. 13-18.
5. Lee, Y.C. and A.Y. Zomaya, *Rescheduling for reliable job completion with the support of clouds.* Future Generation Computer Systems, 2010. **26**(8): p. 1192-1199.
6. Lee, Y.C. and A.Y. Zomaya, *Energy efficient utilization of resources in cloud computing systems.* Journal of Supercomputing, 2010: p. 1-13.
7. Abbasi, Z., Micheal Jonas, Ayan Banerjee, Sandeep Gupta, and Georgios Varsamopoulos., *Evolutionary Green Computing Solutions for Distributed Cyber Physical Systems*. Evolutionary Based Solutions for Green Computing, ed. S.U. Khan, et al. Vol. 432. 2013: Springer. 1-28.
8. Rizvandi, N.B. and A.Y. Zomaya, *A Primarily Survey on Energy Efficiency in Cloud and Distributed Computing Systems.* arXiv preprint 2012.
9. Khan, S.U. and C. Ardil, *Energy Efficient Resource Allocation in Distributed Computing System.* World Academy of Science, Engineering and Technology, 2009(56): p. 667-673.
10. X, W. and W. Y, *Energy-efficient Multi-task Scheduling Based on MapReduce for Cloud Computing.* Seventh Int. Conf. Comput. Intell. Secur., 2011(57-62).
11. Wei, D., Fangming L., Hai, J., Bo, L. and Dan, L., *Harnessing renewable energy in cloud datacenters: opportunities and challenges.* Network, IEEE, 2014. **28**(1): p. 48-55.
12. Hussin, M., Y.C. Lee, and A.Y. Zomaya, *Efficient Energy Management using Adaptive Reinforcement Learning-based Scheduling in Large-Scale Distributed Systems*, in *40th Int'l Conf. on Parallel Processing (ICPP2011)*2011, IEEE Computer Society: Taipei, Taiwan. p. 385-393.
13. Kim, Nakku, Jungwook Cho, and Euiseong Seo, *Energy-credit scheduler: an energy-aware virtual machine scheduler for cloud systems.* Future Generation Computer Systems. **32**(2014): p. 128-137.
14. Beloglazov, A., Rajkumar B., Young C. L. and Albert Y. Z.., *A Taxonomy and Survey of Energy-Efficient Data Centers and Cloud Computing Systems.* Advances In Computers, 2011. **82**(CLOUDS-TR-2010-3): p. 47-111.
15. Barroso, L.A. and U. Holzle, *The Case for Energy-Proportional Computing.* Journal Computer, 2007. **40**(12): p. 33-37.
16. Hussin, M., Y.C. Lee, and A.Y. Zomaya. *Priority-Based Scheduling for Large-Scale Distribute Systems with Energy Awareness*. in *Proc. of the 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*. 2011. Sydney, Australia: IEEE Computer Society
17. Sharifi, M., H. Salimi, and N. M., *Power-efficient distributed scheduling of virtual machines using workload-aware consolidation techniques.* J. Supercomputing, 2011. **61**(1): p. 46-66.
18. Aziz, A. and H. El-Rewini, *Power efficient scheduling heuristics for energy conservation in computational grids.* J. Supercomputing, 2011. **57**(1): p. 65-80.
19. G.Koomey, J., *Estimating Total Power Consumption by Servers In The U.S. and The World*, 2007
20. Lee, Y.C. and A.Y. Zomaya, *Minimizing Energy Consumption for Precedence-constrained Applications Using Dynamic Voltage Scaling*, in *9th IEEE/ACM Int'l Symposium on Cluster Computing and the Grid (CCGRID2009)*2009, IEEE. p. 92-99.
21. Tham, C.-K. and T. Luo, *Sensing-Driven Energy Purchasing in Smart Grid Cyber-Physical System.* IEEE Transaction on Systems, Man and Cybernetics: Systems, 2013. **43**(4): p. 773-784.
22. Minas, L. and B. Ellison, *InfoQ:The Problem of Power Consumption in Servers*, in *Energy Efficiency for Information Technology*2009, Intel PRESS.
23. Li, K., *Energy efficient scheduling of parallel tasks on multiprocessor computers.* Journal of Supercomputing, 2010: p. 1-25.
24. Schreier, P., *An Energy Crisis in HPC*, in *Scientific Computing World : HPC Projects* Dec 2008/Jan 2009.