*Computer Systems Science & Engineering*
DOI:10.32604/csse.2021.014974
***Review***

Tech Science Press

# A Review of Energy-Related Cost Issues and Prediction Models in Cloud Computing Environments

**Mohammad Aldossary**[*]

Department of Computer Science, College of Arts and Science, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
[*]Corresponding Author: Mohammad Aldossary. Email: mm.aldossary@psau.edu.sa

Received: 30 October 2021; Accepted: 28 November 2020

**Abstract:** With the expansion of cloud computing, optimizing the energy efficiency and cost of the cloud paradigm is considered significantly important, since it directly affects providers' revenue and customers' payment. Thus, providing prediction information of the cloud services can be very beneficial for the service providers, as they need to carefully predict their business growths and efficiently manage their resources. To optimize the use of cloud services, predictive mechanisms can be applied to improve resource utilization and reduce energy-related costs. However, such mechanisms need to be provided with energy awareness not only at the level of the Physical Machine (PM) but also at the level of the Virtual Machine (VM) in order to make improved cost decisions. Therefore, this paper presents a comprehensive literature review on the subject of energy-related cost issues and prediction models in cloud computing environments, along with an overall discussion of the closely related works. The outcomes of this research can be used and incorporated by predictive resource management techniques to make improved cost decisions assisted with energy awareness and leverage cloud resources efficiently.

**Keywords:** Cloud computing; cost models; energy efficiency; power consumption; workload prediction; energy prediction; cost estimation

## 1 Introduction

Cloud computing is a significant business model which has revolutionized the Information Technology (IT) industry by providing different services such as Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) for the cloud customers with reasonable prices based on their usage (e.g., *pay-as-you-go* model). However, the widespread adoption of cloud computing and the increasing number of cloud customers has raised the overall operational costs for cloud providers [1–5]. Thus, reducing the operational costs of different cloud services is an active area of research.

The cost mechanisms used by various cloud service providers are directly impacting the adoption of cloud computing within the IT industry. In this regards, the cost mechanisms provided by cloud service providers have become advanced as consumers are paid monthly, weekly, daily, minutely or secondly depending on the services and resources they used [6–8]. However, there are also limitations, since customers are paid for the services and resources they used based on predefined tariffs. Such predefined tariffs do not take into account the variable energy costs [9,10]. With the continued rising and the large fluctuations in electricity prices [11], cloud providers perceive energy consumption as one of the main operating cost factors to manage within their infrastructure [1–3]. Modelling a new cost mechanism for cloud services that can be tailored to the energy costs has therefore drawn a lot of researchers' attention [1–3,12].

In a cloud, each Physical Machine (PM) can run one Virtual Machine (VM) or multiple VMs simultaneously. Such VMs can be homogeneous or heterogeneous based on their features like the number of virtual CPUs (vCPUs) and the memory size, which also consume the energy differently. Thus, these parameters need to be taken into account when modelling and calculating the overall cost of the VMs, along with their power consumption.

The aim of this research is to investigate the energy-related cost issues and prediction models as well as the impact of resource heterogeneity in cloud computing environments. The outcomes of this research can be used and incorporated by *predictive* resource management techniques to make improved cost decisions assisted with energy awareness and leverage cloud resources efficiently.

The remainder of this paper is organized as follows: Section 2 presents the fundamental concepts of cloud computing with a detailed description of its definition, system architecture, services types, deployment types. A detailed description of cloud computing pricing models is also presented. This is followed by positioning the work in the relevant literature, focusing on energy-related cost issues in cloud computing, along with an overall discussion of the closely related works as presented in Section 3. Section 4 discusses the prediction models related to the workload, energy consumption and cost of cloud services, as well as an overall discussion of the closely related works. Section 5 concludes this paper.

## 2 Overview

### *2.1 Cloud Computing*

Cloud computing is defined by the National Institute of Standards and Technology (NIST) as: "*a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction*" p.2, [13].

According to the NIST definition, there are five main characteristics of cloud computing [13]. *On-demand self-service*: the ability of cloud providers to provision computing resources to their customers as needed without requiring human interaction. *Broad network access*: through standard mechanisms, the cloud customers can access their resources over the network. *Resource pooling*: the cloud providers have a pool of computing resources to serve different customers using (e.g., a multi-tenant model). *Rapid elasticity*: the capacity of cloud resources can be more flexible and rapidly provisioned. *Measured service*: the resource utilization is monitored, automatically measured and optimized.

### *2.2 System Architecture*

Zhang et al. [14] categorized the cloud computing architecture into four layers, namely hardware, infrastructure, platform and application layers, as indicated in Fig. 1. At the bottom of this architecture is the *hardware layer* where the cloud physical resources (e.g., routers, servers, switches and cooling systems) are managed within cloud data centers [14]. On top of the hardware comes the *infrastructure layer*, which also known as virtualization layer. The infrastructure layer consists of a pool of virtualized computing resources through the use of virtualization technologies such as KVM [15], Xen [16] and VMware [17]. On top of the infrastructure layer, the operating systems are included in the *platform layer*, which provides the environment to deploy the applications in virtual instances. Finally, the *application layer* sits on top of the architecture which consists of the actual cloud applications.

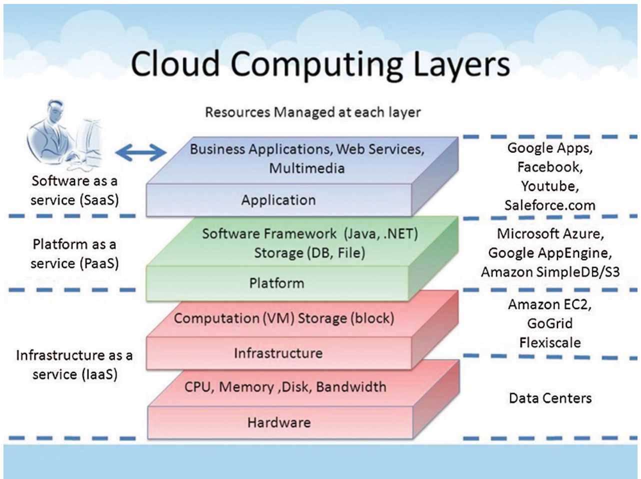
Cloud Computing Layers
Resources Managed at each layer
Software as a service (SaaS)
Business Applications, Web Services, Multimedia
Application
Google Apps, Facebook, Youtube, Saleforce.com
Platform as a service (PaaS)
Software Framework (Java, .NET) Storage (DB, File)
Platform
Microsoft Azure, Google AppEngine, Amazon SimpleDB/S3
Computation (VM) Storage (block)
Infrastructure
Amazon EC2, GoGrid Flexiscale
Infrastructure as a service (IaaS)
CPU, Memory ,Disk, Bandwidth
Hardware
Data Centers

**Figure 1:** Cloud computing architecture [14]

### *2.3 Services Types*

With reference to the cloud architectural layers shown in Fig. 1, there are three main types of cloud services. *Software as a Service (SaaS):* this service provides applications and software programs, in addition to interfaces for the customers. *Platform as a Service (PaaS):* With this type of cloud service, the customer has the ability to deploy and generate cloud applications using programming languages, services, libraries and tools supported by cloud providers. *Infrastructure as a Service (IaaS):* In this service, hardware devices and infrastructure are virtualized and offered as a service (e.g., VMs), which also called instances. Furthermore, *Everything as a Service (XaaS)*: Where X is everything that can be described as a new type of cloud services, such as desktop, network, storage, hardware, security, communication, virtualization, data and business [18].

### *2.4 Deployment Types*

Cloud computing can be deployed through many models. For example, *public cloud:* is owned by a service provider offering services and computational resources to organizations and individuals. A public cloud allows customers access to the cloud through the internet and the customers only pay for the time period that they utilize the service, (e.g., using a pay-per-use model) [14]. *Private cloud:* Which can also be named an internal cloud or corporate cloud, is usually hosted and managed by the company itself. Security is improved in a private cloud as only the company users have access to the provided services. *Hybrid cloud:* This is a composition of private cloud and public cloud. In this type of deployment, a private cloud is connected to one or more external cloud services. *Community cloud:* Is a deployment model that is shared between several organizations in order to meet specific requirements that difficult to achieve in a public cloud (e.g., security requirements, policy, and compliance considerations).

### *2.5 Pricing Models in Cloud Computing*

Cloud service providers offer different types of services to their customers with different pricing models. The strategy of pricing models in clouds can be categorized as 1) *fixed pricing*: When the price of the services doesn't change (flat fees) and determined by the provider, and 2) *variable pricing*: When the price of the services is dynamically changed based on the market supply and demand [19]. Thus, the price of each cloud service will be based on the chosen type of pricing model.

The most popular cloud service providers (e.g., Amazon EC2 [6], Microsoft Azure [7] and Google Cloud [8]) have three common types of pricing models, which are *subscription*, *on-demand* and *auction* pricing models. These pricing models are discussed as follows:

- ***A subscription-based pricing model***: This type of model allows customers to pay a fixed price in advance for a defined period of time, usually monthly or yearly (e.g., reserved instances provided by Microsoft Azure). Typically, customers are paying lower prices with long-term contracts because that can help cloud providers estimate their infrastructure expenses [20]. With this type of pricing model, cloud providers attract more customers' by offering a discount rate and ensuring that their resources will be available at any time they want [19].
- ***A demand-based pricing model***: There are no long-term commitments with this type of pricing strategy, which enables customers to pay service fees on a time-based, usually per hour or second (e.g., pay-as-you-go and on-demand pricing models provided by Google Cloud and Amazon EC2, respectively). Pay-as-you-go model is ideal for companies that cannot pay upfront or estimate the computing resources they needed. The price is set according to the size of the instances and their resources. For example, the instances that do not involve Graphics Processing Units (GPUs) or lots of Central Processing Units (CPUs) or Solid-State Drive (SSD) based storage, will automatically be cheaper since they are not used for high performance [19,21]. Furthermore, a hybrid pricing model is presented by Jelastic plans [22], which is an intermediate model between subscription and on-demand with charged on an hourly basis. In this model, the customers can set a minimum number of resources to be reserved for an application and get a discount rate accordingly, as well as it allows the customers to set maximum limits of resources in case if the application demand increases.
- ***An auction-based pricing model***: The concept of an auction pricing model is based on selling cloud services idle time, which allows customers to bid for services, and cloud providers are entitled to accept or reject the offers. For instance, Amazon EC2 spot instances [23] allow customers to bid on a spare Amazon EC2 computing capacities. Also, customers can view the spot instance price history for the last 90 days to determine which bid price they should offer [23]. Thus, if the customer's bid exceeds or meets the current bid price, the customer can access the resources. Contrarily, if the customer's bid is overridden, the customer gives the resources back. The prices of the auction-based model compared to subscription and on-demand models are significantly lower. However, if a customer loses a bid, these resources can be taken away, which make it not suitable for businesses [19].

In cloud environments, the majority of the costs are derived through resource usage, which can be defined as the resource capacity that required to run applications on the cloud infrastructure. However, not all the costs are related to the resource usage of infrastructure, there are further additional costs. For example, the costs that are associated with software licenses, IT support, cooling and maintenance. These costs are difficult to measure or estimate due to the differences between cloud service providers. Besides, the current pricing models do not provide details of the energy consumed by the offered services. Thus, in order to effectively contribute to the overall business model and offer transparent pricing to the customers, cloud service providers need to take energy consumption into consideration when developing their pricing models [10,24,25]. Therefore, only the costs of the cloud infrastructure that can be calculated through resources along with their energy consumption are considered in the scope of this paper.

## 3 Energy-Related Cost Issues in Cloud Computing

Many cloud service providers such as Amazon [6], Microsoft Azure [7] and Google Cloud [8] have allowed the customers to run their applications in clouds. They therefore have established cost models in order to charge their customers based on the offered services. Although many pricing models in the IaaS are already proposed (e.g., subscription, on-demand, and auction models), there are still inevitable to suffer from wasted payment and resources when using these types of pricing models [10,26,27]. In fact, cost modelling is a critical component of the cloud computing paradigm since it directly affects providers' revenue and customers' payment [26,27]. Thus, designing an appropriate and precise cost model which can make both providers and customers satisfied is considered as a vital concern in a cloud environment.

Furthermore, cloud data centers continue to consume huge amounts of energy and have a major impact on environmental and operational costs caused by this high energy consumption [28]. With the increasing electricity costs for cloud data centers, energy consumption has become one of the major operational cost issues for cloud providers to maintain [29]. In 2013, cloud computing consumed about 684 billion Kilowatt-Hour (kWh) of electricity [30], while the increase in energy consumption is estimated to be around 60% or even more by 2020 [30]. Yet, most of cloud service providers charge their customers for the offered services on a time-based without considering the actual cost of energy consumption [31]. Due to the economic impact of cloud data centers' energy consumption, cloud providers should consider the actual cost of energy consumption when designing their cost models for the offered services [10]. In addition to that, cloud customers cannot affect or know in any way the amount of energy that they consume for running the cloud services. Therefore, they need to be aware of their energy usage, which can help to change their behavior, for example by shutting down, consolidating VMs, or running applications that are energy efficient. In the following subsections, some of the research conducted on cloud cost models to reduce the cost and energy consumption in the IaaS cloud environment are discussed.

### *3.1 Cost Models*

Cloud cost modelling is a challenging issue as the increasing number of business are moving their computation workloads to clouds. Although many public cloud providers are already used the (*pay-as-you-use*) model to charge their customers for the offered services, the customers still usually pay more than what they are actually used [26,27]. Therefore, the work by Belli et al. [32] explored the area of cost models, that allows the customers to optimize their choice of IaaS cloud providers in terms of the offered price. They presented a Cost-Optimized Cloud Application Placement Tool (COCA-PT) based on a Resource Consumption Model (RCM). The main goal of their work is to optimize the placement of customers applications based on the price offered by different cloud providers. However, as mentioned by the authors, the proposed tool is not completed yet and needs a further extension. Moreover, their cost model does not take into account the power consumption consumed by the running applications.

Jin et al. [26] designed a fine-grained fair pricing model to improve the resource utilization and reduce partial usage waste problem. They investigated the optimization of the trade-off between the proposed model and various overheads (e.g., VM maintenance and billing cycle). The model is evaluated using two large-scale production traces (Grid workload archive and Google data center) and the experimental result show that the proposed model can significantly improve social welfare (e.g., increasing provider revenues and reducing customers costs). Although the authors have been focused on the design of precise pricing model that can satisfy both customers and providers, their approach has not shown the impact of the power consumption of the used resources on cloud pricing models.

Moreover, Berndt et al. [31] presented a hybrid IaaS pricing model, tackling the problem of overbooking and dual selling capacity by cloud providers to maintain profitability, which would impact efficiency and cloud adoption. To explain, this pricing model charges on the basis of a flat-rate that guarantees the customers with some level of performance and on a flexible that charges for resource utilization beyond the flat-rate section. Their approach needs only performance measurement on one side and resource utilization measurement on the other, as indicated in their work [31]. Yet, their approach is still limited in the essence that it does not consider the actual cost of energy consumption.

Mao et al. [33] presented a cost-aware auto-scaling mechanism for scheduling tasks in clouds, which called Scaling-Consolidation-Scheduling (SCS). The auto-scaling mechanism takes into consideration the instantiation time that every VM needs to be running, then the Earliest Deadline First (EDF) algorithm is used to schedule tasks on each VM. They primarily focus on minimizing the cost of the VMs and satisfying their performance requirement based on tasks deadline constraints. This is achieved by forcing the tasks to run on the same VM in order to improve performance and save the data transfer cost. They compare the proposed SCS approach with two cost-based approaches, and the results demonstrate that their approach achieved cost-savings of 9.8–40.4% along with improved utilization over other approaches. However, this approach only ensures a reduction in the cost of each VM and does not take into account the trade-off between performance and power consumption of the selected VMs.

Further, a cost-aware super professional executor (Suprex) with auto-scaling mechanism is proposed by Aslanpour et al. in [34]. This approach aims to provide an executor with the capability to isolate the overloaded VM until the billing period is completed, which leads to overcome the challenge of postponed VM start-up and maximize the cost efficiency. The results show that the Suprex executor can reduce the cost of VM by 7%, but in some cases this executor leads to lower resource utilization.

Chard et al. [35] proposed an approach for cost-aware elastics resource provisioning for scientific workloads. This approach monitors a job submission queue and provisions VMs based on pre-defined policies. The authors investigate the impact of workload execution on the total cost of cloud services by using dynamic pricing models (e.g., spot and on-demand instances) provided by Amazon Web Services (AWS) [23], based on different availability zones. They evaluate their approach under realistic conditions based on workload traces through simulation. However, their investigation does not consider the impact of energy consumption on AWS pricing models.

### *3.2 Cost and Energy Consumption Models*

With the expansion of cloud computing, optimizing the energy efficiency of the cloud paradigm at all different layers is considered significantly important, as highlighted by Djemame et al. [36]. The authors have proposed a cloud architecture that enables energy awareness at all layers of the cloud stack and through the cloud application lifecycle. This architecture is an energy efficient solution, capable of self-adaption and aware of the impacts on other quality characteristics such as cost and performance of the applications. An example of a cost model for IaaS providers to align with the energy consumption cost is introduced by Hinz et al. in [10]. They proposed a cost model called Proportional-Shared Virtual Energy (PSVE), which investigates the relationship between energy consumption and VMs workload in a cloud environment. The PSVE model considers the cost of heterogeneous VMs as well as their energy consumption, which is based on the number of allocated virtual CPU to each VM. Also, it consists of two main elements: 1) A cost associated with VMs resources (e.g., CPUs and networks) along with their power consumption, and 2) A shared cost associated with the hypervisor, relatively distributed among VMs. Nevertheless, their model does not consider the actual utilization of the virtual CPUs, only considers the number of allocated virtual CPU to each VM, thus their cost model may not be an accurate, as stated by [24,25]. In this context, the current cost models offered by cloud service providers (e.g., Amazon EC2 [6]) only consider the number of allocated resources to each VM based on the time of usage, and do not consider the utilization ratio of these resources (actual usage). To illustrate that, let's consider two VMs (VM1 and VM2) allocated on the same host and have the same number of virtual CPUs. VM1 and VM2 used 10% and 90% of the CPU utilization, respectively. These ratios of CPUs utilization have different impacts on energy consumption, but both are usually charged the same price, regardless if a VM is using 10% or 90% of its CPU. Consequently, the authors in [24,25] highlighted the need of cloud service providers to offer cost models that fairly charge their customers based on the actual resource usage with consideration of their energy consumption.

Wang et al. [21] argued the importance of having precise cost models for adopting cloud computing. Through their investigation, they found that different system configurations have a significant impact on energy consumption and thus the total cost of cloud services. Consequently, Yousefipour et al. [37] proposed an energy and cost-aware VM consolidation model that aims to minimize the number of active PMs in order to reduce power consumption and cost of heterogeneous cloud data centers. The consolidation process is nearly optimized based on the trade-off between power consumption and cost using a mixed-integer non-linear programming model and a genetic algorithm. The results show that the proposed model is capable of reducing the power consumption and costs when compared to the First Fit (FF), First Fit Decreasing (FFD), and Permutation Pack (PP) algorithms. However, they assume the power consumption is increasing linearly for all PMs, which is not usually the case in the heterogeneous cloud environment, as stated in [38]. Also, this work does not consider the energy consumption overhead incurred by VMs consolidation. Similarly, in [39] authors proposed a cost and energy efficient scheduling algorithm based on Particle Swarm Optimization (PSO). This algorithm aims to optimize execution cost and energy consumption of cloud data centers, considering deadline constraint and time. The proposed algorithm is evaluated using CloudSim [40] based on independent tasks scheduling and compared with honey bee and min-min algorithms. Nevertheless, this work lacks to consider other Quality of Services (QoS) parameters such as application performance variations, load balancing, availability and Service Level Agreement (SLA) violation.

Further, Jung et al. [41] introduced *Mistral*, a holistic framework that balances power usage, the efficiency of applications and transient power costs incurred by the adaptation decisions of the framework. Their approach investigates the problem of dynamic consolidation of homogeneous VMs and focuses on improving the power consumption of the physical host. However, this framework does not optimize the trade-off between all mentioned objectives (power consumption, application performance and costs), only two of these objectives are considered to be optimized at the same time.

### *3.3 Overall Discussion of Energy-Related Cost Issues*

Cost modelling is an important component of the cloud computing paradigm since it directly affects providers' revenue and customers' payment [26,27]. The main aim for cloud providers is to achieve maximum revenue and for cloud customers to achieve the highest service performance at a reasonable price. Current cost models used by cloud service providers (e.g., Amazon EC2 [6] and Microsoft Azure [7]) are based only on the usage of the virtualized resources such as CPU, memory, and disk, and do not consider the variable cost of energy consumed by these resources. With the rising of electricity costs, cloud providers perceive energy consumption as one of the most important operational cost factors to be maintained within their infrastructures [1–3].

In order to properly alleviate the operational cost, cloud service providers can be assisted with cost and energy awareness to enhance their decisions and efficiently manage cloud resources. Section 3 has reviewed the related work on modelling the cost as well as the energy consumption in cloud environments. As discussed in Section 3.1, the work presented in [26,31–33,35] aimed to improve the cost efficiency in cloud environments in order to meet the performance requirements, customers' demands and efficient resource utilization, but not considering the energy consumption of the resources. In Section 3.2, the work presented in [37,41] considered the energy consumption in their models, but their focus is only at the physical level in order to consolidate the VMs and minimize the number of active hosts. Only the work presented in [10] considered the energy consumption at both physical and virtual levels, though this is still limited as their model only consider the number of allocated virtual CPU to each VM.

Thus, there is a clear need to consider energy consumption at VMs level, taking into account the actual utilization of the virtual CPUs in order to obtain a precise cost model. The following Tab. 1 provides a comparison summary of the closely related works on modelling cost and energy consumption for VMs in a cloud environment.

**Table 1:** Summary of existing cost and energy models

| By Criteria | Cost model based on VMs resource utilization consideration | Actual power consumption consideration | |
|---|---|---|---|
| | | PMs level | VMs level |
| [32] | Homogeneous VMs only. | Not considered. | Not considered. |
| [26] | Homogeneous VMs only. | Not considered. | Not considered. |
| [31] | Homogeneous VMs only. | Not considered. | Not considered. |
| [33] | Homogeneous and heterogeneous VMs. | Not considered. | Not considered. |
| [35] | Homogeneous and heterogeneous VMs. | Not considered. | Not considered. |
| [37] | Homogeneous and heterogeneous VMs. | Homogeneous PMs only. | Not considered. |
| [41] | Homogeneous VMs only. | Homogeneous PMs only. | Not considered. |
| [10] | Homogeneous and heterogeneous VMs. | Homogeneous PMs only. | Homogeneous and heterogeneous VMs, but only based on the number of allocated virtual CPUs to each VM. |

## 4 Prediction Models in Cloud Computing

Having discussed the existing work on modelling the cost and energy consumption of cloud services in Sections 3.1 and 3.2, this section discusses the work on predicting the workload, energy consumption and estimating the total cost of the VMs during the service operation.

Providing prediction information of the cloud services ahead of their operation or at the run-time can be very beneficial for the service providers, as they need to carefully predict their business growths and efficiently manage the cloud resources.

To optimize the use of cloud services, *predictive* mechanisms can be applied to improve resource utilization and reduce energy costs while maintaining demands on service performance. However, in order to make improved cost decisions, these mechanisms need to be assisted with energy awareness not only at the level of the PM but also at the level of the VM. Furthermore, estimating the potential costs of cloud services can help cloud service providers deliver suitable services that meet the needs of their customers.

### *4.1 Workload Prediction*

In terms of workload prediction, a number of methods are used in order to predict the workload in cloud environments. For example, an evaluation of commercial cloud services offered by major service providers is provided in [42], where a cloud monitoring tool is used to measure the service performance of a month period for 20 cloud providers. According to the workload data collected from different cloud providers, they applied the Auto-Regressive Integrated Moving Average (ARIMA) and Exponential Smoothing (ETS) to predict the future behavior of service performance. This prediction helps cloud customers and service brokers to select cloud services according to their requirements. The overall performance prediction results show that ARIMA performs better than ETS for predicting service performance. Also, a predictive elastic resource scaling scheme (PRESS) is introduced for cloud systems in [43]. To predict the workloads of PMs and VMs, the approach uses a short-term pattern matching and state-driven approach (Markov chain). This approach is deployed on top of Xen [16], using RUBiS and an application load traces from Google. In their work, only the workload is predicted as a stand-alone application.

Moreover, Huang et al. [44] proposed an elastic resource allocation mechanism for a cloud system, namely Prediction-based Dynamic Resource Scheduling (PDRS). The PDRS is employed to predict the VMs workload fluctuations using the ARIMA model based on the historical workload data. Based on the predictor, they developed dynamic resource allocation algorithms along with VMs live migration in order to reduce the number of active PMs. The results show that this approach is able to realize adaptive resource allocation with an acceptable effect on SLAs and migration overhead. Though this approach is focused on predicting the workload in order to perform the VMs allocation and live migration, without considering the energy overhead due to the migrations of the VMs.

Farahnakian et al. [45] introduced a predictive VM consolidation approach, called Utilization Prediction-aware VM Consolidation (UP-VMC). The UP-VMC aims to optimize three objectives include the number of SLA violations, the number of VM migrations and energy consumption. It considers the current and future PMs and VMs resource utilization in order to migrate VMs into the least number of active PMs, and then switch the idle PM to the sleep mode in order to minimize the energy cost. The future resource utilization (CPU and memory) is predicted using two regression-based prediction models (Linear and K-Nearest Neighbor). The obtained results using Google cluster and PlanetLab workload traces show that the UP-VMC can reduce SLA violations, energy consumption and the number of migrations. However, the experiments conducted on a simulation-based have focused on predicting the PMs and VMs resource utilization and do not consider the prediction of PMs and VMs energy consumption.

Zhang et al. [46] presented a proactive virtual resource management framework, called (PRMRAP), which predicts the amount of resource needed to cope with unexpected workload changes. This approach uses the ARIMA model based on the historical workload data in order to predict the VMs workload changes and the number of resources needed. In this framework, they consider both vertical and horizontal scaling of the VMs, which can reduce time latency for handling the workload changes in a cost-efficient manner. Likewise, Fang et al. [47] presented a novel Resource Prediction and Provisioning Scheme (RPPS), which predicts the workload demands and dynamically adjusts resource provision for cloud applications. This approach takes advantage of the ARIMA model which has high prediction accuracy in order to handle the resource provisioning in a short period of time. They implemented the RPPS model on top of Xen [16] and KVM [15] virtualization platforms, and conducted the experiments in a real cloud data center. The results show that this approach has high prediction accuracy of about 90% and able to scale cloud resources under different situations (e.g., peak and low phases). Further, Yang et al. [48,49] used a Linear Regression Model (LRM) to predict the VMs workload for the next time interval. An auto-scaling mechanism is proposed based on the predicted workload to scale the virtual resources, which combines the real-time scaling and the pre-scaling in order to handle the workload demands. They used the knowledge from workload prediction to select the number of resources needed for scaling, considering both horizontal and vertical scaling. According to the experiment results, this approach is able to predict the VMs workload while lowering the scaling costs and SLA violations. However, all these approaches do not consider or predict VMs energy consumption when performing dynamic resource provisioning (scaling decisions).

Also, it is worth mentioning that the workload prediction modelling requires a quantitative evaluation and statistical analysis in relation to the characteristics of the workloads in terms of their length, pattern, and resource consumption. Thus, modelling the relationships between these different workload characteristics is important in order to achieve accurate and reliable prediction results.

### *4.2 Energy Prediction*

There are many ongoing research projects focusing on the prediction of energy consumption based on resource utilization. For example, Bircher et al. [50] proposed an approach to estimate the power consumption of a complete system using microprocessor performance counters. They developed power models for subsystems (e.g., CPU, memory, disk, and network) on two platforms (server and desktop). Also, synthetic workloads were generated in order to control the utilization of the subsystems. They performed a correlation analysis between the performance counters and the power consumption using linear and polynomial regression techniques. The average error of their models was 14.1% for the memory controller and less than 9% for each subsystem. Similarly, McCullough et al. [51] have evaluated the competence of existing predictive power models based on their accuracy using hardware performance counter for modern hardware architectures. A number of linear and non-linear regression models are compared. For the linear regression models, the results show that these models provide a reliable accuracy with low computational complexity for a single-core scenario. In contrast, the non-linear models provided better accuracy with the multi-core scenario, although they incur a higher computational complexity. However, these approaches are performed on non-virtualized environments and thus do not consider or support the power consumption of the virtual resources.

Smith et al. [52] proposed a power monitoring tool for software-based, called CloudMonitor. The authors argued that such a tool can be used in order to create energy-efficient applications as well as design energy-based cost models. The results show that the power monitoring tool is able to estimate PM power consumption for different applications as long as the physical hardware has the same configuration. However, this tool does not support the heterogeneity of the PMs as well as the estimation of VMs power consumption.

Kistowski et al. [53] introduced a model for predicting the power consumption of physical hosts at run-time. This approach makes use of run-time monitoring data to train the model and then predict the power consumption based on load intensity and performance counters. The authors claimed that this approach can be used with any performance model to optimize the energy efficiency of distributed systems. They evaluated the model using two different web applications deployed in a heterogeneous environment. The results show that this approach can predict the power consumption of a system with an error of 2.21%. Yet, their approach only considers the prediction of the power consumption at the PMs level and does not consider the prediction at the VMs level.

Further, Makaratzis et al. [54] conducted a survey study on energy modelling in cloud simulations. They focused on the energy models that have been proposed for the prediction of the energy consumption of cloud data centers. The most popular cloud simulation frameworks were considered in this survey: CloudSched, CloudSim, DCSim, GDCSim, GreenCloud and iCanCloud. Hence, the experiments were conducted in order to compare these different simulations with their energy models, and the results show that the same tendency prevails for the energy models in all cloud simulation frameworks. However, these simulations along with their energy models do not consider the impact of heterogeneous VMs on the energy consumption in cloud data centers.

Li et al. [55] have built an online power metering model that estimates the power consumption for the PMs and VMs in a cloud environment. The power modelling is performed using a linear regression technique based on the impact of the CPU, memory and disk. The implementation of the model shows that it can achieve an average estimation accuracy of more than 96% with low runtime overhead. Nevertheless, they assumed that all the PMs and VMs are homogeneous, which is very rarely used in cloud environments.

Moreover, Farahnakian et al. [56] introduced a load prediction method, called a Linear Regression-based CPU Utilization Prediction (LiRCUP). This method is used to predict the short-time future CPU utilization of the overloaded and underloaded PMs based on historical data of each PM. Based on this prediction, some VMs are migrated to other hosts in order to avoid SLA violations and reduce energy costs. In order to evaluate this work, the authors implemented the proposed method in the CloudSim and the results show that the proposed method can reduce the energy cost and SLA violation rate. However, this work is focused on predicting the workload and then the energy consumption only at the host level and not considering the workload and energy prediction at the VM level.

Subirats et al. [57] proposed a VM placement algorithm, which is aimed to take the appropriate decisions (e.g., VM replication, migration, cancellation). Generally, mathematical modelling is used to design a CPU utilization predictor in order to predict the energy consumption for different workload types at PMs and VMs levels. Their proposed predictor consists of four prediction models, namely linear regression, moving average, single and double exponential smoothing in order to predict CPU utilization and power consumption of a given VM. Although this work only considers a linear relationship between the CPU utilization and the power consumption, other non-linear relationships such as polynomial and exponential could be considered in order to increase the prediction accuracy.

### *4.3 Cost Estimation*

Estimating the cost of resource provisioning is essential to automatically cope with workload demands. Therefore, Jiang et al. [58] presented an online temporal data mining system, called A Self-Adaptive Prediction (ASAP), which is used to predict the VM demands, and provision resources accordingly. The authors also proposed a cloud prediction cost mechanism, which is used to measure the performance of several prediction models based on historical time series data. The experiments results show that the ASAP is capable to decrease the resource provisioning time of all VMs. Another approach for an efficient auto-scaling is proposed in [59]. They used a second order Auto-Regressive Moving Average (ARMA) model in order to predict the VMs workload and cost for the next time interval based on historical workload data. This look-ahead approach enables early auto-scaling detection, which allows the new VMs to boot (horizontal scaling) before workload increases. The model aims at reducing the use of resources and fulfilling QoS requirements while maintaining low operating costs. However, these two approaches only consider workload prediction for dynamic resource provisioning, and do not consider the energy consumption which would influence the overall cost of the scaling decisions.

Sharma et al. [60] proposed a cost-aware resource provisioning framework for cloud applications, called Kingfisher. It aims to optimize the cost of resource provisioning and reconfiguring using Integer Linear Program (ILP) formulation. Kingfisher exploited both scaling and migration mechanisms to dynamically select appropriate decisions that optimize the cost incurred by customers. In their work, the ARIMA model is employed to estimate the workload in order to capture future workload trends. They implemented the Kingfisher framework using the OpenNebula [61] cloud platform, and the results demonstrate that the Kingfisher has the ability to select the lowest cost of resource provisioning and reconfiguring to meet an application’s requirements. Nevertheless, their approach does not consider the energy consumption overhead when performing the migration and scaling decisions.

Furthermore, Liu et al. [62] designed performance and energy models to estimate VM migration cost based on theoretical analysis and empirical studies on the Xen platform. The theoretical analysis and empirical studies show that the migration-related parameters like VM memory size, memory dirtying rate and network speed are the major factors impacting migration performance in terms of migration time, migration downtime and the total volume of network traffic. Also, they designed a linear regression model and a theoretical model to estimate the energy consumption of the networks during VM migration based on their performance model. The experimental results demonstrate that the proposed models are able to estimate VM migration cost with an estimation accuracy of about 90% based on performance and energy metrics. However, this work does not consider the heterogeneity of the PMs or the VMs when designing their models.

### *4.4 Overall Discussion of Prediction Models*

Cloud service providers can take advantage of prediction models to enhance the efficiency of managing cloud resources. With the unexpected workload demands, cloud service providers should strike a balance between their operating costs, energy consumption and satisfying QoS objectives. Consequently, modelling a predictive mechanism can be beneficial to improve resource utilization and reduce energy-related costs, while maintaining service performance requirements.

Section 4 has reviewed the related work on predicting the workload and energy consumption as well as estimating the total cost of the VMs during the service operation. As discussed in Section 4.1, the work presented in [43–49] aimed to predict the workload in order to improve resource utilization in cloud environments, but without considering the energy consumption of the predicted workloads.

In Section 4.2, the work presented in [52,53,56] considered the prediction of energy consumption in their models, but these approaches only consider the prediction of the power consumption at PMs level and do not consider the prediction at VMs level. Only the work presented in [55,57] considered the prediction of energy consumption at both physical and virtual levels. Though there are still limited as the model in [55] assumed that all the PMs and VMs are homogeneous, whereas, the model in [57] only considers a linear relationship between the CPU utilization and the energy consumption in order to predict the power at the VMs level.

The work presented in [58,59,60] considered the prediction of workload and the estimation of cost for the VMs, but do not consider the energy consumption which would influence the overall cost estimation of cloud services, as discussed in Section 4.3. The only work that considered the prediction of workload and energy consumption as well as the estimation of cost, is presented in [62]. However, this work does not consider the heterogeneity of the PMs or the VMs when designing their models.

Thus, there is still a need for predictive modelling that takes into account the workload, energy consumption and cost not only at the PMs level, but also at the VMs level considering their heterogeneity, in order to make enhanced cost decisions and efficiently manage cloud resources.

The following Tab. 2 provides a comparison summary of the closely related works on prediction models that consider workload, energy consumption and cost for VMs in a cloud environment.

**Table 2:** Summary of prediction models

| By | Criteria | Workload prediction consideration | | Energy prediction consideration | | Cost estimation consideration |
|---|---|---|---|---|---|---|
| | | PMs level | VMs level | PMs level | VMs level | |
| [43,44] | | Homogeneous PMs only. | Homogeneous VMs only. | Not considered. | Not considered. | Not considered. |
| [45] | | Homogeneous and heterogeneous PMs. | Homogeneous and heterogeneous VMs. | Not considered. | Not considered. | Not considered. |
| [46] | | Not considered. | Heterogeneous VMs. | Not considered. | Not considered. | Not considered. |
| [47] | | Homogeneous PMs only. | Not considered. | Not considered. | Not considered. | Not considered. |
| [48,49] | | Not considered. | Heterogeneous VMs. | Not considered. | Not considered. | Not considered. |
| [52] | | Not considered. | Not considered. | Homogeneous PMs only. | Not considered. | Not considered. |
| [53] | | Not considered. | Not considered. | Heterogeneous PMs. | Not considered. | Not considered. |
| [55] | | Not considered. | Not considered. | Homogeneous PMs only. | Homogeneous VMs only. | Not considered. |
| [56] | | Heterogeneous PMs. | Not considered. | Heterogeneous PMs. | Not considered. | Not considered. |
| [57] | | Heterogeneous PMs. | Homogeneous VMs only. | Heterogeneous PMs. | Homogeneous VMs only. | Not considered. |
| [58] | | Heterogeneous PMs. | Heterogeneous VMs. | Not considered. | Not considered. | Based on the resource usage. |
| [59] | | Not considered. | Homogeneous VMs only. | Not considered. | Not considered. | Based on the resource usage. |
| [60] | | Heterogeneous PMs. | Homogeneous VMs only. | Not considered. | Not considered. | Based on the resource usage. |
| [62] | | Homogeneous PMs only. | Homogeneous VMs only. | Homogeneous PMs only. | Homogeneous VMs only. | Based on the resource usage and power consumption cost for homogeneous PMs and VMs. |

## 5 Conclusion

This paper has introduced a comprehensive review on the subject of energy-related cost issues and prediction models in cloud computing environments. Firstly, it has discussed the fundamental aspects of cloud computing including its definition, system architecture, services types, and deployment types, followed by a detailed description of the pricing models in cloud computing. Secondly, it has reviewed the literature on the energy-related cost issues in cloud computing. Thirdly, it has highlighted the prediction models related to predicting the workload, energy consumption and cost of cloud services. This paper has finally concluded with an overall discussion served as potential research directions, along with a comparison summary of the closely related works.

**Funding Statement:** The author received no specific funding for this study.

**Conflicts of Interest:** The author declare that they have no conflicts of interest to report regarding the present study.

## References

[1] T. Mukherjee, K. Dasgupta, G. Jung and H. Lee, "An economic model for green cloud," in *Proc. of the 10th Int. Workshop on Middleware for Grids, Clouds and e-Science*, pp. 1–6, 2012.

[2] X. Zhang, J. Lu and X. Qin, "BFEPM: Best fit energy prediction modeling based on CPU utilization," in *2013 IEEE Eighth Int. Conf. on Networking, Architecture and Storage*, pp. 41–49, 2013.

[3] J. Conejero, O. Rana, P. Burnap, J. Morgan, B. Caminero *et al.,* "Analyzing Hadoop power consumption and impact on application QoS," *Future Generation Computer Systems*, vol. 55, pp. 213–223, 2016.

[4] M. Bagein, J. Barbosa, V. Blanco, I. Brandic, S. Cremer *et al.,* "Energy efficiency for ultrascale systems: Challenges and trends from nesus project," *Supercomputing Frontiers and Innovations*, vol. 2, no. 2, pp. 105–131, 2015.

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

[6] Amazon, "Amazon EC2 pricing." 2018. [Online]. https://aws.amazon.com/ec2/pricing/.

[7] Microsoft, "Microsoft Azure virtual machines pricing." 2018. [Online]. Available: https://azure.microsoft.com/en-gb/pricing/details/virtual-machines/linux/.

[8] Google, "Google Cloud pricing." 2018. [Online]. Available: https://cloud.google.com/pricing/.

[9] A. Narayan, S. Member, S. Rao and S. Member, "Power-aware cloud metering," *IEEE Transactions on Services Computing*, vol. 7, no. 3, pp. 440–451, 2014.

[10] M. Hinz, G. P. Koslovski, C. C. Miers, L. L. Pilla and M. A. Pillon, "A cost model for IaaS clouds based on virtual machine energy consumption," *Journal of Grid Computing*, vol. 16, no. 3, pp. 493–512, 2018.

[11] D. Laganà, C. Mastroianni, M. Meo and D. Renga, "Reducing the operational cost of cloud data centers through renewable energy," *Algorithms*, vol. 11, no. 10, pp. 145, 2018.

[12] W. N. S. Wan Nik, B. B. Zhou, Z. Mohamad and M. A. Mohamed, "Cost and performance-based resource selection scheme for asynchronous replicated system in utility-based computing environment," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 2, pp. 723–735, 2017.

[13] P. M. Mell and T. Grance, "The NIST definition of cloud computing." Gaithersburg, MD, 2011.

[14] Q. Zhang, L. Cheng and R. Boutaba, "Cloud computing: State-of-the-art and research challenges," *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 7–18, 2010.

[15] KVM, "Kernel-based virtual machine." 2018. [Online]. Available: https://www.linux-kvm.org/.

[16] Xen, "Xen project." 2019. [Online]. Available: https://xenproject.org/.

[17] VMware, "VMware cloud." 2019. [Online]. Available: https://www.vmware.com/.

[18] G. Li and M. Wei, "Everything-as-a-service platform for on-demand virtual enterprises," *Information Systems Frontiers*, vol. 16, no. 3, pp. 435–452, 2014.

[19] M. Al-Roomi, S. Al-Ebrahim, S. Buqrais and I. Ahmad, "Cloud computing pricing models: A survey," *International Journal of Grid and Distributed Computing*, vol. 6, no. 5, pp. 93–106, 2013.

[20] J. Fabra, J. Ezpeleta and P. Álvarez, "Reducing the price of resource provisioning using EC2 spot instances with prediction models," *Future Generation Computer Systems*, vol. 96, pp. 348–367, 2019.

[21] H. Wang, Q. Jing, R. Chen and B. He, "Distributed systems meet economics: Pricing in the cloud," in *2nd USENIX Conf. on Hot Topics in Cloud Computing*, pp. 1–6, 2010.

[22] Jelastic, "Cloud Pricing Models," 2019. [Online]. Available: https://jelastic.com/pricing/.

[23] Amazon, "Amazon EC2 spot instances," 2019. [Online]. Available: https://aws.amazon.com/ec2/spot/.

[24] M. Aldossary, I. Alzamil and K. Djemame, "Towards virtual machine energy-aware cost prediction in clouds," in *Int. Conf. on Economics of Grids, Clouds, Systems, and Services*, pp. 119–131, 2017.

[25] M. Aldossary and K. Djemame, "Energy-based cost model of virtual machines in a cloud environment," in *2018 Fifth Int. Sym. on Innovation in Information and Communication Technology (ISIICT)*, pp. 1–8, 2018.

[26] H. Jin, X. Wang, S. Wu, S. Di and X. Shi, "Towards optimized fine-grained pricing of IaaS cloud platform," *IEEE Transactions on Cloud Computing*, vol. 3, no. 4, pp. 436–448, 2015.

[27] S. Mireslami, L. Rakai, M. Wang and B. H. Far, "Dynamic cloud resource allocation considering demand uncertainty," *IEEE Transactions on Cloud Computing*, 2019.

[28] S. Ilager, K. Ramamohanarao and R. Buyya, "ETAS: energy and thermal-aware dynamic virtual machine consolidation in cloud data center with proactive hotspot mitigation," *Concurrency and Computation: Practice and Experience*, vol. 64, no. 2, e5221, 2019.

[29] X. Zhou, K. Li, C. Liu and K. Li, "An experience-based scheme for energy-SLA balance in cloud data centers," *IEEE Access*, vol. 7, pp. 23500–23513, 2019.

[30] G. Cook, T. Dowdall, D. Pomerantz and Y. Wang, *Clicking Clean: How Companies are Creating the Green Internet*. Washington, DC, USA: Greenpeace Inc., pp. 19, 2014.

[31] P. Berndt and A. Maier, "Towards sustainable IaaS pricing," in *Int. Conf. on Grid Economics and Business Models*, pp. 173–184, 2013.

[32] O. Belli, C. Loomis and N. Abdennadher, "Towards a cost-optimized cloud application placement tool," in *2016 IEEE Int. Conf. on Cloud Computing Technology and Science (CloudCom)*, pp. 43–50, 2016.

[33] M. Mao and M. Humphrey, "Auto-scaling to minimize cost and meet application deadlines in cloud workflows," in *SC'11: Proc. of 2011 Int. Conf. for High Performance Computing, Networking, Storage and Analysis*, pp. 1–12, 2011.

[34] M. S. Aslanpour, M. Ghobaei-Arani and A. Nadjaran Toosi, "Auto-scaling web applications in clouds: A cost-aware approach," *Journal of Network and Computer Applications*, vol. 95, pp. 26–41, 2017.

[35] R. Chard, K. Chard, K. Bubendorfer, L. Lacinski, R. Madduri *et al.,* "Cost-aware elastic cloud provisioning for scientific workloads," in *2015 IEEE 8th Int. Conf. on Cloud Computing*, pp. 971–974, 2015.

[36] K. Djemame, R. Bosch, R. Kavanagh, P. Alvarez, J. Ejarque *et al.,* "PaaS-IaaS inter-layer adaptation in an energy-aware cloud environment," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 2, pp. 127–139, 2017.

[37] A. Yousefipour, A. M. Rahmani and M. Jahanshahi, "Energy and cost-aware virtual machine consolidation in cloud computing," *Software Practice and Experience*, vol. 48, no. 10, pp. 1758–1774, 2018.

[38] M. Aldossary, K. Djemame, I. Alzamil, A. Kostopoulos, A. Dimakis *et al.,* "Energy-aware cost prediction and pricing of virtual machines in cloud computing environments," *Future Generation Computer Systems*, vol. 93, pp. 442–459, 2019.

[39] M. Kumar and S. C. Sharma, "PSO-COGENT: Cost and energy efficient scheduling in cloud environment with deadline constraint, Sustainable Computing," *Informatics and Systems*, vol. 19, pp. 147–164, 2018.

[40] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose and R. Buyya, "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and Experience*, vol. 41, no. 1, pp. 23–50, 2011.

[41] G. Jung, M. A. Hiltunen, K. R. Joshi, R. D. Schlichting and C. Pu, “Mistral: Dynamically managing power, performance, and adaptation cost in cloud infrastructures,” in *2010 IEEE 30th Int. Conf. on Distributed Computing Systems*, pp. 62–73, 2010.

[42] S. S. Wagle, M. Guzek and P. Bouvry, “Service performance pattern analysis and prediction of commercially available cloud providers,” in *2016 IEEE Int. Conf. on Cloud Computing Technology and Science*, pp. 26–34, 2016.

[43] Z. Gong, X. Gu and J. Wilkes, “PRESS: Predictive elastic resource scaling for cloud systems,” in *2010 Int. Conf. on Network and Service Management*, pp. 9–16, 2010.

[44] Q. Huang, K. Shuang, P. Xu, J. Li, X. Liu *et al.,* “Prediction-based dynamic resource scheduling for virtualized cloud systems,” *Journal of Networks*, vol. 9, no. 2, pp. 375–383, 2014.

[45] F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, N. T. Hieu *et al.,* “Energy-aware VM consolidation in cloud data centers using utilization prediction model,” *IEEE Transactions on Cloud Computing*, vol. 7, no. 2, pp. 524–536, 2019.

[46] Q. Zhang, H. Chen and Z. Yin, “PRMRAP: A proactive virtual resource management framework in cloud,” in *2017 IEEE Int. Conf. on Edge Computing (EDGE)*, pp. 120–127, 2017.

[47] W. Fang, Z. Lu, J. Wu and Z. Cao, “RPPS: A novel resource prediction and provisioning scheme in cloud data center,” in *2012 IEEE Ninth Int. Conf. on Services Computing*, pp. 609–616, 2012.

[48] J. Yang, C. Liu, Y. Shang, Z. Mao and J. Chen, “Workload predicting-based automatic scaling in service clouds,” in *2013 IEEE Sixth Int. Conf. on Cloud Computing*, pp. 810–815, 2013.

[49] J. Yang, C. Liu, Y. Shang, B. Cheng, Z. Mao *et al.,* “A cost-aware auto-scaling approach using the workload prediction in service clouds,” *Information Systems Frontiers*, vol. 16, no. 1, pp. 7–18, 2014.

[50] W. L. Bircher and L. K. John, “Complete system power estimation using processor performance events,” *IEEE Transactions on Computers*, vol. 61, no. 4, pp. 563–577, 2012.

[51] J. C. McCullough, Y. Agarwal, J. Chandrashekar, S. Kuppuswamy, A. C. Snoeren *et al.,* “Evaluating the effectiveness of model-based power characterization,” in *USENIX Annual Technical Conf.*, pp. 1–14, 2011.

[52] J. W. Smith, A. Khajeh-Hosseini, J. S. Ward and I. Sommerville, “CloudMonitor: Profiling power usage,” in *2012 IEEE Fifth Int. Conf. on Cloud Computing*, pp. 947–948, 2012.

[53] J. von Kistowski, M. Deffner and S. Kounev, “Run-time prediction of power consumption for component deployments,” in *2018 IEEE Int. Conf. on Autonomic Computing (ICAC)*, pp. 151–156, 2018.

[54] A. T. Makaratzis, K. M. Giannoutakis and D. Tzovaras, “Energy modeling in cloud simulation frameworks,” *Future Generation Computer Systems*, vol. 79, pp. 715–725, 2018.

[55] Y. Li, Y. Wang, B. Yin and L. Guan, “An online power metering model for cloud environment,” in *2012 IEEE 11th Int. Sym. on Network Computing and Applications*, pp. 175–180, 2012.

[56] F. Farahnakian, P. Liljeberg and J. Plosila, “LiRCUP: Linear regression based CPU usage prediction algorithm for live migration of virtual machines in data centers,” in *2013 39th Euromicro Conf. on Software Engineering and Advanced Applications*, pp. 357–364, 2013.

[57] J. Subirats and J. Guitart, “Assessing and forecasting energy efficiency on Cloud computing platforms,” *Future Generation Computer Systems*, vol. 45, pp. 70–94, 2015.

[58] Y. Jiang, C. Perng, T. Li and R. Chang, “ASAP: A self-adaptive prediction system for instant cloud resource demand provisioning,” in *2011 IEEE 11th Int. Conf. on Data Mining*, pp. 1104–1109, 2011.

[59] N. Roy, A. Dubey and A. Gokhale, “Efficient autoscaling in the cloud using predictive models for workload forecasting,” in *2011 IEEE 4th Int. Conf. on Cloud Computing*, pp. 500–507, 2011.

[60] U. Sharma, P. Shenoy, S. Sahu and A. Shaikh, “Kingfisher: A system for elastic cost-aware provisioning in the cloud. *Technical Report UM-CS-2010-005*, 2010.

[61] OpenNebula, “The simplest cloud management experience,” 2019. [Online]. Available: https://opennebula.org/.

[62] H. Liu, H. Jin, C. Z. Xu and X. Liao, “Performance and energy modeling for live migration of virtual machines,” *Cluster Computing*, vol. 16, no. 2, pp. 249–264, 2013.