Energy Efficiency in Cloud Computing

Mohamed Deiab, Deena El-Menshawy, Salma El-Abd, Ahmad Mostafa, and M. Samir Abou El-Seoud

***Abstract*—Cloud computing is one of the recent emerging technologies that provides services to consumers in a pay as you go model. Cloud computing offers ITC based services over the internet and the use of virtualization allows it to provide computing resources. Data Centers are the core of cloud computing, which consists of: networked servers, cables, power sources, etc. which host the running applications and store Business information. High performance has always been the most critical concern in cloud data centers, which comes at the cost of energy consumption. The vital challenge is balancing between system performance and power consumption by reducing energy consumption without prejudicial impact on the performance and quality of services delivered. There are many techniques and algorithms proposed to achieve efficient energy utilization in cloud computing, these techniques include: VM Migration, Consolidation and Resources orchestration in cloud computing. This paper provides a survey of approaches and techniques for energy efficiency in cloud computing.**

***Index Terms*—Cloud computing, energy efficiency, resource management, virtualization.**

I. Introduction

The progress of technology and incorporating networks, storage and processing power led to new era of computing, called cloud computing or commonly known as the cloud. Cloud computing is defined as a technological paradigm that allows on-demand access via the internet to a common shared computing resources. It is considered to be a model for supervision, storing and processing data online via the internet [1]. Some cloud computing characteristics include on-demand services, network access by using internet as a medium, shared resources by pooling resources together to be used by multiple clients and scalability by maintaining elasticity of resources. Cloud computing offers different services based on three delivery models, namely:

1) Software as a service (SaaS): this allows users of cloud to access the providers apps (PA) over the internet.
2) Platform as a Service (PaaS): this allows users to deploy their apps on a platform which service provider of cloud (SPC) provides.
3) Infrastructure as a Service (IaaS): this allows users to rent, store, process in an infrastructure provided by SPC.

The rapid growth in mobile devices and the storage needs due to the adoption of cloud data networking are creating huge data traffic due to the emerging issues of data centers and also digital content, media and technology. Energy consumption by the organizations that provide cloud service is continuously increasing. It has been concluded that the amount of energy consumed by the data centers of the cloud service providers is equal to 1.5% of power supplied to an entire city [2]. The data centers for cloud service providers are used for hosting the cloud applications which are normally consuming massive amounts of resources that utilize a huge percentage of electrical energy, which produces growth in operational cost and results in emission of Co2 [3]. Cloud service providers ensure reliability and load balancing for the services provided to the users around the world by keeping servers operating all the time. In order to satisfy the Service Level Agreement (SLA), cloud service providers has to supply power continuously to data centers, which utilizes a huge amount of energy by the data center and subsequently increases the cost of investment. Thus, it has been noticed that high performance has been the sole concern in data center deployments. This demand has been achieved without paying attention to the amount of energy consumed. The key challenge is to balance between system performance and the power consumption. It was found that a huge amount of energy is consumed due to idle and overloaded servers in data center. According to [4], idle servers use 69-97% of total energy in the presence of enabled power management function. This paper will present an overview of the different methodologies to have energy efficiency in cloud by introducing some of the current proposed solutions as servers load balancing, VM virtualization, VM migration and resource allocation.

A. Clouds

In order to meet the rapid-changing business and organization needs, organizations need to devote budget and time to accelerate up their IT infrastructure such as software, hardware and network services. Regardless of the utilization of on-site IT framework, scaling the system could be difficult. And also the organizations are often incapable of achieving an ideal use of IT foundation. Thus, the cloud computing is the proposed solution. According to National Institute of Standards and Technology (NIST), cloud computing is the delivery of IT resources on-demand utilization by providing a pay as you go model for the consumers, while you can self-serve for the services that you need to your own application or any IT infrastructure that you need. A cloud computing service consists of highly utilized resources including software applications or virtual storages that can be used upon user request, consumers can simply connect to the cloud and use the available resources. This causes organizations to stay away from capital consumption for on-premises framework assets and scaling up or downsizing according to business requirements [3]. Cloud computing services can be deployed using three different models a private cloud, public cloud or a hybrid cloud. Private cloud function solely for one organization on a private network and is its highly secure. Public cloud is owned by the cloud service provider and offers the highest level of efficiency and

Manuscript received July 14, 2018; revised October 14, 2018.

The authors are with the British University, Egypt (e-mail: mohamed.deiab@bue.edu.eg, deena.elmenshawy@bue.edu.eg, salma.elabd@bue.edu.eg, ahmad.mostafa@bue.edu.eg, samir.elseoud@bue.edu.eg).

doi: 10.18178/ijmlc.2019.9.1.771

shared resources and hybrid cloud is considered to be a combination of private and public deployment models. In a hybrid cloud, specific resources are run or used in a public cloud and others are run or used on-premises in a private cloud this provides increased efficiency. Fig. 1 illustrates the architecture of cloud computing [3].

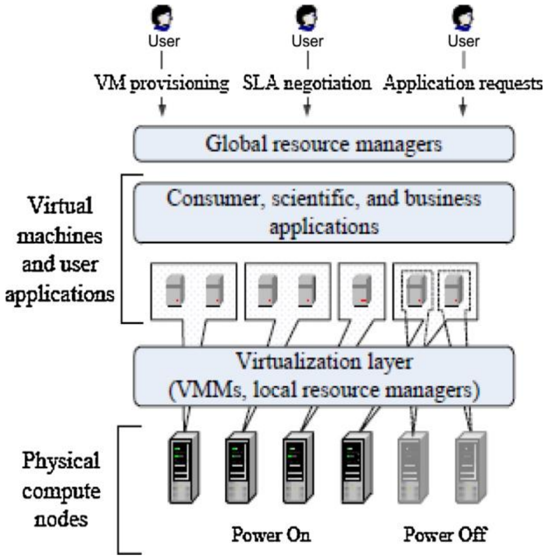
User
User
User
VM provisioning
SLA negotiation
Application requests
Global resource managers
Virtual machines and user applications
Consumer, scientific, and business applications
Virtualization layer (VMMs, local resource managers)
Physical compute nodes
Power On
Power Off

Fig. 1. A high level system architecture of cloud computing [3].

B. Data Centers

Data centers provide an IT backbone for cloud computing. A data center is a technical facility that houses organizations IT operations and equipment where it stores, manages and disseminates its data. A data center houses and networks most critical systems and are vital to business continuity and operations. A Data center is considered to be the heart of cloud computing which contains all the cloud resources including servers, network, cables, etc. on which business information is stored and applications run. Until recently, high performance has been the sole concern in data center distributions and this demand has been satisfied without paying much attention to energy consumption. However, an average data center consumes as much as the consumption of 25,000 houses [3]. As the energy availability decreases and energy cost proportionally increases, the need for shifting the focus for utilizing data center resource management to optimize energy performance while maintaining service performance is becoming a necessity. Thus cloud service providers need to adjust their energy measures to ensure that their profit margin is not dramatically reduced due to high energy costs.

II. Energy Efficient Computing

Energy saving techniques in computing equipment have been classified as static power management (SPM) and Dynamic power management (DPM). SPM and DPM are completely different in categorization, SPM are more energy efficient at single system and supposed to be under the category of hardware level techniques, and since SPM techniques are related to hardware level efficiency, low power consumption circuit designing is an example of this technique. On the other side, DPM are more energy efficient in large systems and supposed to be under the category of level resource management methods. Also, DPM techniques are mostly implemented in software or on network layer, for example protocol design and algorithms. Fig. 2 shows an overview of various energy management schemes in computing equipment. Energy aware scheduling, energy efficient routing, load balancing, virtualization, resource consolidation and migration. Since high availability as well as quality of service and performance guarantee are still ignored which is most required in such distributed environments as the customers pay for their provisioned resources. The customers would not pay or may switch to other similar service providers if either quality of service or expected performance level is not achievable. Energy issues are supposed to be critical and also needs to be managed properly in some environment where mobile cloud computing is involved. Reducing the amount of energy used by applications through green compilers and robust programming can be achieved through application/software level methods [5]. In next sections, Application level and high level resource management techniques are discussed to achieve energy efficiency in single system, clusters, grids and cloud datacenters.

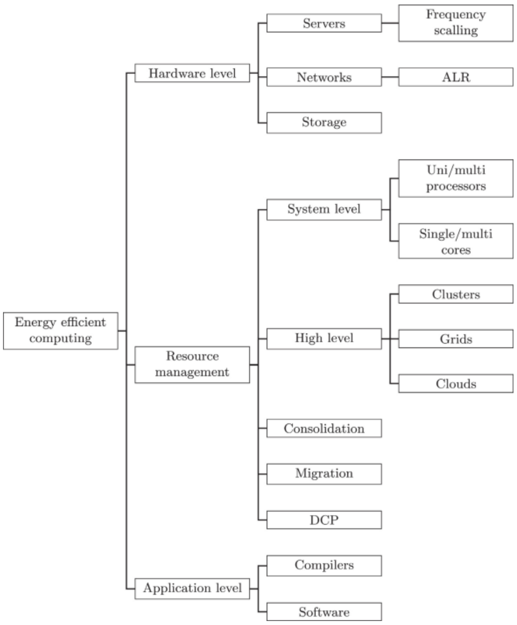

Fig. 2. Overview of various energy management schemes in computing equipment [6].

III. Resource Scheduling Model for Energy Saving in Cloud Computing

Basically the resource model of the cloud data center and the dynamic power model of the physical machine are both built, and afterward a three-dimensional virtual resource scheduling method (TVRSM) is proposed along with related algorithms [7]. The process of virtual resource scheduling in TVRSM is divided into three stages, these stages are virtual resource allocation stage, virtual resource scheduling stage and virtual resource optimization stage. Regarding TVRSM, three algorithms are designed corresponding to the mentioned stages of the virtual resource scheduling. These algorithms are MVBPP based heuristic virtual resource allocation algorithm (HVRAA), multi-dimensional power aware based virtual resource scheduling algorithm

(MP-VRSA) and virtual resource optimization algorithm (VROA).

Initially, the first stage in TVRSM which is virtual resource allocation stage is basically in charge of allocating the requested VMs by the customer to the suitable hosts. This stage is treated as multi-dimensional vector bin packing problem (MVBPP) and the MVBPP based heuristics virtual resource allocation algorithm (HVRAA) is proposed to solve it. In addition, the second stage which is virtual resource scheduling stage is responsible for migrating the VMs from the overload hosts to other hosts with lower resource utilization by using the VM migration technology in order to achieve load balancing of the cloud data centers and also to minimize the amount of violations of Service Level Agreement. The multidimensional power-aware based virtual resource scheduling algorithm (MP-VRSA) is proposed in this stage. Furthermore, the third stage which is virtual resource optimization stage is in charge of migrating the VMs from the hosts with the least resource utilization to other hosts and switch the original hosts to sleep mode, this process can further reduce the energy consumption of the cloud data centers by designing the virtual resource optimization algorithm (VROA). Finally, the authors verified the effectiveness of the proposed method through experimentation. The results prove that the TVRSM is able to efficiently allocate and manage the virtual resources in the cloud data center. And a comparison is made between the proposed methods with other traditional algorithms. The results showed that the TVRSM can effectively reduce the energy consumption of the cloud data center and minimize the amount of violations of Service Level Agreement.

A. Resources Model of Cloud Data Center

In [8], the authors proposed a resource model of the cloud data center, which is shown in Fig. 3, it consisted of M clusters, and each cluster contains N physical machines. Several virtual machines are deployed on each physical machine. According to the resources owned by the virtual machine, each virtual machine can run multiple applications. So the load of each virtual machine results from the applications running on the virtual machine. The node controller runs on each physical machine is responsible for monitoring the resource utilization of each physical machine and control the physical machines status such as setting the physical machine to sleep mode or activating the sleeping physical machine. Also, the node controller sends the management commands to the Hypervisor in order to adjust the resources owned by the local virtual machines. The global resources management node is responsible for scheduling and allocating all the resources owned by the cloud data center. It can manage and monitor all the resources and implement the load balance of the cloud data center. As shown in Fig. 3, each physical machine is characterized by the CPU performance, amount of RAM and network bandwidth, and each physical machine can run multiple virtual machines, and the physical resource owned by each virtual machine consists of CPU, memory capacity and network bandwidth. The physical machines use Network Attached Storage (NAS) instead of having local disks. It uses NAS in order to save data, which can ease the data sharing between all physical machines and enable live migration of virtual machines quickly.

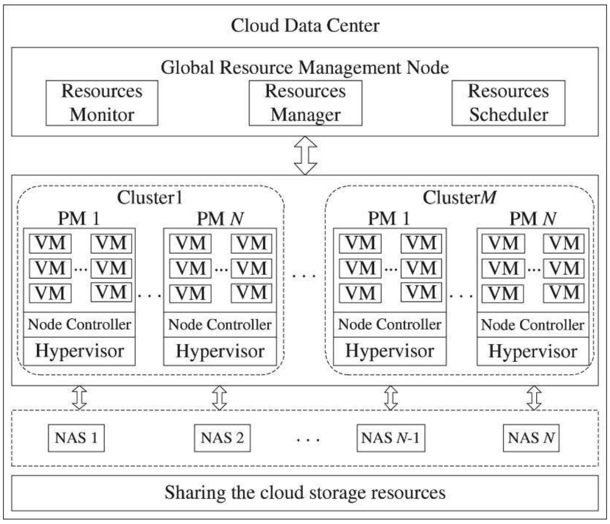

Fig. 3. Resource model of CDC [7].

B. Dynamic Power Model

In [7], the authors have shown that the power consumption by PMs can be described by a linear relationship between the power consumption and CPU utilization. Hence, the power consumption Pi (t) of PM i running on time t can be expressed by the relationship between the CPU utilization ui (t) of PM i on time t and the maximum power consumption PiMax of PM i, as shown in the below formula:

$$p_i(t) = k \,.\, P_{iMax} + (1-k) \,.\, P_{iMax} \,.\, u_i(t) \tag{1}$$

As shown in formula 1, PiMax is considered to be the maximum power consumption of host, Pi(t) is considered to be the power consumption of PM i running on time t, k considered to be the fraction of power consumption when the host is in idle state and ui(t) is the CPU resource utilization of the PM on time t. In the below formula, the CPU utilization of the PM i is defined as the ratio of the total CPU resources requested by the all VMs running on PMi on time t to the all CPU resources owned by the PM I as shown in formula 2.

$$u_i(t) = \sum_{j=1}^{VMS_i} \frac{cpu_{rj}(t)}{CPU_{iTotal}} \tag{2}$$

IV. Resource Scheduling Model for Energy Saving in Cloud Computing

In HVRAA, the main objective of the algorithm is to assign all the VMs requested by customers to the minimum number of PMs. The core idea of HVRAA is as follows: select the VM which has the largest Weight Dot Product (WDP), Select all the VMs that can fit the host and finally, if there is no any fit the current host, then start a new host until all the VMs are assigned into the hosts.

In MP-VRSA, the main objective of this algorithm is to further reduce the energy consumption by identifying and detecting the overloading hosts. The MP-VRSA is composed of four steps: detecting the overloading hosts, choose the VMs that need to be migrated from the overloading hosts, selecting new hosts for the VMs to be migrated and implementing the migration operation for all the overloading hosts. The first step in MP-VRSA is detecting the overloading hosts where the overloading detection strategy is used in order to find the overloading hosts in the CDC to determine whether the VMs running on the host need to be migrated. The below steps are essential for detecting the overloading hosts; setting the utilization threshold and if the CPU utilization of a host exceeds the threshold, then the overloading host can be detected and some VMs running on the host need to be migrated. The second step in MP-VRSA is VM selection strategy, once the host has been detected overload. Maximum Correlation (MC) VM selection strategy is used for selecting VMs to migrate from the overloaded host. The idea of Maximum Correlation (MC) VM selection strategy is that the higher the correlation between the loads of VMs running on a host, the higher the probability of the host overloading. The CPU utilization of VM is considered as the load of VM. So according to this idea, VMs to be migrated that have the highest correlation of the CPU utilization with other VMs are selected. The third step in MP-VRSA is VM placement strategy where the main task of this strategy is to select the suitable host for the VMs that are migrated from the overloading hosts. However, when the VMs are reallocated to other hosts it is bound to make the CPU utilization of the hosts increased. So, the Minimum Power Increasing Strategy (MPIS) is designed in order to place the VMs into hosts quickly and reduce the energy consumption in the CDC. In Virtual Resource Optimization Algorithm (VROA), this algorithm migrates the VMs from the hosts with the least resource utilization to other PMs, and switch the original host to sleep mode. Therefore, it can reduce the energy consumption of data centers. VROA consists of three main steps; after the virtual resource scheduling step is finished, the VROA will select the host PM lowest with the lowest CPU utilization and attempts to migrate the VMs to other hosts. Then, the system will set host PM lowest to sleep mode when the VMs migrate to other hosts successfully. Finally, If any of the VMs on host PM lowest cannot be migrated, then the host is kept active and all the migration of VMs are canceled.

V. Energy-Aware Virtual Machine Migration for Cloud Computing

The proposed technique proposes another methodology for maintaining energy efficiency in cloud computing, by migrating the maximally loaded virtual machines to the least loaded active machine, while maintaining system performance by performing a live migration of the virtual machines to ensure that all the running applications will not get disconnected during migration. The proposed technique introduces a new methodology for improving resource utilization levels based upon the bio-inspired Firefly optimization technique to achieve energy efficiency in cloud data centers. The achievability of the proposed technique has been shown by executing the results by using the CloudSim simulator.

A. Firefly Optimization (FFO) Algorithm

The Firefly Optimization (FFO) algorithm has been introduced by Xin-She Yang in the late 2007 and 2008 at Cambridge University [9]–[11]. It was implemented upon the fireflies flashing characteristics and behavior, the characteristics have been introduced as follows:

1) One firefly is attracted to the other fireflies regardless of their sex as all fireflies are unisex
2) The attractiveness is proportionate to the brightness, thus they both decrease as their distance increases and for any two flashing fireflies, the less bright one will be attracted near the brighter one
3) The brightness of a firefly is calculated using the objective function to be optimized

The (FFO-EVMM) Technique introduces the idea of migrating the most loaded VM from an active node which satisfies minimum criteria for energy consumption, to another active node that consumes the least energy. The technique is implemented in four main parts as shown in Fig. 4:

1) Selection of source node
2) Selection of VMs
3) Selection of destination node
4) Distance updated values.

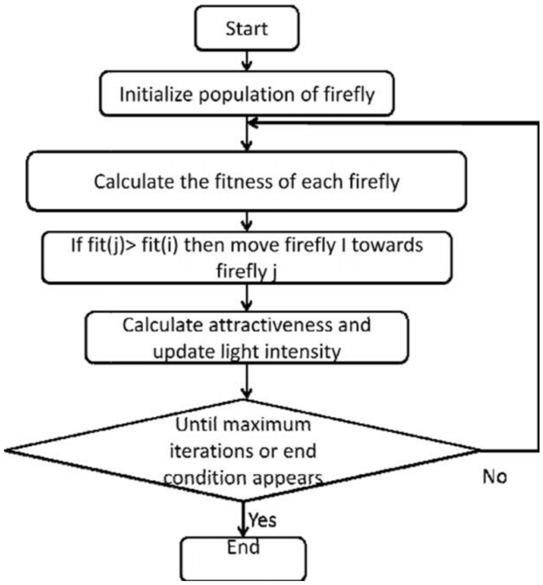
Start
Initialize population of firefly
Calculate the fitness of each firefly
If fit(j)> fit(i) then move firefly I towards firefly j
Calculate attractiveness and update light intensity
Until maximum iterations or end condition appears
No
Yes
End

Fig. 4. Flow chart for FFO algorithm [12].

VI. Results and Analysis

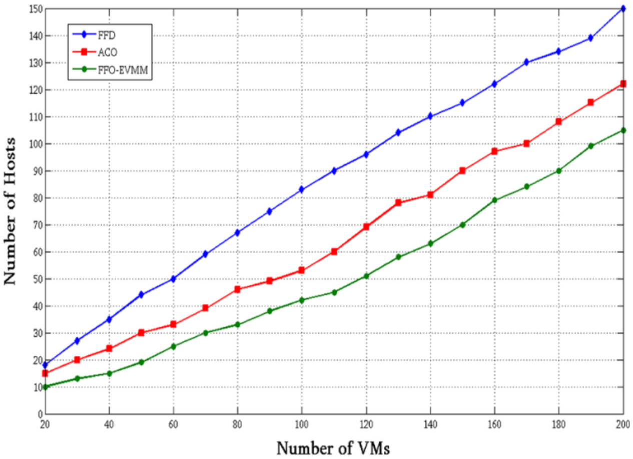

Fig. 5. VMs vs hosts [12].

The statistical results for the proposed FFO-EVMM algorithm were compared with the ACO-based and

FFD-based algorithms, using the CloudSim toolkit simulator which has captured that that FFO-EVMM technique runs less number of active hosts and performs less number of virtual machines migration in comparison to ACO and FFD-based algorithms.

As it's shown with the less number of running hosts and live migrations, FFO-EVMM requires lesser energy demand comparing to FFD and ACO algorithms as it has been noticed from Fig. 5 and 6.

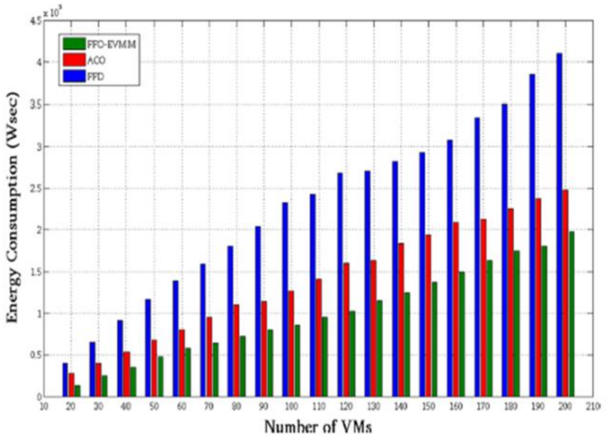

Fig. 6. VMs vs energy consumption [12].

VII. Conclusion

Cloud computing is considered one of the most crucial technologies that provide services to consumers in a pay as you go model. It offers ITC based services over the internet and the utilization of virtualization allows it to provide computing resources. Data centers are the core of cloud computing that store business information and host the running applications. High performance has always been the sole concern of all in data centers. This concern has been managed without considering energy consumption and performance. The challenge is to balance between power consumption and system performance. Many techniques and algorithms have been proposed to achieve adequate energy utilization in cloud data centers. This paper provided a survey of recent approaches and techniques for energy efficiency in cloud computing.

References

[1] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud com-puting: Vision, hype, and reality for delivering it services as computing utilities," in *Proc. 10th IEEE International Conference on High Performance Computing and Communications*, 2008, pp. 5–13.

[2] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397–1420, 2012

[3] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya, "A taxonomy and survey of energy-efficient data centers and cloud computing systems," *Advances in Computers*, Elsevier, 2011, vol. 82, pp. 47–111.

[4] N. R. D. Council. (2014). Scaling up energy efficiency across the data center industry. [Online]. Available: https://www.infrastructureusa.org

[5] F. Fakhar, B. Javed, R. Rasool, O. Malik, and K. Zulfiqar, "Software level green computing for large scale systems," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 1, no. 1, p. 4, 2012.

[6] M. Zakarya and L. Gillam, "Energy efficient computing, clusters, grids and clouds: A taxonomy and survey," *Sustainable Computing: Informatics and Systems*, vol. 14, pp. 13–33, 2017.

[7] W. Zhu, Y. Zhuang, and L. Zhang, "A three-dimensional virtual resource scheduling method for energy saving in cloud computing," *Future Generation Computer Systems*, vol. 69, pp. 66–74, 2017.

[8] R. Jhawar, V. Piuri, and P. Samarati, "Supporting security requirements for resource management in cloud computing," in *Proc. 2012 IEEE 15th International Conference on Computational Science and Engineering (CSE)*, IEEE, 2012, pp. 170–177.

[9] X.-S. Yang, *Nature-Inspired Metaheuristic Algorithms*, Luniver press, 2010.

[10] Firefly algorithms for multimodal optimization, presented at International Symposium on Stochastic Algorithms, Springer, 2009, pp. 169–178.

[11] X.-S. Yang and X. He, "Firefly algorithm: Recent advances and appli-cations," *International Journal of Swarm Intelligence*, vol. 1, no. 1, pp. 36–50, 2013.

[12] N. J. Kansal and I. Chana, "Energy-aware virtual machine migration for cloud computing-a firefly optimization approach," *Journal of Grid Computing*, vol. 14, no. 2, pp. 327–345, 2016.

M. Samir Abou El-Seoud received his BSc degree in physics, electronics and mathematics from Cairo University in 1967, his higher diplom in computing from Technical University of Darmstadt (TUD) / Germany in 1975 and his doctor of science from the same University (TUD) in 1979. Currently, his research interests are focused on computer aided learning, parallel algorithms, mobile applications, augmented reality, cloud computing, IoT, numerical scientific computations and computational fluid mechanics. Professor El-Seoud helds different academic positions at TUD Germany. Letest full-professor in 1987. outside Germany professor El-Seoud spent different years as a full-professor of computer science at SQU – Oman, Qatar University, and PSUT-Jordan and acted as a head of computer science for many years. At industrial institutions, Professor El-Seoud worked as Scientific Advisor and Consultant for the GTZ in Germany and was responsible for establishing a postgraduate program leading to M.Sc. degree in Computations at Colombo University / Sri-Lanka (2001 – 2003). He also worked as Application Consultant at Automatic Data Processing Inc., Division Network Services in Frankfurt/Germany (1979 – 1980). Professor El-Seoud joined The British University in Egypt (BUE) in 2012. Currently, he is Basic Science Coordinator at the Faculty of Informatics and Computer Science (ICS) at BUE. Professor El-Seoud has more than 150 publications in international proceedings and international reputable journals.

Ahmad Mostafa is an assistant professor at The British University in Egypt. He graduated with his Ph.D. from the the Center for distributed and mobile computing at the University of Cincinnati, Ohio. His research interests include routing, energy conservation and localization in wireless sensor networks, VoIP and physical implementation over wireless mesh network, QoS in vehicular networks as well as heterogeneous networks. His research experience is reinforced by strong passion for education as he has taught various courses at the British University in Egypt, the University of Cincinnati and other universities in the US over the past eight years.