ISSN 2278-3091
Volume 4, No.4, July - August 2015
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse01442015.pdf

A Survey on Resource Management in Cloud Computing Environment

S.K.Sonkar [1], Dr.M.U.Kharat[2]
[1] Research Scholar, Department of Computer Engineering, KKWagh IEE&R Nashik, SP Pune University, India.
sonkar83@gmail.com
[2] Head & Professor, Department of Computer Engineering METIOE Nashik, SP Pune University, India.
mukharat@rediffmail.com

ABSTRACT

Nowadays cloud computing is very popular for offering services over the Internet. The very important advantage of cloud computing is the ability to provision resources on demand. This avoids the problems of over-provisioning and under-provisioning which are commonly seen with organizations that have widely variable requirements due to increase/decrease, seasonal high and low workload etc. Resource allocation policies decide the amount of resource to be allocated to a particular or set of virtual machines (VMs). This resource allotment policy can also update the dynamically. For implementing prioritization, it require to provide more resource to a specific virtual machine, compared to other virtual machine

The resources offered may include memory consumption, storage, CPU processing power, IT services, and so on. Many of the touted gain in cloud computing that comes from resource multiplexing through virtualization technology.

Keywords: cloud computing**,** Resource Allocation, VM, virtualization.

1. INTRODUCTION

Cloud Computing in which datacenters hardware and system software provides the application service over the internet. The datacenter hardware and software is what we call a cloud.

Cloud Computing (CC) provides the illusion of unlimited computing resources available on demand. This eliminates the need for cloud computing users to plan far ahead of provisioning. The elimination of an up-front commitment by Cloud users, thereby allowing companies to start small and increase hardware resources only when there is increase in their needs. One of the other features of CC is ability to pay for use of computing resources on a short-term basis as needed and release them as work is over. Thereby rewarding conservation by letting machines and storage go when they are no longer useful [1].

Public clouds, private cloud, are cloud deployment models. When cloud is made available in a pay as you go manner to the general public then it is called public cloud. The public cloud providers are Amazon web services, Google AppEngine, Microsoft azure etc. The term private cloud refers to internal datacenters of a business or other organization, not made available to the general public [2].

There are three different types of cloud service models i.e. are Software as a service (SaaS), Platform as a Service (Paas), and Infrastructure as a service (IaaS). When the application delivered as a service over the internet and hardware and the system software in datacenter providing those services, then those services are called software as a service. In PaaS User can deploy his own applications created by using programming languages and the tools that are supported by provider. The user does not need to manage cloud infrastructure but he has to control over deployed applications. Where as in IaaS user provided with infrastructure such as storage, network and other computing resources. User is able to deploy and run arbitrary software by using these resources.

For cloud applications, virtualization is one of very important technology, which consist of two features that make it ideal for cloud computing, first is service partitioning and second is isolation[3]. With partitioning, virtualization is able to support many applications and operating systems to share the same physical device while isolation enables each guest virtual machine to be protected from system crashes or viruses in the other virtual machines. Virtualization abstracts the physical infrastructure through a virtual machine monitor (VMM) or hypervisor, which maps virtual machine to physical hardware. VMM enables multiple virtual machines or guest operating systems to share a single physical machine securely and fairly. VMM shows the virtual machine under the illusion that it has its own physical device.

Resource allocation policies decide the amount of resource to be allocated to a particular or set of virtual machines. This resource allotment policy can also update the dynamically. For implementing prioritization, we can provide more resource to a specific virtual machine, compared to other virtual machine. In next section we will see diff. Resource allocation strategies.

2. LITERATURE SURVEY - RESOUCE ALLOCATION STRATEGIES

T. Wood et al. [4] give the Black and Grey box strategies with BG algorithm. Author uses Xen hypervisor and finds with Nucleus and monitoring engine, Grey-box

enables proactive decision making. While it has the limitation as, Black-box is limited to reactive decision making and BG algorithm requires more number of migrations.

A. Singh et al. [5] introduces the integrated server storage virtualization (Vector dot algorithm) using Configuration and performance manager. This scheme has a smaller amount of complication but its forecasting is not believable. Because of uneven distribution of remaining resource makes it hard to be fully utilized in the future.
Zhen Xiao [6] gives the strategy for dynamic resource allocation with Skewness and load prediction algorithm. He uses Xen hypervisor Usher controller. The merits in his system are no overheads, high performance. It requires less number of migrations and residual resource is friendly to virtual machines. It improves the scheduling effectiveness. The demerit of the system is it is not cost effective.

The benefit of shared space of cloud infrastructure explained by L. Qiang et al. [7] in which author proposed resource allocation strategy using feedback control theory, for suitable management of virtualized resources, which is based on virtual machine (VM). In this VM-based architecture all hardware resources are combined into common shared space in cloud computing infrastructure so that hosted application can right to use the required resources as per there need to meet Service Level Objective (SLOs) of application. The adaptive manager use in this architecture is multi-input multi-output (MIMO) resource manager, which consist of 3 controllers: CPU controller, memory controller and I/O controller, its goal is control multiple virtualized resources utilization to achieve SLOs of application by using control inputs per-VM CPU, memory and I/O allocation.

Utility functions provide a natural and advantageous framework for achieving self-optimization in distributed autonomic computing systems explained by walsh et al. [8]. Author present a distributed architecture, implemented in a realistic prototype data center that demonstrates how utility functions can enable a collection of autonomic elements to continually optimize the use of computational resources in a dynamic, heterogeneous environment. The architecture consists of a two-level structure of autonomic elements that supports elasticity, modularity, and self-management. Each individual autonomic element manages application resource usage to optimize local service-level utility functions, and a global arbiter maps resources among application environments based on resource-level utility functions obtained from the managers of the applications. The utility function scheme is suitable for handling realistic, fluctuating Web-based transactional workloads running on a Linux cluster.

Resource provision based on updated actual task executed explained by Jiayin Li et al. [9] which proposes an adaptive resource allocation algorithm for the cloud system with preempt able tasks in which algorithms adjust the resource provision adaptively based on the updated of the actual task executions. Author proposed Adaptive list scheduling (ALS) and adaptive min-min scheduling (AMMS) algorithms and used for task scheduling which includes static task scheduling, for static resource allocation, is generated offline. Online adaptive method is use for re-evaluating the remaining static resource allotment repeatedly with predefined frequency. For every re-evaluation process, the schedulers are re-calculating the finish time of their respective submitted tasks, not the tasks that are assign to that cloud. So this method is suitable for static resource allocation.

The dynamic resource allocation using distributed multiple criteria decisions in computing cloud explained by Yazir Y.O.et al. [10]. In it author contribution is tow-fold, first distributed architecture is adopted, in which resource management is separated into independent tasks, each of which is performed by Autonomous Node Agents (NA) in ac cycle of three activities: (I) VM Placement, in it suitable physical machine (PM) is found which is capable of running given VM and then assigned VM to that PM, (II) Monitoring, in it total resources use by hosted VM are monitored by NA, (III) In VM Selection, if local accommodation is not possible, a VM need to migrate at another PM and process loops back to into placement. Second using PROMETHEE method, node Agent carry out configuration in parallel through multiple criteria decision analysis. This scheme is most suitable for large data centers as compared with centralized approaches.

Nowadays distributed computing systems solves rising demand of computing and memory. In the distributed systems specifically resource allocation is one of the most important challenges while the clients have Service Level Agreements (SLAs) and the whole profit in the system depends on how the system can meet these SLAs. This issue was solved by solved by Goudarzi et al. [11] which optimizes the total profit gained from the multidimensional SLA contracts for multi-tire application. In this scheme higher level of entire profit is provided by using force-directed resource assignment (FRA) heuristic algorithm, in this case primary solution is based on provided solution for profit higher level problem. Then, distribution rates are set and local optimization step is use for improving resource sharing. Resource consolidation method is applied lastly to consolidate resources to determine the active (ON) servers and further optimize the resource obligation. As concluding this method is suitable for improving resource sharing and to optimize the resource assignment.

Use of steady state timing models, tafi. et al. [12] presents information of cloud HPC resource arrangement. In which author proposed quantitative application dependent instrumentation scheme to inspect several important dimensions of a program's scalability. Sequential and parallel timing model with program instrumentations can reveal architecture exact deliverable performances that are difficult to measure otherwise. These models are introduces to connect

several dimensions to time domain and application speed up model is use to tie these models in same equation. This provides ability to explore multiple dimension of program quantitatively to gain non--trivial insight. Authors use Amazon EC2 as a target processing environment.

Provisioning of computing, storage, and networking resources in order to satisfy requests generated by remote end-users is achieved through Cloud services. Very fast Internet access and multi-core Virtual Machines (VMs) permit today the provisioning of diversified and enriched types of services in Cloud environment. In this issue Aoun R. et.al [13], consider several types of basic services and show how their orchestration may lead to the provisioning of more sophisticated services. For this purpose, they define four types of requests that cover the wide spectrum of possible services. Which devise the resource provisioning problem as a Mixed Integer Linear Program (MILP). It assumes that the underlying infrastructure is based on a set of end-to-end connections with guaranteed sustainable bandwidth such as Carrier-Grade Ethernet (CGE) circuits. Author investigates the impact of two innovative services on resource allocation carried out by a Cloud Service Provider (CSP). These services matches with distributed data storage and to multicast data transfer. For the former service, it requires to consider the possibility of splitting a storage request onto different remote storage nodes. The final service targets to allocate a similar data sequence from one server towards numerous remote nodes assuming a limited number of network nodes have multicast capacities. These two novel services provide a gain of 7% in terms of accepted requests when applied to the 18-node NSFnet backbone network.

In recent times, Internet-based distributed, multitenant [14] applications connective to internal business applications, known as software as a service (SaaS) are gaining popularity.

Aversa et al. [15] present and assess an implementation of a prototype scalable web server containing of a load-balanced cluster of hosts that collectively accept TCP service connections. The system IP addresses are advertised using round robin DNS (RR-DNS) system, allowing any system to receive requests from any client. As soon as client attempts to establish a TCP connection with one of the hosts, a decision is made as to whether or not the connection should be redirected to a different host-namely, the host with the lowest number of established connections. Authors use the low-overhead Distributed Packet Rewriting (DPR) technique to redirect TCP connections. In this prototype, each host keeps information about the remaining hosts in the system. Load record is handled using periodic multicast amongst the cluster hosts. Performance measurements suggest that this method outperforms both pure RR-DNS and the stateless DPR solutions.

Scalability is very important to the success of several enterprises that currently involved in doing business on the Web and in providing information that may vary drastically from one time to another. Managing enough resources just to meet peak requirements can be costly. Cloud computing gives a powerful computing model that allows users to access resources on-demand. Chieu T.C. et al. [16] express a new architecture for the dynamic scaling of Web applications based on thresholds in a virtualized cloud computing environment. They illustrate scaling approach with a front-end load-balancer for routing and balancing user requests to Web applications deployed on Web servers installed in virtual machine instances. For automated provisioning of virtual machine resources based on threshold number of active sessions is achieved by introducing dynamic scaling algorithm. At any instance ability of the cloud to quickly provision and dynamically assign resources to users will be discussed. This work has demonstrated the compelling benefits of the cloud which is capable of handling sudden load surges, delivering IT resources on-demands to users, and maintaining higher resource utilization, thus reducing infrastructure and management costs

3. CONCLUSION

Cloud computing can solve complex set of tasks in shorter time by proper resource utilization. To make the cloud to work efficiently, best resource allocation strategies have to be employed. Utilization of resources is one of the most important tasks in cloud computing environment where the user's jobs are scheduled to different machines. Virtualization provides an efficient solution to the objectives of the cloud computing paradigm by facilitating creation of Virtual Machines (VMs) over the underlying physical servers, leading to improved resource utilization and abstraction.

ACKNOWLEDGEMENT

FA would like to thanks KKWagh IEE&R for providing facilities to make successful work of this paper.

REFERENCES

1. M. Armbrust et al., "Above the clouds: A berkeley view of cloud computing," University of California, Berkeley, Tech. Rep., Feb 2009.
2. Abhinandan S. Prasad, Member, IEEE, and Shrisha Rao,"A Mechanism Design Approach to Resource Procurement in Cloud Computing," IEEE Transactions On Computers, Vol. 63, No. 1, Pp.17-30, January 2014.
3. En-Hao Chang, Chen-Chieh Wang, Chien-Te Liu, Kuan-Chung Chen, Student Member, IEEE, and Chung-Ho Chen, Member, IEEE," Virtualization Technology for TCP/IP Offload Engine," IEEE Transactions on Cloud computing, Vol. 2, no. 2, pp.117-129, April-June 2014.
4. T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Black-box and gray-box strategies for virtual machine migration," in Proc. Of the Symposium on Networked

Systems Design and Implementation (NSDI'07), Apr. 2007.
5. A. Singh, M. Korupolu, and D. Mohapatra, "Server-storage virtualization: integration and load balancing in data centers," in Proc. of the ACM/IEEE conference on Supercomputing, 2008.
6. Zhen Xiao, Senior Member, IEEE, Weijia Song, and Qi Chen, "Dynamic resource allocation using virtual machine in cloud computing environment," IEEE Transaction on Parallel and Distributed Systems, vol.24, no.6, pp.1107-1117, June 2013.
7. L. Qiang, Q. Hao, L. Xiao and Z. Li "Adaptive Management of Virtualized Resources in Cloud Computing Using Feedback Control," in First International Conference on Information Science and Engineering, pp. 99-102 ,April 2010.
8. W. E. Walsh, G. Tesauro, J. O. Kephart, and R. Das, "Utility Functions in Autonomic Systems," inICAC '04: Proceedings of the First International Conference on Autonomic Computing. IEEE Computer Society, pp. 70–77, 2004.
9. Jiayin Li, Meikang Qiu, Jian-Wei Niu, Yu Chen, Zhong Ming, "Adaptive Resource Allocation for Preemptable Jobs in Cloud Systems," in 10th International Conference on Intelligent System Design and Application, pp. 31-36,Jan. 2011.
10. Yazir Y.O., Matthews C., Farahbod R., Neville S., Guitouni A., Ganti S., Coady Y., "Dynamic resource allocation based on distributed multiple criteria decisions in computing cloud," in 3rd International Conference on Cloud Computing, pp. 91-98, Aug. 2010.
11. Goudarzi H., Pedram M., "Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems," in IEEE International Conference on Cloud Computing, pp. 324-331, Sep. 2011.
12. Shi J.Y., Taifi M., Khreishah A.,"Resource Planning for Parallel Processing in the Cloud", in IEEE 13th International Conference on High Performance and Computing, pp. 828-833, Nov. 2011.
13. Aoun R., Doumith E.A., Gagnaire M., "Resource Provisioning for Enriched Services in Cloud Environment", IEEE Second International Conference on Cloud Computing Technology and Science, pp. 296-303,Feb. 2011.
14. F. Chong, G. Carraro, and R. Wolter,"Multi-Tenant DataArchitecture," Microsoft Corporation, 2006.
15. Aversa and A. Bestavros. "Load Balancing a Cluster of Web Servers using Distributed Packet Rewriting", Proceedings of the 19th IEEE International Performance, Computing, and Communication Conference, Phoenix, AZ, Feb. 2000, pp. 24-29
16. Chieu T.C., Mohindra A., Karve A.A., Segal A., "Dynamic Scaling of Web Applications in a Virtualized Cloud Computing Environment", in IEEE International Conference on e-Business Engineering, Dec. 2009, pp. 281-286.

Authors Profile

Mr. S. K. Sonkar Completed the Bachelor and Master degree in Computer science and Engineering from SRTMU Nanded. He is currently pursuing the Ph.D. degree in computer engineering from University of Pune. He is presently Research Scholar of KKWagh IEE&R Nashik, India. His current research interests include Computer Network and Cloud Computing.

Dr. M.U.Kharat Completed PHD (Computer Science & Engg.) From Devi Ahilya University, Indore, India. Presently he is working at MET's IOE, Nashik, Maharashtra, India, as Professor & Head Computer Engineering Department. He has presented papers at National and International conferences and also published papers in National and International Journals on various aspects of Computer Engineering and Networks. He has worked in various capacities in academic institutions at the level of Professor, Head of Computer Engineering Department, and Principal. His areas of interest include, Computer Networks, Distributed System and the Cloud Computing.