Proceedings Paper:
Alzamil, I and Djemame, K orcid.org/0000-0001-5811-5263 (2017) Energy Prediction for Cloud Workload Patterns. In: Bañares, J, Tserpes, K and Altmann, J, (eds.) GECON 2016: Economics of Grids, Clouds, Systems, and Services. 13th International Conference on Economics of Grids, Clouds, Systems and Services (GECON'2016), 20-22 Sep 2016, Athens, Greece. Lecture Notes in Computer Science (10382). Springer , Cham, Switzerland , pp. 160-174. ISBN 978-3-319-61919-4

https://doi.org/10.1007/978-3-319-61920-0_12

Energy Prediction for Cloud Workload Patterns

Ibrahim Alzamil[1, 2] and Karim Djemame[1]

[1] School of Computing, University of Leeds, Leeds, UK
{sc11iaa, K.Djemame}@leeds.ac.uk
[2] College of Science and Humanities, Majmaah University, Alghat, KSA
I.Alzamil@mu.edu.sa

Abstract. The excessive use of energy consumption in Cloud infrastructures has become one of the major cost factors for Cloud providers to maintain. In order to enhance the energy efficiency of Cloud resources, proactive and reactive management tools are used. However, these tools need to be supported with energy-awareness not only at the physical machine (PM) level but also at virtual machine (VM) level in order to enhance decision-making. This paper introduces an energy-aware profiling model to identify energy consumption for heterogeneous and homogeneous VMs running on the same PM and presents an energy-aware prediction framework to forecast future VMs energy consumption. This framework first predicts the VMs' workload based on historical workload patterns using Autoregressive Integrated Moving Average (ARIMA) model. The predicted VM workload is then correlated to the physical resources within this framework in order to get the predicted VM energy consumption. Compared with actual results obtained in a real Cloud testbed, the predicted results show that this energy-aware prediction framework can get up to 2.58 Mean Percentage Error (MPE) for the VM workload prediction, and up to -4.47 MPE for the VM energy prediction based on periodic workload pattern.

Keywords: Cloud Computing, Energy Efficiency, Energy-Aware Profiling, Energy Prediction, Workload Prediction, Cloud Workload Patterns

1 Introduction

With the wide adoption of Cloud Computing, energy consumption has become one of the main issues for Cloud providers to maintain. A Cloud infrastructure along with its cooling resources consume a large amount of energy in order to operate, which may cause ecological and economic issues. The ICT industry is responsible for about 2 percent of the global CO2 emission, which is similar to the amount caused by the aviation industry, as stated by Gartner [1]. For economic aspects, a data centre may consume about 100 times more energy compared to a typical office with the same size [2]. In terms of maintenance, Cloud providers consider energy consumption as one of the largest cost factors [3] with a big impact on the operational cost of a Cloud infrastructure [4]. Therefore, various energy efficient techniques have been introduced recently to help the Cloud providers reduce the energy consumption cost of their infrastructure,

which can then lead to reducing the cost of operational expenditure (OPEX) and having less impact on the environment.

The impact of energy consumption is not only dependent on the efficiency of the physical resources, but also on the efficiency of the tools deployed to manage these resources as well as the efficiency of the applications running on these resources [5]. Different methods have been used to efficiently manage the Cloud resources, all of which can be based on certain thresholds, called reactive, or based on prediction, called proactive. For example, once exceeding a certain threshold, 80% of CPU utilisation, some actions take place by reactive methods to add more resources and avoid performance degradation. With prediction, proactive methods have the advantage of taking some actions at earlier stages to avoid getting that threshold and maintain the expected performance. To enable such optimisation and the energy efficient design of Cloud applications, the applications' designers and developers should be provided with energy-aware information to support their programming decisions. Also, the deployment tools should incorporate energy-ware information to make energy-efficient decisions when deploying these applications on the Cloud resources. As discussed in [6], having appropriate tools for energy monitoring and profiling is essential to get better energy-awareness and then help for energy optimisation at all layers of such large-scale system. Also, predicting the workload of a Virtual Machine (VM) can help make effective deployment strategies and energy efficient resource allocation methods [7]. Thus, managing the Cloud paradigm in all different levels and reducing the energy consumption has been an active area of research as it can result in reduction of OPEX costs for the Cloud providers.

Cloud applications can experience different workload patterns based on the users' usage behaviours, and these workload patterns are depicted by the utilisation of the resources hosting these applications. As stated in [8], there are mainly five Cloud workload patterns, namely: static workload experiencing the same and stable resource utilisation over a period of time; periodic workload experiencing repeated resource utilisation peaks in time intervals; continuously changing workload experiencing a resource utilisation that continuously decreases or increases over time; unpredicted workload experiencing a resource utilisation randomly over time, and once-in-a-life-time workload experiencing a resource utilisation peak once over time. These different workload patterns consume energy differently based on the resources they utilise. Thus, it is important to have reactive and proactive methods to efficiently manage these resources when being utilised. In order to do that, current energy usage for physical and virtual resources has to be profiled so that such reactive methods can rely on it. Consequently, future energy usage can be predicted so that such proactive methods make use of. Energy consumption can be directly measured at the Physical Machine (PM) level, but it is difficult and not directly measured at the VM level. Thus, enabling energy-awareness at different levels is a key aspect towards efficiently managing the Cloud paradigm.

In our previous work [9], we proposed and implemented a system architecture to enable energy-aware profiling for Cloud infrastructure resources at both physical and virtual levels. In this paper, we extend our work to consider heterogeneity when profiling energy consumption for different sizes of VMs running on a single PM. Also, we extend this architecture to enable energy prediction for the VMs requested to run Cloud

applications by considering previously profiled and stored data as well as the incoming workload's characteristics before service deployment. The outcomes of this work can help and add value to enhance the energy efficiency of Cloud environment by feeding into other deployment models or scheduling strategies to enable energy efficient management of Cloud resources, which can lead to lowering the cost of OPEX for Cloud providers. This paper's main contributions are:

- Energy-aware profiling model that enables energy-awareness for homogeneous and heterogeneous VMs in Clouds.
- Energy-aware prediction framework that forecasts the future energy usage for VMs prior to deployment.

This paper is structured as follows: a discussion of the related work is summarised in the next section. Section 3 presents the system architecture followed by a discussion of the energy-aware profiling model for attributing PM's power consumption to heterogeneous and homogeneous VMs, and a discussion of the energy-aware prediction framework for forecasting the future VMs' energy usage before their deployment. Section 4 discusses the experimental set up followed by results and evaluation in Section 5. Finally, Section 6 concludes this paper and discusses future work.

2 Related Work

Djemame et al [10] emphasised the importance of optimising the energy efficiency of the Cloud paradigm at different layers and proposed an architecture that addresses energy efficiency at all Cloud layers and all through Cloud application life-cycle. Monitoring and profiling as well as forecasting the energy consumption is a key step towards enhancing and optimising the energy efficiency in the Cloud paradigm. However, VMs' energy consumption cannot be measured and profiled directly as they do not have direct hardware interfaces. Therefore, their energy information can be indirectly identified via modelling the energy consumed by the servers in which they are hosted [9, 11–13].

Further, uncertainty issues associated with the Cloud environment makes it more difficult to do such prediction, like predicting job runtime. Tchernykh et al [14] have emphasised the difficulty of dealing with uncertainty in Cloud environment especially since its workload can change dramatically over time. So, they have reviewed and classified the uncertainty issues associated with a Cloud environment and discussed some approaches to mitigate them. For example, some looked at the historical data of applications to predict the runtime job of similar applications to be executed [15].

Tchernykh et al [16] have presented an experimental study for several online scheduling strategies in a Cloud environment with different workloads. In the experimental results, they used and analysed eight allocation strategies based on three group categories, namely, knowledge-free, energy-aware, and speed-aware. The energy model used in their work simply considers summing up the machine's idle power and the extra variable power, which depends on the workload. However, they do not consider the workload in their model when calculating the variable power consumption. Also, the workload used in their work is based on HPC jobs for parallel and grid environments

and not precisely on real Cloud environments that should also consider the complexity of virtualisation aspects.

Some work focuses on predicting power consumption based on historical data while others use performance counters, which are queried from chips or OS. But, relying on performance counters would not work appropriately in heterogeneous environments with different server's characteristics, as argued by Zhang et al [17]. Therefore, they presented a best fit energy prediction model BFEPM that flexibly selects the best model for a given server based on a series of equations that consider only CPU utilisation [17]. Dargie [18] proposed a stochastic model to estimate the power consumption for a multi-core processor based on the CPU utilisation workload and found out that the relationship between the workload and power is best estimated using a linear function in a dual-core processor and using a quadratic function in a single-core processor. Further, Fan et al [19] have introduced a framework to estimate the power consumption of servers based on CPU utilisation only and argued with their results that the power consumption correlates well with the CPU usage. As their framework produced accurate results, they argued that it is not necessary to use more complex signals, like hardware performance counters, to model power usage. Their work also indicates that the activity of other system components, other than CPU, may have either small effect on power usage or their activity correlates well with the CPU activity.

In terms of future prediction based on historical data, estimating the energy consumption of a Cloud application prior to deployment on VMs would require understanding the characteristics of the underlying physical resources, like idle power consumption and variable power under different workload, and the projected virtual resources usage, as stated in [20]. Thus, it is essential to get the predicted VMs' workload first in order to get their predicted energy. Some work has predicted future workload in a Cloud environment based on Autoregressive Integrated Moving Average (ARIMA) model [21–24]; nonetheless, their objectives do not consider predicting the energy consumption. For example, Calheiros et al [24] introduced a Cloud workload prediction module based on the ARIMA model to proactively and dynamically provision resources. They define their workload as the expected number of requests received by the users, which are then mapped to predict the number of VMs needed to execute users' requests and meet the Quality of Service (QoS).

Compared with the work presented in this paper, ARIMA model is used to predict the VM workload, defined as VM CPU utilisation, which is then mapped within the energy-aware prediction framework to get the forecasted VM energy consumption for the next time interval. Then, having predicted the VM workload and its energy consumption, other methods can rely on this information to help introduce a proactive resource provisioning and scheduling that aim to not only utilise resources efficiently and meet the demands, but also consider the energy efficiency aspects as well. This can drive towards a cost reduction of the energy consumption and OPEX for Cloud service providers.

3 Energy-Aware Profiling and Prediction

Enabling energy-awareness in the Cloud paradigm is a key step towards optimising its energy efficiency. An energy-aware profiling model is introduced for Cloud infrastructures where the service operation takes place in order to understand how the energy has been consumed; this profiled information can then be used to help the software developers and reactive management tools make energy-efficient decisions when optimising the applications and efficiently managing the Cloud resources. Also, an energy-aware prediction framework is proposed to predict the energy consumption of VMs, requested to execute the application, prior to service deployment, which can help and facilitate such proactive deployment tools with energy-awareness to efficiently manage the Cloud resources. The overall system architecture of this work will be discussed in the next subsection, followed by a detailed discussion of the energy-aware profiling and prediction within this architecture.

3.1 System Architecture

The system architecture is aimed at enabling energy-awareness at the deployment and operational levels of the Cloud paradigm. As depicted in Figure 1, this architecture consists of a number of components, mainly, the Resource Monitoring Unit (RMU), Energy-aware Profiling Unit (EPU), Reporting and Analysis Unit, and Energy-aware PREdiction Unit (EPREU). The highlighted components, EPU and EPREU, are the main focus of this paper.

Starting at the bottom layer when the Cloud infrastructure is operating to run the Cloud services, the resources' usage and physical energy consumption along with the number of assigned VMs to each PM are dynamically collected by RMU. EPU has an appropriate energy model that takes as input the monitored data from RMU and outputs

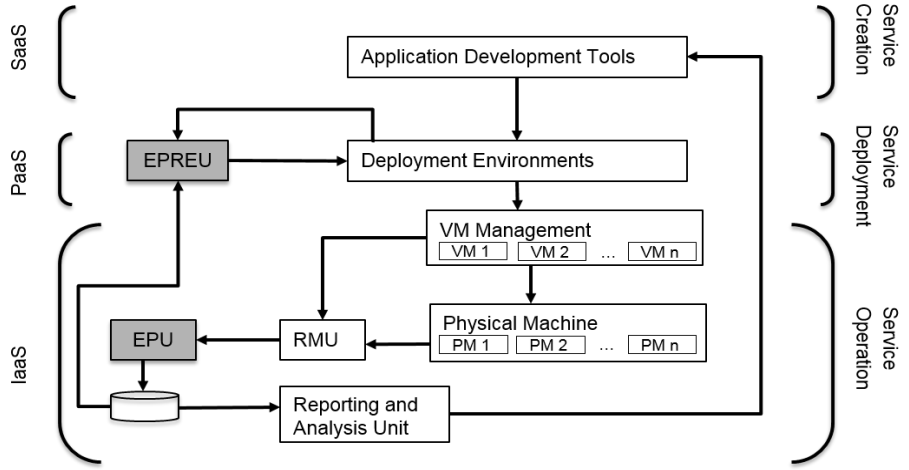

Fig. 1. System architecture

the attribution of the energy consumption to each VM based on the energy consumption of their physical hosts. Then, EPU profiles and populates these measurements to a knowledge database, which can be further used by the Reporting and Analysis Unit to provide energy-aware reports to the application developers to help them learn how their applications consume energy and make such energy-efficient decisions accordingly to optimise their applications. Also, these measurements can be very useful for such resource management tools by enhancing their energy-awareness and making energy-efficient decisions when, for example, scheduling the tasks and balancing the workload. Further, this energy-related information of VMs, which can be used by different customers and run on the same PM, can help the service providers introduce a new pricing mechanism that charge the customers based not only on their IT resources usage, but on their energy usage as well.

Moving up to the middle layer when the Cloud services are about to be deployed, EPREU has a framework consisting of a number of models that predict the energy consumption of VMs prior to service deployment by considering the type of these VMs and their historical data. The predicted energy consumption for VMs can help other deployment strategies make energy-efficient decisions proactively.

3.2 Energy-Aware Profiling Model

The energy consumption of PMs can be directly measured and mainly consists of two parts, idle and active. The idle energy is consumed when the PM is turned on but not running any workload. The active energy is the extra energy added to the idle when the PM is busy and running some workload. As the case with the PM, the total energy consumption of the VM equals its idle energy consumption plus its active energy consumption. Yet, the energy consumption of VMs is difficult to identify and not directly measured.

In our previous work [9], we introduced an energy-aware profiling model that attributes the PM's energy consumption to VMs. It attributes the PM's idle energy evenly among the number of VMs running on it, and attributes the active energy based on VM CPU utilisation mechanism. This model enables a fair attribution of a PM's energy consumption to homogeneous VMs.

In this paper, we extend our work and introduce a new energy-aware profiling model that fairly attributes the energy consumption to homogeneous and heterogeneous VMs running on the same PM. This new model works by fairly attributing the PM's idle energy to VMs based on the number of Virtual CPUs (VCPUs) assigned to each VM, and the active energy to VMs based on the VM CPU utilisation mechanism as well as the number of VCPUs assigned to each VM.

As shown in Equation 1, VM_{xPwr} is the power consumption of the targeted VM; $PM_{IdlePwr}$ is the idle power consumption of the PM where the VMs are hosted; $VM_{x_{VCPU}}$ and VM_{xUtil} are the number of assigned VCPUs and the CPU utilisation of that VM; VM_{Count} is the number of VMs running on the same PM; $VM_{y_{VCPU}}$ and VM_{yUtil} are the number of assigned VCPUs and the CPU utilisation of a member of the VMs set hosted by the same PM, and the active power consumption of the PM is the total PM's power PM_{Pwr} minus its idle power.

$$VM_{xPwr} = PM_{IdlePwr} \times \frac{VM_{x_{VCPU}}}{\sum_{y=1}^{VM_{Count}} VM_{y_{VCPU}}} + (PM_{Pwr} - PM_{IdlePwr}) \times \frac{VM_{xUtil} \times VM_{xVCPU}}{\sum_{y=1}^{VM_{Count}} (VM_{yUtil} \times VM_{yVCPU})} \quad (1)$$

Hence, the new energy-aware profiling model can now fairly attribute the idle and active energy consumption of a PM to the same or different sizes of VMs in terms of the allocated VCPUs for each VM. For instance, when both a small VM with 1 VCPU and a large VM with 3 VCPUs are being fully utilized on the same PM, the large VM would have triple the value in terms of energy consumption as compared to the small VM; so that the energy consumption can be fairly attributed based on the actual physical resources used by each VM.

3.3 Energy-Aware Prediction Framework

As measuring the current energy consumption is difficult and cannot be performed directly at the VM level, predicting the future energy consumption is even more difficult at this level because it would rely on the estimated PM's energy to be used. Therefore, an energy-aware prediction framework that aims to forecast the energy consumption for the new VMs prior to service deployment is presented. This framework includes a model that first predicts the workload at the VM level. After that, this predicted VM workload is correlated to physical workload in order to estimate the new PM energy consumption, from which the predicted VM energy consumption would be based on. As depicted in Figure 2, this energy-aware prediction framework includes four main steps in order to forecast the VMs' energy consumption.

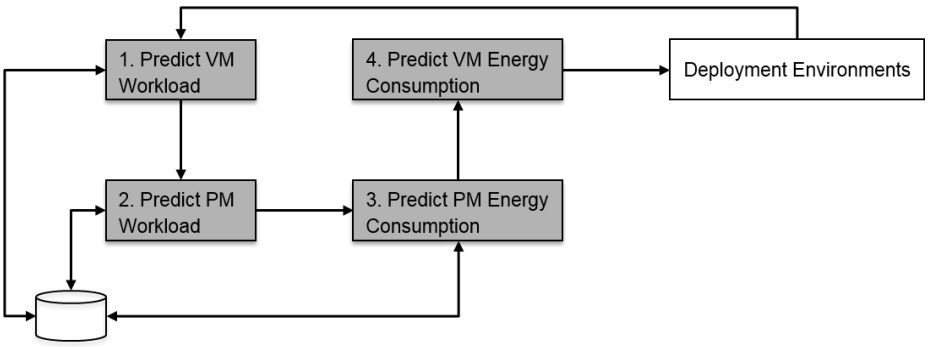

Fig. 2. Energy-aware prediction framework

Step 1: this framework starts by receiving from the deployment environment prerequisite information, which is the requested number of VMs along with their capacity in terms of VCPUs to execute the application, before such deployment process takes place. Then, by using the ARIMA model, the VM workload, which is VM CPU utilisation, is predicted based on historical static and periodic workload patterns.

The ARIMA model is a time series prediction model that has been used widely in different domains, including finance, owing to its sophistication and accuracy; further details about the ARIMA model can be found in [25]. Unlike other prediction methods, like sample average, ARIMA takes multiple inputs as historical observations and outputs multiple future observations depicting the seasonal trend. It can be used for seasonal or non-seasonal time-series data. The type of seasonal ARIMA model is used in this work as the targeted workload patterns are reoccurring and showing seasonality in time intervals. In order to use the ARIMA model for predicting the VM workload in our work, the historical time series workload data has to be stationary, otherwise Box and Cox transformation [26] and data differencing methods are used to make these data stationary. The model selection can be automatically processed in R package [27] using the **auto.arima** function, which selects the best fit model of ARIMA based on Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) value.

Step 2: once the VM's workload is predicted, the next step is to understand how this workload would be reflected on the physical resources and predict the new PM's workload, which is PM CPU utilisation, with consideration of its current workload as the PM may be running another VM already. Therefore, the relationship between the number of VCPUs and the PM's CPU utilisation is characterised for each PM in the Leeds Cloud testbed (this testbed is discussed in Section 4). For instance, Figure 3a shows a linear relation between the number of VCPUs and CPU utilisation for a single physical host. Thus, using this relation equation can help estimate the new increment of PM's CPU utilisation based on the used ratio of the requested VCPUs for the VM, $VM_{xReqVCPUs}$, identified by the predicted VM CPU utilisation, $VM_{xPredUtil}$. This new increment of PM's utilisation would be also added to the current PM's CPU utilisation, $PM_{xCurrUtil}$, in order to identify the new total of the predicted PM's CPU utilisation, $PM_{xPredUtil}$, as described in Equation 2. The PM's idle CPU utilisation, $PM_{xIdleUtil}$, is subtracted from the current because the relation equation already considers this idle value.

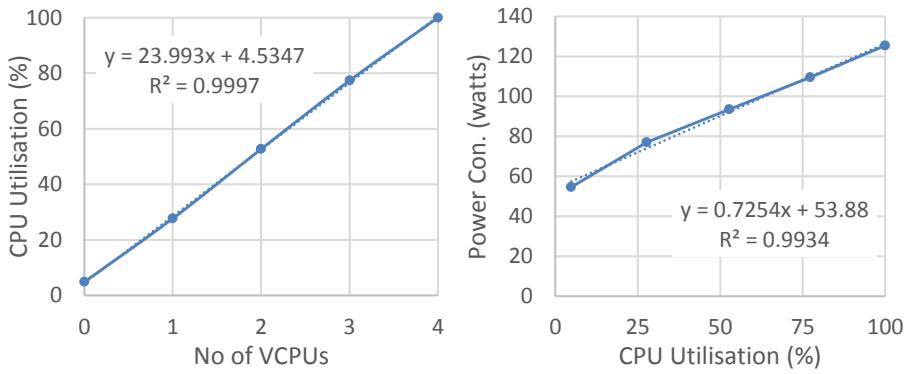

Fig. 3. (a) On the left: Number of VCPUs vs CPU utilisation for a single host. (b) On the right: CPU utilisation vs power consumption for a single host.

$$PM_{xPredUtil} = \left(23.993 \times \left(VM_{xReqVCPUs} \times \frac{VM_{xPredUtil}}{100}\right) + 4.5347\right) + (PM_{xCurrUtil} - PM_{xIdleUtil}) \quad (2)$$

Step 3: after predicting the PM's workload, the next step is to predict the PM's energy consumption based on the correlation of this predicted workload with PM energy consumption. For example, Figure 3b shows a linear relation between the power consumption and the CPU utilisation on the same physical host.

Considering this relation, Equation 3 is used to predict the PM's power consumption, $PM_{xPredPwr}$, based on the predicted PM's CPU utilisation.

$$PM_{xPredPwr} = 0.7254 \times PM_{xPredUtil} + 53.88 \quad (3)$$

Step 4: the final step within this framework is to profile and attribute the predicted PM's energy consumption to the new requested VM and to the VMs already running on that physical host based on the energy-aware profiling model introduced in Section 3.2. Hence, the energy consumption for the new VM prior to deployment will be predicted for the next interval time using Equation 1, but substituting the $VM_{x_{VCPU}}$ with $VM_{xReq_{VCPU}}$, PM_{Pwr} with $PM_{xPredPwr}$, and VM_{xUtil} with $VM_{xPredUtil}$.

4 Experimental Set Up

This section describes the environment and the details of the experiments conducted in order to evaluate the work presented in this paper. In terms of the environment, the experiments have been conducted on the Leeds Cloud testbed, discussed in details in [9]. Briefly, this testbed includes a cluster of commodity Dell servers, and one of these servers with a four core X3430 Intel Xeon CPU was used. The server has a WattsUp meter [28] attached to directly measure the energy consumption and push it to Zabbix [29], which is also used for resources usage monitoring purposes. This testbed currently uses OpenNebula [30] version 4.10 as the Virtual Infrastructure Manager (VIM), and KVM [31] hypervisor for the Virtual Machine Manager (VMM).

In terms of the experiments' design, the aim is to evaluate that the new energy-aware profiling model presented in this paper is capable of fairly attributing the PM's energy consumption to homogeneous and heterogeneous VMs. Thus, one scenario is designed to show how the energy consumption would be attributed when two small VMs with 1 VCPU for each are running on the same PM, and another scenario is designed to show how the energy consumption would be attributed when a small VM with 1 VCPU and a large VM with 3 VCPUs are running on the same PM. Secondly, the aim is also to evaluate that the energy-aware prediction framework is capable of predicting the energy consumption of the VM prior to service deployment based on historical static and periodic workload. Thus, a number of direct experiments have been conducted on the testbed to synthetically generate static and periodic workload by stressing the CPU on different types of VMs, like a small VM with 1 VCPU and a large VM with 3 VCPUs. The generated workload of each VM type has four time intervals of 30 minutes each. The first three intervals will be used as the historical data set for prediction, and the last

interval will be used as the testing data set to evaluate the predicted results. The prediction process starts by firstly predicting the VM workload offline using the **auto.arima** function in R package [27] and then completing the cycle of this framework and considering the correlation between the physical and virtual resources to predict energy consumption of the VM prior to deployment on a single PM. This single PM is expected to host this VM only, so this VM would have the same energy consumption as the PM.

5 Results Discussion and Evaluation

Starting with evaluating the capability of the energy-aware profiling, Figures 4a and 4b show the results of attributing the PM's energy consumption to two homogeneous and heterogeneous VMs. The first part of Figures 4a and 4b shows the attribution of the PM's idle energy when the VMs are running but not generating any workload, and the second part shows the attribution of the PM's total energy when the VMs are running the same workload at 80% of CPU utilisation.

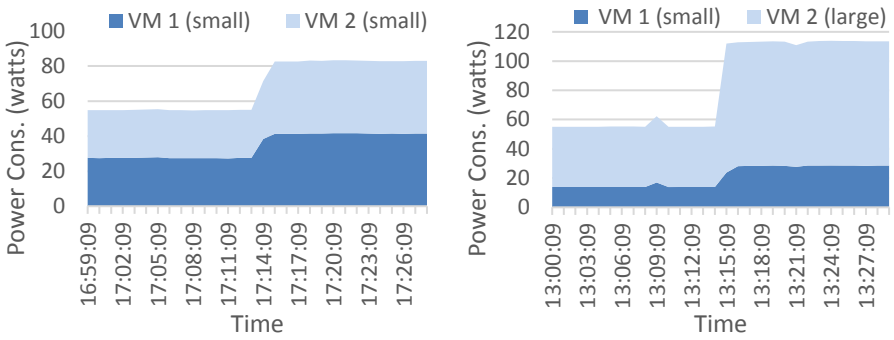

Fig. 4. Energy consumption of a single host attributed to two homogeneous VMs shown on the left (a) and to two heterogeneous VMs shown on the right (b).

Figure 4a shows the results of attributing the PM's energy consumption to two homogeneous small VMs, each with 1 VCPU. Based on the results shown on Figure 4a, both of the VMs have the same energy consumption as they are homogeneous and have the same usage of the actual physical resources. Figure 4b shows the attribution of PM's energy consumption to heterogeneous VMs, one small with 1 VCPU and another large with 3 VCPUs. As having triple the size in terms of VCPUs, the large VM's energy consumption during the idle and active states is three times larger than the energy consumption of the small VM. Overall, the results show that the energy-aware profiling model is capable of fairly attributing PM's energy consumption to homogeneous and heterogeneous VMs based on their utilisation and size, which reflect the actual physical resources' usage.

In terms of evaluating the energy-aware prediction framework, Figure 5 presents the predicted results for a large VM based on a historical static workload pattern at 80% of CPU utilisation, and Figure 6 presents the predicted results for a large VM based on a historical periodic workload pattern with two utilisation peaks.

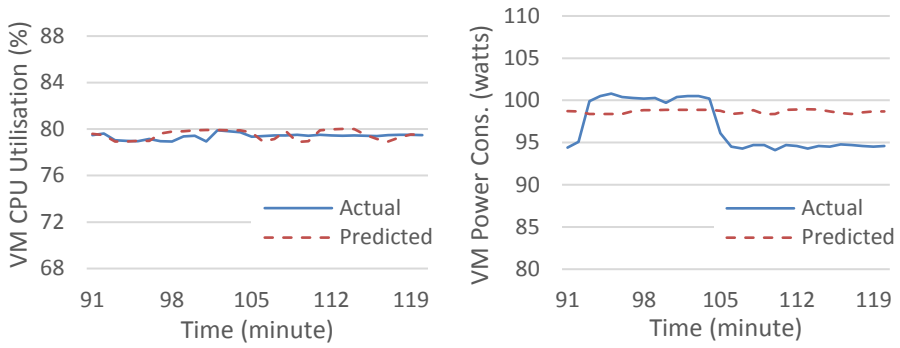

Fig. 5. Prediction results for a large VM based on static workload pattern. (a) On the left: results of workload prediction. (b) On the right: results of energy prediction.

For the prediction based on the historical static workload pattern, Figure 5a shows the results of the predicted versus the actual VM workload. Figure 5b shows the results of the predicted versus the actual VM energy consumption over a time period.

As discussed previously, the VM workload prediction within the proposed framework uses the ARIMA model to forecast the next 30 minute period of workload, as shown in Figure 5a, based on three historical intervals of workload. Overall, the predicted VM workload results closely match the actual workload owing to the sophistication of the ARIMA model. Based on this predicted workload, the VM energy consumption is predicted using the remaining models, as previously discussed, within the proposed framework (see Section 3.3). Figure 5b shows the predicted VM energy consumption results, which have a small variation as compared to the actual energy consumption. The reason of this variation is because there is an accumulation of error from the previous steps within the framework, especially when correlating the PM CPU utilisation to PM power consumption. As seen on Figure 5b, the actual energy consumption increases in the first part of the interval; this may be due to the thermal energy, which is not captured in this work, causing the machine's fan to run faster and thus leading to an increase of PM energy, which is then attributed to the VM. Despite this accumulation of error, the proposed framework can predict the VM energy consumption accurately.

In terms of prediction accuracy, a number of metrics, as summarised in Table 1, are used to evaluate the predicted VM workload and energy consumption based on static

Table 1. Prediction accuracy for a large VM based on static workload pattern

Accuracy Metric	Predicted VM Workload	Predicted VM Energy Consumption
Mean Error (ME)	-0.11	-1.75
Root Mean Squared Error (RMSE)	0.42	3.28
Mean Absolute Error (MAE)	0.33	3.04
Mean Percentage Error (MPE)	-0.14	-1.89
Mean Absolute Percentage Error (MAPE)	0.42	3.17

workload. As previously discussed in Section 4, the actual data of the VM workload and energy consumption are used as the testing data set for evaluation purposes.

As shown in Table 1, the accuracy of the predicted VM workload is very high as its metrics' values are close to zero. The predicted VM energy consumption is less accurate as compared with the predicted VM workload, but still achieves a good prediction accuracy, with -1.89 of MPE. The reason of the predicted VM energy consumption being less accurate than the predicted workload when compared to the actual data is due to the accumulated error when correlating this VM workload to physical resources.

In terms of prediction based on the historical periodic workload pattern, Figure 6a shows the results of the predicted versus the actual VM workload. Figure 6b shows the results of the predicted versus the actual VM energy consumption over a period of time.

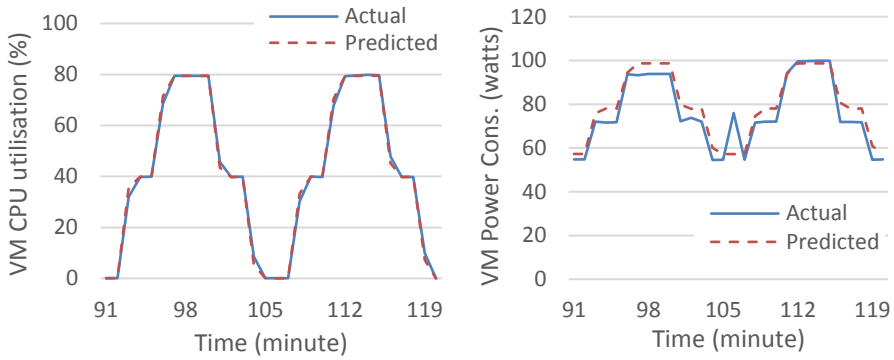

Fig. 6. Prediction results for a large VM based on periodic workload pattern. (a) On the left: results of workload prediction. (b) On the right: results of energy prediction.

Despite the periodic utilisation peaks, the predicted VM workload results are closely matched with the actual results, which reflect the capability of the ARIMA model to capture the historical seasonal trend and give a very accurate prediction accordingly. The proposed framework is also capable of predicting the energy consumption of the VM with only a small variation as compared to the actual. As shown in Figure 6b, the actual VM energy consumption in the middle of the interval has a small peak, which was not followed by the predicted VM energy consumption. This is again can be due to the thermal energy which is not considered in the proposed framework.

Table 2. Prediction accuracy for a large VM based on periodic workload pattern

Accuracy Metric	Predicted VM Workload	Predicted VM Energy Consumption
ME	-0.02	-3.04
RMSE	1.51	5.76
MAE	0.81	4.61
MPE	2.58	-4.47
MAPE	5.30	6.43

For evaluating the accuracy of the predicted VM workload and energy consumption based on periodic workload, different accuracy metrics are used, as shown in Table 2.

Despite the high variation of the workload utilisation in the periodic pattern, the accuracy metrics, as shown in Table 2, indicate that the predicted VM workload achieves a good accuracy, with 2.58 of MPE. As previously discussed, the accumulated error when correlating the predicted VM workload to the physical resources in order to get the energy affects the accuracy of the predicted VM energy consumption. Therefore, the predicted VM energy consumption is less accurate as compared with the predicted VM workload, but still achieves a good prediction accuracy, with -4.47 of MPE.

6 Conclusion and Future Work

This paper has presented and evaluated a new energy-aware profiling model that enables a fair attribution of a PM's energy consumption to homogeneous and heterogeneous VMs based on their utilisation and size, which reflect the physical resource usage by each VM. Also, it has proposed an energy-aware prediction framework to forecast the energy consumption of the VM prior to service deployment. A number of direct experiments were conducted on the Leeds Cloud testbed to evaluate the capability of the energy prediction. Overall, the results show that the proposed energy-aware prediction framework is capable of forecasting the energy consumption for the VM with a good prediction accuracy for static and periodic Cloud workload patterns.

The application of the proposed work is providing energy-awareness which can be used and incorporated by other reactive and proactive management tools to make enhanced energy-aware decisions and efficiently manage the Cloud resources, leading towards a reduction of energy consumption, and therefore lowering the cost of OPEX for Cloud providers and having less impact on the environment.

In future work, we aim to facilitate the proposed prediction framework and make an online modeller on the Leeds testbed to make the prediction process dynamic. Also, we will consider the scalability aspects with different prediction scenarios to further show the capability of the proposed work, like predicting the energy usage for a number of VMs to be run on a single or multiple PMs already hosting other running VMs, and predicting the energy usage for these VMs to run all together. Further, we aim to consider the thermal energy and its impact on the energy consumption. With the evolving technologies of containers, further work will investigate the applicability of using this research in that context and consider attributing the system's energy consumption to container instances instead of VM instances.

References

1. Gartner: Gartner Estimates ICT Industry Accounts for 2 Percent of Global CO2 Emissions, http://www.gartner.com/newsroom/id/503867.
2. Scheihing, P.: Creating Energy-Efficient Data Centers. In: Data Center Facilities and Engineering Conference. , Washington, DC, 18th May. (2007).

3. Mukherjee, T., Dasgupta, K., Gujar, S., Jung, G., Lee, H.: An economic model for green cloud. Proc. 10th Int. Work. Middlew. Grids, Clouds e-Science - MGC '12. 1–6 (2012).
4. Conejero, J., Rana, O., Burnap, P., Morgan, J., Caminero, B., Carrión, C.: Analyzing Hadoop power consumption and impact on application QoS. Futur. Gener. Comput. Syst. 55, 213–223 (2016).
5. Beloglazov, A., Buyya, R., Lee, Y.C., Zomaya, A.Y.: A Taxonomy and Survey of Energy-Efficient Data Centers and Cloud Computing Systems. CoRR. abs/1007.0, (2010).
6. Bagein, M., Barbosa, J., Blanco, V., Brandic, I., Cremer, S., Karatza, H.D., Lefevre, L., Mastelic, T., Oleksiak, A.: Energy Efficiency for Ultrascale Systems: Challenges and Trends from Nesus Project. Supercomput. Front. Innov. 2, 105–131 (2015).
7. Jheng, J.-J., Tseng, F.-H., Chao, H.-C., Chou, L.-D.: A novel VM workload prediction using Grey Forecasting model in cloud data center. In: Information Networking (ICOIN), 2014 International Conference on. pp. 40–45 (2014).
8. Fehling, C., Leymann, F., Retter, R., Schupeck, W., Arbitter, P.: Cloud computing patterns. Springer (2014).
9. Alzamil, I., Djemame, K., Armstrong, D., Kavanagh, R.: Energy-Aware Profiling for Cloud Computing Environments. Electron. Notes Theor. Comput. Sci. 318, 91–108 (2015).
10. Djemame, K., Armstrong, D., Kavanagh, R., Juan Ferrer, A., Garcia Perez, D., Antona, D., Deprez, J.-C., Ponsard, C., Ortiz, D., Macías Lloret, M., Guitart Fernández, J., Lordan Gomis, F.-J., Ejarque, J., Sirvent Pardell, R., Badia Sala, R.M., Kammer, M., Kao, O., Agiatzidou, E., Dimakis, A., Courcoubetis, C., Blasi, L.: Energy efficiency embedded service lifecycle: Towards an energy efficient cloud computing architecture. In: Joint Workshop Proceedings of the 2nd International Conference on ICT for Sustainability 2014. pp. 1–6. CEUR-WS.org (2014).
11. Kavanagh, R., Armstrong, D., Djemame, K.: Towards an Energy-Aware Cloud Architecture for Smart Grids. In: 12th International Conference on Economics of Grids, Clouds, Systems, and Services. pp. 1–14. , Cluj-Napoca, Romania (2015).
12. Gu, C., Huang, H., Jia, X.: Power Metering for Virtual Machine in Cloud Computing-Challenges and Opportunities. IEEE Access. 2, 1106–1116 (2014).
13. Yang, H., Zhao, Q., Luan, Z., Qian, D.: iMeter: An integrated VM power model based on performance profiling. Futur. Gener. Comput. Syst. 36, 267–286 (2014).
14. Tchernykh, A., Schwiegelsohn, U., Alexandrov, V., Talbi, E.: Towards Understanding Uncertainty in Cloud Computing Resource Provisioning. Procedia Comput. Sci. 51, 1772–1781 (2015).
15. Ramírez-Alcaraz, J.M., Tchernykh, A., Yahyapour, R., Schwiegelshohn, U., Quezada-Pina, A., González-García, J.L., Hirales-Carbajal, A.: Job Allocation Strategies with User Run Time Estimates for Online Scheduling in Hierarchical Grids. J. Grid Comput. 9, 95–116 (2011).
16. Tchernykh, A., Lozano, L., Schwiegelshohn, U., Bouvry, P., Pecero, J.E., Nesmachnow, S., Drozdov, A.Y.: Online Bi-Objective Scheduling for IaaS Clouds Ensuring Quality of Service. J. Grid Comput. 14, 5–22 (2016).
17. Zhang, X., Lu, J., Qin, X.: BFEPM: Best Fit Energy Prediction Modeling Based on CPU Utilization. 2013 IEEE Eighth Int. Conf. Networking, Archit. Storage. 41–49 (2013).
18. Dargie, W.: A stochastic model for estimating the power consumption of a processor. Comput. IEEE Trans. 64, 1311–1322 (2015).
19. Fan, X., Weber, W.-D., Barroso, L.A.: Power Provisioning for a Warehouse-sized Computer. In: Proceedings of the 34th Annual International Symposium on Computer Architecture. pp. 13–23. ACM, New York, NY, USA (2007).

20. Armstrong, D., Kavanagh, R., Djemame, K.: ASCETiC Project: D2.2.2 Architecture Specification - Version 2. (2014).
21. Fang, W., Lu, Z., Wu, J., Cao, Z.: RPPS: A Novel Resource Prediction and Provisioning Scheme in Cloud Data Center. In: Services Computing (SCC), 2012 IEEE Ninth International Conference on. pp. 609–616 (2012).
22. Han, Y., Chan, J., Leckie, C.: Analysing Virtual Machine Usage in Cloud Computing. In: Services (SERVICES), 2013 IEEE Ninth World Congress on. pp. 370–377 (2013).
23. Huang, Q., Su, S., Xu, S., Li, J., Xu, P., Shuang, K.: Migration-Based Elastic Consolidation Scheduling in Cloud Data Center. In: Distributed Computing Systems Workshops (ICDCSW), 2013 IEEE 33rd International Conference on. pp. 93–97 (2013).
24. Calheiros, R.N., Masoumi, E., Ranjan, R., Buyya, R.: Workload prediction using ARIMA model and its impact on cloud applications' QoS. IEEE Trans. Cloud Comput. 3, 449–458 (2015).
25. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control. John Wiley & Sons, Hoboken, N. J. (2008).
26. Box, G.E.P., Cox, D.R.: An analysis of transformations. J. R. Stat. Soc. Ser. B. 211–252 (1964).
27. R Core Team: R: A Language and Environment for Statistical Computing, https://www.r-project.org/.
28. Watts Up? Plug Load Meters, www.wattsupmeters.com.
29. ZABBIX: The Enterprise-Class Monitoring Solution for Everyone, http://www.zabbix.com/.
30. Moreno-Vozmediano, R., Montero, R.S., Llorente, I.M.: IaaS cloud architecture: From virtualized datacenters to federated cloud infrastructures. Computer (Long. Beach. Calif). 65–72 (2012).
31. KVM -: Kernel-based Virtual Machine, http://www.linux-kvm.org/.