Cloud Computing Systems Exploration over Workload Prediction Factor in Distributed Applications

Ankit Sharma
Department of Computer Science and Engineering
Sant Longowal Institute of Engineering and Technology
Deemed University, Longowal-148106,Sangrur, India
E-mail:ankit28sharma@gmail.com

Dr. Vinod Kumar Verma*
Department of Computer Science and Engineering
Sant Longowal Institute of Engineering and Technology
Deemed University, Longowal -148106 , Sangrur, India
E-mail: vinod5881@gmail.com

***Abstract*—**This paper highlights the different techniques of workload prediction in cloud computing. Cloud computing resources have a special kind of arrangement in which resources are made available on demand to the customers. Today, most of the organizations are using cloud computing that results in reduction of the operational cost. Cloud computing also reduces the overhead of any organization due to implementation of many hardware and software platforms. These services are being provided by cloud provider on the basis of pay per use. There are lots of cloud service providers in the modern era. In this competitive era, every cloud provider works to provide better services to the customer. To fulfill the customer's requirements, dynamic provisioning can serve the purpose in cloud system where resources can be released and allocated on later stage as per needs. That's why resource scaling becomes a great challenge for the cloud providers. There are many approaches to scale the number of instances of any resource. Two main approaches namely: proactive and reactive are used in cloud systems. Reactive approach reacts at later stage while proactive approach predicts resources in advance. Cloud provider needs to predict the number of resources in advance that an application is intended to use. Historical data and patterns can be used for the workload prediction. The benefit of the proactive approach lies in advance number of instances of a resource available for the future use. This results in improved performance for the cloud systems**.**

Keywords- *workload prediction; resource allocation;, dynamic provisioning; reactive and proactive approach; virtual machines*

***** Corresponding Author

*****

I. INTRODUCTION

Cloud computing developed a new idea to use IT resources. These resources have to be used in efficient manner so that while maintaining quality of service (QOS), maximum benefit can be achieved in terms of investment and power consumption. Provider has to manage resources in order to provide good quality of services while investing minimum amount of money. To achieve this target, many policies are suggested in the past. Cloud provider upgrade amount of resources by obeying rules based on policies so that over and underutilization of resources can be avoided. Proactive and reactive approaches are the two main approaches of resources scaling. In reactive approach, a fixed value is selected as threshold parameter. As the resource usage crosses the fixed value, amount of resources is incremented by a certain number. In practice, computer hardware has some booting time and this is also a matter of concern because it is not negligible to ignore. To compensate the booting time, resources have to be predicted in advance and it will more often be dependent on the nature of services. So, workload characterization and classification is performed on historical data. Cloud environment consists of data center, virtual machines, and high speed network lines which carry data and many other electronics components that deal with computing environment. Cloud provider has to keep many things in mind while setting up the cloud environment. The major factors are power consumption, heat emission, application performance, and cost of the system. Workload prediction plays an important role in these areas. Prediction shows clear picture to the provider of system implementation. Quality of service of the application is also an important factor that will provide a way to set up and manage cloud resources. After predicting, provider can achieve better performance and optimize the use of its resources. Prediction algorithm mainly depends on historical data and frequent patterns that have occurred in past analysis. Resource provisioning has effects in many areas like operational cost, quality of service, policy of the cloud provider, so it has a high impact while setting up a cloud system. Rest of the paper will be devoted towards analysis of cloud computing and its association with workload prediction. In next section of paper, features of cloud computing, method of resource provisioning, prediction algorithm, various tools and factors affecting implementation of cloud system have been discussed.

II. FEATURES OF CLOUD SYSTEM

Main features of cloud computing are virtualization, dynamic provisioning, quality of service, resource accessibility, on demand pay ability, resources scaling that are illustrated as below:

A. Virtualization

Virtualization is an essential feature of cloud computing where many instances of the resource can be made on real hardware and software platform [1]. Creation of virtual instances requires a series of operations to be executed. Therefore, certain time is required to complete the operations. In this

amount of time, services of cloud provider may be affected. To solve these issues, resources have to be predicted in advance. After prediction, many instances in advance can be created to provide better quality of service. Then the purpose of virtualization may be fulfilled, and real look and feel can be provided by the cloud provider. Cloud computing provides virtualization in terms of hardware and software. Both take time for booting at initial stage making the prediction necessary in virtualization. Every computer component takes power to operate and virtualization saves power consumption of a system due to optimized use of hardware system. According to the study, the power consumption of cloud data centers has increased by 56 percent from 2005 to 2010 [2].An average size data center consumes as much energy as 25,000 households [3]. This is the reason of addressing virtual issue, managing virtualization and consolidating approach like virtual machine (VM).

B. Scalability

As the cloud system provides services to the user, lots of instances of the resource are required. They differ in number according to users, types of application and nature of services. To provide better quality of service, management of resources is required. There is also a tradeoff between number of resources and quality of service. In order to achieve optimum set up, resource provisioning is used. Scaling can be performed at any level as platform in terms of resources, users etc [4]. Combined effect should be calculated to estimate number of required resourcesin the cloud system.

C. On Demand Payability

Cloud providers charge for their services. There are many policies such as hourly basis, usage of resources etc. Provider has to avail best service in optimum cost to survive in such a competitive market. To achieve this target, resource provisioning plays an important role. Scaling mechanism should be carefully designed while maintaining SLA rules. From client's point of view, suitable policy should be chosen so that work can be completed in minimum amount of time.

D. Remote Accessibility

Cloud resources can be accessed remotely from any location. Many a times**,** resources have to move from one host to another. Resource provisioning becomes the necessity for the best placement of resources so that the workload on servers can be managed. Under and over utilized server should be clearly identified and load should be balanced amongst them. In parallel, service providers must take care about the standard of application. Many a times, cloud application lags in number of features from desktop application. To design a compatible application, resources are used in high amount like network bandwidth, storage. In such situations, resources must be used in an optimum manner.

E. Quality of service

Quality of service is a major concern for cloud providers. Popularity of a service is mainly dependent on quality and demand of application. Nowadays, memory is cheap but not free and speed of CPU is fast but not infinite. So, to provide high quality of service, resources must be used in an efficient manner.

III. CLOUD SYSTEM FACTORS AND RELATED ISSUE

Virtual machine migration is one of the related areas come under resources provisioning. The virtual machine should be migrating in nature to balance the workload between different physical hosts. Migration criteria can be due to many factors such as workload, location of resources where service has to be provided. Sometimes power consumption and utilization rate of server may also decide the migration policy. Power consumption in cloud computing is also a major issue. Heat emission by the cloud component has to maintain in range, otherwise it can affect environment and service quality. Security is a necessary aspect of cloud computing. Cloud providers must maintain security mechanism so that protection can be achieved from threats. Virtual resources replica should be maintained that will help in case of failure of the system. The state of an application can be achieved by these phenomena. A proper log file must be saved so that in later stage, recovery can be done.

A. Prediction Factors

Cloud computing has different sets of users. There are three types of services in cloud computing that are IaaS (Infrastructure as a Service), PaaS (Platform as a Service), SaaS (Software as a Service). Usage of areas is completely different. But mainly number of users and usage per user will be the two important factors that will affect statistics of workload prediction. Number of users can be calculated by using any mathematical model. Best model should be selected as per the suitability of the cloud environment [5]. Workload prediction also includes classification and characterization. Historical data can give some idea about the usage of any service by cloud customers. According to this, resources can be managed and one can identify the focused area by which SLA can be verified. Criteria of satisfying SLA can be different; sometimes it can be execution speed, response time, memory capacity, or storage.

B. Workload Analysis

Workload of the cloud is closely dependent on the type of service, service popularity, service rate at different times, audience of service, environmental factors and any event getting occurred. There are many issues in which one cannot automate the system. System provisioning is being controlled manually. As an example, popularity of an application can increase very rapidly, internet technology can increase the number of users, and pricing policy of an application can scale users. These are the factors that can affect workload. These situations can be handled automatically up to a level. Beyond the level, manual provisioning is done.

C. Nature of Workload

Workload patterns can be classified according to their behavior. Five types of workloads [6] have been shown in Fig 1.The one that has constant number of requests is known as static workload. Rapidly increasing nature workload comes under rapidly growing workload. Periodic workload follows a pattern over time. On-and-off workload represents the work to be processed periodically or occasionally. Unpredictable workloads are special class of the periodic workloads as they need elasticity which is not predictable. This class of workload represents the constantly fluctuating loads. Method of dealing

 ISSN: 2321-8169
837 – 841

is specific according to the pattern of workload. Sometimes provider has to use reactive approach where nature of data is not known, otherwise proactive approach can be followed according to the repetitive pattern of data. Many times hybrid approach is also followed according to condition [7]. This combines the features of both approaches. The amount by which resources scale after following reactive approach may change according to the given criteria in cloud system.

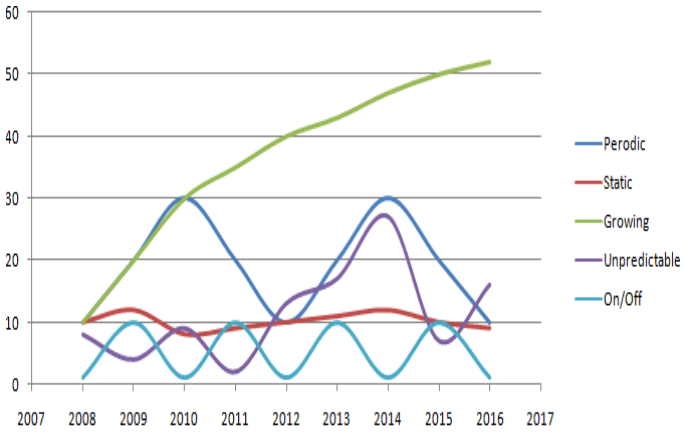

Figure 1: Nature of Workload in Last decades

They can be different in number and the nature depends upon the application. As an example, online gaming requires space on primary memory and bandwidth over network, but not so much space in the secondary memory. Effective scaling scheme should be automatic and more accurate in order to achieve high quality of service in cloud computing system, as scaling factor will be different in various scenarios. After allocating resources, their utilization percentage should be taken into account at regular intervals to achieve more effective scheme.

IV. Resource Scaling Methods

Scaling of resources can be done in various ways. Two main methods are proactive and reactive. In proactive approach, prediction has to be done and various implementations of resources are carried out for future use. Reactive approach follows a threshold value. After exceeding that value, a series of actions is implemented. The new resources can only be instantly used if virtual machine is available instantaneously. However in practice, VM provisioning takes few minutes. Thus, the reactive approach may fail to ensure QoS due to the VM provisioning delay. That is the biggest disadvantage of reactive approach. An alternative of this approach is to use a proactive approach by forecasting future resource demand values based on demand history. Prediction requires a lot of effort in advance. Many techniques have been proposed for forecasting future values based on historical data. Time series analysis is one of them. Below table contains various model names that can forecast time future value over time series analysis.

TABLE I. Models Summarization

Model Name	*Equation*	*Subhead Constant*	*Subhead*
Autoregressive AR(1))	$X_t=c+\theta x_{t-1}+\varepsilon_t$	C=constant , ε_t=white noise	Autoregressive AR(1))
Moving average MA(1)	$X_t=u+\varepsilon_t+\theta\ \varepsilon_{t-1}$	u =Mean, θ =model parameter	Moving average MA(1)
Simple exponential smoothing:	$\bar{Y}=\alpha y_t+(1-\alpha)y_t$	α=Smoothing constant	Simple exponential smoothing:
ARIMA	autoregressive (AR(p))+ integrated (I(q)),	----------	ARIMA
Neural network auto regression	----------	---------	Neural network auto regression

Apart from these models, support vector machine and neural network can also be used for forecasting future values. A shortcoming of proactive approach is that it is mainly dependent on historical data, so it can not reflect changes in real scenario that occur frequently in the cloud system. Many a times, prediction suffers from large number of errors due to uncertainty in the cloud system. In such special conditions, provider can scale resources using reactive approach. It occurs generally where unexpected high spikes are identified due to any reason [8]. To solve this problem, a hybrid approach is followed by the cloud service provider. If such a situation is periodic then provider can follow patterns based on past data and allocate number of instances of a resource to provide better service. A nonlinear workload is much difficult to handle in comparison to linear workload. Special algorithms are used to provide better services in such cases.

V. Prediction Methodology

Pattern matching is one of the widely used techniques in computer science. In paper [9], the author calculates the longest length of matching pattern. On the basis of matching, prediction is done. Prediction standalone is not sufficient for providing better services in the real workload. An integrated approach is Data grouping [1] can be implemented on the basis of classification. Same nature data has higher probability to process at the same time. One of the major advantages of this policy is that data can be available locally to the server resulting in high quality of service to the user in optimum cost. There is always a trade-off between SLA and the amount of resources being used. To satisfy SLA, virtual machines having workload higher than threshold value can be identified and estimation can be done to track the possible number of extra virtual machines for the workload balance. Periodic check is

performed to identify the overloaded and under loaded virtual machines. In case of under loading, servers for which sum of workload is less than the threshold value are searched and workload is transferred to a single server [1]. Many models are available to forecast the future values based on the historical data. Each model has specific area, dataset in which more accurate result can be obtained. In the reference [5], accuracy of many methods based on dataset was proposed and shown according to the amount of past data accuracy. Some have better performance in small traces of data and others on large traces. Once the virtual machine is used to perform the task, a lower and upper limit can be decided for the virtual machine to handle workload on that particular VM. Once such scheme is followed, a period is found in which machine has been work loaded near lower limit and usage of resources is not optimized in such a period. To get optimum usage of shared hosting proposed methodology is implemented where virtual machine can be shared between applications [10].VM can be added or removed fractionally but not fully. The benefit of this approach is that a large amount of web applications can be hosted. Utilization logs of virtual machine are maintained and overloaded and under loaded servers are identified. Different policies are implemented in order to balance the workload. SLA is one of the greatest factors in the cloud and plays an important role in setting up any system. There are many situations where provider has to deny the request of user in cloud computing. In reference [11], the authors have suggested a method for user where a decision has to be made. The user should get service and which set of user should be denied without compromising the SLA criteria. Resources have to be used in different manner to provide service to the customer. One of the approaches used in cloud system is buffer set of resources ready to use [8]. This leads to high cost and power consumption in cloud system. So, the prediction algorithm should be accurate as much as possible. In real world, workload is not stationary and linear model is not able to predict resource accurately. In such cases, different policies can be implemented. Feature selection of workload [8] should be identified and based on these features, predicting policy should be chosen. Real time workload is generally non stationary and heterogeneous in nature. A key matrix, quality factors and nature of application are used for the evaluation of key matrices and based on real matrices, system is periodically updated.

A. *Forecasting Methods*

The time series is a sequence of measurements of the same variable(s) made over time. To solve time series model various methods have been proposed.

- Autoregressive model (AR): An AR model is one in which y_t depends only on its own past values $y_{t\text{-}1}$, $y_{t.2}$,$y_{t\text{-}3}$,$y_{t\text{-}4}$ etc. Equation can be written as

$$y_t = f(y_{t\text{-}1}, y_{t.2}, y_{t\text{-}3}, y_{t\text{-}4}, \ldots\ldots, \varepsilon_t) \quad (1)$$

- Moving average model (MA): An AR model is one in which y_t depends only on the random error terms which follow a white noise process. Equation can be written as

$$y_t = f(\varepsilon_{t\text{-}1}, \varepsilon_{t\text{-}2}, \varepsilon_{t\text{-}3}, \varepsilon_{t\text{-}4}, \ldots\ldots\ldots\ldots\ldots\ldots) \quad (2)$$

- Autoregressive moving average model (ARMA): It is a mixture of both time series and moving average model. ARMA model requires stationary process. A stationary process is one in which mean and variance do not change over time and process does not have trends. Equation can be written as

$$y_t = \beta_0 + \beta_1 y_{t\text{-}1} + \beta_2 y_{t\text{-}2} + \beta_3 y_{t\text{-}3} + \ldots\ldots\ldots \beta_p y_{t\text{-}p} + \varepsilon_t + \phi_1 \varepsilon_{t\text{-}1} + \phi 1 \varepsilon_{t\text{-}2} + \phi_3 \varepsilon_{t\text{-}3} + \ldots\ldots \phi_q \varepsilon_{t\text{-}q} \quad (3)$$

- Autoregressive integrated moving average model (ARIMA): ARIMA model is a generalization of the autoregressive moving average (ARMA) model.

B. *Tools and methodologies*

To make real scenario, various implementation tools can be used. The table 2 describes the tools that can generate workload and have ability to store and process information for simulation purpose in cloud system. Hardware specification is also mentioned in table 2 on which simulation are performed. Real traces are used for the testing purpose.

TABLE II. CLOUD SIMULATION TOOLS SUMMARIZATION

Name	Tools and source of data
Juan [1]	Last FM
Carlos Vazquez [5]	R statistical programming [41] version 3.1.3, Lenovo Yoga Pro 2 laptop, Intel Core i7-4500U CPU @ 1.80 GHz-2.40GHz, 8GB of RAM, 512GB SSD, and a Windows 8.1 OS
Ali Yadavar Nikravesh [6]	Java implementation of TPC-W as the benchmark , Amazon EC2 are used as the infrastructure. Multi-Layer Perceptron and Support Vector Machine, WEKA tool is used to carry out the prediction task.
Mahesh Balaji [8]	Amazon AWS t2.micro instance, Time and Expense Management System (TEMS), HP JMeter, AWS Elastic Load balancer, MYSQL DB
Eddy [9]	Animoto , LCG, NorduGrid, SHARCNET
Adnan [10]	synthetic load (real traces not available)
Yao-Chung Chang [12]	NCHC public environment constructed from hadoop, 3194 users with a total number of 160,447 jobs having already been run
Chunhong[13]	Google cluster traces
Rodrigo [14]	real traces of requests from the Wikimedia Foundation, CloudSim as a simulator
Adnan [15]	ARVUE, CRAMP, and ACVAS. Squid proxy server access logs is used as a traces obtained from the IRCache project

VI. PERFORMANCE ASSESSMENT

The performance of a cloud system can be viewed from many aspects. From client point of view, it is related to the quality of service in the given cost. The provider can consider performance on the basis of the profit of the system. Performance can be improved by applying a better prediction algorithm along with suitable policy used for resource provisioning. There is always a difference between the estimated number of resources and actual statistics. This can be termed as error. The buffer system can be one of the solutions to such problems. A specific amount of system should be kept prepared in advance which may be used under certain conditions. Workload nature can heavily affect the performance and operational cost of the system. A policy must be implemented in the dynamic fashion. Many issues affecting performance like management, migration policy of virtual

resources can also be incorporated for the enhancement in cloud systems. In this paper, issues related to scaling are addressed for prediction and threshold value in reactive approach. Errors measured can be root mean square error, mean absolute error, mean absolute percentage error and mean absolute scaled error.

VII. CONCLUSION

Resource provisioning seems an important factor in the field of cloud computing. Resources should be deployed in such an efficient manner that maximizes the gain to be achieved in terms of cost. Lot of techniques has been purposed for the resources deployment. Every technique focused on specific area in which more accuracy can be achieved. Virtual machine management has many issues namely: booting time and migration. Resources need to be prepared in advance so that booting time can be avoided during execution. There remains a considerable separation between the predicted value and the real value. The amount of resources should be less as much as possible. SLA is one of the issues which should be carefully designed. According to SLA, the policy should be implemented. Issue related to resource availability, execution speed, security, storage amount and failure condition are addressed in SLA. To deal with many of the issues, resource provisioning is required. Resources are the backbone of any cloud system. Resources can be of any type such as storage, bandwidth, CPU etc. Methods of resource provisioning generally depend on the type and nature of the application. We must establish a relation between the estimated workload and the required resources. In calculation, the nature of workload should be involved. This gives cloud provider an idea of how much resources are required. From power consumption point of view, virtual machine can be kept in standby mode if workload on that particular machine is low from a fixed determined value. Workload management in cloud computing deeply affects cloud system and it is the key factor while setting up any cloud system. The reason is the entire system of cloud touches workload at any point.

REFERENCES

[1] Juan M. Tirado, Daniel Higuero, Florin Isaila, Jesus Carretero (2011). Predictive Data Grouping and Placement forCloud-based Elastic Server Infrastructures. 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 285-294.

[2] YonghongLuo, Shuren Zhou (2014).Power Consumption Optimization Strategy of Cloud Workflow Scheduling Based on SLA. Wseas Transactions on Systems. pp.368-367.

[3] Mohammad AlaulHaque Monil1,Rashedur M. Rahman (2016).VM consolidation approach based onheuristics, fuzzy logic, and migrationcontrol.Journal of Cloud Computing:Advances, Systems and Applications. pp.1-18.

[4] Maram Mohammed Falatah, Omar Abdullah Batarfi (2014).Cloud Scalability Considerations. International Journal of Computer Science & Engineering Survey. pp. 37-47.

[5] Carlos Vazquez, Ram Krishnan, Eugene John.Time Series Forecasting of Cloud Data Center Workloads forDynamic Resource Provisioning. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, vol: 6, number: 3, pp. 87-110.

[6] Ali YadavarNikravesh, Samuel A. Ajila, Chung-Horng Lung (2017).An Autonomic Prediction Suite for Cloud Resource Provisioning. Journal of Cloud Computing:Advances, Systems and Applications. pp.1-20.

[7] Dr.K.Chitra, B.Jeeva Rani (2014).DES: Dynamic and Elastic Scalability in Cloud Computing Database Architecture. International Journal of Advanced Computer Science and Applications,pp.173-175.

[8] Mahesh Balaji, Ch. Aswani Kumar, G. Subrahmanya VRK Rao (2016). Predictive Cloud Resource Management Framework for Enterprise Workloads. Journal of King Saud University - Computer and InformationSciences.

[9] Eddy Caron, Frederic Desprez, Adrian Muresan(2010). Forecasting for Grid and Cloud Computing On-Demand Resources Based on Pattern Matching.2nd IEEE International Conference on Cloud Computing Technology and Science, pp. 456-463.

[10] Adnan Ashraf, Benjamin Byholm, Ivan Porres (2012). Cost-Efficient Resource Allocation forMultiple Web Applications with Proactive Scaling.IEEE 4th International Conference on Cloud Computing Technology and Science, pp.581-586.

[11] H. Morshedlou, and M.R. Meybodi (2014).Decreasing Impact of SLA Violations: AProactive Resource Allocation Approach forCloud Computing Environments. IEEE Transactions on Cloud Computing. pp.1-12.

[12] Yao-Chung Chang, Ruay-Shiung Chang, Feng-Wei Chuang (2014). A Predictive method for Workload Forecasting in the Cloud Environment. Advanced Technologies, Embedded and Multimedia for Human-centric Computing, Lecture Notes in Electrical Engineering 260.

[13] ChunhongLiua, ChuanchangLiua, YanleiShanga, ShipingChenb, Bo Chenga, JunliangChen (2016). An Adaptive Prediction Approach Based on Workload PatternDiscrimination in the Cloud. Journal of Network and Computer Applications. pp.1-26.

[14] Rodrigo N. Calheiros, EnayatMasoumi, Rajiv Ranjan, RajkumarBuyya (2014). Workload Prediction Using ARIMA Model and Its Impact on Cloud Applications' QoS. IEEE Transactions on Cloud Computing, Vol. 3. pp.450-458.

[15] Adnan Ashraf, Benjamin Byholm and Ivan Porres (2016).Prediction-based VM provisioning andadmission control for multi-tier webapplications. Journal of Cloud Computing:Advances, Systems and Applications. pp.1-21.