ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 8, October 2012

A Study on Autonomic Placement and Resource Management with Cloud Workloads

T.Ambika[1], S.Ananthi[2] , Dr.S.Karthik [3]

Abstract— **With the advent of Cloud computing, with hosting and delivering the demanded services, enormous benefits were reaped by its users as capital expenditure on the computing resources is reduced to a very large extent. As cloud has its rising popularity for its services, the count of customers is also escalating at a higher level resulting in the need for framing novel techniques for managing the available resources with the providers effectively and also at the benefit of the customers such that their requirements are met. The workloads arrived at the cloud providers are of heterogeneous in nature. Hence it is necessary to arrive at automated decisions with the different types of workloads, may it be interactive or non interactive workload prevailing with an organization. This paper aims at making a detailed study of different approaches with autonomic placement and efficient and dynamic resource management techniques with different types of workloads in cloud.**

Index Terms—**Cloud computing, autonomic computing, Resource management, Workload Management.**

I.INTRODUCTION

The enormously growing computing power requirement of the customers has a serious impact over the cloud providers. As the cloud providers are concerned, they are to face many a number of challenges. They have to provide services to its customers by meeting the SLAs as well as attain a profitable state for themselves. The prominent among the challenges face by the cloud providers is the efficient resource provisioning in an automated manner along with the fluctuations in the arrival of applications [1], which can also be expected with heterogeneous nature.

Manuscript received Sep 15, 2012.

Ambika.T[1]*, PG Scholar, Department of Computer Science, SNS College of technology, Coimbatore.*

Anathi.S[2]*, Asst. Prof, Department of Computer Science, SNS College of technology, Coimbatore.*

Dr.S.Karthik[3]*, Dean, Department of Computer Science, SNS College of technology, Coimbatore.*

A. Resource Management

The Cloud resources are no different from the resources of any enterprise own resource environment, except that they reside in a remote place, that is with the providers. Ideally, to have a complete view of the cloud computing resources one may use today or may want to use in the future. These resources are to be effectively allocated to the customer requirements. Various issues are to be faced on making such resource allocation by the cloud providers.

The resource management decisions can be made in either of the two ways.

I.Proactive manner

II.Reactive manner

Where in proactive method, the performance model is used for making periodic predictions and allocating resources based on it and in reactive method, dynamic reaction in resource allocation is done immediately.

B. Types of Workloads

Besides the resource management issues such as power-aware resource allocation, SLA oriented resource allocation etc., there comes the vital necessity of making workload aware resource provisioning. The Workload can be stated as an abstraction of the actual work at an instance which is to be performed.

The workloads may be interactive with user inputs or non-interactive where the jobs may be interdependent in its processing and hence of long-running nature, say batch jobs. In [2], the author clearly specifies the possibility of different types of workloads in an enterprise. Transactional web workloads or long-running batch jobs, both may have different processing requirements. Transactional web workloads require flow control and dynamic application placement while the long-running batch jobs may require job scheduling with different job scheduling techniques [3]. They are mapped to a set of resources depending upon its performance requirements for its execution. Heterogeneous workloads are those where different types of workloads say batch jobs or web applications which are dealt at a time based on their arrivals. Such heterogeneous jobs could be consolidated within a single machine for more effective resource utilization.

On resource provisioning resource requirement estimation must be calculated such that resource utilization

is made effective. Underestimating the resource requirements leads to the condition where the user requirements cannot be aptly met and therefore the SLAs cannot be met. On the other hand, overestimating the resource requirements leads to a miserable condition where underutilization of resources takes place. This would reduce the providers revenue to a larger extend. Hence resources are to be managed in such a way that both the cloud service providers and the consumers have a win-win situation [4].

C. Autonomic Computing

On considering the features of cloud autonomic computing plays a prominent role. It refers to a state where self managing characteristics are adapted. It includes the conditions where adapting to unpredictable changes occurs. On such conditions the intrinsic complexity faced must be hidden with its users. This autonomic computing is facilitated with the usage of autonomic components.

The rest of the paper gives a detailed view with the different approaches in autonomic placement, resource and workload management with the cloud enterprises.

II. Related Study

The challenges faced by cloud computing such as resource management with dynamic allocation, workload prediction and management are discussed with different approaches and with various parameters.

A. Dynamic Workload Placement

The resources are allocated dynamically on the arrival of workloads according to its requirements.

1) Utility based autonomic execution:

In [5], the authors proposed a utility based approach for adaptive workload execution. The workloads considered are the queries and workflows. The autonomic workload mapper uses a single utility function at a time choosing one of the two properties, response time or profit for autonomic workflow execution. And for the autonomic query execution, the admission controller, a query scheduler and execution controller are used and one of the two properties, response time or number of QoS targets met are chosen. They coordinate adaptation at various granularities and addresses context-specific optimization. Utility based optimization is adapted in [7] also where the concentration was jointly on capacity allocation, load balancing and energy saving policies while meeting availability constraints. And the optimization procedure concentrates on increase or decrease of working frequency upon varying loads dynamically. Also for effective resource usage server migration and server state transition depending on the oncoming loads is considered.

2) Static vs. dynamic allocation:

The overhead of dynamic allocation scheme is compared with static allocation in both system capacity and application performance in [6] with different virtualization technologies. Comparison was done with different combinations of workloads and virtualization techniques but with homogeneous environment.

3) Online clustering approach:

In [8] the authors provide a decentralized robust online clustering approach for cloud workloads where cluster based VM provisioning is done. For high-performance workloads the quadratic response surface model is used to analyze the behavior.

4) Ant colony optimization:

A workload placement problem is modeled as an instance of Multi-Dimensional Bin Packing problem and uses the artificial swarm intelligence over it. The solution is proposed with ant colony optimization. This provides the dynamic placement of workloads with current load. This technique proves to poses an energy-aware resource utilization mechanism [14].

B. Workloads

1) Homogeneous workloads:

The statistical analysis of workloads on data-intensive clusters which consists of the workloads of parallel single CPU tasks are studied in [9]. The new patterns of job arrivals are discussed, which includes Pseudo-periodicity, long range dependence and grid based bag-of-tasks. They are much necessary for making workload modeling and performance predictions. The job run time or memory sequences on clusters are auto correlated and cross correlations are made significant.

2) Heterogeneous workloads:

The performance study of network I/O workloads is discussed in [12]. The different workloads which demand for the CPU or I/O resources are studied here. Further analysis is made on various factors that affect the performance throughput and resource sharing process. The performance on co-location of applications requiring CPU and network resources i.e., the heterogeneous applications are measured along with performance measurement of co-locating homogeneous applications. Comparison was also made with different allocation and scheduling strategies. A detailed study in [2] is done for consolidation of heterogeneous workloads in the same machine. The performance goals for different workloads vary, say for long running batch jobs it is the completion time of each job and in case of transactional workload the goals are defined in terms of response time and throughput. So the performance for different workloads are measured accordingly and with relative performance calculation the actual performance achieved is calculated and compared with the required

performance of the workload such that the SLAs could also met.

3) Rank based workloads:

The level of the workloads, the least or the maximum workload level is considered such that the rank is assigned accordingly as regular state and critical state, which is obviously a peak workload state, is categorized in [10]. For the regular state the local data centre is sufficient while for the critical state the cloud environment is advised. Here the author also gives a minimum migration algorithm for reducing the overhead involved in the process.

4) Statistical analysis:

The statistical workload analysis and replay for mapreduce is used in [11] for measuring the performance trade-offs with cloud. It creates the real time workload for analysing the latency and utilization trade offs in mapreduce.

III. Conclusion

With the ever increasing cloud requirements and enormous transactions with such virtualized data centers, an efficient management of various cloud resources for large amount of workloads is vital. Moreover it is necessary for dynamic resource allocation of workloads can be made possible by making perfect predictions on workload arrival. Since the cloud users are increasing tremendously, its essential to adapt power aware resource management and service providing mechanisms. This paper studies various approaches for resource and workload management strategies in a detailed manner where the cloud resources are effectively managed for different workloads of homogeneous or heterogeneous in nature at a time. This must be done under the condition that the service level agreements of the cloud consumers are met. Moreover on such circumstances the cloud providers' profit is also considered.

REFERENCES

[1] Qi Zhang, Lu Cheng, Raouf Boutaba, "Cloud computing: state-of-the-art and research challenges", *J Internet Serv Appl* (2010) 1: 7–18

[2] David Carrera, Malgorzata Steinder, Ian Whalley, Jordi Torres, Eduard Ayguade, "Autonomic Placement of Mixed Batch and Transactional Workloads", *IEEE Transactions on Parallel and Distributed Systems*, Vol.23, No.2, February 2012

[3] Malgorzata Steinder, Ian Whalley, David Carrera, Ilona Gaweda, David Chess, "Server virtualization in autonomic management of heterogeneous workloads", *Proc. IEEE/IFIP 10th symp Integrated Management (IM '07)*, 2007

[4] Marios D. Dikaiakos , George Pallis, Dimitrios Katsaros , Pankaj Mehra , Athena Vakali , "Distributed Internet Computing for IT and Scientific Research", *IEEE computer society*, September/October 2009

[5] Norman W. Paton, Marcelo A. T. de Arag˜ao, Kevin Lee, Alvaro A. A. Fernandes, Rizos Sakellariou, "Optimizing Utility in Cloud Computing through Autonomic Workload Execution", *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2009

[6] Z. Wang, X. Zhu, P. Padala, and S. Singhal, "Capacity and Performance Overhead in Dynamic Resource Allocation to Virtual Containers," *Proc. IFIP/IEEE 10th Int'l Symp. Integrated Network Management* (IM '07), pp. 149-158, May 2007

[7] Bernardetta addis, Danilo Ardagna and Barbara Panicucci politecmio di milano, Li zhang, "Autonomic management of cloud service centres with availability guarantees", *IEEE 3rd Int. Conf. on cloud computing*, 2010

[8] Andres Quiroz, Hyunjoo Kim, Manish Parashar, Nathan Gnanasambandam, Naveen Sharma, "Towards Autonomic Workload Provisioning for Enterprise Grids and Clouds", *10th IEEE/ACM International Conference on Grid Computing*, 2009

[9] Hui Li, Lex Wolters, "Towards a Better Understanding of Workload Dynamics on Data-Intensive Clusters and Grids", *IEEE*, 2007

[10] Snehil Sharma, Abhishek Mathur, Shailendra Shrivastava, "ESRWF: Extreme State-Rank based Workload Factoring for Integrated Cloud Computing Model", *International Journal of Electronics and Computer Science Engineering*, ISSN 2277-1956/V1N3-1340-1349

[11] Yanpei Chen,Archana Sulochana Ganapathi, Rean Griffith, Randy H. Katz, "Towards Understanding Cloud Performance Tradeoffs Using Statistical Workload Analysis and Replay", http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-81.html

[12] Yiduo Mei, Ling Liu, Xing Pu, Sankaran Sivathanu, and Xiaoshe Dong, "Performance Analysis of Network I/O Workloads in Virtualized Data Centers", *IEEE Transactions on Service Computing*, 2011

[13] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, "Market-Oriented Cloud Computing: Vision, Hype and Reality for Delivering IT Sevices as Computing Utilities" *HPCC*, 2008

[14] Eugen Feller, Louis Rilling, Christine Morin," Energy-Aware Ant Colony Based Workload Placement in Clouds", *INIRIA* Version-1, May 2011

[15] Kannan S, Kalaikumaran T, Karthik S, Arunachalam VP, "Ant Colony Optimization for Routing in Mobile Ad-Hoc Networks", *International Journal of Soft Computing* 5 (6), 223-228, 2010

[16] Karthik S, Arunachalam V.P, Ravichandran T, "A Comparative Study of Various IP Trace back Strategies and Simulation of IP Trace Back", *Asian Journal of Information security page* 454-458; ISSN: 1682-3915, 2008

[17] Karthik S, Arunachalam V.P, Ravichandran T, "Multi Directional Geographical Traceback with n Directions Generalization" , *Journal of Computer Science*, page646-651, 1549-3636; Science publications

[18] Karthik S, Arunachalam V.P, Ravichandran T,, " A Novel Direction Ratio Sampling Algorithm (DRSA) Approach for Multi Directional Geographical Traceback", *International Journal of Computer Science and Security*, Vol.3 issue 3, 2010 ,pp.272-279

[19] Karthik S, Kannan S, Maragatham T, Arunachalam V.P,"A Study of Attacks, Attack Detection and Prevention Methods in Proactive and Reactive Routing Protocols", *International Journal of Business Management*, Vol.5,No.3,2011,pp.178-183

[20] Kannan S, Karthik S, Arunachalam V.P,"An Investigation on Performance Analysis and Comparison of Proactive and Reactive Routing Protocols in Mobile Ad-Hoc Network", *International Journal of Soft Computing* 5(5),194-199

Ambika.T received her MCA degree at Institute of Road and Transport Technology, Erode. She is currently pursuing M.E degree in Computer Science and Engineering at SNS College of Technology, Coimbatore. Her areas of interests are Network Secutity and Cloud Computing.

Ms.S.Ananthi is presently Assistant professor, Department of Computer science and Engineering, SNS college of Technology, affiliated to Anna University- Coimbatore, Tamil Nadu, India. She received her MCA degree in 2003 from bharathidasan University, Trichy and received her Master of Computer Science and Engineering degree from SNS college of Technology, affiliated to Anna University ,Coimbatore, India.

Professor Dr.S.Karthik is presently Professor & Dean in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University- Coimbatore, Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Ann University of Technology, Coimbatore. His research interests include network security, web services and wireless systems. In particular, he is currently working in a research group developing new Internet security architectures and active defense systems against DDoS attacks. Dr.S.Karthik published more than 35 papers in refereed international journals and 25 papers in conferences and has been involved many international conferences as Technical Chair and tutorial presenter. He is an active member of IEEE, ISTE, IAENG, IACSIT and Indian Computer Society.

.