Volume 5, Issue 2, February 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Power Consumption and Revenue Growth in a Large-Scale Utility Cloud

Prasanthi Boyapati, Praveen Kumar Dudiki
Assistant Professor, Department of Computer Science& Engineering,
RVR&JC CE, Chowdavarm, Guntur,
AP, India

Abstract*— Modern data centres that operate under the cloud computing model are hosting a variety of applications ranging from those that run for a few seconds (e.g., serving requests of Web applications such as ecommerce and social network portals) to those that run for longer periods of time (e.g., simulations or large dataset processing) on shared hardware platforms. The need to manage multiple applications in a data centre creates the challenge of on-demand resource provisioning and allocation in response to time-varying workloads. Normally, data centre resources are statically allocated to applications based on peak load characteristics in order to maintain isolation and provide perfor- mance guarantees. Until recently, high performance has been the sole concern in data centre deployments, and this demand has been fulfilled without paying much attention to energy consumption. Energy costs of powering a typical data centre doubles every five years. Because energy costs are increasing while availability dwindles, there is a need to shift focus from optimizing data centre resource management for pure performance alone to optimizing for energy efficiency while maintaining high service-level performance.

***Index Terms*—Cloud Computing, Energy Characterization, Cloud Computing Simulation, Energy modeling in cloud**

I. INTRODUCTION

CLOUD computing environments are large-scale heterogeneous systems that are required to meet Quality of Service requirements demanded by consumers in Order to fulfil diverse business objectives. Such system Characteristics result in a diversity of Cloud efficiency in terms of user behaviour, task execution length and energy utilization patterns. In this context, Energy is defined as: "The amount of power consumed to, or done by, a Client, workgroup, server, or system in a given time period" and consists of two components: tasks and users. Tasks are defined as the basic unit of computation Assigned or performed in the Cloud, and a user is defined as the actor responsible for creating and configuring the volume of tasks to be computed. In order to further enhance the effectiveness of managing Cloud Computing environments there are two critical requirements. The first is that such environments require extensive and continuous analyses in order to understand and quantify the characteristics of system components. The Second is the exploitation of the parameters derived from such analyses in order to develop simulation models which accurately reflect the operational conditions Analysis and simulation of Cloud tasks and users significantly benefits both providers and researchers, as it enables a more in-depth understanding of the entire system as well as offering a practical way to improve data centre functionality.

For providers, it enables a method to enhance resource management mechanisms to effectively leverage the diversity of users and tasks to increase the productivity and QoS of their systems. For example, exploiting task heterogeneity to reduce performance interference of physical servers or analyzing the correlation of failures to power consumption. For researchers, simulation of Cloud workload enables evaluation of theoretical mechanisms supported by the characteristics of Cloud data centres.

Ideally such simulation parameters are derived from the empirical analysis of large-scale production Cloud data centres. Failure to do so results in misleading assumptions about the degree of workload diversity that exists within the Cloud and the creation of unrealistic Simulation parameters. This consequently results in limitations to the usefulness and accuracy of simulation Parameters. However, deriving such analyses is challenging in two specific areas. The first and most critical problem is that there are few available data sources pertaining to large-scale production utility Clouds, due to business and confidentiality concerns. This is a particular challenge in academia, which relies on the very few publicly available Cloud tracelogs. The second problem is analysis and simulation of realistic workloads; this is due to the massive size and complexity of data that a typical production. Cloud can generate in terms of sheer volume of users and server events as well as recording resource utilization of tasks.

Recently, there has been initial work from the analysis of limited Cloud traces from Google [2], [3] and Yahoo! [4] in an effort to provide mechanisms to analyze and characterize workload patterns. However, such efforts are predominately constrained to traces of short observational periods [5] and coarse-grain statistics [6] which are not sufficient to characterize the workload diversity of Cloud environments. In addition, there have been a number of approaches that analyze the diversity of workload by classifying tasks according to critical characteristics [7], [8], [9].

However, none of these provide a comprehensive study of the diversity of users and tasks, or provide a model containing sufficient details about the model parameters obtained from the analyses in order to be of practical use to researchers.

The objective of this paper is to present an in-depth Energy efficient analysis of cloud environment and its diversity in a large scale production Cloud computing data centre. Additionally, this work aims to provide a validated simulation model that includes parameters of tasks and users to be made available for other researchers to use. The analysis is conducted using the data from the second version of the Google Cloud tracelog [3], [10], which contains over 25 million tasks, submitted by 930 users over the observational period of a month There are three core contributions within this work:

1. An in-depth statistical analysis of the characteristics of energy diversity within a large-scale production Cloud. The analysis was performed over the entire tracelog time span as well as a number of observational periods to investigate patterns of diversity for both users and tasks within the system.
2. An extensive analysis of distribution parameters derived from the energy analysis that can be applied to simulation tools by other researchers.
3. A comprehensive validation of the simulation model based on empirical and statistical methods. A significant contribution of the simulation model provided is that it does not just replay the data within the tracelog. Instead, it creates patterns that randomly fluctuate based on realistic parameters. This is important in order to emulate dynamic environments and to avoid just statically reproducing the behaviour from a specific period of time.

A secondary contribution of this paper is presenting practical applications of the model obtained to identify sources of inefficiencies and enhance resource-management and energy usage in virtualized Cloud environments. This paper applies the methodology of analysis introduced in our previous approach [9], but is substantially different in a number of ways. First, this paper focuses specifically on a substantial analysis of Cloud diversity for tasks and users. Additionally, we analyze the entire tracelog time span and three additional observational periods, instead of just two days—which limited the original approach's applicability, as it could potentially omit crucial behaviour within the overall Cloud environment. Furthermore, extensive analysis and parameter details are provided for user and task distributions. The remainder of this paper is organized as follows: Section 2 presents the background; Section 3 discusses related work; Section 4 details the methodology used. Section 5 presents the cluster and distribution analysis of task and user diversity. Section 6 presents the validation of the model simulation. Section 7 describes the improvements to the model based on the validation results. Section 8 discusses practical applications of the results obtained with in this paper. Sections 9 and 10 discuss the conclusions and further research directions of this work, respectively.

II. BACKGROUND

2.1 Diversity Energy Patterns in Cloud

According to the NIST [11], the Cloud computing model has the following five essential characteristics: on-demand self service, resource pooling, broad network access, rapid elasticity and measured service. These characteristics create highly dynamic environments where customers from different contexts co-exist submitting workloads with diverse resource requirements at anytime. Workloads by them selves have properties or attributes that describe their behaviour. These attributes are normally expressed by the type and amount of resources consumed and other attributes that could dictate where a specific workload can or cannot be executed. For example, security requirements, geographical location, or specific hardware constraints such as processor architecture, number of cores or Ethernet speed among others described in [13]. As discussed in [14], as more and more customers adopt

Cloud platforms to fulfil their IT requirements, Cloud providers need to be prepared to manage highly heterogeneous workloads that are served on the top of shared infrastructure. Workloads can be broadly classified according to the fundamental resources that they consume in terms of CPU, memory and storage-bound workloads [15]. Moreover, depending on the interaction with the end-users, they can also be classified as latency sensitive and batch workloads [16]. Common examples of workloads running in multi-tenant Cloud data centres according to [17] include Business Intelligence, scientific high-performance computing, gaming and simulation.

2.2 Importance of Energy Models in Cloud

Models abstract reality to aid researchers and providers in understanding system environments in order to develop or enhance such systems. Workload models enable a way to actually study Cloud environments and the effect of workload variability on the performance and productivity of the overall system. Specifically, they support researchers and providers in further understanding the actual status and conditions of the Cloud system and identify Key Performance Indicators (KPI) necessary to improve operational parameters. Such models can be used in a number of research domains including resource optimization, security, dependability and energy-efficiency. In order to produce realistic models, it is critical to derive their components and parameters from real-world production tracelogs. This leads to capturing the intrinsic diversity and dynamism of all co-existing components within the system as well as their interactions. Moreover, realistic workload models enable the simulation of Cloud environments whilst being able to control selected variables to study emergent system-wide behaviour, as well as support the estimation of accurate forecasting under dynamic system conditions to improve QoS offered to users. This supports the enhancement of Cloud Management Systems (CMSs) as it allows providers to experiment with hypothetical scenarios and assess their decisions as a result of changes within the Cloud environment(i.e., Capacity planning for increased system size, alteration of the workload scheduling algorithm, performance tradeoffs, and service pricing models).

III. RELATED WORK

The analysis of workload patterns for Cloud computing environments has been addressed previously [5], [6], [7], [8], [9], [18], [19], [20], [21], [22]. In this section, the most relevant approaches are described; their limitations and gaps are also discussed. Wang et al. [22] present an approach to characterize the workloads of Cloud computing Hadoop ecosystems, based on an analysis of the first version of the Google tracelog [2].

The main objective of this work is to obtain coarse-grain statistical data about jobs and tasks to classify them by duration. This characteristic limits the work's application to the study of timing problems, and makes it unsuitable to analyze other Cloud computing issues related to resource usage patterns. Additionally, the analysis focuses on tasks and ignores the relationship with the users, a crucial component in Cloud workload as discussed previously. Zhang et al. [5] present a study to evaluate whether the mean values for task waiting time, CPU, Memory, and disk consumption are suitable to accurately represent the performance characteristics of real traces. The data used in their study is not publicly available and consists of the historical traces of six Google compute clusters spanning five days of operation. The evaluation conducted suggests that mean values of runtime task resource consumption is a promising way to describe overall task resource usage. However, it does not describe how the boundaries for task classification were made and how members behave. Mishra et al. [7] describe an approach to develop Cloud computing workload classifications based on task resource consumption patterns. The analyzed data consist of records from five Google clusters over four days. The proposed approach identifies workload characteristics, constructs the task classification, identifies the qualitative boundaries of each cluster and then reduces the number of clusters by merging adjacent clusters. This approach is useful to create the classification of tasks, but does not perform an analysis of the characteristics of the formed clusters in order to derive a detailed workload model. Finally, it is entirely focused on task modelling, neglecting user patterns. Kuvulya et al. [6] present a statistical analysis of MapReduce traces. The analysis is based on ten months of MapReduce logs from the M45 supercomputing cluster [4]. Here, the authors present a set of coarse-grain statistical characteristics of the data related to resource utilization, job patterns, and source of failures. This work provides a detailed-
description of the distributions followed by the job completion times, but only provides very general information about the resource consumption and user behavioural patterns. Similar to [22], this characteristic limits the proposed approach mainly to the study of timing problems. Aggarwal et al. [8] describe an approach to characterize Hadoop jobs. The analysis is performed on a data set spanning 24 hours from one of Yahoo!'s production clusters comprising of 11,686 jobs. This data set features metrics generated by the Hadoop framework. The main objective of this work is to group jobs with similar characteristics using clustering to analyze the resulting centroids. This work only focuses on the usage of the storage system, neglecting other critical resources such as CPU and Memory. Our previous work [9] provides an approach for characterizing Cloud energy based on user and task patterns using the second version of the Google tracelog; it presents coarse-grain statistical properties of the tracelog, and classifies tasks and users using statistical mechanisms to select the number of clusters. A concise analysis of the clusters is performed as well as best fit distributions for each. Finally, the derived analysis parameters are simulated and compared against the empirical data for validation. This work has a number of limitations; the analysis performed is confined to only two days as opposed to the entire tracelog time span, resulting in the potential omission of crucial system environment behaviour. Also, the cluster analysis and intra-cluster analysis do not contain sufficient detail to quantify the diversity of workload, instead presenting high-level observations. Furthermore, there is insufficient detail about the parameter distributions used; more detail is necessary in order for other researchers to simulate the workload obtained. Finally, the validation of the simulated model against that of the empirical data is based only on a visual match of the patterns from one single execution, and does not consider more rigorous statistical techniques. From the analysis of the related work it is clear that there are few available production tracelogs to analyze workload patterns in Cloud environments. Previous analyses present gaps that need to be addressed in order to achieve more realistic workload patterns. It is imperative to analyze large data samples as performed by [5], [6], [9]. Small operational time frames as those used in [7], [8], [22] could lead to unrealistic models. Second, analyses need to explore more than coarse-grain statistics and cluster centroids. To capture the patterns of clustered individuals it is also necessary to conduct analysis of the parameters and study the trends of each cluster characteristic. Although previously approaches offer some insights about workload characteristics, they do not provide a structured model which can be used for conducting simulations. Finally, the workload is always driven by the users, therefore realistic workload models must include user behavioural patterns linked to tasks. The approaches previously described completely focus on tasks, neglecting the impact of user behaviour on the overall environment workload.

IV. METHODOLOGY

The methodology, analysis and subsequent simulation within this paper was applied to the second version of the Google Cloud tracelog [3], [10] which contains over 12,000 servers, 25 million tasks and 930 users over the period of a month. The tracelog includes detailed data such as submission patterns, resource requests of users and resource consumption of tasks within the system.The methodology is divided into two distinct steps: The first is defining the model that will be used for simulating the Cloud workload from the derived data set analysis. As stated previously, users are responsible for driving the volume and behaviour of tasks in terms of requested resources and the volume of task submission. Therefore, three important characteristics that define this behaviour within the tracelog are referred to as parameters that are fundamental to describe the user behaviour: the submission rate a, and requested amount of CPU b and Memory f. The submission rate is the quotient of dividing the number of submissions by the tracelog time span and is presented as task submissions per hour. Requested CPU and memory are represented as normalized resources requested by users taken directly from the task events log within the tracelog.

V. ANALYSIS OF DIVERSITY

This section presents the analysis of user and task characteristics within the tracelog after performing the k-means clustering algorithm on the entire tracelog time span as described in Section 4. Specifically, we are interested in quantifying and characterizing the diversity of user and task behaviour that exists within the system environment. The analysis is divided into two sections; cluster analysis and distribution analysis. The cluster analysis discusses the characteristics and behaviour of the k-clusters and studies the statistical properties of each parameter within the clusters for users and tasks, including the Mean, Standard Deviation and Coefficient of Variation (Cv). The distribution analysis consists of analyzing the inner data distributions for each of the components within each cluster parameter for tasks and users.

This required fitting the data to the closest theoretical distribution using a Goodness of Fit (GoF) test to obtain the parameters of their Probabilistic Distribution Functions (PDF). The data of each cluster is fitted to a parametric distribution by using the Anderson-Darling (AD) GoF statistical test. The theoretical distribution with the lowest ADvalue is selected to represent the data distribution of each cluster. The objective is to use the PDFs of the parameters in the workload model described in Equations (3) and (4). A number of assumptions for the distribution analysis can be found in [9]. The main alteration to the methodology in
order to improve the accuracy of the model is to consider the amount of CPU and memory requested by users instead of the proportions of overestimation and underestimation of resources. This is because the overestimation is an approximated value, whilst the amount of requested resources
is a factual value which produces more accurate results.

Moreover, for both the cluster and distribution analysis we have also investigated the variance of task and user clusters and parameters over a number of observational periods. The reason for this is to inspect patterns that exist within the data and to explore the degree of variance over the system lifespan. As a result, this analysis comprises of four observational periods; the entire month trace, Day 2, Day 18 and Day 26. The latter three observational periods were selected for two reasons: First, they represent observational periods of low task length, high submission rate and an average of these two parameters respectively. Second, the periods are temporally far apart, and provide insight into system diversity at different system states.

5.1 Cluster Analysis

Fig. 1 illustrates the k-clusters partitioning that satisfies f (k) < 0.85 for users across observational periods. It can be observed from Fig. 1a that the majority of users across the entire month request similar portions of CPU and memory, and exhibit similar submission rates. Furthermore, there are three specific users that have a substantially

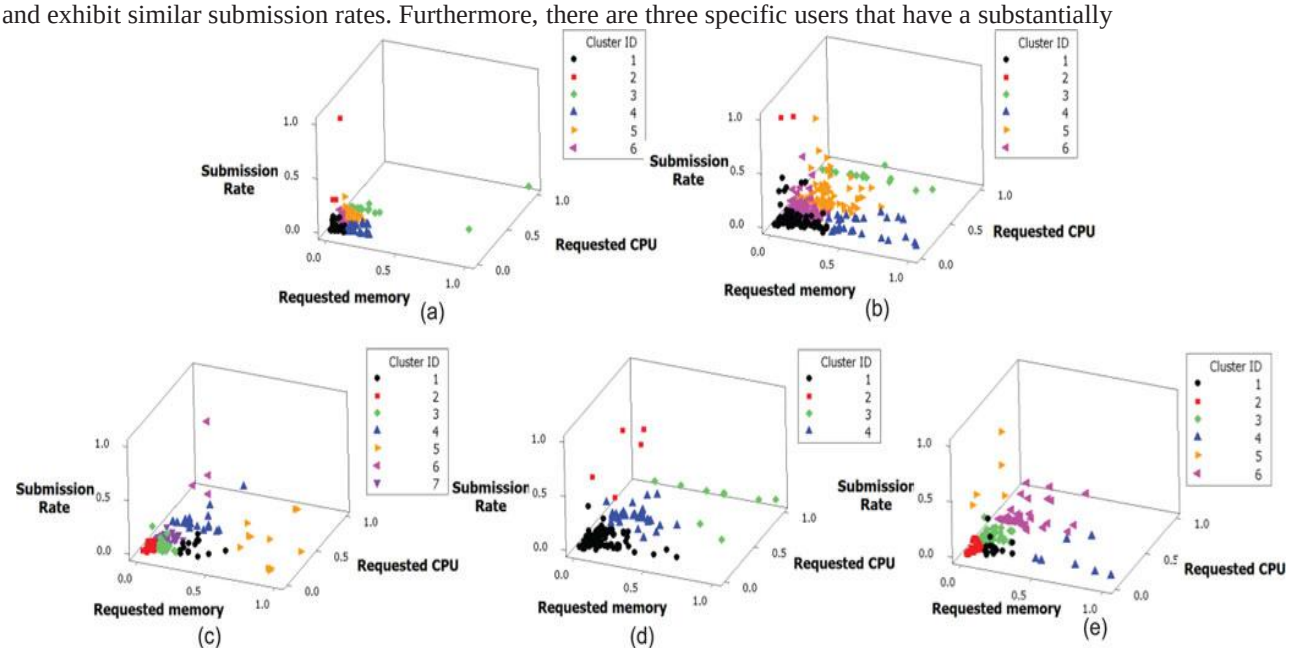

Fig. 1. Clusterization for users (a) entire month (b) entire month (omitting outliers) (c) Day 2, (d) Day 18, and (e) Day 26.

high submission rate and request larger amounts of CPU and memory as shown in clusters 2 (U2) and 3 (U3), respectively. When omitting these three users from the cluster analysis in Fig. 1b it is clearer to observe that clusters characteristics are similar across additional observational periods as demonstrated in Figs. 1c, 1d, and 1e, with a substantial amount of users exhibiting a similar submission rates and resource request patterns. Table 2 shows the statistical properties of each parameter for the defined clusters for the entire tracelog period. It is observable that users follow different resource utilization and submission patterns. For example, U2 contains 0.71 percent of the total user population and has an incredibly high submission rate in comparison to other clusters. Another example is that U3 has the highest average requested CPU and memory, but has the lowest submission rate, indicating this type of user infrequently submits more resource intensive tasks.

We observe that requested CPU and memory across most clusters exhibits low variance, with an average Cv of 0.42 and 0.79 respectively (U3 requested memory appears to have higher variance due to the strong influence of three specific users discussed above). The parameter submission rate exhibits highly variant behaviour across all user clusters, with an average Cv of 1.97. U2 is the only user cluster whose Cv submission rate is less than 1, which is most likely due to the cluster population size of 3. There are three reasons for the above observations. First, as reported in previous works [9] the Cloud data centre environment is naturally heterogeneous in workload due to user behaviour. Second, requested resources by users are possibly a reflection of the application and system domain boundaries. For example, applications deployed or invoked within the Cloud environment have pre-defined resource requests to meet the demands of user QoS. Third, the submission rate is outside the boundaries of the system and is entirely driven by users; Such behaviour is reflective of the definition of Cloud computing, which provides the illusion of infinite resource to users [25], allowing them to submit as many tasks as required without conscious thought about system limitations. The k-clusters for tasks across all observational periods, and demonstrates that it was possible to define three clusters for all observational periods where $f(k) < 0.85$. It is observable that the cluster shapes are visually similar across all observational periods, with cluster 3 (T3) containing the lowest values for CPU, memory and length while T2 exhibits more variant behaviour. Moreover, T2 composes less than 2 percent of the total task population and T3 contains over 70 percent of the task population across all time periods as shown in Table 3. In addition, we observe that the proportions of tasks within the clusters stay relatively constant. In comparison to the heterogeneity of user clusters, task patterns appear to be more uniform across different observational periods. Table 4 presents the statistical properties of the task parameters length, CPU and Memory utilization for all clusters across the four observational periods. It is possible to make a more balanced comparison of task clusters over different time periods in contrast to user clusters due to the observed stability. Similar to the characteristic of user submission rate, we observe that task length is highly heterogeneous across all clusters and observational periods with an average Cv of 2.36, indicating high variation between values. This is due to the same reasons as for the variability that exists for user submission rates; task length is a parameter that is outside the boundaries of the system environment and is entirely dependent on the demands of the user (i.e., Users will execute tasks of different execution length to meet their QoS demands). CPU and memory are less variable due to application domain constraints imposed by the system environment, reflected by an average Cv value of 0.93 and 0.83 for CPU and memory utilization respectively. These results highlight two important findings. First, when quantifying the diversity of the Cloud environment, it appears that parameters that are outside the boundaries of the system environment introduce the highest level of heterogeneity. This is demonstrated by the parameters user submission rate and task execution length exhibiting highly variant behaviour in comparison to CPU and memory requests and utilization for users and tasks, respectively. Second, the diversity of workload imposed by these two parameters introduces potential challenges to workload prediction; for this case, where the parameters are highly variable and dynamic; the expiration time of historical data seems to be considerably shorter. Therefore, there exists the need for adaptive and evolving mechanisms that allow providers to obtain more accurate predictions.

VI. MODEL SIMULATION

Data centres are not only expensive to maintain, they are also unfriendly to the environment. Carbon emissions due to data centres worldwide are now more than the emissions of both Argentina and the Netherlands [118]. High energy costs and huge carbon footprints are incurred due to the massive amount of electricity needed to power and cool the numerous servers hosted in these data centres. Cloud service providers need to adopt measures to ensure that their profit margins are not dramatically reduced due to high energy costs. According to Amazon's estimate, the energy-related costs of its data centres amount to 42% of the total budget, which includes both direct power consumption and the cooling infrastructure amortized over a 15-year period. As a result, companies such as Google, Microsoft, and Yahoo! Are building large data centres in barren desert lands ur-rounding the Columbia River in the United States to exploit cheap hydro electric power. There is also increasing pressure from government's world wide to reduce carbon footprints, which have a significant impact on climate change. To address these concerns, leading IT vendors have recently formed a global consortium, called The Green Grid, to promote energy efficiency for data centres and minimize their impact on the environment. Pike Research forecasts that data centre energy expenditures world wide will reduce from $23.3 billion in 2010 to $16.0 billion in 2020, as well as causing a 28% reduction in green house gas (GHG) emissions from 2010 levels as a result of the adoption of the cloud computing model for delivering IT service. Lowering the energy usage of data centres is a challenging and complex issue because computing applications and data are growing so quickly that larger servers and disks are needed to process them fast enough within the required time period. This cloud computing is envisioned to achieve not only efficient processing and utilization of computing infrastructure but also minimize energy consumption. This is essential for ensuring that the future growth of cloud computing is sustainable. Cloud computing, with increasingly pervasive front-end client devices such as iPhones interacting with back-end data centres, will cause an enormous escalation in energy usage. To address this problem, data centre resources need to be managed in an energy-efficient manner to drive green cloud computing. In particular, cloud resources need to be allocated not only to satisfy QoS requirements specified by users via service-level agreements (SLAs) but also to reduce energy usage. This can be achieved by applying market-based utility models to accept user requests that can be fulfilled to enhance revenue along with energy-efficient utilization of cloud infrastructure. In order to characterize and analyse the performance of similar large-scale Cloud data centres under a projected set of operating conditions, we implemented the task and user model parameters described previously as an extension to the CloudSim framework [26], [27], [28], [29]. CloudSim is a Java based framework that enables the simulation of complete Cloud Computing environments [27]. It provides abstraction of

all the elements within the Cloud computing model and the interaction among them. However, as with any other simulation software, the quality and accuracy of the results entirely depends on how accurately the introduced parameters reflect the analysed system in reality. The following subsections describe the implemented workload generator and the conducted simulation validation. The model components and their relationship are formalized
in Equations (1) to (6).

$U = \{ u1,u2,u3,u4,.....,u_i \}$ (1)

$T = \{ t1,t2,t3,t4,.....,t_i \}$ (2)

$ui = \{f(\alpha),f(\beta),f(\varphi)\}$ (3)

$ti = \{f(x),f(\gamma),f(pie)\}$ (4)

$E(u_i)=u_iP(u_i)$ (5)

.

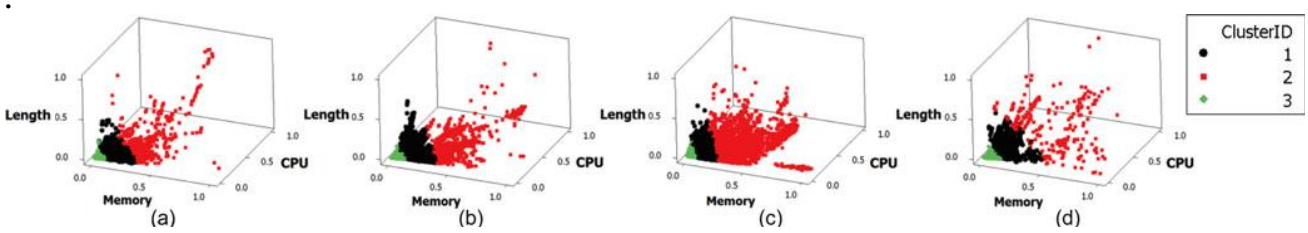

Fig. 2. Clusterization for tasks (a) entire month, (b) Day 2, (c) Day 18, and (d) Day 26.

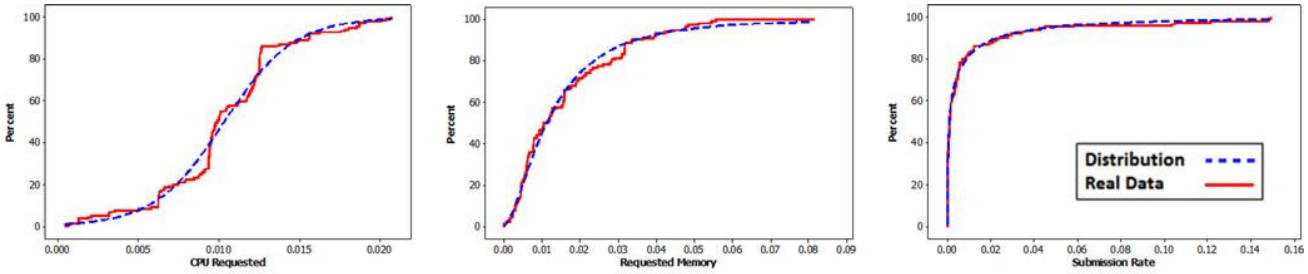

Fig. 3. CDF of user cluster U1 (a) CPU requested, (b) memory requested, and (c) submission rate.

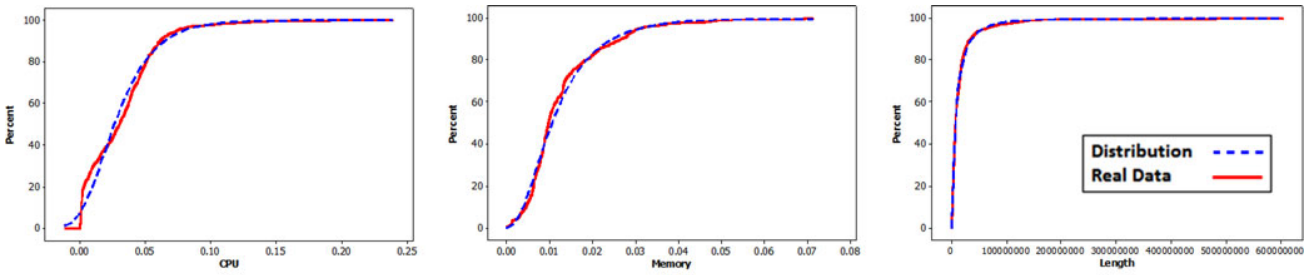

Fig. 4. CDF of task cluster T1 (a) CPU, (b) memory, and (c) submission rate.

6.1 Workload and Environment Generator

The workload and environment generator is composed of six modules: The Profile Manager, Data centre Generator, Customer Generator, Task Generator and Environment Coordinator. The user and task profiles describe respectively the user and task types identified during the clustering process and encapsulate the outlined behavioural patterns derived during the cluster and distribution analysis. The server profiles describe the capacities and characteristics of the data centre hosts according to the data within the trace log. These characteristics as well as the proportion of servers from each type are listed in Table 8. The profiles manager loads each element description making them available to the generators. The User Generator creates the CloudSim user instances and connects them with a specific profile determined by their associated probabilities as described in Equation (5). The Task Generator creates the CloudSim task instances and connects them with a specific task profile determined by the conditional probability in Equation (6). Each one of the user and task characteristics defined such as submission rate, length and resource consumption described in the model are obtained by sampling the inverse CDFs of the distributions in Equations (3) and (4). Finally, the Environment Coordinator controls the interactions between the three generators and the CloudSim framework that executes the simulation with the created instances.

6.2 Simulation Configuration

We have executed a model simulation of a data centre composed of 12,000 servers with 160 customers submitting tasks during 24 hours a total of five iterations. The user and task profiles are configured using the statistical

parameters derived for the entire month analysis as described in Tables 5 and 6. The profiles of the simulated servers are outlined from the tracelog as presented in Table 8 where the values of CPU and memory are normalized. The normalization is a scaling relative to the largest capacity of the resource on any server in the trace which is 1.0.

6.3 Simulation Validation

Model validation is defined as the "substantiation that a computerized model with its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model" [3]. In the case of the historical data of trace-driven models where the analyst does not have access to the real system or to a different dataset sample from the same system, a common validation technique consists of using a portion of the available data to construct the model and the remaining data to determine whether the model behaves as the real system does. This is typically addressed by sampling the analyzed tracelog where both the input and the actual system response must be collected from the same period of time [31]. According to Sargent [30], there are two basic approaches in comparing the simulation model to the behaviour of the real system. The first consists of using graphs to empirically evaluate the outputs and the second involves the application of statistical hypothesis tests to make an objective decision. To validate our model simulation we use both techniques; the proportions of categorical data such as task, user and server types as well as tasks priorities are contrasted empirically by plotting comparative charts and evaluating the absolute error between the average output from the simulations and the data in the real system. Additionally we analyze the variability of results and their corresponding confidence interval (CI). On the other hand, continuous data such as the user and task resource request and consumption patterns are compared statistically using the Wilcox Mann-Whitney test (WMW) [32], [33]. WMW is one of the most powerful non-parametric tests for comparing two populations.

According to Mauger [34], "it is based on the test of the null hypothesis that the distributions of two populations, although unspecified, are equal, against the alternative hypothesis that the distributions have the same shape but are shifted, so the outcomes of one population tends to be larger than the other". It is commonly applied instead of the two-sample t-test when the analyzed data does not follow a normal distribution as is the case of the outlined user and tasks patterns. Additionally, in order to verify the consistency of the WMW test, we have applied the Fisher's Method [35]; a meta-analysis technique to combine p-values from different and independent tests which have the same null hypothesis. The objective is to verify whether the rejections are statistically significant given the variances reported, or are consistent with the results of the other simulations.

6.4 Validation Results

The results from our simulation experiments demonstrate the accuracy of the derived model to represent the operational characteristics of the workload within the Cloud computing data centre for the analyzed scenario. the proportion of components (users, tasks, task priorities and servers) created during the simulations which are contrasted against the observations from the real system. Comparing the average simulation outputs with the real values, it is possible to observe that simulated proportions of fundamental elements consistently match the proportions of the elements in the actual system. From the detailed results presented in Table 9, it can be observed that while the proportions of tasks do not significantly fluctuate, the proportions of users and servers across different simulation executions present a higher variability. This is mainly produced by a very small population of specific clusters.

VII. IMPROVEMENTS OF CPU CONSUMPTION PATTERNS

This makes fitting such data sets with a single theoretical distribution unsuitable and creates significant gaps between the simulated and real data as observed. To improve the accuracy of our model, we applied "multi-peak histogram analysis for region splitting" [38] and fitted the derived dataset sub-regions to new parametrical distributions. Essentially, the data is ranked and presented in a histogram, which is split based on the lowest points of the different valleys created by the multimodal distribution. To identify the peaks and valleys of a given multimodal data set, we smooth the histogram by applying the LOWESS [36] (Locally-Weighted Scatter plot Smoother) technique using the Minitab statistical package [37]. Then, the derived sub-regions are fitted to new parametrical distributions following the same process described in Section 5.2. Consequently, the CPU utilization patterns of the affected clusters comprise a combination of different distributions which are sampled by the model simulator based on the proportional size of the derived sub-regions. The distribution parameters and sizes of the obtained sub-regions The results of this process are illustrated in Fig. 9 where it can be observed that the split distributions improve the fitting between the simulated and real datasets. The p-values of the WMW test for both clusters are sufficiently statistically strong to support the equality of patterns. This reduces the error for execution time from 8.07 to 0.42 percent and from 5.91 to 0.13 percent for T2 and T3, respectively.

VIII. APPLICATION OF WORK

The workload model presented in this paper enables researchers to simulate request and consumption patterns considering parameters and patterns statistically close to those observed from a production environment. This is critical in order to improve resources utilization, reduce energy waste and in general terms support the design of accurate forecast mechanisms under dynamic conditions to improve the QoS offered to customers. Specifically, we use the proposed model to support the design and evaluation of two energy-aware mechanisms for Cloud computing environments. The first is a resource over allocation mechanism that considers customers' resource request patterns and the actual resource utilization imposed by their submitted tasks. Taking into account these parameters from the proposed

model it is possible to estimate the resource overestimation patterns. The main idea is to exploit the resource overestimation patterns of each user type in order to smartly over allocate resources to the physical servers. This reduces the waste produced by frequent overestimations and increases data centre availability. Consequently, it creates the opportunity to host additional Virtual Machines in the same computing infrastructure, improving its energy-efficiency [39].The second mechanism considers the relationship between Virtual Machine interference due to competition for resources and energy-efficiency. The core idea is to collocate different types of workloads based on the level of interference that they create, to reduce resultant overhead and thus improve the energy-efficiency of the data centre. By considering the resource consumption patterns of each task type we estimate the level of interference and energy efficiency decrement when they are co-located in a physical server. We classify incoming tasks based on their resource usage patterns, pre-select the hosting servers based on resources constraints, and make the final allocation decision based on the current servers' performance interference level [40]. In both cases the proposed workload model and the parameters derived from the presented analysis are used to emulate the user and tasks patterns required by the energy-aware algorithms. The model integrates the relationship between user demand and the actual resource usage—essential in both scenarios where the aim is to achieve a balance between resource request and utilization in order to reduce resource waste. Another important benefit of our approach is that as values of customer and task parameters are represented as proportions of resources requested or consumed, they areagnostic of underlying hardware characteristics. Therefore, the proposed model can be used to evaluate the performance of different data centre configurations under the same workload. Furthermore, the comprehensive analysis at cluster and intra-cluster level, the workload model that integrates user and tasks patterns and the applicability of the model independently of the hardware characteristics represent unique advances in comparison with the related work previously discussed in Section 3. Additionally, the proposed model supports the assessment of resource management mechanisms such as those recently presented in [41], [42] and [43] with parameters from a large-scale production Cloud environment.

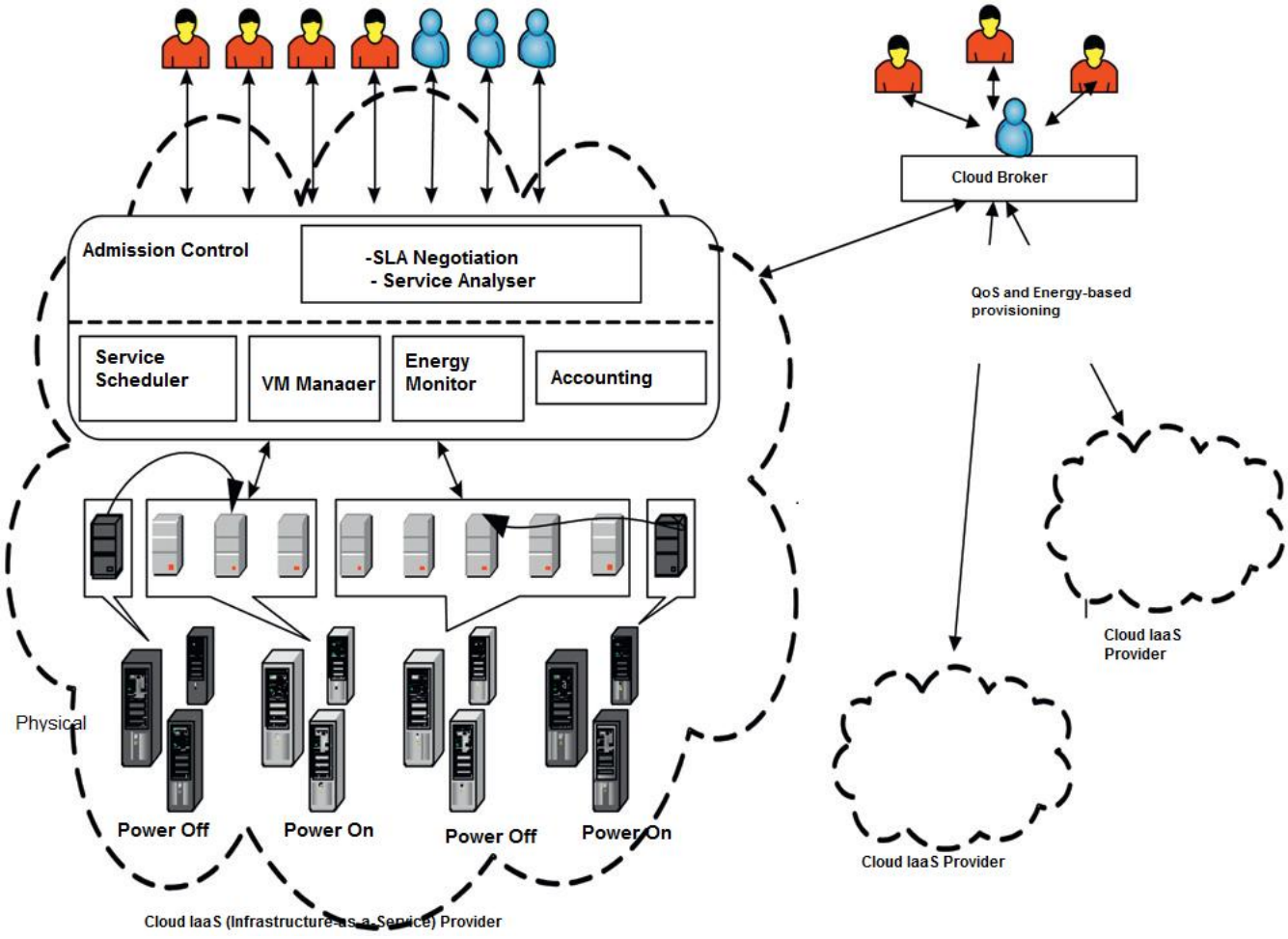

Fig. 5. Cloud Iaas (Infrastructure-as-a-Service) Provider in consumption of power

IX. CONCLUSIONS

This paper presents an analysis that quantifies the diversity of Cloud workloads and derives a workload model from a large-scale production Cloud data centre. The presented analysis and model captures the characteristics and behavioural patterns of user and task variability across the entire system as well as different observational periods. The derived model is implemented using the CloudSim framework and extensively validated through empirical comparison and statistical tests. From the observations presented within this work and the results obtained from the simulations, a number of conclusions can be made. These are as follows:

1. Workload in Cloud data centres is driven not only by tasks characteristics but also by user behavioural patterns. Related approaches on workload analysis are focused on parameters such as the duration and the resources consumed by tasks. However, as observed from the presented analysis, in some scenarios specific types of users impose a strong influence on the overall Cloud workload. Therefore, comprehensive workload models must consider both tasks and users in order to reflect realistic conditions.
2. User patterns tend to be significantly more diverse than task patterns across different observational periods. Depending on the type of service offered, providers can control the type of tasks and the environment in which they are running (i.e., SaaS and PaaS). This can create more "stable" tasks patterns over the time. On the other hand, user patterns tend to change according to needs derived from their own business objectives which are completely out of the boundaries of Cloud providers. This creates new challenges on workload prediction mechanisms that need to evolve and adapt according to such dynamic characteristics.
3. Describing Cloud analyses is an important first step, but providing the parameters and characteristics derived from these analyses is critical. This supports the development and validation of simulation models as presented in this work. Such simulations can support the evaluation of new operational policies, new system designs, and support the decision-making process as result of changes in the Cloud environment.
4. Workload models can be exploited to improve diverse and critical operational parameters. This paper has presented two examples of how the derived model can be used to improve performance and energy efficiency by exploiting the diversity of users and tasks. In addition, the workload model can be used to improve parameters such as security, dependability, and economics.

X. FUTURE WORK

Future research directions includes extending the model to include tasks constraints based on server characteristics; this will allows us to analyze the impact of hardware heterogeneity on workload behaviour. Other extensions include analyzing the workload from the jobs perspective specifically modelling the behaviour and relationship of users and submitted jobs, accurately emulating and analyzing workload energy consumption and reliability enabling further research into energy-efficiency, resource optimization and failure-analysis in the Cloud environment. Finally, it is important to enable a collaboration link with the CloudSim group in order to integrate the proposed workload generator as an add-in of the current framework implementation allowing it to be made publicly available.

ACKNOWLEDGMENTS

The work was supported by CONACyT (No. 213247), the National Basic Research Program of China (973) (No. 2011CB302602), and the UK EPSRC WRG platform project (No. EP/F057644/1).

REFERENCES

[1] R. Buyya, R. Ranjan, and R. N. Calheiros, "InterCloud: Utility-oriented federation of cloud computing environments for scaling of application services," Proc. 10th Int. Conf. Algorithms Archit. Parallel Process., 2010, pp. 13–31.
[2] Google. Google Cluster Data V1 (2010). [Online] Available: http:// code.google.com/p/googleclusterdata/wiki/TraceVersion1
[3] Google. Google Cluster Data V2 (2011). [Online] Available: http:// code.google.com/p/googleclusterdata/wiki/ClusterData2011_1
[4] Yahoo. Yahoo! M45 Supercomputing Project. (2007). [Online]. Available: http://research.yahoo.com/node/1884
[5] Q. Zhang, J. Hellerstein, and R. Boutaba, "Characterizing task usage shapes in Google compute clusters," in Proc. 5th Int. Workshop Large Scale Distrib. Syst. Middleware, 2011, pp. 2–8.
[6] S. Kavulya, J. Tan, R. Gandhi, and P. Narasimhan, "An analysis of traces from a production MapReduce cluster," in Proc. IEEE/ACT Int. Conf. Cluster, Cloud Grid Comput., 2010, pp. 94–103.
[7] A. K. Mishra, J. Hellerstein, W. Cirne, and C. R. Das, "Towards cloud backend workloads: Insights from Google compute clusters," ACM SIGMETRICS Perform. Eval. Rev., vol. 37, pp. 34–41, 2010.
[8] S. Aggarwal, S. Phadke, and M. Bhandarkar, "Characterization of Hadoop jobs using unsupervised learning," in Proc. 2nd Int. Conf. Cloud Comput. Technol. Sci., 2010, pp. 748–753.
[9] I. Solis Moreno, P. Garraghan, P. Townend, and J. Xu, "An approach for characterizing workloads in Google cloud to derive realistic resource utilization models," in Proc. IEEE Int. Symp. Serv. Oriented Syst. Eng., 2013, pp. 49–60.
[10] C. Reiss, J. Wilkes, and J. Hellerstein, "Google Cluster-Usage Traces: Format þ Schema," Google Inc., Mountain View, CA, USA, White Paper, 2011.
[11] P. Mell and T. Grance, "The NIST definition of cloud computing," NIST Spec. Publication, vol. 800, p. 145, 2011.
[12] M. A. El-Refaey and M. A. Rizkaa, "Virtual systems workload characterization: An overview," in Proc. IEEE Int. Workshops Enabling Technol. Infrastructures Collaborative Enterprises, 2009, pp. 72–77.
[13] B. Sharma, V. Chudnovsky, J. Hellerstein, R. Rifaat, and C. R. Das, "Modeling and synthesizing task placement constraints in Google compute clusters," in Proc.ACMSymp. Cloud Comput, 2011, pp. 1–14.

[14] J. Zhan, L. Wang, W. Shi, S. Gong, and X. Zang, "PhoenixCloud: Provisioning resources for heterogeneous workloads in cloud computing," arXiv preprint arXiv:1006, vol. 1401, 2010.

[15] V. Vasudevan, D. Andersen, M. Kaminsky, L. Tan, J. Franklin, and I. Moraru, "Energy-efficient cluster computing with FAWN: Workloads and implications," in Proc. Int. Conf. Energy-Efficient Comput. Netw., 2010, pp. 195–204.

[16] T. N. B. Doung, X. Li, R. S. M. Goh, X. Tang, and W. Cai, "QoSaware revenue-cost optimization for latency-sensitive services in IaaS clouds," in Proc. IEEE/ACM Int. Symp. Distrib. Simul. Real Time Appl., 2012, pp. 11–18.

[17] IBM, "Get more out of cloud with a structured workload analysis," White Paper IAW03006-USEN-00, 2011.

[18] A. Bahga and V. K. Madisetti, "Synthetic workload generation for cloud computing applications," J. Softw. Eng. Appl., vol. 4, pp. 396–410, 2011.

[25] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley view of cloud computing," Univ. California, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2009-28, Feb. 2009.

[26] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and simulation of scalable cloud computing environments and the CloudSim toolkit: Challenges and opportunities," in Proc. Intl Conf. High Perform. Comput. Simul., 2009, pp. 1–11.

[27] 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," Softw. Practice Experience, vol. 41, pp. 23–50, 2010.

[28] S. K. Garg and R. Buyya, "NetworkCloudSim: Modelling parallel applications in cloud simulations," in Proc. IEEE Intl. Conf. Utility Cloud Comput., 2011, pp. 105–113.

[29] B. Wickremasinghe, R. N. Calheiros, and R. Buyya, "CloudAnalyst: A CloudSim-based visual Modeller for analysing cloud computing environments and applications," in Proc. IEEE Intl. Conf. Adv. Inf. Netw. Appl., 2010, pp. 446–452.

[30] R. G. Sargent, "Verification and validation of simulation models," in Proc.Conf, Winter Simul., 2010, pp. 166–183.

[31] O. Balci and R. G. Sargent, "Some examples of simulation model validation using hypothesis testing," Proc. Conf. Winter Simul., vol. 2, pp. 621–629, 1982.

[32] D. Brown and P. Rothery, "Models in biology: Mathematics, statistics and computing," Proc. 14th Conf. Winter Simul, 1993.

[19] A. Beitch, B. Liu, T. Yung, R. Griffith, A. Fox, and D. A. Patterson, "Rain: A workload generation toolkit for cloud computing applications," Elect. Eng. Comput. Sci. Univ. California, Berkeley, CA, USA, White Paper UCB/EECS-2010-14, 2010.

[20] Y. Chen, A. S. Ganapathi, R. Griffith, and R. H. Katz, "Analysis and lessons from a publicly available Google cluster trace," USA, EECS Dept., Univ. California, Berkeley, CA, UCB/EECS-2010-95., Jun. 2010.

[21] J. W. Smith and I. Sommerville, "Workload classification & software energy measurement for efficient scheduling on private cloud platforms," presented at the ACMSOCC, Cascais, Portugal, 2011.

[22] G. Wang, A. R. Butt, H. Monti, and K. Gupta, "Towards synthesizing realistic workload traces for studying the Hadoop ecosystem," in Proc. IEEE Int. Symp. Modeling, Anal. Simul. Comput. Telecommun. Syst., 2011, pp. 400–408.

[23] R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Trans. Neural Netw., vol. 16, pp. 645–678, 2005.

[24] D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of K in K-means clustering," Proc. Inst. Mech. Eng., Part C: J. Mech. Eng. Sci., vol. 219, pp. 103–119, 2005.

[33] A. Gold, "Understanding the Mann-Whitney Test," J. Property Tax Assessment Admin., vol. 4, pp. 55–57, 2007.

[34] D. T. Mauger and G. L. Kauffman Jr, "82 - statistical analysis—specific statistical tests: Indications for use," Surgical Research W. S. Wiley and W. W. Douglas, eds., San Diego, CA, USA, Academic, 2001, pp. 1201–1215.

[35] D. A. S. Fraser, A. K. M. Saleh, and K. Ji, "Combining p-values: A definitive process," J. Statist. Res., vol. 44, pp. 15–29, 2010.

[36] D. Borcard and P. Legendre, "Exploratory data analysis,"in Numerical Ecology, New York, NY, USA, Springer, pp. 9–30, 2011.

[37] Minitab, Version: Release 16 (2010). MINITAB statistical software [Online]. Available: http://www.minitab.com.

[38] S. Pal and P. Bhattacharyya, "Multipeak histogram analysis in region splitting: A regularization problem," in Proc. IEEE Comput. Digit. Tech., 1991, vol. 138, pp. 285–288.

[39] I. Solis Moreno and J. Xu, "Neural network-based overallocation for improved energy-efficiency in real-time cloud environments," in Proc. IEEE Int. Symp. Object/Compon./Serv.-Oriented Real-Time Distrib. Comput., 2012, pp. 119–126.

[40] I. Solis Moreno, R. Yang, J. Xu, and T. Wo, "Improved energy-efficiency in cloud datacentres with interference-aware virtual machine placement," in Proc. IEEE Int. Symp. Auton. Decentralized Syst., 2013, pp. 1–8.

[41] X. Lu, H. Wang, J. Wang, J. Xu, and D. Li, "Internet-based virtual computing environment: Beyond the data centre as a computer," Future Generation Comput. Syst., vol. 29, pp. 309–322, 2013.

[42] M. Kesavan, I. Ahmad, O. Krieger, R. Soundararajan, A. Gavrilovska and K. Schwan, "Practical compute capacity management for virtualized datacentres," IEEE Trans. Cloud Comput., vol. 1, no. 1, pp. 88–100, Jan.-Jun. 2013.

[43] J. Doyle, R. Shorten, and D. O'Mahony, "Stratus: Load balancing the cloud for carbon emissions control," IEEE Trans. Cloud Comput., vol. 1, no. 1, pp. 116–128, Jan.-Jun. 2013.

