Pertanika J. Sci. & Technol. 31 (5): 2559 - 2575 (2023)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Workload Characterization and Classification: A Step Towards Better Resource Utilization in a Cloud Data Center

Avita Katal[1]*, Susheela Dahiya[2] and Tanupriya Choudhury[1,2,3]

[1]School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, 248007, India
[2]Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, 248002, India
[3]Symbiosis Institute of Technology, Symbiosis International University, Pune, Maharashtra, 412115, India

ABSTRACT

Advancements in virtualization technology have led to better utilization of existing infrastructure. It allows numerous virtual machines with different workloads to coexist on the same physical server, resulting in a pool of server resources. It is critical to understand enterprise workloads to correctly create and configure existing and future support in such pools. Managing resources in a cloud data center is one of the most difficult tasks. The dynamic nature of the cloud environment, as well as the high level of uncertainty, has created these challenges. These applications' diverse Quality of Service (QoS) requirements make data center management difficult. Accurate forecasting of future resource demand is required to meet QoS needs and ensure better resource utilization. Consequently, data center workload modeling and categorization are needed to meet software quality solutions cost-effectively. This paper uses traces of Bitbrain's data to characterize and categorize workload. Clustering (K Means and Gaussian mixture model) and Classification strategies (K Nearest Neighbors, Logistic Regression, Decision Trees, Random Forest, and Support Vector Machine) characterize and model the workload traces. K Means shows better results as compared to GMM when compared to the Calinski Harabasz index and Davies-Bouldin score. The results showed that the Decision Tree achieves the maximum accuracy of 99.18%, followed by K Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM) Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Back Propagation Neural Networks.

Keywords: Classification, cloud data center, clustering, Gaussian mixture model, K Means, workload

ARTICLE INFO

Article history:
Received: 16 August 2022
Accepted: 14 November 2022
Published: 27 July 2023

DOI: https://doi.org/10.47836/pjst.31.5.27

E-mail addresses:
avita207@gmail.com (Avita Katal)
susheela.iitr@gmail.com (Susheela Dahiya)
tanupriya1986@gmail.com (Tanupriya Choudhury)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Data centers are undergoing rapid evolution in the age of virtualization, and new technologies like containerization are evolving rapidly. However, with the growth of cloud and serverless computing, the development of predictive analytics, edge computing, the arrival of 5G, and the COVID-19 pandemic has swept the entire globe, making almost everything online. People's online activities have increased, leading to data generation and resource utilization difficulties for data centers. The virtual machine (VM), as a key component of the cloud environment, is typically responsible for performing and maintaining the operating system's operation and storage and ensuring the operating system's normal operation (OS). The cloud platform is becoming more prominent and complex as it grows. Consequently, concerns regarding competitive segmentation of the platform's underlying hardware have arisen. Any VM behavior that is out of the ordinary can disrupt routine operations, resulting in a major loss for the organization, lowering computing capabilities, or even preventing the effective implementation and practice of cloud computing.

Cloud platforms are in high demand to host a variety of workloads, particularly web applications that require high Service Level Agreements (SLAs) agreed between the Cloud Service Provider (CSP) and the customer. In terms of accessibility, dependability, and efficiency, these services necessitate a diverse set of Quality of Service (QoS) requirements. Workloads that are typically transferred to cloud systems need resources such as memory, CPU, network bandwidth, and storage. Depending upon the resource these workloads use, more than others categorize them as specific resource-intensive workloads. The actual resource consumption of these workloads is often lower than the resources they have demanded. The service providers profit from this behavior by offering more resources at cheaper prices than the actual amount of resources they have, reliant on the fact that most customers' applications will not operate at maximum capacity. CSP exploits the dynamic provisioning characteristic of the cloud to provide on-demand performance. Recognizing workload behavior in a cloud data center is vital because it enables elastically scaling up and down provisioned services critical to its capabilities.

Workload characterization forecasts resource needs, making capacity management, allocation, and resource deployments more effective. The workload is typically characterized using one of two methodologies: trace-based (Abrahao & Zhang, 2004) or model-based (Delimitrou & Kozyrakis, 2011; Huang & Feng, 2009; Moro et al., 2009). The model-based technique is favored over the trace-based procedure since it is unconcerned about the operating platform upon which the trace was documented. Trace-based strategies have a limited number of production and quality traces which necessitates regular tinkering of workload characteristics to make them consistent with a new data center environment, making them less efficient than model-based strategies. Most workloads in cloud data centers are a combination of disparate applications (Mishra

et al., 2010). It is not easy to create a truly united approach to estimate the future usage of resources in these disparate application areas. These responsibilities show various behavior regarding periodicity, co-relation, and repeating trends. Workload classification requires a more thorough understanding of workload behavior and properties. However, few studies have been conducted on workload characterization due to the lack of open-source traces. Different scheduling models can be implemented by identifying workloads that heavily utilize shared resources.

The categorization and characterization of cloud workloads is an important research topic for better understanding workloads and managing cloud resources efficiently. There are many studies on workload characteristics. Google Cluster Trace (GCT) (Reiss et al., 2012), Bit Brains Trace (BBT) (Shen et al., 2015), Alibaba (https://github.com/alibaba/clusterdata) Yahoo trace (https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67&guccounter=1) and Wikipedia (http://www.wikibench.eu/?page_id=60) are four most common workload traces available in this domain. Workload qualities for data centers need to be statistically combined and evaluated to predict the future resource demand of the data center. Many researchers have employed statistical methods like Pearson Coefficient of Correlation (PCC), standard deviation, and mean techniques related to and existing methods (Birke et al., 2014)

Calzarossa et al. (2016) created a list of standard internet workloads, workloads from social networks, streaming platforms, mobile applications, and cloud computing infrastructure facilities. The workloads' characteristics were covered, and historical workload patterns like periodicity were considered an important distinguishing feature of cloud network workloads. With time-series analysis, Ali-Eldin et al. (2014) explored the time series of Wikipedia's workload and discovered that it is completely predictable and has strong seasonal variation. Self-similarity and burstiness are two of the main workload characteristics, according to Yin et al. (2015), so they developed a workload generator for cloud computing that is bursty and self-similar. Wang et al. (2015) examined workload process statistics. Combined optimization-based modeling of slow time-scale workload with stochastic modeling of fast time-scale workload is done to anticipate the value of dynamic resizing.

For proactive workload management, Zhang et al. (2014) created a service for workload factoring. It used a data item detection method to detach the application workload's two naturally distinct components, flash crowd, and base workload. In order to adapt to changing application data popularity, it evaluated incoming traffic based not just on quantity but also on data content. Recognizing and forecasting patterns in cloud workloads is a difficult problem that Patel et al. (2015) address. They presented a resource usage-based clustering approach to identify periodic tasks. Non-periodic tasks' resource consumption was depicted as a time series. Panneerselvam et al. (2014) classified cloud workloads in terms of workload patterns. They divided workloads into five categories: unpredictable,

static, continuously changing, periodic, and once-in-a-lifetime workloads, and used this classification to compare the performance of Markov modeling and Bayesian modeling.

Understanding workload characteristics is beneficial for enterprise data centers. Due to the scarcity of open-source workload traces, only a few attempts at characterizing cloud data center workloads have been made so far. The following are some of the most well-known research projects: The authors analyzed the BBT dataset representing business-critical workloads (Shen et al., 2015). Statistical methods such as standard deviation and mean, PCC, Autocorrelation Function, Peak Mean Ratio, and Coefficient of Variation were used to characterize the workload. The analysis was carried out using basic statistical and time pattern analysis. The findings from the study are as follows: (1) there is a strong correlation between demanded memory and CPU utilization, (2) Memory and CPU utilizations are easy to predict over short periods, and (3) disc and network utilization follow patterns, implying that prediction granularity is measured in days. The authors' proposed methodology has a major limitation in terms of trustworthiness. Errors in analysis are common in many fields of applied statistical data. The tools used for data gathering are another major disadvantage that puts the dataset's validity into question.

Zhang et al. (2011) presented a task usage shape classification that precisely reproduces the technical specifications of historical data on average job wait time and machine resource utilization. They utilized real-time data from Google and found that merely simulating the mean job usage can gain considerable precision in replicating resource utilization and task wait time. One major drawback is that the results are very complex and produce complex characterization of task shape classification. Rasheduzzaman et al. (2014) examined the production workload trace (version 2) by Google and utilized K-means clustering to group similar jobs together. They demonstrated a simple method for establishing workload attributes, knowledge, and insights for workload performance on cluster machines. The authors did not use the complete trace to perform the analysis, which led to the discrepancy in the results. The GCT dataset was used to classify workloads by Shekhawat et al. (2018). The authors used the K-means algorithm to generate task clusters after first identifying workload aspects such as low, high, and medium. To locate coordinate clusters, breakpoints within workload parameters were identified to find coordinate clusters.

Finally, utilizing coefficient of variation principles, the total number of clusters was minimized by merging nearby clusters. As per the key characteristics of the workload, the execution length of tasks was bimodal. Most tasks were short, and a few long-duration tasks had high demands of memory and CPU. Moro et al. (2009) introduced an innovative method to assess the execution workload performed by a computer precisely. Their proposed method directly utilized the memory reference sequence generated during program execution. The memory reference sequences were treated as sequences of floating-point numbers and subjected to analysis using signal-processing techniques. Spectral analysis

was employed during the feature extraction phase, while Ergodic Continuous Hidden Markov Models (ECHMMs) were used in the pattern matching phase. The effectiveness of the proposed algorithms was evaluated through trace-driven simulations utilizing the SPEC 2000 workloads. Ismaeel and Miri (2019) provide a real-time VM provisioning system that uses effective and unique clustering, time-series prediction, and placement algorithms to lower the energy consumed in a cloud data center.

It considers user behavior and previous VM utilization to anticipate the number of VMs required. It enhances the consolidation process while consuming the least amount of energy. On average, the results show an improvement of up to 80%. They have used only one day of data, which is a major drawback. Cheng et al. (2018) characterized the batch instance workloads based on: CPU utilization, memory utilization, and job timeframe into three different categories. The authors determined the arrival pattern for applications. The primary strength lies in the author's use of the traces of Alibaba's data center and workload categorization based on resource utilization. Mishra et al. (2010) proposed a multi-level task categorization technique and explained task categorization requests for capacity management and job scheduling. By monitoring resource usage by task class, task classification allows users to predict application expansion.

The authors use well-known statistical clustering techniques to implement proper research methods: (1) assess the workload aspects, (2) use an off-the-shelf method like k-means to construct task clusters, (3) evaluate the break marks of qualitative cartesian coordinates inside workload elements, and (4) combine adjoining task clusters to decrease the number of variables predictions. Their methodology yields eight workloads when applied to several Google compute clusters. They demonstrated that, for the same compute cluster, the features of each workload in relation to the number of tasks and resources consumed are coherent across days. In contrast, the medium-grain characterization detects discrepancies in workload features among clusters where such distinctions are predicted. They did not consider the job constraints, and they did not take the entire dataset for analysis which is the major drawback of their proposed approach. Ismaeel et al. (2019) introduced a new methodological process for selecting the appropriate task clustering approach in data centers based on validation indices and result correlation.

They developed an effective pre-processing strategy, reducing the big data challenge to a compact 2D matrix of independent jobs using CPU and memory requirements. Shekhawat et al. (2018) proposed a technique for classifying and characterizing data center workloads based on resource utilization. For workload classification, seven distinct machine-learning techniques have been used and compared. Workload distribution is approximated for GCT and BBT datasets using various application components. Finally, the authors have presented an approach for assessing the relevance of various categorization attributes. The authors have not considered and compared the results with any other clustering algorithm.

Due to the rapid advancement of hardware and resource management strategies, cloud data centers handle many application types. Recognizing the workload features and understanding the data centers is intrinsically critical for CSP to continue adopting cloud technology. In cloud data centers, proper resource management is crucial for predicting the future requirements of resources. The first step to better resource management is to characterize and classify the workload before placing it on the virtual machines. It also helps to meet the required or agreed QoS.

In this paper, the following research questions (RQs) are addressed:

RQ1: How workload categorization helps in better resource management in cloud data centers?

RQ2: How do K Means and the Gaussian Mixture Model aid in grouping various workloads?

RQ3: Which clustering algorithm performs better for the characterization of workload?

RQ4: How classification of workload helps in understanding the workload type, and which classification algorithm achieves maximum accuracy?

This paper uses K Means and Gaussian Mixture Model clustering algorithms to properly cluster the Bitbrains trace (Fast Storage) dataset to characterize the workload. Initially, the feature significance analysis is done to understand the important features. Feature selection becomes important in developing robust and efficient classification models while reducing training time (Onan & KorukoGlu, 2017). Both clustering algorithms are compared based on the Calinski-Harabasz index and Davies-Bouldin score coefficients. Following clustering, several ML methods are used to classify the data. The models built using these methods are also compared in terms of accuracy.

MATERIALS AND METHODS

Dataset Characteristics

Bitbrains' distributed data center is a managed hosting and business computing powerhouse. The dataset provides performance information for 1,750 virtual machines. Among the company's clients are numerous major banks, credit card companies, insurers, and others. Towers Watson and Algorithmics are two application manufacturers that host solvency applications on Bitbrains. These programs are typically used for accounting information completed after a fiscal quarter. Each file of the dataset includes the performance metrics for a single VM. These files are classified into two types: fastStorage and Rnd. FastStorage is the first trace, with 1,250 virtual machines (VMs) linked to Storage Area Network (SAN) storage devices. The second trace, Rnd, has 500 virtual machines (VMs) connected to either a fast SAN or a much slower Network Attached Storage (NAS) device. Because storage connected to the fastStorage devices is more effective, the fastStorage trace contains a greater fraction of server-side and computation units than the Rnd trace. In the Rnd trace,

on the other side, we notice a greater share of administration computers, which only require minimum storage and less frequent accesses. In the Rnd trace, on the other hand, we notice a greater share of management machines, which only require minimum storage and less frequent usage. The Fast Storage directory is divided into three sub-directories based on the month the data was collected. A row-based format is used in each file. Each row represents a performance metric observation.

Clustering Algorithms

Grouping data items using a similarity metric is known as clustering. Clustering can be hierarchical, partitional, complete, partial, overlapping, fuzzy, or exclusive. The partitional clustering technique, K-Means, divides data objects into non-overlapping groups (Onan, 2019). K-Means clusters are based on prototypes when the cluster is symbolized by a prototype and all nodes in the cluster are near it. Two common prototypes are centroid and medoid. Gaussian Mixture Models (GMM) clusters are based on density. A cluster based on density is a collection of high-density objects surrounded by low-density areas. Identifying and summarizing properties of interest is made easier by grouping data into clusters. Similar utilization trends of workload can help develop capacity strategies for meeting future resource requirements while preserving the SLA for operating services.

Classification Algorithms

The classification algorithm is a supervised learning approach that uses training data to determine the type of new observations. The software trains from a dataset or observations and then classifies additional observations. As a supervised learning approach, the classification algorithm employs labeled input data comprising input and output. The following classification algorithms have been used in this paper: Random Forest (RF), Logistic Regression (LR), K Nearest Neighbor (KNN), Support Vector Machine (SVM) and Decision Tree (DT), Multi-Layer Perceptron (MLP) and Back Propagation Neural Network.

Methodology

Figure 1 shows the methodology followed to characterize and categorize the workload. Generally, the workload in the data centers consists of different attributes. Some attributes are more important when it comes to characterizing the workload. Understanding the distribution of attributes in terms of significance is critical for determining the type of workload. It is important as, during the classification stage, the significance of the attribute determines how much weight has to be given to it or a group of attributes. The main purpose of performing a significance assessment is to order the attributes in terms of predictive power. The decision Tree Classifier algorithm has been used to perform the significance analysis. The top four attributes having the highest significance have been taken for further

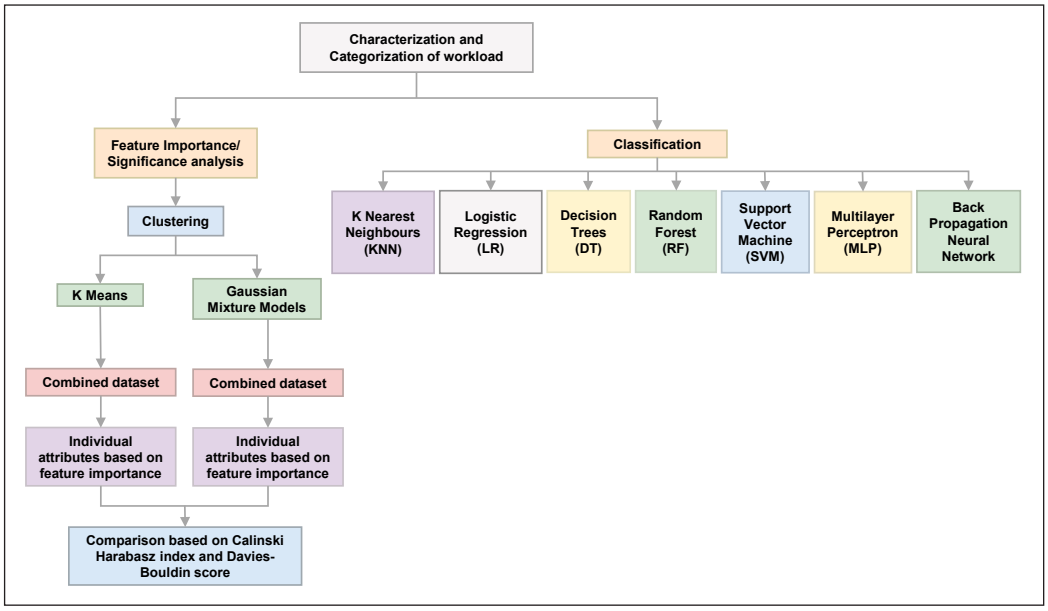

Figure 1. Methodology for categorization and characterization of workload

analysis. In the BBT dataset, memory usage [KB], disk read throughput, CPU usage [MHZ], network transmitted throughput, disk write throughput, and network received throughput have high percentages in attribute significance analysis. The disk writes throughput and disk read throughput are combined to form a single attribute named disk usage.

Similarly, network transmitted and received throughput are combined to form a single attribute named network usage. Normalization of the BBT dataset is done using min-max normalization. The K means technique was applied to the combined data set of highly significant attributes. The technique was further applied to each attribute to calculate K. The value of the number of clusters, K, is determined using the elbow criterion. If c represents the clusters obtained for CPU usage [MHZ], m for Memory usage [KB], d for Disk usage, and n represents clusters acquired for Network usage. The product of c, m, d, and n yields the number of possible workloads in the dataset. As a result, the frequency of various workloads is calculated. This analysis yielded the dataset's workload distribution.

GMM clustering is applied to all attributes first, followed by individual attributes. The outcomes of both algorithms are compared using parameters such as the Calinski-Harabasz index(CHI) and the Davies-Bouldin Index (DBI). CHI is also known as the Variance Ratio Criterion. A higher CHI score denotes a model with more defined clusters. The index is the ratio of all clusters' total between-cluster and within-cluster variance. The score is greater when clusters are large and well-spaced, which correlates to a classic cluster idea. The score is rapidly computed. When DBI is used to evaluate the model, a lower DBI indicates a model with greater cluster separation. This index represents clusters' average

similarity, which is defined as a criterion that relates cluster distance to several clusters. It distinguishes between clusters that are both remote and small. The Davies–Bouldin criterion is based on a weighted average of distances "within-cluster" and "between-cluster." The lowest possible score is zero. Closer to zero indicates a better partition. Davies-Bouldin scores are easier to calculate because they only use point-wise distances, and the index is solely based on amounts and characteristics inherent in the dataset.

Following the characterization of the dataset, classification is done using different classification algorithms. The classification of workloads is an important step in workload analysis. The models' accuracy is crucial for workload analysis, resource usage prediction, and provisioning. The classification accuracy analysis aids us in determining which algorithm is best for a given data center workload. The before-mentioned algorithms are applied to determine the accuracy.

RESULTS AND DISCUSSION

Firstly, the feature importance or attribute significance analysis uses a decision tree algorithm. The results are shown in Figure 2. The attributes are represented on the x-axis, while the value of coefficients is depicted on the y-axis. The importance of CPU usage [MHZ] is highest for the BBT dataset. Network and disk usage dominates the second and third position in terms of significance analysis. Memory usage has low significance among all the attributes.

The elbow method to determine the value of K is applied, and the graph is plotted between K and inertia, where the inertia value indicates how far apart the points in a cluster are. The graph is shown in Figure 3. The value of the x-axis depicts the K values, and the value of the Y-axis depicts the inertia value. At K = 4, it produces an elbow, indicating that the BBT fastStorage dataset has four types of workloads.

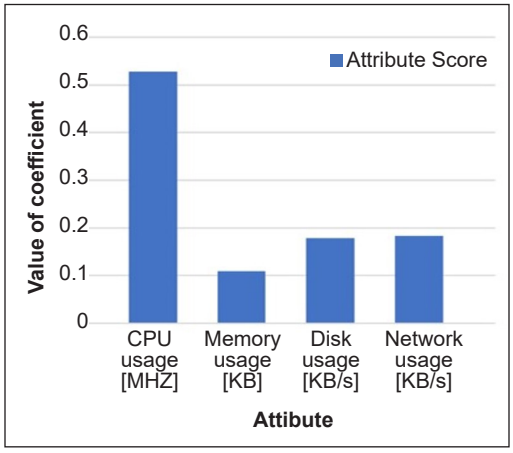

Figure 2. Attribute score of different attributes

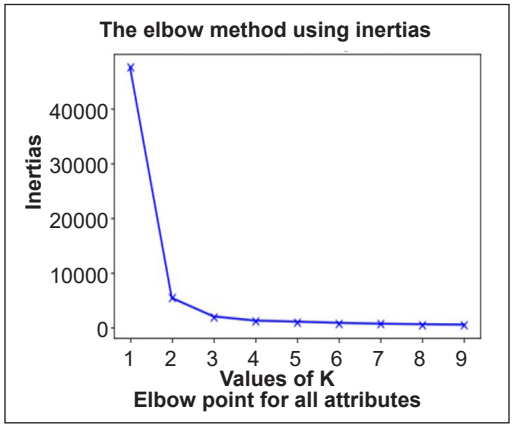

Figure 3. Value of K vs. inertias on highly significant attributes

The plots of the clustering results based on each attribute individually, such as CPU usage, memory usage, disk usage, and network usage, are depicted in Figures 4, 5, 6, and 7, respectively. It can be concluded from these graphs that for CPU usage, two different workloads have identified that is LOW (C_L) and HIGH (C_H); for memory, three different workloads have observed that is LOW(M_L), MEDIUM(M_M), and HIGH(M_H); two workloads have identified for LOW disk usage (D_L) and HIGH(D_H), and two workloads have identified for LOW network usage (N_L) and HIGH(N_H).

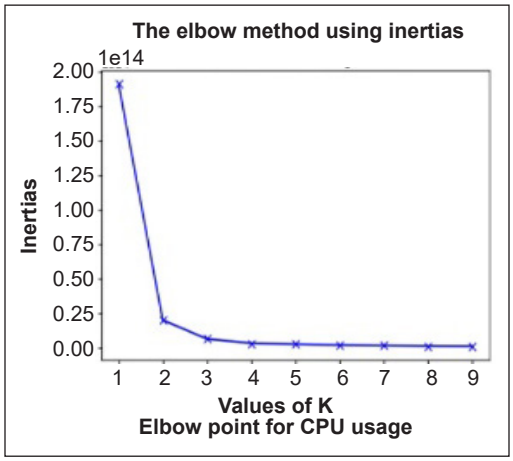

Figure 4. K vs. Inertia (CPU usage)

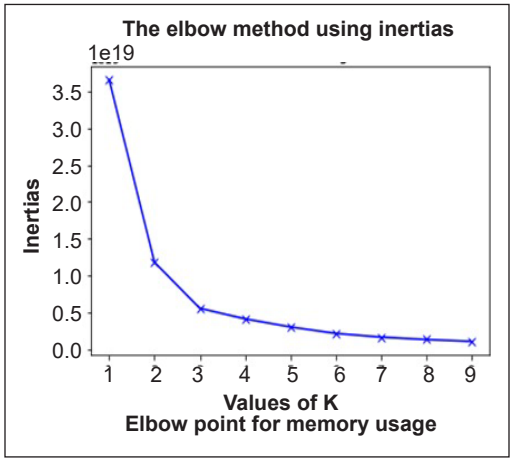

Figure 5. K vs. Inertia (Memory usage)

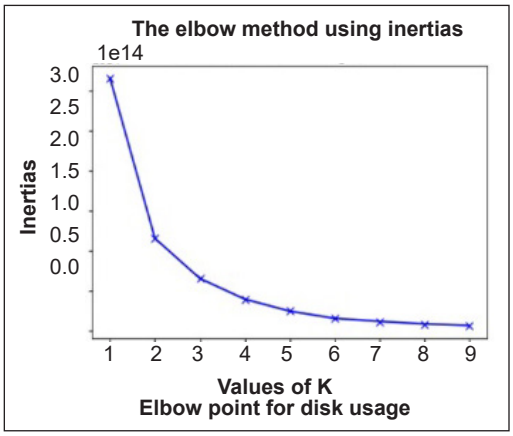

Figure 6. K vs. Inertia (Disk usage)

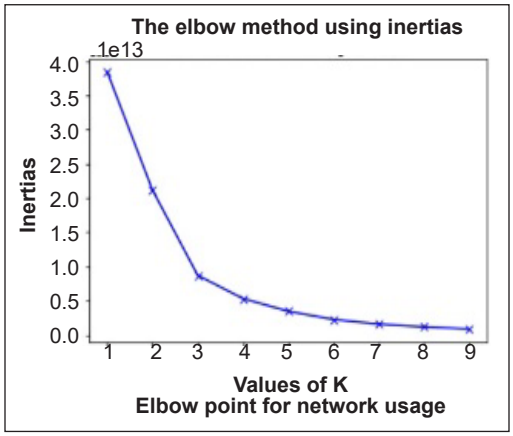

Figure 7. K vs. Inertia (Network usage)

The 24 distinct workload combinations are determined once the elbow point is calculated using K Means clustering. The percentage of tasks can be identified by calculating the number of tasks in each combination. Table 1 is formatted as [CPU usage] [Memory usage][Disk usage][Network usage].

Table 1 shows that most tasks (93.38%) have modest resource utilization. These processes used less CPU, memory, storage space, and network bandwidth. These virtual machines are made up of short administrative chores and application inquiries. The workload then consists of 3.41% of jobs with high CPU, medium memory, and low disk and network use. Typically, these virtual machines are utilized to run CPU-intensive consumer applications.

Similarly, GMM clustering is applied to the BBT dataset. The Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) graphs are plotted initially on the entire dataset. It can be concluded from the graph that there are seven types of workloads in the entire dataset. On the x-axis of Figure 8, the number of clusters or components is depicted, whereas, on Y-axis, the score of AIC and BIC is depicted. The value of AIC and BIC is the same for each analysis; therefore, both the lines overlap, which results in the depiction of one line in all graphs.

Once the clustering is done on the combined dataset, the GMM is applied to the individual attributes. Clustering based on individual attributes, CPU usage, memory usage, disk usage, and network usage yielded the results (Figures 9, 10, 11, and 12). It can be concluded that CPU and memory usage have five types of workloads, followed by disk and network usage having four types.

Once the number of components is identified for each attribute, 600 workload combinations are formed. There are some

Table 1
Percentage of tasks in each cluster (K Means)

Type of workload	Number of tasks	Percentage of tasks
$[C_L][M_L][D_L][N_L]$	10479078	93.380000
$[C_L][M_L][D_L][N_H]$	6242	0.050000
$[C_L][M_L][D_H][N_L]$	10823	0.090000
$[C_L][M_L][D_H][N_H]$	3150	0.020000
$[C_L][M_M][D_L][N_L]$	127939	1.140000
$[C_L][M_M][D_L][N_H]$	1492	0.01
$[C_L][M_M][D_H][_NL]$	1823	0.010000
$[C_L][D_M][M_H][N_H]$	401	0.000000
$[C_L][M_H][D_L][N_L]$	28650	0.250000
$[C_L][M_H][D_L][N_H]$	147	0.000000
$[C_L][M_H][D_H][N_L]$	3873	0.030000
$[C_L][M_H][D_H][N_H]$	9	0.000080
$[C_H][M_L][D_L][N_L]$	88450	0.780000
$[C_H][M_L][D_L][N_H]$	74	0.000600
$[C_H][M_L][D_H][N_L]$	4	0.000030
$[C_H][M_L][D_H][N_H]$	59	0.000500
$[C_H][M_M][D_L][N_L]$	382850	3.410000
$[C_H][M_M][D_L][N_H]$	50	0.000400
$[C_H][M_M][D_H][N_L]$	90	0.000800
$[CH][M_M][D_H][N_H]$	6	0.000050
$[C_H][M_H][D_L][N_L]$	83166	0.740000
$[C_H][M_H][D_L][N_H]$	14	0.000100
$[C_H][M_H][D_H][N_L]$	3406	0.030000
$[C_H][M_H][D_H][N_H]$	3	0.000020

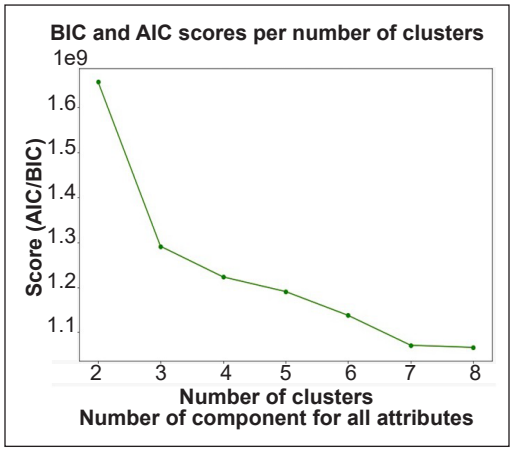

Figure 8. AIC and BIC plot for all attributes

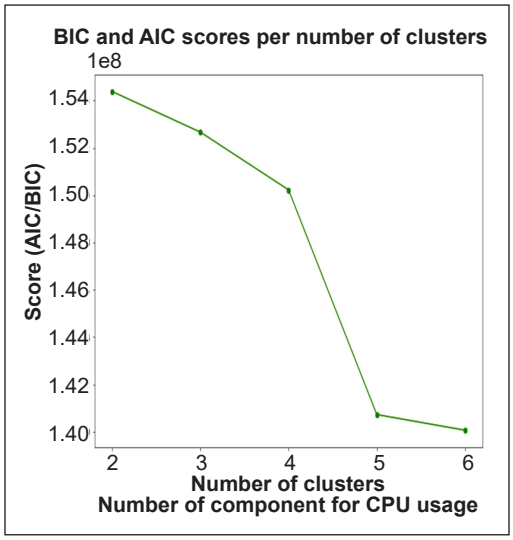

Figure 9. AIC and BIC plot (CPU usage)

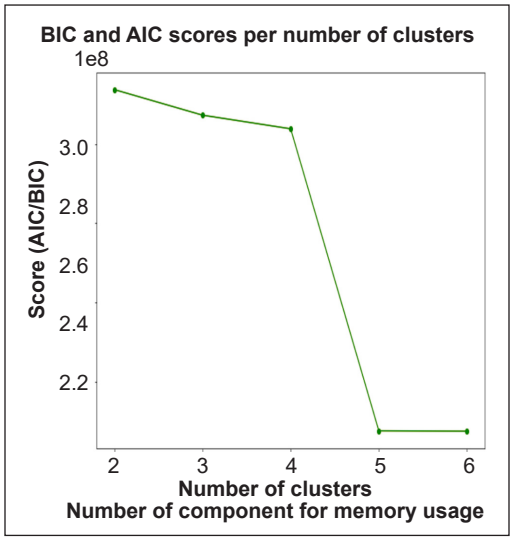

Figure 10. AIC and BIC plot (Memory usage)

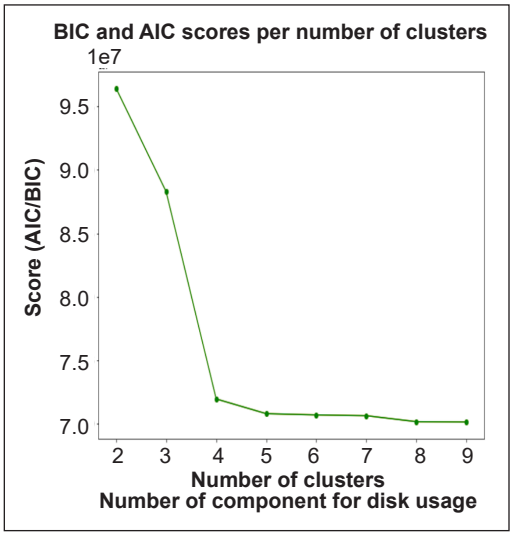

Figure 11. AIC and BIC plot (Disk usage)

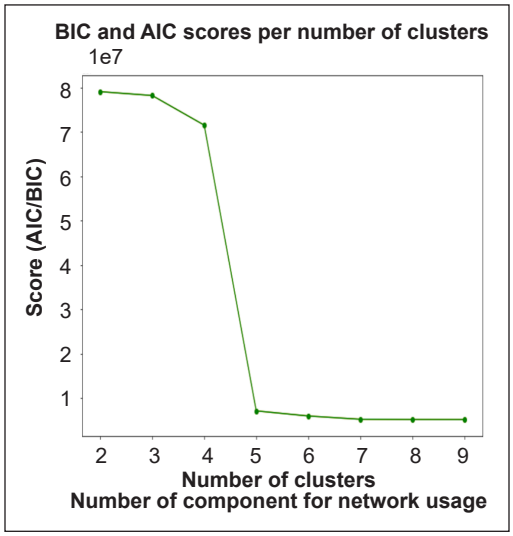

Figure 12. AIC and BIC plot (Network usage)

combinations in all clusters that are not identified. Therefore, 336 combinations are taken, which consist of workloads from all four attributes. Table 2 depicts the various clusters formed by the GMM.

K means, and GMM performance is evaluated based on the Calinski Harabasz index and Davies-Bouldin score. Table 3 shows the values of different parameters by applying K Means and GMM.

It can be concluded that K means it outperforms the GMM on all mentioned parameters (Table 3). The Davies-Bouldin score of K Means is 0.37, much better than that of

3.77 GMM. Closer the value to zero, the better the cluster separation. The Calinski Harabasz Index of K means 54500062.89, whereas, for GMM, it is 12027922.51. The greater the value of the Calinski Harabasz Index score better is the density and cluster separation.

An experimental assessment was done to approximate classification accuracy for various machine-learning algorithms to explore how workload distribution affects each model. The analysis was carried out using the BBT fastStorage dataset. The training and testing data ratio is kept at 70 and 30, respectively. Table 4 displays the accuracy results, AUC ROC score, and Precision for different algorithms.

Cloud computing guarantees high throughput, adaptability, and cost-effectiveness to address evolving processing

Table 3
Scores for performance evaluation parameters

Parameter	K Means	Gaussian Mixture Models
Calinski Harabasz Index	54500062.89	12027922.51
Davies-Bouldin score	0.37	3.77

Table 2
Percentage of tasks in each cluster (GMM)

Types of workloads (CPU, DISK, NETWORK, MEMORY)				Percentage of tasks
C_{VL}	D_{VL}	N_{VL}	M_{VL}	15.10235435
C_{VL}	D_{VL}	N_{VL}	M_{VH}	16.61488353
C_{VL}	D_{VL}	N_{L}	M_{VH}	1.29211
C_{VL}	D_{VL}	N_{H}	M_{VL}	0.805227325
C_{VL}	D_{VL}	N_{H}	M_{VH}	6.928157693
C_{VL}	D_{VH}	N_{VL}	M_{VH}	2.761330624
C_{VL}	D_{VH}	N_{H}	M_{M}	0.545286852
C_{VL}	D_{VH}	N_{H}	M_{VH}	4.255609617
C_{L}	D_{VL}	N_{H}	M_{H}	1.256375982
C_{M}	D_{L}	N_{L}	M_{M}	0.614936998
C_{M}	D_{L}	N_{L}	M_{H}	0.627697874
C_{M}	D_{VH}	N_{H}	M_{M}	0.495963214
C_{H}	D_{VL}	N_{VL}	M_{VH}	1.344632768
C_{H}	D_{VL}	N_{H}	M_{VL}	1.736004919
C_{H}	D_{VL}	N_{H}	M_{M}	0.64278458
C_{H}	D_{VL}	N_{H}	M_{VH}	1.911226363
C_{H}	D_{L}	N_{L}	M_{M}	0.912589781
C_{H}	D_{L}	N_{H}	M_{M}	1.64271329
C_{H}	D_{L}	N_{H}	M_{VH}	0.655527634
C_{H}	D_{VH}	N_{VL}	M_{VL}	0.598780944
C_{H}	D_{VH}	N_{VL}	M_{M}	1.03473596
C_{H}	D_{VH}	N_{VL}	M_{VH}	3.523222656
C_{H}	D_{VH}	N_{L}	M_{M}	0.821481402
C_{H}	D_{VH}	N_{H}	M_{VL}	1.349453742
C_{H}	D_{VH}	N_{H}	M_{M}	5.648532321
C_{H}	D_{VH}	N_{H}	M_{VH}	9.933228181
C_{VH}	D_{VL}	N_{H}	M_{H}	0.637268531

Table 4
Accuracy percentage of classification algorithms

Algorithms	Accuracy in %	AUC ROC Score	Precision
K Nearest Neighbors (KNN)	98.79	0.963	0.96
Logistic Regression (LR)	79.19	0.941	0.93
Decision Trees (DT)	99.18	0.976	0.97
Random Forest (RF)	97.80	0.958	0.95
Support Vector Machine (SVM)	84.34	0.922	0.91
Multi-Layer Perceptron	79.72	0.825	0.82
Back Propagation Neural Network	80.00	0.847	0.84

necessities. As the quantity of data continues to grow at an appalling rate, many firms are turning to data centers to make effective choices and achieve a competitive edge. The cloud-computing model is used for a multitude of applications. These applications differ in their characteristics and have varying demands on the Physical Machines' resources (PMs). The requirements of database applications (which perform intensive read and write operations on discs, for example) differ from those of a scientific computing application (which demands significant computing power from the CPU). To effectively configure cloud resources, network managers must be able to characterize and predict the workload on VMs. Clustering the tasks into groups or clusters is feasible based on the different demands of dissimilar tasks of cloud applications.

The clustering process can identify characterizations that can improve the efficiency of historical workload traces over a wide range of critical performance parameters, such as increasing the utilization of PMs hosted in cloud data centers. Any workload classification should consider the usage of resources, job progression, and other issues that necessitate adherence to service-level agreements (SLAs). Analytical models (Bennani & Menascé, 2005; Bodnarchuk & Bunt, 1991) and performance metrics (Bienia et al., 2008), (Jackson et al., 2010) are used in workload characterization. Analytical models are the models of mathematics having a closed-form solution, which indicates that the solution to the equations used to explain system changes may be expressed mathematically as an analytic function.

A performance model is used to characterize the fundamental elements of how a planned or existing system performs in terms of usage of resources, demand for resources, and delays induced by processing or physical restrictions. To maximize earnings, cloud providers aim to get as many new requests as feasible; conversely, they must encounter QoS constraints in line with the appropriate SLA with end-users. One needs efficient resource provisioning mechanisms to achieve this goal. Users typically have sporadic access to cloud resources, and workloads will fluctuate. Workload fluctuation causes under-provisioning and over-provisioning issues, wasting resources and time. One solution is forecasting workload based on past consumption behaviors and the present state of cloud resources. The trends equate user requests with cloud resources depending on the type of requirement.

CONCLUSION

The utilization and prominence of cloud computing as among the most well-known internet-based inventions for supplying computational power and infrastructural facilities to IT organizations for executing/hosting cloud workloads is expanding every day and is anticipated to expand even further. Consumers upload heterogeneous cloud workloads to the cloud through internet services, banking applications, online payment processing assistance, portable computing assistance, and graphics-based services, with varying QoS parameters

in the form of SLA. The significance of workload characterization and classification in cloud data centers is discussed in this paper. The workload of types is clustered using two different clustering techniques. Workload distribution is accomplished by combining distinct workload pairings in both clustering modes. After clustering is completed, the performance is examined to determine which clustering works best.

Most tasks, according to the K Means algorithm (93.38%), have low resource utilization (CPU, memory, storage space, and network bandwidth). Short administrative tasks and application inquiries make up these virtual machines. GMM shows a maximum of 16.61% of the tasks consume CPU (Very Low), Disk (Very Low), Network (Very Low), and Memory (Very High) resources. However, the results demonstrate that K means beats in Calinski Harabasz and the Davies-Bouldin Index. After clustering, classification is carried through by utilizing several classification techniques. The decision tree shows a maximum accuracy of 99.18%. Compared to the existing study, this work explains different clustering and classification strategies for cloud data center workloads.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude and appreciation to the University of Petroleum and Energy Studies, India, for providing invaluable resources and opportunity to conduct the research. This work has been undertaken as part of first author's doctoral studies.

REFERENCES

Abrahao, B., & Zhang, A. (2004) *Characterizing application workloads on CPU utilization for utility computing* (HPL-2004-157). Hewlett-Packard Company. https://www.hpl.hp.com/techreports/2004/HPL-2004-157.html

Ali-Eldin, A., Rezaie, A., Mehta, A., Razroev, S., Luna, S. S. de, Seleznjev, O., Tordsson, J., & Elmroth, E. (2014, March 11-14). *How will your workload look like in 6 years? Analyzing Wikimedia's workload*. [Paper presentation]. 2014 IEEE International Conference on Cloud Engineering, Boston, USA. https://doi.org/10.1109/IC2E.2014.50

Bennani, M. N., & Menascé, D. A. (2005, June 13-16). *Resource allocation for autonomic data centers using analytic performance models*. [Paper presentation]. Second International Conference on Autonomic Computing, ICAC'05. Seattle, USA. https://doi.org/10.1109/ICAC.2005.50

Bienia, C., Kumar, S., Singh, J. P., & Li, K. (2008, October 25-29). *The PARSEC benchmark suite: Characterization and architectural implications*. [Paper presentation]. Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques. Toronto, Canada. https://doi.org/10.1145/1454115.1454128

Birke, R., Chen, L. Y., & Smirni, E. (2014, May 5-9). *Multi-resource characterization and their (in) dependencies in production datacenters*. [Paper presentation]. IEEE/IFIP Network Operations and Management Symposium (NOMS), Krakow, Poland. https://doi.org/10.1109/NOMS.2014.6838300

Bodnarchuk, R., & Bunt, R. (1991, May 21-24). *A synthetic workload model for a distributed system file server*. [Paper presentation]. Proceedings of the 1991 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems, California, USA. https://doi.org/10.1145/107971.107978

Calzarossa, M. C., Massari, L., & Tessera, D. (2016). Workload characterization. *ACM Computing Surveys (CSUR)*, *48*(3), 1-43. https://doi.org/10.1145/2856127

Cheng, Y., Chai, Z., & Anwar, A. (2018, August 27-28). *Characterizing co-located datacenter workloads: An Alibaba case study*. [Paper presentation]. Proceedings of the 9th Asia-Pacific Workshop on Systems, Jeju, Korea. https://doi.org/10.1145/3265723.3265742

Delimitrou, C., & Kozyrakis, C. (2011, June 20-24). *Cross-examination of datacenter workload modeling techniques*. [Paper presentation]. International Conference on Distributed Computing Systems Workshops, Minneapolis, USA. https://doi.org/10.1109/ICDCSW.2011.45

Huang, S., & Feng, W. (2009, May 18-21). *Energy-efficient cluster computing via accurate workload characterization*. [Paper presentation]. 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, Shanghai, China. https://doi.org/10.1109/CCGRID.2009.88

Ismaeel, S., Al-Khazraji, A., & Miri, A. (2019, April 15-17). *An efficient workload clustering framework for large-scale data centers*. [Paper presentation]. 8th International Conference on Modeling Simulation and Applied Optimization, Manama, Bahrain. https://doi.org/10.1109/ICMSAO.2019.8880305

Ismaeel, S., & Miri, A. (2019, January 7-9). *Real-time energy-conserving VM-provisioning framework for cloud-data centers*. [Paper presentation]. IEEE 9th Annual Computing and Communication Workshop and Conference, Las Vegas, USA. https://doi.org/10.1109/CCWC.2019.8666614

Jackson, K. R., Ramakrishnan, L., Muriki, K., Canon, S., Cholia, S., Shalf, J., Wasserman, H. J., & Wright, N. J. (2010, November 30 – December 3). *Performance analysis of high performance computing applications on the Amazon Web Services cloud*. [Paper presentation]. IEEE Second International Conference on Cloud Computing Technology and Science, Indianapolis, USA. https://doi.org/10.1109/CLOUDCOM.2010.69

Mishra, A. K., Hellerstein, J. L., Cirne, W., & Das, C. R. (2010). Towards characterizing cloud backend workloads. *ACM SIGMETRICS Performance Evaluation Review*, *37*(4), 34-41. https://doi.org/10.1145/1773394.1773400

Moro, A., Mumolo, E., & Nolich, M. (2009, September 16-18). *Ergodic continuous hidden markov models for workload characterization*. [Paper presentation]. Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis, Salzburg, Austria. https://doi.org/10.1109/ISPA.2009.5297771

Onan, A. (2019). Consensus Clustering-based undersampling approach to imbalanced learning. *Scientific Programming, 2019,* 1-14. https://doi.org/10.1155/2019/5901087

Onan, A., & KorukoGlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25-38. https://doi.org/10.1177/0165551515613226

Panneerselvam, J., Liu, L., Antonopoulos, N., & Bo, Y. (2014, December 8-11). *Workload analysis for the scope of user demand prediction model evaluations in cloud environments*. [Paper presentation]. IEEE/ACM 7th International Conference on Utility and Cloud Computing, London, United Kingdom. https://doi.org/10.1109/UCC.2014.144

Patel, J., Jindal, V., Yen, I. L., Bastani, F., Xu, J., & Garraghan, P. (2015, March 25-27). *Workload estimation for improving resource management decisions in the cloud*. [Paper presentation]. IEEE 12th International Symposium on Autonomous Decentralized Systems, Taichung, Taiwan. https://doi.org/10.1109/ISADS.2015.17

Rasheduzzaman, M., Islam, M. A., Islam, T., Hossain, T., & Rahman, R. M. (2014, February 21-22). *Task shape classification and workload characterization of google cluster trace*. [Paper presentation]. IEEE International Advance Computing Conference, Gurgaon, India. https://doi.org/10.1109/IADCC.2014.6779441

Reiss, C., Tumanov, A., Tumanov, A., Ganger G. R., & Katz, R. (2012). *Towards understanding heterogeneous clouds at scale: Google trace analysis*. ResearchGate. https://www.researchgate.net/publication/265531801_Towards_Understanding_Heterogeneous_Clouds_at_Scale_Google_Trace_Analysis

Shekhawat, V. S., Gautam, A., & Thakrar, A. (2018, December 1-2). *Datacenter workload classification and characterization: An empirical approach.* [Paper presentation]. IEEE 13th International Conference on Industrial and Information Systems, Rupnagar, India. https://doi.org/10.1109/ICIINFS.2018.8721402

Shen, S., van Beek, V., & Iosup, A. (2015, May 4-7). *Statistical characterization of business-critical workloads hosted in cloud datacenters.* [Paper presentation]. IEEE/ACM 15th International Symposium on Cluster, Cloud, and Grid Computing, Shenzhen, China. https://doi.org/10.1109/CCGRID.2015.60

Wang, K., Lin, M., Ciucu, F., Wierman, A., & Lin, C. (2015). Characterizing the impact of the workload on the value of dynamic resizing in data centers. *Performance Evaluation*, *85-86*, 1-18. https://doi.org/10.1016/J.PEVA.2014.12.001

Yin, J., Lu, X., Zhao, X., Chen, H., & Liu, X. (2015). BURSE: A bursty and self-similar workload generator for cloud computing. *IEEE Transactions on Parallel and Distributed Systems*, *26*(3), 668-680. https://doi.org/10.1109/TPDS.2014.2315204

Zhang, H., Jiang, G., Yoshihira, K., & Chen, H. (2014). Proactive workload management in hybrid cloud computing. *IEEE Transactions on Network and Service Management, 11*(1), 90-100. https://doi.org/10.1109/TNSM.2013.122313.130448

Zhang, Q., Hellerstein, J., & Boutaba, R. (2011) *Characterizing task usage shapes in Google compute clusters.* Google Research. https://research.google/pubs/pub37201/