IJCSN - International Journal of Computer Science and Network, Volume 7, Issue 2, April 2018
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

A Survey Based on Dynamic Resource Allocation Techniques in Cloud Computing

[1] Mukta Chaturvedi; [2] Dr. Hemant Kumar Garg

[1] Research Scholar
Career Point university, Kota

[2] Lecturer (Selection Grade), Dept. of Computer Engg.,
Govt. Women Polytechnic College,
Gandhi Nagar, Jaipur, India.

Abstract- Cloud computing is becoming one of the most escalating technologies that have got huge potentials in enterprises and markets. Clouds can makes possible to access any applications and associated data from anywhere. Companies are able to rent resources from cloud for storage and many other computational purposes so that their infrastructure cost can be reduced significantly. On the other hand it can make use of company wide access to applications, based on pay-as-you-go model. Hence there is no need to take licenses for individual products. However one of the major pitfalls in cloud computing is to optimizing the resources. Because of the uniqueness of the model, resource allocation is performed to minimizing the costs. The other challenges of resource allocation are to fulfill the customer demands as requirements. In this paper, various resource allocation strategies and their challenges are discussed in detail. It is supposed that this paper will be beneficiary both cloud users and researchers in overcoming the challenges faced.

Keywords- *Cloud Computing; Cloud Services; Resource Allocation; Infrastructure.*

1. Introduction

Cloud computing permit customers to scale up and down their resources based on requirements. Cloud computing tools make the resources as a single point of access to the client and cost is pay per practice. Cloud computing is a computing technology where a collection of resources are connected in private and public networks and to provide these dynamically scalable infrastructure for purpose. Cloud computing is a service oriented and it is not an application oriented. It provides the virtualized resources to the cloud users. Cloud computing provide dynamic provisioning and allocate machines to store data and add or remove the machines according to the demands. Cloud computing platforms such as Microsoft, Amazon, Google, IBM. Cloud computing can be used for sharing resources without the information of the infrastructure and makes it possible to access the applications and its related data from anywhere at any time [1].

2. Resource allocation

In cloud computing, resource allocation is the process of allocating resources to the desirable cloud applications. Cloud resources can be provisioned on demand in a fine-grained and multiplexed approach. In cloud the resource allocation is depends on the infrastructure as a service (IaaS) [2].In cloud platforms, resource allocation takes place at two stages:

- The load balancer assigns the requested instances to physical computers as application is uploaded to the cloud to balance the computational load of multiple applications across physical computers
- When an application accepts various requests, these requests should be assigned to a specific application request to balance the computational load.

Resource allocation techniques should satisfy the following condition:

- Resource contention arises when more than one application try to access the similar resource at the same time.
- Resource fragmentation arises while the resources are inaccessible. Due to fragmentation some resources cannot be allocated to the needed application while they are available.
- Insufficiency of resources arises when there are less resources and the demand for resources is high.
- The multiple applications needed different types of resources such as CPU, memory, I/O devices and the technique should satisfy that request
- Over provisioning of resources arises when the application gets surplus resources than the demanded one

3. Resource allocation in cloud computing

In cloud computing, resource allocation (RA) is a field that is used in many computing areas such as data center management, operating systems, and grid computing. RA deals with the sharing of available resources between cloud users and applications in an economic and effective way. It is one of the difficult tasks in cloud computing based on the IaaS. Furthermore, RA for IaaS in cloud computing is cost effective because users do not need to install and update hardware or software to access the applications, its flexibility allows access applications and data on any system in the world, and there are no limits of the medium or usage site [3].

In addition, there are two major processes of RA via cloud computing

3.1Static Allocation

Static Allocation schemes assign fixed resources to the cloud user or application. In this case, the cloud user should know the number of resource instances needed for the application and what resources are requested and should aim to confirm the application's peak load requests. But the limitation for static allocation is usually affected by the over-utilization or under-utilization of computing resources based on the normal workload of the application. This is not cost-effective and is related to insufficient use of the resource during off-peak periods.

3.2Dynamic Allocation

Dynamic Allocation schemes provide cloud resources on the fly when the cloud user or application is requested, specifically to avoid over-utilization and under-utilization of resources. A possible drawback when needed resources are requested on the fly is that they might not be accessible. Thus, the service supplier must allocate resources from different participating cloud data centers.

Resource allocation strategy (RAS) is related to combining cloud provider functions for utilizing and assigning scarce resources within the boundaries of the cloud system in order to suit the demand of the cloud application [4].

As cloud computing has its characteristics, the RAS should avoid the following situations as much as possible:

1) ***Resource contention:*** This situation occurs when several users and applications attempt to allocate the same resource at the same time.
2) ***Resource fragmentation:*** This occurs when applications is not able to assign resources due to less isolated resources.
3) ***Scarcity:*** This occurs when requirements for the resources are large and there are fewer resources viz requests for memory, I/O devices, CPUs, and the techniques serve the demand**.**
4) ***Over provisioning:*** This occurs when the users and applications attain more resources than those that are demanded to fit the quality of service (QoS) requirements**.**
5) ***Under provisioning:*** This occurs when the users and applications attain fewer resources than those demanded to fit the QoS requirements**.**

From the perspective of cloud users, RA should be achieved at a lesser cost and in as soon as possible. However, it is impractical to forecast the lively of user demands, nature of users, and application demands. Therefore, resource diversity, limited resources, locality restrictions, dynamic nature of resource requests, and environmental requirements necessitate an efficient and dynamic RAS that is suitable for cloud environments. Since the dynamic and uncertainty of resource demand and supply are unpredictable, different strategies for dynamic resource allocation are suggested [5].

4. Research issues in dynamic resource allocation techniques

In this paper we have investigated some of the techniques in cloud environment.

4.1Dynamic Optimization of Multi-Attribute Resource Allocation in Self-Organizing Clouds

The existing system creates the extra messages for a single request. The proposed system using the SOC achieves the maximized resource consumption and it also delivers best execution efficacy.

4.2SOC

SOC connect a large number of desktop computers on the internet by P2P network. Each participating computer behaves as a resource provider and resource consumer [6].

- SOC having four main problems
  - Locating a qualified node to satisfy a user task's resource demand with surrounded wait.
  - To optimize a task's execution time by determining the optimal shares of the multi-attribute resources to allocate to the tasks with various QoS constraints, such as the expected

- Execution time Algorithm:
- Dynamic optimal proportional share Multi range query protocol

4.3DOPS

This scheme is used to redistribute available resources among running tasks vigorously.

- **Slice handler**: It is activated to equally scale the amount of resources.
- **Event handler**: It is used for resource redistribution on the events of task arrival and completion.

Multi Range Query Protocol: This algorithm used to locate qualified nodes in the SOC environment; we design a fully-decentralized range query protocol, namely pointer-gossiping CAN (PG-CAN), DOPS to find the qualified resources with minimized contention among requesters based on task's demand. It is unique in that for each task, there is only one query message propagated in the network during the entire discovery [7].

Range query protocol proactively diffuses resource indexes over the network and randomly route query messages among nodes to locate qualified ones that satisfy tasks' minimal demands. To avoid possibly uneven load distribution and abrupt resource over-utilization caused by un-coordinated node selection process from autonomous Participants.

4.4Virtualization technology

Cloud computing is based on the virtualization technology. Virtualization technology is used to allocate the data center resources dynamically based on the application demands.

Virtualization having two types,

- Para virtualization
- Full virtualization

4.5Live migration

Virtual machine live migration technology makes it possible to mapping between the virtual machines (VMs) and the physical machines (PMs) while applications are running. Live migration increase the resource utilization and provide the better performance result.

a. Cloud environment provide the four types of cloud.
   - Public cloud
   - Private cloud
   - Hybrid cloud
   - Community cloud

b. Cloud computing offers three types of services
   - Software as a service(Saas)
   - Platform as a service(Paas)
   - Infrastructure as a service(IaaS)

5. Models for dynamic RA in cloud computing

The quality and cost of the services in cloud computing are depends on their RA process, and in the in which resource provider assign the resource to the clients. There are various RA techniques and proposed models that are used in the area of cloud computing. Here we present some of the dynamic RA techniques, classifying them with the strategy that they use to allocate resources. The result of any optimal RAS must consider some parameters such as latency, throughput, and response time. In this paper, we present some of the commonly used strategies such as service level agreement-based, utility-based, market-based, and priority-based strategies.

5.1SLA-Based Dynamic RA Models

The SLA is an agreement that specifies the QoS between the service provider and the service consumer, and it includes the service price with the level of QoS adjusted by the price of the service. Most of the RA models in cloud computing environments focus on satisfying the agreed specifications of the SLA for the cloud user. Some other models' strategies in RA focus on achieving the objectives of the cloud provider, which could negatively affect some of the users' requirements and the level of QoS provided. One such model is proposed by Popovici They investigated the QoS parameters such as offered load and price on the SaaS provider's side but did not consider the user's side.

For a multi-cloud environment, Soodeh Farokhi developed a framework for resource allocation in a multi-cloud system from the perspective of the SaaS level, agreed SLA, and service provider conditions. The proposed model utilizes a selection engine, construction engine, and SLA violation detection and monitoring with the use of the service provider's QoS parameters.

There are few models that focus on both the cloud provider and consumer perspectives. One such model was proposed by Wu it focuses on the QoS parameters of both the SaaS provider and the consumer through proposed RA algorithms aimed at minimizing SLA violations and infrastructure costs, as well as controlling the dynamic change of customers, by specifying customer demands to the infrastructure level aspect and managing dissimilarity of virtual machines (VMs). The two proposed algorithms perform well by decreasing costs by about 50% with fewer VMs and optimizing the means to avoid SLA violations.

Also, another work proposed by Lee addresses the issue of profit basis on service request scheduling in cloud computing by taking into account the purposes of both the consumers and service providers.

Zhu et al. [8] proposed architecture to solve virtualized RA problems for multi-tier applications. Their model improved overall performance, reduced the cost, maximized the profit for SaaS providers, and aimed to meet the user's performance requirements.

EARA [9] is an efficient agent-based resource allocation framework designed by Kumar et al. EARA uses agent computing as several agents collect available resource information to allocate it for user requests based on the signed SLA agreement and, therefore, balancing the performance and controlling the cost; however, this model only considers the SaaS level of cloud environments.

Buyya proposed a market-oriented RA scheme by integrating both the customer-driven service management and provider-driven risk management to promote SLA-based RA. Nevertheless, it requires a market maker and a market registry to bring the consumer and providers together and to publish the cloud services and discover their providers.

Pawar proposed a priority-based allocation model by considering various SLA parameters such as bandwidth, memory, and execution time. Using a preemption mechanism associated with the benefit of parallel processing, their model improved utilization, especially in a resource contention situation.

5.2Market-Based Dynamic RA Models

Using the market economy to manage RA has been studied extensively in the past, and several researchers have investigated the economic aspects of cloud computing from different points of view. To deal with dynamically fluctuating resource demands, market-driven RA has been proposed, and it has been implemented by many public IaaS providers such as Amazon In this environment, the cloud provider could follow the commodity market approach or auction-based mechanisms, with the main goal of achieving maximum revenue while minimizing the cost.

Zama proposed a combinatorial auction-based mechanism for resource management in clouds. Their algorithm is based on the users' valuation concept, which is that each user desires a specific bundle of VM instances and bids on it. It represents efficient allocation and high profits for the provider, but it still allows users to pay a minimum cost.

In addition, Zhang introduced a mechanism for spot markets, which addressed the problem of allocating resources for different VM types in Amazon spot instances using the model predictive control (MPC) algorithm. The proposed model insures high revenue for the providers over time by changing the price depending on the level of demand, meeting customer expectations, and minimizing energy consumption; however, future or forward markets are not included in the model.

From there, Fujiwara improved the market-based allocation by developing a double-sided combinatorial auction-based model, which allows both the users and providers to trade their current and future services in the spot and forward market.

There are other market-based models, such the RAS-M model, that define the equilibrium theory and use the GA-based price adjusted algorithm. This model is efficient and improves utilization and profit; nevertheless, it only considers the physical level of the cloud environment and is limited to CPU resources only.

Lin proposed an RA model for clouds based on a sealed-bid auction, where the users submit their bids to the cloud service providers who collect the bids and determine the price. This mechanism provides efficient allocation of resources, but no profit maximization is ensured due to its truth telling property [9].

5.3Utilization-Based RA Models

In order to overcome the under-utilization of resources that results from allocating fixed resources to applications and services, the main approach of methods that fall into this category is to dynamically manage VMs to maximize utilization of resources and minimize costs. A model that adjusts the VMs according to an application's actual needs has been developed by which is based on the threshold. The proposed algorithm uses monitoring and predicting the needs of cloud applications, which leads to increased resource utilization and decreased costs.

Yin focus on RA at the application level. The authors proposed a multi-dimensional RA (MDRA) schema using a framework of application allocation to minimize the cost of the data center by assigning small-sized nodes to the processors of users' programs.

Simulated annealing-based RA has been performed by Pandit using a bin packing algorithm with multi-parameters to decrease the unallocated part of resource parameters. The proposed model has improved utilization of cloud resources at the multi-level in the cloud system and has decreased the cost.

The topology aware resource allocation (TARA) schema was introduced by Lee that deals with unconcerned of the hosted application's demands for IaaS system. They proposed a prediction engine and genetic algorithm-based search for minimum latency and proper confidence. The authors showed that the TARA experiment could result in a decrease in job compilation time by up to 59%.

Li and Qiu suggested an adaptive RA algorithm for cloud a computing model with preempt able jobs. The authors defined two algorithms, adaptive list scheduling (ALS) and adaptive min-min scheduling (AMMS), which apply to task scheduling [10], and they proved that the proposed algorithm is effective and efficient for use with resources. Younge proposed a resource management model to improve job scheduling using a green cloud framework. Their model maximizes utilization as it reduces performance overload and energy consumption and provides an overall efficiency for data-centers in the cloud computing environment. An online optimization for scheduling preempt able tasks on IaaS cloud system models was proposed by J. Li using a min-min algorithm, and the scheduling process is based on feedback information about actual task execution. Their solution proved to reduce RA execution time and energy consumption; however, the strength of their model is based on the reliability of the feedback information.

Rammohan and Baburaj developed an RA in a cloud computing model based on the interference-aware resource allocation (IARA) technique, providing optimal energy consumption, and it is practical for a resource-constrained environment and supports special hardware.

There are a number of other models. Some of these models mainly focus on improving resource utilization, such as the ones by Minarolli., Buyya and Ergu.

6. Research challenges

The research on RA in cloud systems is still at an early stage. Several existing issues have not been fully addressed while new challenges keep emerging. Some of the challenging research issues are given as follows:

1) ***Migration of VM:*** This migration problem occurs due to the need of the user to switch to another provider in order to get better data storage.
2) ***Control:*** There often is a lack of control mechanism over the resources as they are rented by the users from the remote server.
3) ***Energy Efficiency:*** Due to the emergence of huge data centers that have various computing operations, there is a need for energy efficient allocation. These centers lead to the release of large quantities of carbon emission.
4) ***The Scheduling of Parallel Jobs:*** Parallel jobs in the field of computing increase the job that is serving. There are two types of jobs: dependent and independent. The first type must be done very carefully. These jobs include communication issues. Independent jobs can be performed using several VMs at the same time.
5) ***Reduction of Cost and Maximizing of Resources:*** It is important to handle the constraints that must be met in the allocation of resources in terms of cloud operating costs and to maximize the use of all resources. In other words, the service provider must provide users with low-cost services.
6) ***Maintaining High Availability:*** The availability of resources in the cloud must be guaranteed in case there is a job with long running computations that can take many hours. Thus, there is a need for some techniques to automatically handle any interruption or unavailability in resources and switch the jobs to an available resource. Moreover, these techniques should support the transparency property by which the user cannot observe the unavailability or any failure problem.
7) ***Elasticity:*** In the cloud, elasticity refers to what extent resource requirements can be handled dynamically. Demand for resources may increase over time, and the cloud should automatically detect the size of these demands to be met and the necessary resources required to meet them.

7. Findings and future research directions

A great deal of research has been done, and many solutions have been presented in the area of cloud computing in respect to the RA problem; however, there are still some issues and challenges that need further research, and an optimal solution that is practical for most cloud environments has still not been found.

Some of the findings based on our literature of previous studies are as follows:

1) There is a need for reducing the user's SLA violations when maximizing the RA utilization because most of the models affect the QoS in order to reduce the cost and keep high utility algorithms.
2) There is a need for an RA framework that is practical for different cloud environments in order to ease the complexity of allocation in heterogeneous clouds.
3) There is a need for RA to minimize the cost for cloud consumers and maximize the profit for cloud providers. It is very important for cloud providers to offer efficient utilization and management of the limited amount of resources available.
4) There is a need to consider load balance in cloud resources and scheduling the workload in optimal ways in order to satisfy the QoS requirements of users and maximize profit by enhancing the use of resources.

The future direction of research into resource allocation in cloud computing should address each of the above-named challenges and try to implement best practices models.

Table shows the comparative study between various resource allocation techniques in cloud environment and their advantage and parameter results.

Table 1: Shows comparison of the dynamic resource allocation technique

TITLE	ADVANTAGE	PARAMETER RESULT
Dynamic Optimization of Multi-Attribute Resource Allocation In Self organizing Cloud.	Locating qualify nodes and optimize task execution time.	Throughput Ratio: 60% increased.
Priority Based Resource Allocation Model for Cloud Computing.	Resource wastage is minimized.	Parameters: No. of users, Time to run, No. of processor, job type, User type.
Dynamic Resource Allocation Using Virtual Machine for Cloud computing Environment.	Server overload is Decreased.	Migration of VM for resource Requirement.
Survey on Resource Allocation Strategies In Cloud Computing(2013).	It maintain the SLA and also manage the QoS.	Strategies used: Virtual machine, SLA, utility.
Heterogeneity Aware Resource Allocation In Cloud.	Provide the fairness among Jobs when multiple jobs are submitted.	The result is based On the Instance Type Ex:m1.small.
Dynamic Resource Allocation for Parallel Data Processing in cloud.	Overload is avoided.	Gain utility: Preemptive>Non-Preemptive Penalty: Non Preemptive> Preemptive.
Efficient Idle Desktop Consolidation with Partial VM Migration.	This migrates only working set of an idle VM.	This can deliver the 85%to 104% of the energy saving Compare full VM. migration.
Survey on Resource Allocation in Cloud Computing.	This avoids the resource Contention and scarcity of resources.	Technique: Topology Aware resource allocation.
Dynamic Resource Allocation for Spot Market in Cloud.	Total revenue is maximized.	Income:15173. 28 Loss:1083.63 NetIncome:140 89.65.
Heuristic Based Resource Allocation Using Virtual Machine Migration: A Cloud Computing Perspective.	Less SLA Violation and less performance degradation.	Average: SLA violation is reduced to 17.64 to 16.44.

8. Conclusion

This paper addresses the theoretic study of various dynamic resource allocation techniques in cloud environment. These models are classified based on their strategies, and a discussion of their strengths and limitations is supported by a comparison table. Finally, research directions and findings from our literature review are included, and hopefully they will help in motivating future research to determine optimal RA solutions for cloud environments. The detail description of the techniques is summarized and also summarizes the advantages with parameters of the various techniques in cloud computing environment.

References

[1] Ronak Patel, Sanjay Patel" Survey on Resource Allocation Strategies in Cloud Computing" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 2, Feb- 2013.

[2] K C Gouda, Radhika T V, Akshatha M," Priority based resource allocation model for Cloud computing" International Journal of Science, Engineering and

Technology Research (IJSETR)Volume 2, Issue 1, January 2013.

[3] Gunho Leey, Byung-Gon Chunz, Randy H. Katzy," Heterogeneity-Aware Resource Allocation and Scheduling in the Cloud", IJERT-2012.

[4] Nilton Bilay, Eyal de Laray, Kaustubh Joshi_, H. Andr´es Lagar-Cavilla ,Matti Hiltunen_ and Mahadev Satyanarayananz," Efficient Idle Desktop Consolidation with Partial VM Migration", Journal of computer application-2012.

[5] Venkatesa Kumar, V. And S. Palaniswami," A Dynamic Resource Allocation Method for Parallel data processing in Cloud Computing", Journal of Computer Science 8 (5): 780-788, 2012.

[6] Sheng Di and Cho-Li Wang," Dynamic Optimization of Multi-Attribute Resource Allocation in Self-Organizing Clouds", IEEE Transactions on parallel and distributed systems, - 2013.

[7] V.Vinothina, Dr.R.Sridaran, Dr.padmavathiganapathi," A Survey on Resource Allocation Strategies in Cloud Computing "International Journal of Advanced Computer Science and Applications, Vol. 3, No.6, 2012.

[8] Qi Zhang, Eren G¨urses, Raouf Boutaba, Jin Xiao," Dynamic Resource Allocation for Spot Markets in Clouds", Journal of computer science-2012.

[9] Zhen Xiao, Senior Member, IEEE, Weijia Song, and Qi Chen," Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment", IEEE Transactions on parallel and distributed systems, vol. 24, no. 6, June 2013.

[10] Ts`epomofolo, R Suchithra," Heuristic Based Resource Allocation Using Virtual Machine Migration: A Cloud Computing Perspective", International Refereed Journal of Engineering and Science (IRJES) Volume 2, Issue 5(May 2013)

Authors –

Mrs. Mukta Chaturvedi received her M.C.A. degree from Punjab Technical University, Jalandhar and A level certificate course from DOEACC, New Delhi. She is pursuing her Ph.D. from Career Point University, Kota. Her research interest in cloud computing. She has published a research papers in International Journal.

Dr. Hemant Kumar Garg received the Bachelors degree in Computer Engineering from the University of Amravti of Amravati, India in 1991, and the M.Tech. Degree in Computer Science from the BIT, Ranchi, India in 2002, and Ph.D. degree in Computer Engineering from JNU, Jaipur, India. In 1994, he joined the Computer Engineering Department, Department of Technical Education, Govt. of Rajasthan, Rajasthan State, INDIA, where he is currently working as a Lecturer (Selection Scale). He published a book on "Electronic Communication and Data Communication". Dr. Garg is a Council Member of Computer Division, Rajasthan Chapter, Institution of Engineers, India. He also has many publications in National and International Journals and conferences. His areas of interests include Ad Hoc networking; cloud Computing and its Applications.