Resource Allocation in Cloud Computing

Sudeepa R, Dr. H S Guruprasad

Abstract— **Cloud computing is an on-demand service because it offers dynamic flexible resource allocation for reliable and guaranteed services in pay as-you-use manner. Due to the ever increasing demands of the users for services or resources, it becomes difficult to allocate resources accurately to the user demands in order to satisfy their requests and also to take care of the Service Level Agreements (SLA) provided by the service providers. This paper discusses various Resource Allocation techniques.**

***Index Terms*—SLA, Service provider, Resource Allocation**

I. Introduction

Cloud is a group of computers or servers which are interconnected together to provide resources to the clients. It emerges as a brand new computing paradigm that aims to supply reliable, custom-made and QoS (Quality of Service) warranted computing dynamic environments for the end customers. The main problems related to cloud computing are the network bandwidth, response time, minimum delay in data transfer and minimum transfer cost for data. In cloud computing the resource allocation plays an important role in the performance of the entire system and also the level of customer satisfaction provided by the system. However, while providing the utmost customer satisfaction, the service provider ought to make sure of the profits to him also. So the resource allocation should be economical on both views i.e. on the end user and the service provider perspective. So as to get such a system the new technologies insist that the system should be with minimum SLA (Service Level Agreements) violation.

There are numerous advantages of cloud computing, the most basic ones being lower costs, re-provisioning of resources and remote accessibility. Cloud computing lowers cost by avoiding the capital expenditure by the company in renting the physical infrastructure from a third party provider. Due to the flexible nature of cloud computing, we can quickly access more resources from cloud providers when we need to expand our business. The remote accessibility enables us to access the cloud services from anywhere at any time. To gain the maximum degree of the above mentioned benefits, the services offered in terms of resources should be allocated optimally to the applications running in the cloud.

The cloud computing users get good quality services from their service providers with an affordable cost. The quality and cost of the services are based on their source allocation process in the particular service environment. The provider should assign the resource to the clients in an optimal way. There are so many resource allocation models that are used in cloud computing area. Each of this models use certain methods and algorithms for this purpose. The following is the survey done on the prior works of resource allocation models of the cloud computing environment. The main concentration is on resource allocation methodologies.

II. LITERATURE SURVEY

In cloud computing, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. In [1], Vinothina et. al. discuss Resource Allocation Strategy (RAS) as an integrating cloud provider activity for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application. Also, a summary of the classification of RAS and its impacts in cloud system is given. In [7], Sowmya Koneru et. al. Focus on increasing the efficacy of the scheduling algorithm for the real time Cloud Computing services. The RR scheduling Algorithm utilizes the Turnaround time Utility efficiently by differentiating it into a gain function and a loss function for a single task and also used to maximize the efficiency gain. An overall improvement in the resource utilization and reduction in the processing cost is shown. In [11], Thangaraj et. al. allocate the resource for Infrastructure as a Service (IaaS) based on predefined allocation policies. Here, the focus is on deadline sensitive policy to allocate the resources in a successful manner by reducing the request rejections by Haizea. Haizea is an open source resource lease manager, and can act as a scheduler for open source cloud toolkit Nebula. In [15], Siva theja et. al. considered a stochastic model for load balancing and scheduling in cloud computing clusters, where jobs arrive according to a stochastic process and request resources like CPU, memory and storage space. It takes the performance of JSQ (Join-the shortest-queue) routing and two-choice routing algorithms with MaxWeight scheduling policy have throughput optimal. The results show that these algorithms are queue length optimal in the heavy traffic limit. In [16], Ikki Fujiwara et. al. propose a market-based resource allocation mechanism that allows participants to trade their services by means of a double-sided combinational auction. An efficient public market place for cloud computing environments is given. It proposes a market-based resource allocation mechanism to allocate services to participants in effective manner. It orders a combination of services for workflows and co-allocations and enables participants to trade future and current services in the forward market and the spot market. In [17], Kazuki et. al. Mention the cloud resource allocation guidelines for limit to electric power capacity available in some area. It proposes an optimal resource allocation ability and bandwidth as well as electric power capacity. Optimal allocation means that the number of requests that can be processed is maximized, and the power consumed by a request is minimized. This paper presents an algorithm that attempts to reduce total electric power

Manuscript received April 03, 2014

Sudeepa R, PG Scholar, Dept. of ISE, BMSCE, Bangalore
Dr. H S Guruprasad, Professor and Head, Dept. of ISE, BMSCE, Bangalore

consumption by aggregating request processing of multiple areas.

Makhlouf et. al. [21] deal cloud environment using game theory. This paper explored to maximize the revenue and satisfy users by maximizing their utilities. A theoretical model based on Stackelberg game is proposed and a Stackelberg/Nash equilibrium solution is found. Zhenzhong et. al. [22] proposes resources allocation method named Statistic based Load Balance (SLB) to solve the problem of load imbalance in cloud Environment. This approach consists of a data analysis of on-line historical performance for forecasting the resource demand of each VM and an algorithm for the purpose of load balancing is given. In [23], Praveen et. al. address two main issues using multi-valued distributed hash tables for efficient resource discovery and the resource allocation is optimal based on the Nash equilibrium obtained from the auction game model. In [25], Zuling et. al. Introduce a new cloud resource allocating algorithm called CRAA/FA (Cloud Resource Allocating Algorithm via Fitness-enabled Auction), which creates a market for cloud resources and makes the resource agents and service agents bargain in that market.

In [8], Qi Zhang et. al. present a cloud management framework that dynamically allocates data center resources to spot markets to best match customer demand in terms of both supply and price in order to maximize the providers revenue and customer satisfactions while minimizing energy cost. In [12], Gihun et. al. provides a framework to dynamically place a new virtual machine (VM) to an appropriate physical machine (PM) that would achieve the best performance, guarantee the maximized utilization level, and prevent performance degradation of the data center on cloud computing environments. It also proposes resource management architecture to perform a location aware VM placement and dynamic resource utilization management. In [19], Makhlouf et. al. propose a Minimum Cost maximum Flow (MCMF) algorithm for optimal dynamic placement of virtual resources in data centers and cloud infrastructures to serve multiple users and also time varying demands and workloads. This paper also proposes a modified Bin-Packing algorithm which is used to benchmark the minimum cost maximum flow solution. The MCMF close to optimal placement and Bin-Packing algorithm find a viable solution to the dynamic resource allocation problem in clouds. In [20], Sheng et. al. formulates the resource allocation problem in a VM-multiplexing resource allocation scheme to manage decentralized resources to achieve maximized resource utilization using the proportional share model (PSM), and also delivers adaptively optimal execution efficiency. This paper proposes a novel scheme (DOPS) for virtual resource allocation on a Self-organizing cloud (SOC), and the three key contributions are, Optimization of task’s resource allocation under user’s budget, Maximized resource utilization based on PSM and Lightweight resource query protocol with low contention.

In [4], Gihun et. al. propose allocation of consumer resource to a proper data center using adaptive resource allocation model based on geographical location of consumer and the workload of data center in cloud computing environment. This model is tested using agent based test bed, and the proposed model shows a better response time for allocation than related resource allocation models. In [9], Gopalakrishna et. al. propose Adaptive Resonance Theory-2 which identifies and solves the pattern of incoming request problem by auto classifications and organizes pre-allocation strategies in a predictive way. This technique can reduce the cost per task completion in a cloud environment. In [13], Nilolaus et. al. present a new approach to self-adaptive resource allocation in virtualized environments based on online architecture-level performance models. The use of such models helps to predict the effects of changes in user workloads, as well as to predict the effects of respective reconfiguration actions, undertaken to avoid SLA violations or inefficient resource usage. In [14], Xavier et. al. discuss working towards Reinforcement learning to automatic resource allocation approach for performing resource allocation in cloud computing. The resource allocation problem for cloud computing is discussed which also deal with the problem in the Q-learning framework. The paper also presents the implementation of the workflow meant to bring Reinforcement Learning to real cloud computing infrastructures. In [24], Saeed et. al. presented a security-aware approach for resource allocation in clouds framework that allows for effective enforcement of defense-in-depth for cloud VMs. Modeling the cloud provider’s constraints and customer’s requirements as a constraint satisfaction problem (CSP), which can be solved using Satisfy ability Modulo Theories (SMT) solvers to reducing risk and improving manageability in cloud.

In [5], Chih et. al. propose a Map-Reduce-Merge framework, a naming and configuring scheme that extends Map-Reduce to processing heterogeneous datasets simultaneously in cloud computing. Map-Reduce-Merge framework can be used to implement many relational operators, particularly joins and other database operations. In [6], Warneke et. al. discuss various possible opportunities and challenges for efficient parallel data processing in clouds. Here, a user framework called Nephele is proposed which is the first data processing framework to explicitly exploit the dynamic resource allocation offered by both task scheduling and execution. This paper presents a performance comparison to the well-established data processing framework Hadoop. In [10], Rajkumar et. al. presents a Rule Based Resource Manager which increases the scalability of private cloud on-demand and reduces the cost. Also, time is set for public cloud and private cloud to fulfill the request and provide the services in time. On the basis of resource utilization and cost in hybrid cloud environment, it evaluated the performance of Resource Manager. In [18], Radhya et. al. present a search algorithm in the optimization module of the Virtual Design Advisor (VDA) which improves the efficiency for resource partition and allocation in the cloud. The proposed algorithm in this paper is called Greedy Particle Swarm Optimization (GPSO), which is a hybrid between particle swarm optimization and greedy. Using experiments with TPC-H benchmark on PostgreSQL database, it is presented that GPSO algorithm was able to escape from local minima and reduced the cost as compared to the greedy algorithm. In [2], Bo yin et. al. present a mechanism for the Cloud computing service provider to allocate resources based on the client’s SLA (Service Level Agreement). Force directed search algorithm is proposed which is the solution for SLA based resource allocation problem for multi-tier applications in cloud computing. In [3], Hadi et. al. present two common ways to optimize resource utilization. One is at the application level when applications are arriving and other is in the period of applications running. This multi-dimensional resource allocation (MDRA) scheme dynamically allocates the virtual resources among the cloud computing applications to reduce cost by using fewer nodes to process applications.

III. CONCLUSION

Cloud computing is an emerging technology and various researches have been carried out in order to solve the challenges faced by cloud. There are several challenges that cloud is facing, out of which a major challenge being the resource allocation techniques. This paper provides an overview of different resource allocation techniques.

ACKNOWLEDGMENT

The authors would like to acknowledge and thank Technical Education Quality Improvement Programme [TEQIP] Phase 2, BMS College of Engineering and SPFU [State Project Facilitation Unit], Karnataka for supporting the research work.

REFERENCES

[1] V.Vinothina, Dr.R.Sridaran, Dr.Padmavathi Ganapathi, "Resource Allocation Strategies in Cloud Computing", International Journal of Advanced Computer Science and Applications [IJACSA], Vol. 3, No.6, 2012. ISSN: 2158-107X (Print), DOI: 10.14569/issn.2156-5570.

[2] Bo Yin, Ying Wang, Luoming Meng, Xuesong Qiu, "A Multi-dimensional Resource Allocation Algorithm in Cloud Computing", Journal of Information and Computational Science, 2012, pp 3021-3028.

[3] Hadi Goudarzi, Massoud Pedram, "Multi-dimensional SLA based Resource Allocation for Multi-tier Cloud Computing Systems", IEEE International Conference on Cloud Computing (CLOUD), 4 - 9 July 2011, Washington DC USA, pp 324-331, Print ISBN: 978-1-4577-0836-7, DOI: 10.1109/ CLOUD.2011.106.

[4] Gihun Jung, Kwang Mong Sim, "Agent-based Adaptive Resource Allocation on the Cloud Computing Environment", 40th International Conference on Parallel Processing Workshops[ICPPW], Taipei City, 13-16 Sept. 2011, pp 345-341, DOI 10.1109/ICPPW.2011.18.

[5] H. chih Yang, A. Dasdan, R.-L. Hsiao, D.S. Parker, "Map-Reduce-Merge: Simplified Relational Data Processing on Large Clusters, "International Conference on Management and Data [SIGMOD 07], June 12 - 14 2007, Beijing, China, ACM 978-1-59593-686-8/07/0006.

[6] D. Warneke, O. Kao, "Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud", IEEE Transactions on Parallel and Distributed Systems, Vol. 22, No. 6, pp 985 - 997, June 2011, DOI: http://doi.ieee computersociety.org/10.1109/TPDS.2011.65.

[7] Sowmya Koneru, V N Rajesh Uddandi, Satheesh Kavuri, "Resource Allocation Method using Scheduling methods for Parallel Data Processing in Cloud", International Journal of Computer Science and Information Technologies[IJCSIT], Vol. 3(4), 2012, pp 4625 - 4628 4625, ISSN: 0975-9646.

[8] Qi Zhang, Quanyan Zhu, Raouf Boutaba, "Dynamic Resource Allocation for Spot Markets in Cloud Computing Environments", Fourth IEEE International Conference on Utility and Cloud Computing, Melbourne Australia, 5 - 8 Dec 2011, pp 178 - 185, ISBN: 978-0-7695-4592-9, DOI: http://doi.ieeecomputersociety.org/10.1109/UCC.2011.33.

[9] Dr.T R Gopalakrishnan, P Jayarekha, "Pre-allocation Strategies of Computational Resources in Cloud Computing using Adaptive Resonance Theory-2", International Journal on cloud computing: Services and Architecture(IJCCSA), Vol.1, No.2, August 2011, ISSN: 2231 - 5853 DOI : 10.5121/ijccsa.2011.1203.

[10] Rajkamal Kaur Grewal, Pushpendra Kumar Pateriya, "A Rule based Approach for Effective Resource Provisioning In Hybrid Cloud Environment", International Journal of Computer Science and Informatics ISSN (Print): 2231 - 5292, Vol - 1, Issue -4, 201..

[11] Thangaraj P, Soundarrajan S, Mythili A, "Resource allocation policy for IaaS in Cloud computing", International Journal of Computer Science and Management Research, Vol 2, Issue 2, pp 1645 - 1649, February 2013, ISSN 2278-733X.

[12] Gihun Jungand, Kwang Mong Sim, "Location-Aware Dynamic Resource Allocation Model for Cloud Computing Environment", International Conference on Information and Computer applications (ICICA 2012), pp 37 - 41, IPCSIT, Vol. 24, IACSIT Press, Singapore.

[13] Nikolaus Huber, Fabian Brosig, Samuel Kounev, "Model-based Self-Adaptive Resource Allocation in Virtualized Environments", 6th International Symposium on Software engineering for Adaptive and Self Managing Systems, SEAMS '11, May 23-24, 2011, Waikiki, Honolulu, HI, USA, ISBN: 978-1-4503-0575-4, DOI: 10.1145/1988008.1988021.

[14] Xavier Dutreilhy, Sergey Kirgizov, Olga Melekhova, Jacques Malenfant, Nicolas Rivierrey, Isis Truckz, "Using Reinforcement Learning for Autonomic Resource Allocation in Clouds: Towards a Fully Automated Workflow", ICAS 2011: The Seventh International Conference on Autonomic and Autonomous Systems, IARIA, 2011, pp 67 - 74, ISBN: 978-1-61208-134-2.

[15] Siva Theja Maguluri, R Srikant, Lei Ying, "Heavy Traffic Optimal Resource Allocation Algorithms for Cloud Computing Clusters", 24th International Teletraffic Congress, Article No. 25, ISBN: 978-1-4503-1896-9.

[16] Ikki Fujiwara, Kento Aida, "Applying Double-sided Combinational Auctions to Resource Allocation in Cloud Computing", 2010 10th Annual International Symposium on Applications and the Internet, ISSN: 978-0-7695-4107-5/10, DOI 10.1109/ SAINT.2010.93.

[17] Kazuki Mochizuk, Shinichi Kuribayashi, "Evaluation of optimal resource allocation method for cloud computing environments with limited electric power capacity", 2011 International Conference on Network-Based Information Systems, ISSN: 978-0-7695-4458-8/11, DOI 10.1109 /NBiS.2011.11.

[18] Radhya Sahal, Sherif M. Khattab, Fatma A. Omara, "GPSO: An Improved Search Algorithm for Resource Allocation in Cloud Databases", International Conference on Computer Systems and Applications [AICCSA], Ifrane, 27-30 May 2013, pp 1 - 8, ISSN: 2161-5322, DOI: 10.1109/ AICCSA.2013.6616472.

[19] Makhlouf Hadji, Djamal Zeghlache, "Minimum Cost Maximum Flow Algorithm for Dynamic Resource Allocation in Clouds", Fifth International Conference on Cloud Computing, 2012, ISSN: 978-0-7695-4755-8/12, DOI 10.1109/ CLOUD.2012.36.

[20] Sheng Di, Cho-Li Wang, "Dynamic Optimization of Multi-attribute Resource Allocation in Self-Organizing clouds", IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 3, March 2013, ISSN: 1045-9219/13_ 2013.

[21] Makhlouf Hadji, Wajdi Louati, Djamal Zeghlache, "Constrained Pricing for Cloud Resource Allocation", IEEE International Symposium on Network Computing and Applications, 2011, ISSN: 978-0-7695-4489-2/11, DOI: 10.1109/ NCA.2011.64.

[22] Zhenzhong Zhang, Haiyan Wang, Limin Xiao, Li Ruan, "A Statistical based Resource Allocation Scheme in Cloud", International Conference on Cloud and Service Computing, 2011, ISSN: 978-1-4577-1637-9/11.

[23] Praveen Khethavath, Johnson Thomas, Eric Chan-Tin, Hong Liu, "Introducing a Distributed Cloud Architecture with Efficient Resource Discovery and Optimal Resource Allocation", Ninth World Congress on Services, ISSN: 978-0-7695-5024-4/13, DOI 10.1109/ SERVICES.2013.68.

[24] Saeed Al-Haj, Ehab Al-Shaer, HariGovind V. Ramasamy, "Security-Aware Resource Allocation in Clouds", IEEE 10th International Conference on Services Computing, 2013, ISSN: 978-0-7695-5026-8/13, DOI 10.1109/ SCC.2013.36.

[25] Zuling Kang, Hongbing Wang, "A Novel Approach to Allocate Cloud Resource with Different Performance Traits", IEEE 10th International Conference on Services Computing, 2013, ISSN: 978-0-7695-5026-8/13, DOI 10.1109/ SCC.2013.109.

Dr. H S Guruprasad is working as Professor and Head, Information Science Department at BMS College of Engineering, Bangalore. He has twenty four years of teaching experience. He has been awarded with Rashtriya Gaurav award in 2012. His research areas are Network Communications, algorithms, Cloud Computing and Sensor Networks.

Sudeepa R is pursuing M.tech in Computer Network and Engineering at BMS College of Engineering, Bangalore, Karnataka. His area of interest Cloud computing and Computer Networking.