International Journal of Engineering & Technology, 7 (3.24) (2018) 517-521

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Review on Optimization Approaches in Cloud Computing Service

V.E. Jayanthi[1*], R. Divya[2], M. Jagannath[3]

[1]Department of Electronics and Communication Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu, India
[2]Department of Computer Science and Engineering, NPR College of Engineering and Technology, Natham, Tamil Nadu, India
[3]School of Electronics Engineering, Vellore Institute of Technology (VIT), Chennai, Tamil Nadu, India

Abstract

Cloud has become the best revenue generator tool in the business world. The Cloud Service Provider (CSP) gives importance to the big data arriving in the cloud. Many methodologies are currently available for cloud storage, retrieval and processing. But a discrete solution for all the problems is not possible due to the volume, velocity and variety of data arriving in cloud. Certain best and proved algorithms can be used based on resource utilization, cost pricing, load balancing for the effective utilizations of cloud. The profit based optimization becomes the goal of the CSP. This draws the attention in finding the major factors that influence cloud computing services. In this paper, a detailed survey of optimization techniques for various key factors of cloud are analyzed and the result obtained in each technique is consolidated, tabulated and compared.

Keywords: *Cloud service provider; Deadlock;Dynamic resource provisioning;Virtual machine migration; Starvation.*

1. Introduction

Cloud has become a basic amenity of daily life for organizations based on business, gaming, software, file sharing, billing, web-hosting and software development [1]. Because of the enormous resource management with elastic nature and self-service manner, the cloud is considered to be a fragile platform to handle. The arising demand of resources in all fields increase the number of cloud service providers (CSP) in market. Creating resources, effectively managing them, protecting them from threats are the major duties of CSP. They create efficient methods in achieving profit through it [2,3]. Many investors are making their investment in cloud seeing its faster development.

Even though cloud was initially invented in the 1960s, it was introduced for industrial purpose in India in 2006. Once it was practiced in the business areas, many issues in security, service level agreement and its infrastructure started to follow. Research in cloud gained attention after it and started increasing every year towards better results each time. Each year many new cloud deployment models are developed for each dormant factor in cloud. The virtualization environment of cloud, the hardware architecture, self-organizing and optimizing nature of it creates new challenges and makes the research still active. The problems of cloud are easy to solve using one crisp solution due to the fact that the issues are interlinked with one another. Increasing the number of servers may be a solution for effective speed and throughput but on the contrary affecting the profit, investment and proper resource utilization [4]. Enhancing the performance with the resources available would be the possible way of solving cloud problems. Minimizing the resources available is nothing but optimization. Certain adaptation in the underlying factors and parameters of cloud would produce efficient utilization [5].

Optimization is a mathematical model based on decision making. The CSP is the decision maker providing the inputs, taking control over the factors and optimizing it. The cloud price data, number of data centers and VM has to be decided by the CSP based on his detailed history from Amazon web services, Google trace data, planet lab etc. User provides their requirement in numerical values of input. The factors of cloud to be optimized are controlled by the CSP and defined as control variables. The type of cloud service and the number of users are used as decision variables. The CSP should take care of various resources of cloud like bandwidth, memory, storage, number of servers, processing speed, latency, power, cost, virtual machine down time and migration time [4].

Once the factors of the cloud to be managed are defined as control variables by the CSP, it becomes easy in optimizing the cloud environment by maximizing or minimizing the desired profit function with decision thresholds and solved for different input data sets from the cloud user [5,6]. The CSP have hundreds of servers with identical or different memory size, storage space, CPU speed, and bandwidth. They aim profit in providing service. They have to make their resources completely utilized for better profit.

Optimization techniques are meant for effective utilization of cloud resources. This also optimizes the cost at the provider side. The arriving load on the provider side has to be examined and hacking avoidance and resource over utilization. The optimization technique faces many challenges.It needs to deal with the varying loads on the cloud.

It needs to scale with the increased number of users at a time [7]. In a short interval of time, it should make fast decisions in service allocation, and deal with additional problems like hackers, over-utilization, under-utilization, deadlocks, distributed database, replication etc. Thus optimization techniques are widely extended for three factors of cloud: Resource based optimization, Cost based optimization and Load based optimization. Resource based optimization takes care of the user side resource and ways of utilizing them at the maximum profit without deadlock in handling multiple users [8,9]. Cost based optimization also deals with the techniques in maximizing the profit using different pricing techniques and cost estimation techniques. Load based optimization takes care of balancing the resources using decided parameters for hacking avoidance and fairness in service. The existing methods, methodologies adopted in the literatures are elaborated in Table 1.

Table 1: Summary of methods and methodologies adopted.

Method	Methodology	Inference
• Zoutendijk’s feasible direction method • Gradient projection method • Penalty method	• Karush Kuhn Tucker condition and convergent step size rules are adopted • Changing penalty co-efficient in each iteration until the convergence of the expected result occurred	• Properties are difficult to satisfy • Suitable in cases where objectives do not conflict with each other
• Apriori method	• Global criterion: Sum of the squares of the relative deviations of the individual objective function has to be reduced from the feasible ideal solutions	• Used in cases where user is able to mention his decision variables and goals clearly
• Lexicographic method	• Ranking of objectives and constraints is done	
• Weighted min-max method	• Backtracking is followed • Each iteration involves the process of minimizing the defined function involving the maximal parameters thereby yielding the final effective solution	
• Weighted product method	• Each parameter is multiplied by weight ratio depending on the priority	
• Goal programming method	• Degree of attainment of the goals has determined with the existing resources • Goals are connected with priority with priority factor	
• Bounded objective method	• Optima of respective objective function existing is made to coincide by strong duality condition	
• Posteriori method (Genetic Algorithm)	• Evaluation, selection, cross over, mutation are involved in the iteration process	• Used in cases where user has no preplans on the goals and constraints • User decides it later based on the multiple optimal solutions generated.
• Normal boundary intersection method	• Multi-objective optimization is compressed in to beta problem and then resolved as weighted single optimization problem	
• Normal constraint method	• Pareto filter is used for finding the best optimal points considering the mentioned constraints	
• Multi-objective particle swarm optimization	• Parameters are the particles and position and velocity connected with them are changed in each iteration to reach the desired solution	
• Zionts-Wallenius method	• Interactive method where the iteration proceeds with the choice of feasible solution or with the option of continuing the iteration	• Used in the cases where user specifies no conditions in the start of the iteration and interact as the algorithm executes and finds effective solution later
• Satisfying trade-off method	• Min-Max approach is reduced by a local approach using simulated annealing method	
• NIMBUS method	• Aspiration levels, upper bounds and weighting co-efficient are evaluated and new alternatives are found in each iteration • Most desired one is selected for the following iteration until the user is satisfied	

This paper is organized as follows. Section 2 introduces the resource optimization methods which contribute to the maximum profit in CSP. Proper resource utilization is the correct way of resource scheduling taking into account the varying cloud requests. Section 3 explains the cloud cost maximization techniques for profit. It includes the pricing models, power and energy of cloud resources. Load balancing on the CSP side according to the arriving cloud requests and their observation on running cloud requests is elaborated in Section 4. In Section 5, we conclude this paper.

2. Resource Optimization Methods

Resource management and allotment plays an important role in achieving profit for CSP. It is the process where CSP distributes the existing resources to the requested cloud users over the internet. The method deals with the effective resource utilization at the provider side and prevention of starvation at the user side [10]. Resource optimization solves the problem encountered during allocation. This optimization makes the resource allocation process easier by pre-defined threshold values for the user requirements before allocation. This makes certain that the resource at the CSP side remains secured. The cloud request contains the amount of resource required and their service required period to the CSP. The CSP has to run their resource optimization algorithm to determine the sequence of serving the cloud requests so that both starvation and under-utilization are not encountered [11]. Resource optimization should prevent certain abnormalities in resource allocation and are listed below:

- Resource conflict arises when two or more users try to access the same resource at the same time exceeding its threshold
- Insufficiency of resources may occur when there are inadequate resources and also when resources not allocated in proper manner.
- Resource splinter situation evolves when the resources are separate in remote servers and not able to allocate together immediately at the time to the cloud request.
- Over-provisioning is the process of user being allocated resources in excess to the requested one. This in turn also called over-utilization of resources at the CSP point of view.
- Under-provisioning is the scenario of user receiving resources in smaller amount than his demand. It is called the under-utilization of resources at the CSP side.

Nguyen et al. [9] presented Parallel Deadlock Detection Algorithm (PDDA) using Resource Allocation Graph (RAG) for IAAS heterogeneity cloud. It prevented resource starvation using RAG. The time and space complexity of the algorithm was found to be good for the cloud scenario. It was implemented in cloud sim and outperformed the optimal time algorithm. Yu et al. [4] suggested coalition based game model for profit attainment involving best resource utilization technique at the provider side. The profit and utility factor was calculated to be good than the normal schemes on checking it with normal schemes data. Shi et al. [12] suggested for VM auction algorithm based dynamic resource provisioning at the provider side. It provided enhanced profit on evaluation using Google-cluster data for different cloud users, data centers and no: of iterations. The trace driven simulation was run for 3 data centers and 6 types ofVM.Cao and Zegura [13] suggested utility based switch algorithm for efficient utilization of bandwidth in attaining profit. It resulted in enhanced fairness index and quality of service when checked for different bandwidth injava code. The algorithm surpassed the standard bandwidth allocation schemes and yielded better results but it mainly works on the utility function of the arriving functions for its processing.

Portaluri et al. [1] presented genetic algorithm for server scheduling resulting in less power consumption of servers. The algorithm was run in Intel 17 with 8GB RAM, Ubuntu OS in IJ Metal frame work. The results showed power efficiency with the arriving tasks and completion time. It can be extended to consider the internal communication cost, electricity cost, data center load cost etc., for better efficiency. The cloud request differs in each type of cloud service. Some request resources in duration for rent. Some users rent resources in storage. In static resource allotment method, the resources are fixed and rented to the users on cost. In dynamic scheduling, the resources are allocated on demand and they are priced based on usage of each resource per hour [14]. The resource optimization strategy must assure a deadlock free environment with better speed, bandwidth, throughput, response time, completion time, profit, VM provisioning and less power consumption. Thus managing and allocating resources in cloud is crucial.

3. Cost Optimization Methods

CSP put a lot of investment depending on their cloud service type followed. They aim at serving the users demand at minimal cost and expect profit in return. Profit of CSP can also be achieved by minimizing cost. This can be implemented by using proper pricing models, scheduling based on geographical location of servers, virtual machine migration method saving energy [15]. The pricing models have to done in an efficient manner to satisfy the user demands. The pricing methodology changes for each type of CSP [16]. The static cost provisioning makes the cost fixed for users. Dynamic pricing methodology changes the cost at various hours based on user demand. The scheduling of resources between users plays a vital part in achieving profit. The instances of resources to be allocated for better profit have to be pre-determined. The server selection has to be done considering their geographical location so that speed, transfer rate and thereby throughput can be maximized. Energy saving comes as the next major factor in achieving profit [17,18]. The cost of electricity, working and cooling power for servers has to be reduced for saving energy [19]. Different cooling and power saving techniques are being adopted to achieve it. The brief overview of the above discussed parameters in cost optimization methodologies is detailed in Table 2.

Table 2: Cost optimization methodologies adopted in the existing literature.

Study	Issues Addressed	Algorithm and Dataset Used	Performance Metric	Inference
[3]	Energy reduction Maximizing profit	Profit driven online resource allocation framework Dataset: Google data traces	Profit Average CPU utilization Cluster energy	Evaluated results in Google traces for heuristic max, min and random values Only CPU utilization is concentrated Other resources like memory, bandwidth are also to be incorporated
[5]	Cost efficiency	Co-Efficient and Reliable Resource Allocation algorithm (CERR) Dataset: Amazon EC2 Instances [22]	Cost Reliability CERR rate	Outperforms the Max-Min algorithm, MIN_MIN algorithm and FCFS algorithm

[6]	Task scheduling Resource allocation Profit	Multi-objective optimization Ant colony optimization Resource cost model Make span and budget cost as constraints Dataset: 100 hosts, 10 Virtual Machines Compared with Original colony algorithm Heuristic algorithm Min-Min algorithm FCFS algorithm	Make span Cost Dead violation rate Resource utilization	Outperformed the Min-Min algorithm even at worst case Better effectiveness than FCFS algorithm
[18]	Energy consumption Congestion or hot spots	Data center Energy Efficient Network Scheduling Algorithm (DENS) Dynamic Voltage and Frequency Scaling (DVFS) Dataset: Three Tier Data center topology 1536 servers 32 racks 48 servers//rack 1 GE internal link 10 GE topology link Propagation Delay 10ns	Power consumption Uplink traffic load	Outperforms the round robin scheduler and green scheduler in reduced power consumption Suitable for three tier architecture and should be checked for other data center architectures
[20]	Electricity cost involved in cloud data centers	Energy efficient algorithm Dataset: Internet traffic archive, CLARKNET-HTTP, NASA –HTTP, UC BERKELEY IP	Normalized cost Maintenance cost Electricity cost	Outperforms the traditional cooling technique.
[21]	Cost optimization	Reserved instances optimizer with hill climbing algorithm based profit function Compared with the theoretical values of the inventory model Dataset: Industrial data	Demand trace and profit function	Outperforms the heuristic methods, machine learning techniques Risk analysis has to be checked

The parameters of the cost optimization can be prioritized based on its significance in achieving profit. Each parameter contributes in achieving profit in its own way depending on the cloud service [23]. The cloud service provider concentrates on optimizing the parameter based on their network topology and architecture. The priority factors considered in cost optimizationmethod is depicted in Figure 1.

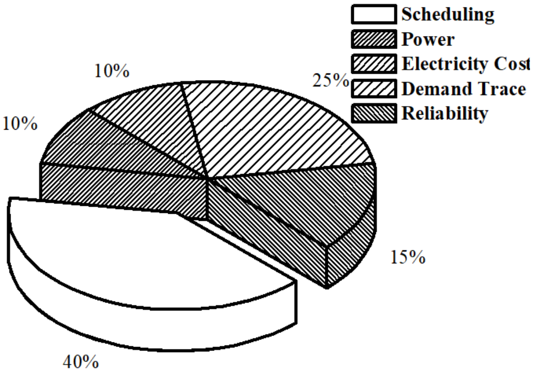

Fig. 1: Priority factors in cost optimization approach.

4. Load Optimization Methods

Load balancing has to be concentrated by CSP for its efficient run. If not done, it leads to resource over utilization and deadlock. In cloud, the arriving requests are enormous and vary from time to time. Hence some load balancing technique has to be adopted. Virtualization solves these issues and becomes the powerhouse of cloud computing [24,25]. It is the process of separating resources, balancing load and system maintenance in order to serve big data applications. Virtual machine migration is the hidden method of moving OS instances from the existing physical server to the needed server. It is done to balance the incoming enormous load on cloud.

In virtual machine migration scheme, resources are consumed on both host and guest side causing a migration overhead [26]. The distance between them has also to be considered. By making effective use of virtual machine migration, profit and optimization gets achieved in cloud, while handling large data [27]. Even in small load cases, the resources have to be balanced and monitored properly to avoid deadlock and hacking. Many load balancing methods have been developed based on the CSP. Yu et al. [28] discussed about stochastic load balancing with hot spot virtual machine migration.The stochastic algorithm outperforms other virtual migration schemes in three tier architecture. The algorithm had run well on CloudSim with the VM utilization trace from planed lab and Google trace data. But the migration schemes have to be checked for all kind of probability distribution and network topology in cloud.

Jaiganesh and Vincent Antony Kumar [29] designed a fuzzy logic model for Data Center Load Efficiency (DCLE) Monitoring. It was a model for monitoring CPU, bandwidth, and memory. It was implemented for physical machine with limited users. The DCLE meter has to be validated for different virtual machine scenario handling big data in cloud. The CSP may serve software applications, operating systems, servers etc. Load balancing also becomes an important issue in security point of view. The hackers try to utilize maximum of a particular resource to hack the system. Hence resource utilization has to be examined to avoid complete systemcrack. CSP must concentrate on this factor for efficient running of their service.

The profit can be achieved by using multi-objective optimization of resource, cost and load using different parameters for each factor. It can be inferred that the cooperative optimization of all the three factors will result in better throughput. Based on the type of the cloud service, the number of users and the network topology each factor and its parameters play its role in providing profit.

5. Conclusion

The optimization technique is the most thriving one in cloud optimization. Many types of optimization models have been developed so far. With respect to the cloud environment, resource, cost and load based optimization models have been developed. The resource optimization helps the cloud provider in effective utilization of resources by considering the type of scheduling and the nature of service. The resources can be protected before allocation by putting certain constraints to avoid over and under provisioning and deadlock of resources and requests in case of heterogeneous databases. This leads to the better profit in providers. Cost can be optimized by introducing pricing models and prior estimation of resource allocation factor leading to profit. The load at the provider side needs to be balanced by using virtual migration techniques reducing the down time and the migration time considering the network topology. This way of effective utilization of incoming load leads to the balanced cloud system avoiding hackers and earning high profit. Thus optimization techniques can be evaluated by various factors like profit, resource utilization factor, traffic load speed, transfer rate, down time, deadlock violation rate etc.

References

[1] G. Portaluri, S. Giordano, D. Kliazovich, B. Dorronsoro, A power efficient genetic algorithm for resource allocation in cloud computing data centers, In Proceedings of the IEEE 3rd International Conference on Cloud Networking (CloudNet), 2014, pp. 58-63.

[2] H. Goudarzi, M. Pedram, Maximizing profit in cloud computing system via resource allocation, In Proceedings of the International Conference on Distributed Computing Systems Workshops, 2011, pp. 1-6.

[3] M. Dabbagh, B. Hamdaoui, M. Guizani, A. Rayes, Exploiting task elasticity and price heterogeneity for maximizing cloud computing profits, IEEE Transactions on Emerging Topics in Computing, Vol. 6, No. 1, 2018, 85-96.

[4] R. Yu, J. Ding, S. Maharjan, S. Gjessing, Y. Zhang, D. Tsang, Decentralized and optimal resource cooperation in geo-distributed mobile cloud computing, IEEE Transactions on Emerging Topics in Computing, Vol. 6, No. 1, 2015, pp. 72-84.

[5] H. Chen, F. Wang, N. Helian, A cost-efficient and reliable resource allocation model based on cellular automaton entropy for cloud project scheduling, International Journal of Advanced Computer Science and Applications, Vol. 4, No. 4, 2013, pp. 7-14.

[6] L. Zuo, L. Shu, S. Dong, C. Zhu, T. Hara, A multi-objective optimization scheduling method based on the ant colony algorithm in cloud computing, IEEE Access, Vol. 3, 2015. pp. 2687-2699.

[7] G. Zhang, J. Lu, Y. Gao, Multi-level Decision Making: Models, Methods and Applications, Springer Publishing Company, 2015, ISBN 978-3-662-46058-0.

[8] M.A. Vouk, Cloud computing: Issues, research and implementations, In Proceedings of the 30th International Conference on Information Technology Interfaces, 2008, pp. 31-40.

[9] H.H.C. Nguyen, H.V. Dang, N.M.N. Pham, V.S. Le, T.T. Nguyen, Deadlock detection for resource allocation in heterogeneous distributed platforms. In: Unger H., Meesad P., Boonkrong S. (Eds.) Recent Advances in Information and Communication Technology, Advances in Intelligent Systems and Computing, Vol. 361. Springer, Cham, 2015.

[10] L. Wu, S.K. Garg, R. Buyya, SLA-based resource allocation for a software as a service provider in cloud computing environments, In Proceedings of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Los Angeles, USA, 2011, pp. 195-204.

[11] R.N. Calheiros, R. Ranjan, A. Beloglazov, C.A.F.D. Rose, and R. Buyya, CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms, Software – Practice and Experience, Vol. 41, 2011, pp. 23-50.

[12] W. Shi, L. Zhang, C. Wu, Z. Li, F.C.M. Lau, An online auction framework for dynamic resource provisioning in cloud computing, IEEE/ACM Transactions on Networking, Vol. 24, No. 4, 2016, pp. 2060-2073.

[13] Z. Cao, E.W. Zegura, Utility max-min: an application-oriented bandwidth allocation scheme, In Proceeding of the IEEE INFOCOM International Conference on Computer Communications, Vol. 2, 1999, pp. 793-801.

[14] Y.O. Yazir, C. Matthews, R. Farahbod, S. Neville, A. Guitouni, S. Ganti, Y. Coady, Dynamic resource allocation in computing clouds using distributed multiple criteria decision analysis, In Proceedings of the IEEE 3rd International Conference on Cloud Computing, 2010, pp. 91-98.

[15] W. Voorsluys, J. Broberg, S. Venugopal, R. Buyya, Cost of virtual machine live migration in clouds: A performance evaluation, In: Jaatun M., Zhao G., and C. Rong C. (Eds.), Cloud Computing, Lecture Notes in Computer Science, Springer Berlin Heidelberg, Vol. 5931, 2009, pp. 254-265.

[16] Greenberg, J. Hamilton, D. Maltz, and P. Patel, The cost of a cloud: research problems in data center networks, ACM SIGCOMM Computer Communication Review, Vol. 39, No. 1, 2009, pp. 68-73.

[17] A. Beloglazov, R. Buyya, Energy efficient resource management in virtualized cloud data centers, In Proceedings of the IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010, pp.826-831.

[18] D. Kliazovich, P. Bouvry, S. Khan, Dens: Data center energy-efficient network-aware scheduling, In Proceedings of the Green Computing and Communications (GreenCom), IEEE/ACM International Conference on Cyber, Physical and Social Computing (CPSCom), 2010, pp. 69-75.

[19] K.H. Kim, A. Beloglazov, R. Buyya, Power-aware provisioning of cloud resources for real-time services, In Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science, Urbana Champaign, Illinois, 2009, pp. 1-6.

[20] S. Chen, S. Irving, L. Peng, Operational cost optimization for cloud computing data centers using renewable energy, IEEE Systems Journal, Vol. 10, No. 4, 2016, pp. 1447-1458.

[21] A. Nodari, Cost Optimization in Cloud Computing, Master’s Thesis, Aalto University, 2015.

[22] Amazon EC2, http://aws.amazon.com/ec2/ [Accessed July 09, 2018].

[23] M. Andreolini, S. Casolari, M. Colajanni, and M. Messori, Dynamic load management of virtual machines in cloud architectures, In: Avresky D.R., Diaz M., Bode A., Ciciani B., Dekel E. (Eds.) Cloud Computing, Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Springer, Berlin, Heidelberg, Vol. 34, 2009.

[24] D. Bruneo, S. Distefeno, Quantitative Assessment on Distributed Systems: Methods and Techniques, Wiley Scrivener Publishing, 2015, ISBN: 978-1-118-59521-3.

[25] S. Kaur, V. Pandey, A survey of virtual machine migration techniques in cloud computing, Computer Engineering and Intelligent Systems, Vol. 6, No. 7, 2015, pp. 28-34.

[26] P. Kaur, A. Rani, Virtual machine migration in cloud computing, International Journal of Grid Distribution Computing, Vol. 8, No. 5, 2015, pp. 337-342.

[27] K. Sato, H. Sato, S. Matsuoka, A model-based algorithm for optimizing I/O intensive applications in clouds using VM-based migration, In Proceedings of the 9th IEEE/ACM Conference on Cluster Computing and the Grid, 2014, pp. 466-471.

[28] L. Yu, L. Chen, Z. Cai, H. Shen, Y. Liang, Y. Pan, Stochastic load balancing for virtual resource management in datacenters, IEEE Transactions on Cloud Computing, Vol. PP, No. 99, 2016, pp. 1-1.

[29] M. Jaiganesh, A. Vincent Antony Kumar, B3: Fuzzy-based data center load optimization in cloud computing, Mathematical Problems in Engineering, Vol. 2013, Article ID 612182, pp. 1-11.