Volume 8, Issue 6
December 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Data Driven Resource Provisioning for Efficient Utilization of Cloud Resources

[1]Seema Chowhan, [2]Ajay Kumar and [3]Shailiaja Shirwaikar
[1]Department of Computer Science, Baburaoji Gholap College, Pune, India
[2]Jayawant Institute of computer Application, Pune, India
[3]Department of Computer Science, Savitribai Phule Pune University, Pune, India
E-mail: [1]ssc_chow@yahoo.com, [2]ajay19_61@rediffmail.com, [3]scshirwaikar@gmail.com

ABSTRACT

Cloud computing is Internet based technology in which cloud provider provides the operating system, application software on the top of hardware as resources to its customer and delivers it as a service to the end users over the Internet. Cloud makes it possible to access application and data any time anywhere full filling Quality of service parameters (QoS) like performance, availability and reliability. Service provider supplies infrastructure including hardware and software as a Virtual Machine. The Challenge for a customer is to identify requirements at different workloads and choose an appropriate set of VMs that give a satisfying performance to its end users in terms of response time. To overcome this, this paper presents a data driven architectural framework for efficient resource utilization of cloud resources for online data, offline data and simulated data.

Keywords: *Service Level Agreement, Quality of Service, Virtual Machines, Resource Provisioning*

1. INTRODUCTION

Over the past years, computing technology has undergone a series of platform, software and environment changes .We access evolutionary changes in machine architecture, operating system platform, and network connectivity and application workload. Instead of using centralized computers to solve computational problem, a parallel and distributed system uses multiple computers to solve large scale problem over the Internet. Billions of people use the Internet everyday life. As a result, supercomputer site and large Data-centers must provide high performance computing service to huge numbers of Internet users concurrently, because of this high demand. An Internet cloud of resources can be either centralized or a distributed computing system. The cloud applies parallel or distributed computing or both. Clouds can build with physical and virtualized resources over large data centers that are centralized or distributed. The load balancing is an important issue to be addressed by cloud provider.

Load balancing at resource provisioning includes managing of huge bundle of resources that makes up the cloud and providing these resources to customer as per their requirement. Cloud provider has the herculean task of satisfying the **elastic demand** of its users with an optimal investment in infrastructure. From cloud provider's perspective, load balancing is an important activity to achieve efficient utilization of resources. SLA based resource provisioning involves maintaining appropriate levels of various resources while meeting QoS attributes or minimization of SLA violations. Data is not only important for designing templates but also important in developing resource provisioning and utilization strategy and decision making. With data one can find behavior, pattern and deviation, so there is need for data driven framework for efficient utilization of cloud resources. In this paper, a data driven architectural framework is proposed for resource provisioning that systematically builds a knowledge base and uses it to meet SLA and satisfying QoS. This frame work is applied and tested with on-line, off-line and simulated data.

Section 2 describes the prior work done on load balancing and resource provisioning. Section 3 describes data driven architectural frame work for resource provisioning. In Section 4, research solutions are describe with proposed methods and techniques. Section 5 conclude the research.

2. RELATED WORK

Cloud provider has to create an illusion of availability of unlimited computing resources to the end users on limited hardware and unpredicted request loads. The challenges for cloud computing provider is to allocate resources as per Service Level Agreements (SLAs) and performance of cloud system should be stable in any dynamic changes of workload as per SLA specified without effecting quality of service (QoS). By Byun et al [1], SLAs specify the resources and quality levels required for the execution of job in order to minimize the cost from user perspective and to maximize the resource utilization from provider's perspective. In such systems Quality of service parameters are availability, reliability, response time and throughput in contractual documents agreed between provider and customer called SLA [1-5]. QoS parameters play an important role in ranking service providers. With the rapid development of cloud computing, wide variety of web applications in network are moving to the cloud platform which makes high performance load balancing important. Load balancing distributes workload traffic amongst multiple computing resources to balance the load. Load Balancing is gaining critical importance in a cloud computing environment. Efficient load balancing scheme ensures efficient resource utilization by provisioning of resources to cloud users on demand basis in pay-as- you-go manner. Load Balancing can be carried out both at resource provisioning level which is heavily dependent on the Service Level Agreement (SLA) and also at the resource utilization level. Load balancing at Resource provisioning involves the distribution of resources to different cloud users without increasing wasted capacities and yet maintaining required Quality of service. The consumer of the service may provide one or more Service level objectives depending on application specific requirements but the cloud provider has to translate them into low level technical attributes that can be monitored and controlled to achieve the higher level objectives.

Efficient resource provisioning policies allow sharing of resources in Data Center to enhance the cloud performance. Resource provisioning that maintains quality of service with optimum resource utilization is one of the challenge. It is a multidimensional problem that can have issue based solution in the form of a set of services that helps to allocate and negotiate service level agreements (SLA) and design. Several research work have explored resource allocation using static and dynamic load balancing algorithms. Static load balancing algorithms like Round robin [6], Min-Min and Max-Min [7] are commonly used. Zhang [8] presented efficient load balancing mechanism for under-load and over-load situations using ant colony and complex network theory. Radojevic and Zagar [9] proposed Central load balancing decision model (CLBDM) with improvement on round robin algorithm which in turns makes use of session switching. Map Reduce-based load balancing technique for entity resolution was proposed by [10]. Nishant et al [11] recommended Ant Colony Optimization algorithm for distributing the workload among nodes at cloud for balancing the load.

Dynamic load balancing algorithms are more accurate and could result in more efficient load balancing. In [12] an algorithm was proposed to get a suitable execution sequence for workflow activities and their recoverability by the adaptive scheduling algorithm (ASA) which considers resource allocation constraints and dynamic topology changes. The dynamic scheduling Earliest Deadline First (EDF) algorithm which is used in real time system scheduling for multiprocessor system for efficient load balancing [13]. Zhong et al [14] proposed resource scheduling strategy in cloud computing using genetic algorithm. In [15] make use of the algorithms namely round robin, equally spread current execution and Throttled load balancing algorithm to distribute the load across VM instance to check the performance parameter time and cost. SLA is an important document for load balancing at resource provisioning level. The higher level attributes such as availability, reliability and low level attributes such as response time, throughput, latency time, downtime per week, Mean time to Repair (MTTR), Mean time between failures (MTBF) etc. are QoS parameters [16]. QoS of a cloud service fluctuates drastically at small timescales, due to network traffic conditions, cloud platform loads. The need is to continuously measure, monitor and take majors to maximize SLA parameters. Researchers has applied machine learning regression, decision tree, collaborative filtering approach, regression based hybrid CF algorithm and reduction modeling technique to overcome the above problems at resource provisioning level.

3. ARCHITECTURAL FRAMEWORK MODEL

Resource provisioning decisions are based on the predictive or deductive reasoning based on data collected and stored. The data can be on-line or off-line or simulated. It is a multidimensional problem that can have issue based solutions in the form of service utilizing few of the parameters from the knowledge base. The Architectural framework model for resource provisioning is presented in Figure 1. The middle-ware handles all the transactions between the stack holders and the Resource Layer (RL). The resource layer comprises of virtual machines as allocation unit on the top of the physical machines. The middleware components are knowledge Base, Service Repository, SLA base, broker SLA resource allocator and load balancer (Resource utilization).

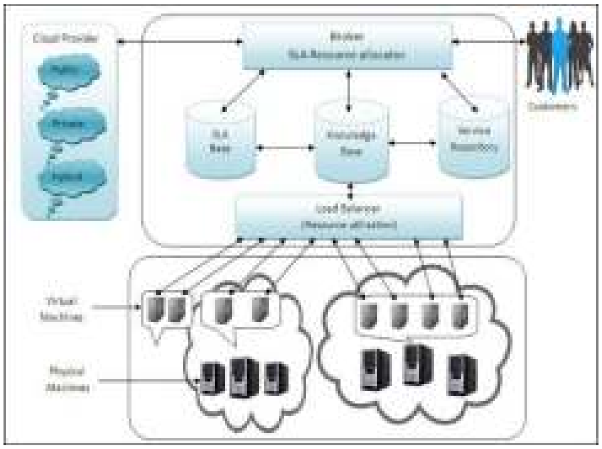

Figure 1: Architectural framework for data driven resource provisioning

3.1 SLA Base

SLA base stores negotiated service level agreement of different services used by customers. Penalties and rewards on the basis of SLA agreed between customer and provider are also stored in SLA base.

3.2 Knowledge Base

Knowledge base stores data continuously extracted about various QoS parameters. This data can be further analyzed and utilized for various provisioning level decisions. The knowledge base can form the basis of any such Resource management scheme and the issues and challenges that are to be addressed when implementing a solution to balance the load in a highly dynamic, elastic, cloud environment. Knowledge base comprises of SLA properties including the provider predefined QoS parameters and the customer specified QoS parameters .QoS parameters which are continuously monitored for better performance are

- **Response Time:** It represents the time taken by the provider to process a particular customer request. For example, An SLA violation occurs when the actual response time is longer than it was defined in the SLA [4].

- **Service Processing Time:** It is defined as the time taken to process an operation of a service. For example, how long it takes to generate a report, or save a transaction record [4].
- **Service Initiation Time:** How long it takes to initialize the service, which includes the VM initiation time and application deployment and installation time [4].
- **Data Transfer Time:** How long it takes to transfer one GB record from one VM to another. This depends on the network bandwidth [4].
- **Throughput:** It is used to calculate the no. of tasks whose execution has been completed [4].
- **Performance:** It is used to check the efficiency of the system. Performance and statistics of user requirements [4].
- **Resource Utilization Parameters:** Memory utilization ratio, theCPU utilization ratio, the network bandwidth are some such resource utilization parameters which are important when taking decision to minimize resource utilization.

The data about above parameters continuously extracted on-line can be used for constructing a predictive model. QoS parameters which are derived statically from simulation experiments and generated data are

- **Request Type:** This defines the type of customer request, which may be a first time rent or a service upgrade request. First time rent means the customer is renting a new service from cloud provider. To downgrade a service first the customer needs to terminate the current contract and then processing of this downgrade request will be treated as a new request [4].
- **Contract Length:** How long the customer is going to use the software service [4].
- **Number of Accounts:** The actual number of user accounts that a customer wants to create. The maximum number of accounts is related to and restricted by the type of product edition [4].
- **Number of Records:** The average number of records that a customer is able to create for each account during a transaction and this may impact the data [4].
- **Penalty Conditions:** For each SLA violation the provider needs to pay a penalty, which is based on the delay in the response time to the customer. For each request type there is a different penalty. Penalty rate is the monetary cost incurred to the provider for unit time delay in serving the customer request [4].
- **VM types:** The type of VM image that can be initiated. For instance, there may be three types of VMs large, medium, and small. The three types of VMs have different capability to serve different numbers of accounts and records since different requests may consume different memory and storage [4].
- **VM Price:** How much it costs for the provider to use a VM for the customer request per hour. It includes the physical equipment, power, network and administration cost [4].
- **Capacity Parameters:** Nodes capacity, number of CPUs, the CPU processing speed, memory size are some indicator of Data-Center capacity.

The data about above parameters are extracted through simulation experiments and can be used for designing VM templates, SLA clause and algorithms for resource provisioning and utilization.

3.3 Service Repository

There are various essential services that are required for using predictive and deductive reasoning approach. A service that periodically maps the idle capacities into a set of SLA templates and support user requirements that automatically balance the load.

- A service can design VM template with different configuration of small, medium and large VMs.
- A service can derive SLA clause for designed template.
- A service can derive pricing for these SLA templates so that customers find them attractive.
- A service can implement resource provisioning and utilization strategies.
- A service can monitor QoS parameters.
- A service can devise predictive modeling of SLA parameters. Predictive modeling is a collection of techniques that create or extract a model in the form of a mathematical relationship between a set of features from the training data, validate its efficacy by measuring the error or deviation on test data and use it to predict the values of certain features when certain other features are known in the current or future data.

The middle-ware will selectively choose the combination of services depending on the requirement to optimize resource provisioning. All such services will form the services Repository.

3.4 Broker SLA resource allocator

Resource allocator agent does resource provisioning as requested by other agent and also allocation of resources. The dynamism of cloud, requiring continuous monitoring of requests and resources, handling of ever changing requirements, schedules and prices, selecting appropriate services and plans to meet overall objectives of the cloud, all this suggests use of autonomous agents for managing cloud. Due to the naturally decentralized architecture, this paradigm provides appropriate concepts for realizing systems that offer inherently non-functional requirements such as scalability, robustness and failure tolerance in cloud. An agent is an autonomous software system that reacts pro-actively to changes in the environment and interacts with other agents, persistently pursuing its goals. The multi-agent system has a set of agents that interact together to resolve a common problem by using the resources and the knowledge base of each agent.

3.5 Load Balancer (Resource Utilization)

Load balancing is a technique to achieve efficient resource utilization by distributing the workload amongst one or more resources. The resources includes CPU and network links. With advancement in technology load balancing involves virtual machine management. Load balancer is also responsible for extracting various QoS parameters like response time, throughput and store it in knowledge base.

4. RESEARCH SOLUTIONS

In this research work researcher has used data driven architectural frame work model for resource provisioning by mixing historical, predictive and live information about resources for dynamic re-planning and provisioning of resources. Cloud provider should give priority to customers need and satisfaction since they pay for accessing services for cloud data centers. A single solution can be difficult and not smart solution to design complete system. Multiple solution strategies are the best. To explore working of middle ware three separate methods are proposed and used to attain the objective of research that are an analytical model to monitor performance of VM server, Predictive Modeling for efficient management of Service Level Agreement Parameters in Cloud Services and Template-Based Resource Provisioning (TBRP) method for efficient resource Utilization in Cloud Data-Center.

- An analytical model has been presented for customer to understand customer requirement for provisioning resources for web-based application which is hosted on cloud. As a case study a document management system designed on Amazon public cloud is used for running different services and to monitor the performance of VM server for different CPU usage parameters [17].

- An Agent based SLA-management has proposed, where the agent uses predictive approach for predicting SLA violations and taking appropriate actions by interacting with other agents. The design is based on a case study on available datasets containing measurements on web services of SLA parameters such as response time and throughput. Simple statistics based predictive methods like Last-State Based Method (LSAM), Simple Moving Average Method (SMAM), and Weighted Moving Average Method (WMAM). In addition Regression analysis with Multivariate Linear Regression with five point summary (RFPS) and mean and variance (RMV) are applied to estimate the parameters and minimize the error function. For choosing the right Prediction Model a comparative analysis of prediction error for different algorithms are used for deciding the right prediction model. Clustering is also used to further improve prediction accuracy [18].

- Template based resource provisioning and Utilization (TBRP) method with its procedure has been proposed in [19]. TBRP method considers different workload scenarios and various experiments to extract QoS parameters. It considers design of fixed and varying capacity VM template with its cost and completion time SLA clause. This method explores performance of TBRP method with fixed and varying capacity VM templates on varying load to overcome the problem of under-provisioning and over-provisioning of resources as per QoS specified in SLA. In addition effectiveness of TBRP method is also measured for different capacity Data [20].

5. CONCLUSION

Cloud computing describes a progression of Internet utilization, from computers, to people and now computing as utility anytime anywhere, allowing for many new applications and services using cloud environment. With the increase in cloud users and the dynamic and elastic demands, Resource provisioning that maintains Quality of Service with optimum resource utilization is a challenge. It is a multidimensional problem that can have issue based solutions in the form of a service utilizing few of the parameters from the knowledge base as shown in Figure 1. A service that periodically maps the idle capacities into a set of SLA templates can support user requirements that automatically balance the load. A service can derive pricing for these SLA templates so that customers find them attractive. All such services will form the services Repository. The middle-ware will selectively choose the combination of services depending on the requirement to optimize resource provisioning. Data driven architectural frame work provides efficient utilization of cloud resources, maximize resource utilization and minimize SLA violation.

REFERENCES

[1] M. Byun, E. K., Kee, Y. S., Kim, J. S., & Maeng, S. (2011). Cost optimized provisioning of elastic resources for application workflows. *Future Generation Computer Systems*, *27*(8), 1011-1026.

[2] Bianco, P., Lewis, G. A., and Merson, P. 2008. Service level agreements in service-oriented architecture environments. Tech. rep., CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST.

[3] John, M., Gurpreet, S., Steven, W., Venticinque, S., Rak, M., David, H., Gerry, M., DI MARTINO, B., Le Roux, Y., John, M., et al. 2012. Practical guide to cloud service level agreements.

[4] Wu, L., Garg, S. K., and Buyya, R. 2011. Sla-based resource allocation for software as a service provider (saas) in cloud computing environments. In Proceedings of the 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. IEEE Computer Society, 195–204.

[5] Liu, F., Tong, J., Mao, J., Bohn, R., Messina, J., Badger, L., and Leaf, D. 2011. Nist cloud computing reference architecture. NIST special publication 500, 2011, 1–28.

[6] Sotomayor, B., Montero, R. S., Llorente, I. M., and Foster, I. 2009. Virtual infrastructure management in private and hybrid clouds. IEEE Internet computing 13, 5, 14–22.

[7] Etminani, K. and Naghibzadeh, M. 2007. A min-min max-min selective algorihtm for grid task scheduling. In 2007 3rd IEEE/IFIP International Conference in Central Asia on Internet. IEEE, 1–7.

[8] Zhang, Z. and Zhang, X. 2010. A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation. In 2010 The 2nd International Conference on Industrial Mechatronics and Automation. Vol. 2. IEEE, 240–243.

[9] Radojevi´c, B. and ˇZagar, M. 2011. Analysis of issues with load balancing algorithms in hosted (cloud) environments. In MIPRO, 2011 Proceedings of the 34th International Convention. IEEE, 416–420.

[10] Kolb, L., Thor, A., and Rahm, E. 2012. Load balancing for map reduce based entity resolution. In 2012 IEEE 28th international conference on data engineering. IEEE, 618–629.

[11] Nishant, K., Sharma, P., Krishna, V., Gupta, C., Singh, K. P., Rastogi, R., et al. 2012. Load balancing of nodes in cloud using ant colony optimization. In 2012 UKSim 14th International Conference on Computer Modelling and Simulation. IEEE, 3–8.

[12] Avanes, A. and Freytag, J.-C. 2008. Adaptive workflow scheduling under resource allocation constraints and network dynamics. Proceedings of the VLDB Endowment 1, 2, 1631–1637.

[13] Singh, J. 2010. An algorithm to reduce the time complexity of earliest deadline first scheduling algorithm in real-time system. arXiv preprint arXiv:1101.0056. 9, 34

[14] Zhong, H., Tao, K., and Zhang, X. 2010. An approach to optimized resource scheduling algorithm for open-source cloud systems. In 2010 Fifth Annual China Grid Conference. IEEE, 124–129.

[15] Sharma, T. and Banga, V. K. 2013. Efficient and enhanced algorithm in cloud computing. International Journal of Soft Computing and Engineering (IJSCE) ISSN, 2231–2307.

[16] Tang, B. and Tang, M. 2014. Bayesian model-based prediction of service level agreement violations for cloud services. In TASE. 170–176.

[17] Chowhan, S. S., Kumar. A, & Shirwaikar, S. (2019) .Performance Analysis of services in Cloud Computing. Journal of Emerging Technologies and Innovative Research (JETIR), Vol.–6, Issue-5 251-258

[18] Chowhan, S. S., Shirwaikar, S., & Kumar, A. (2016). Predictive Modeling of Service Level Agreement Parameters for Cloud Services. International Journal of Next-Generation Computing, 7(2), 115-129.

[19] Chowhan, S. S., Kumar. A, & Shirwaikar, S. (2019) .Template-Based Efficient Resource Provisioning and Utilization in Cloud Data-Center. International Journal of Computer Sciences and Engineering, Vol.-7, Issue-1 463-477.

[20] Chowhan, S. S., Kumar. A, & Shirwaikar, S. (2019) .Measuring effectiveness of TBRP method for different capacity Data Centers. IPASJ International Journal of Information Technology (IIJIT), Vol.-7, Issue-5, 6-16.

AUTHOR PROFILES

Ms. Seema Chowhan is working as a faculty and head in subject of computer science in Baburaoji Gholap College Pune, India affiliated to Savitribai Phule Pune University, Pune. She has 18+ years of experience in teaching UG and PG courses. She has completed M.Phil (CS).Her research interests include Cloud Computing and Networking.

Dr. Ajay Kumar experience covers more than 26 years of teaching and 6 years of Industrial experience as IT Technical Director and Senior Software project manager. He has an outstanding academic career completed B.Sc. App. Sc. (Electrical) in 1988, M.Sc. App. Sc. (Computer Science-Engineering and Technology) in1992 and PhD in1995. Presently, working as Director at JSPMs Jayawant Technical Campus, Pune (Affiliated to Pune University). His research areas are Computer Networks, Wireless and Mobile Computing, Cloud computing, Information and Network Security. There are 74 publications at National and International Journals and Conferences and also worked as expert, appointed by C-DAC to find Patent-ability of Patent Applications in ICT area. Six commercial projects are completed by him for various companies/ Institutions. He holds variety of imperative position like Examiner, Member of Board of Studies for Computer and IT, Expert at UGC.

Dr. Shailaja Shirwaikar has a PhD in Mathematics of Mumbai University, India and worked as Associate Professor at Department of Computer Science, Nowrosjee Wadia College affiliated to Savitribai Phule Pune University, Pune for last 27 years. Her research interests include Soft Computing, Big Data Analytics, Software Engineering, machine learning and Cloud Computing.