Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 5, May -2015

An Extensive Analysis on Resource Allocation in Cloud Computing Environment: A Survey

Abhijitsinh Parmar[1], Rutvik Mehta[2]

Department of CSE, [1]PG Student, [2]Asst. Professor, Parul Institute of Eng.& Technology, Vadodara, Gujarat

Abstract– *Cloud computing becomes a popular internet technology in last some years. It's providing facility as a services and renting resource based model is widely accepted in enterprises and markets. One of the most pressing issues in cloud computing for IaaS is the resource management. In Cloud computing multiple cloud users can request number of cloud services simultaneously. So there must be a provision that all resources are made available to requesting user in efficient manner to satisfy their need. Because of the uniqueness of the model, resource allocation is performed with the objective of minimizing the costs associated with it. The other challenges of resource allocation are meeting customer demands and application requirements. This paper focuses on one of the important resource management technique: resource allocation and provide detail description of its various strategies to allocate resources and challenges related to it.*

Keywords- *Cloud Computing, Cloud Services, Resource Allocation, Virtual Machine, VM Allocation.*

I. INTRODUCTION

Cloud Computing is the latest term encapsulating the delivery of computing resources as a service. It is the current iteration of utility computing and returns to the model of 'renting' resources. Cloud computing has appeared as an accepted computing model for processing very large volume of data. The terms Leveraging cloud computing is today, the de facto means of deploying internet scale systems and much of the internet is tethered to a small number of cloud providers. The advancement of cloud computing is therefore intrinsic to the development of the next generation of internet. Cloud computing emerges as a base of all computing directly or indirectly. Due to its attractive advantages and popular services it becomes quite popular nowadays.

Cloud computing providers offer their services according to three fundamental models Infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) where IaaS is the most basic and each higher model abstracts from the details of the lower models. Platform -as-a-service in the Cloud is defined as a set of software and product development tools hosted on the provider's infrastructure. Developers create applications on the provider's platform over the Internet. PaaS providers may use APIs, website portals or gateway software installed on the customer's computer. In the software-as-a-service Cloud model, the vendor supplies the hardware infrastructure, the software product and interacts with the user through a front-end portal. Cloud computing nowadays becomes quite popular among a community of cloud users by offering a variety of resources. Cloud computing platforms, such as those provided by Microsoft, Amazon, Google, IBM, and Hewlett-Packard, let developers deploy applications across computers hosted by a central organization. These applications can access a large network of computing resources that are deployed and managed by a cloud computing provider.

Cloud Computing has a numerous advantages like Pay as per usage – users have to pay only whatever they consume, Zero upfront investment – No need of infrastructure establishment, No worry of Maintenance, Highly automated, Flexibility and scalability – Whenever demand is high cloud service is scalable, Mobility – can be used from anywhere by using internet.

II. RESOURCE MANAGEMENT

Resource management is one of the hot topic of cloud computing research nowadays. It includes following issues Resource provisioning, Resource allocation, Resource adaption, resource mapping, resource modelling, Resource estimation out of all this Resource allocation is most affecting issue. Basically resource allocation means distribution of resources economically among competing groups of people or programs. Resource allocation has a significant impact in cloud computing, especially in pay-per-use deployments where the number of resources are charged to application providers. The issue here is to allocate proper resources to perform the computation with minimal time and infrastructure cost. Proper

resources are to be selected for specific applications in IaaS. In Cloud Computing VM allocation also referred as problem of resource management which is part of load balancing.

In cloud platforms, resource allocation takes place at two levels. First, when an application is uploaded to the cloud, the load balancer assigns the requested instances to physical computers, attempting to balance the computational load of multiple applications across physical computers. Second, when an application receives multiple incoming requests, these requests should be each assigned to a specific application instance to balance the computational load across a set of instances of the same application. For example, Amazon EC2 uses elastic load balancing (ELB) to control how incoming requests are handled. Application designers can direct requests to instances in specific availability zones, to specific instances, or to instances demonstrating the shortest response times.

III. VM ALLOCATION

VM is considered as a main resource in cloud environment. The main objective of resource allocation is to plan sufficient capacity of all resource by spreading the load on processors and maximizing their utilization while minimizing the total task execution time.

However one of the major pitfalls in cloud computing is related to optimizing the resources being allocated. Resource allocation is performed with the objective of minimizing the costs associated with it. The other challenges in resource allocation are meeting customer demands and application requirements. In the current cloud computing environment there is numerous of application, consist of millions of module, these application serve from large quantity of users and the user request becomes dynamic. The resource provisioning was done by considering Service Level Agreements (SLA) and with the help of parallel processing using different types of scheduling heuristic.

VM allocation and task scheduling for cloud is a three folded problem which requires: (1) to decide when a VM should be allocated (2) allocating an appropriate physical machine (PM) for it - a problem related to bin packing and (3) scheduling tasks on the VM depending on various client and application given objectives. Usually the provider is controlling the second stage while the first and third are left to the client's decision. The common difficulty between all of them seems to be the balancing between efficiency goals and QoS parameters. [ ]

Most commercial clouds leave the client decide when to provision VMs. Their concern is primarily related to VM to PM assignment and use simple allocation methods based on Round Robin (Amazon EC2), least connections and weighted least connections (Rackspace2). Other simple policies include Least-Load or Rotating-Scheduling.

VM allocation can be done basically two way based on location and state of system.[9]

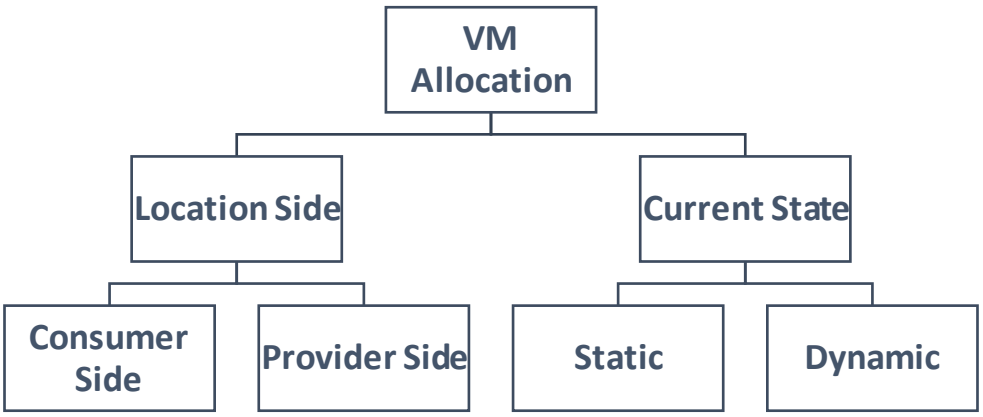

Figure 1. VM Allocation Strategies

3.1 Consumer Side VM allocation

In Consumer Side VM allocation consumer's application load (Users Requests) is distributed across VM. Here application load is generally generated due to the end users who access application and make a requests to use these applications. So here load is considered as end users requests.

Here if demand is low then some VMs are deactivated and user's requests are routes to other capable active VMs, which will reduce the VM usage and cost. Same way if demand is high then extra VM is created and remaining extra user's requests are routes to newly created VM so all these things are possible due to the use of the virtualization technology. By this way applications user's requests are assigns to VMs at run-time to maximize VM usage. If the VM allocation rule is fixed, there are parameters that could be optimized. For example, in a utilization based VM allocation rule, low utilization threshold and high utilization threshold values need to be identified.

3.2 Provider Side VM allocation

Provider Side VM allocation is generally process of allocating VMs on data center. That's why sometimes it is also known as hardware VM allocation or client VM allocation. It can also define as a process of resource management. Provider Side VM allocation can be carried out through various approaches like VM Migration, Initial Placement, VM consolidation, SLA Management, Cost Estimation, etc.

3.3 Static VM allocation

In these types of allocation load (Users Requests) is allocated to VMs in the system based on the prior knowledge about system.

The performance of the VMs is determined at the time of request arrival. Static VM allocation are non-preemptive and therefore each machine has at least one task assigned for itself. Here Non-Preemptive in the sense once request is assigns to VM then it stick to that VM it can't transfer to other VM. Its objective is to minimize the execution time and delay (Response Time) and limit communication overhead.

Static VM allocation generally do not consider dynamic changes of these attributes at run-time. Furthermore these algorithms cannot adapt to load changes during run-time. These schemes consider only the workload that is already assigned to every server, but not the remaining resource capacity server, thus may not be applicable in the case of heterogeneous systems since in such systems, some of the servers might be overloaded while others may have sufficient capacity to handle more user requests.

3.4 Dynamic VM allocation

In Dynamic VM allocation decision made for VM allocation based on dynamically changing state of the system. Which considers the different attributes of the system capabilities, network bandwidth and run-time properties collected as the selected system process the requests. These type of algorithms allows requests assign and reassign it from over utilized machine to underutilized machine dynamically. This means dynamic VM allocation is preemptive which helps in improving the overall performance of the system by migrating the load dynamically.

These algorithms require constant monitoring of the physical machines and request execution progress and are usually harder to implement. Though, they are more accurate and could result in more efficient VM allocation.

IV. SIGNIFICANCE AND QoS PARAMETERS

Managing VM Resource is a crucial task in making such an innovative technology to a larger consultation. Resource Management is done by various cloud provider for dynamic workload of client's requirement. So, VM allocation is part of resource management which is also known as load balancing. VM allocation affect to both Cloud provider and Clients so it's important concern that to take care of both Clients requirements and provider side heuristics. For Optimal VM allocation following criteria needs to be consider seriously,

a) **Resource contention:** situation arises when two applications try to access the same resource at the same time.
b) **Scarcity of resources:** arises when there are limited resources.
c) **Resource fragmentation:** situation arises when the resources are isolated. [There will be enough resources but not able to allocate to the needed application.]
d) **Over-provisioning** of resources arises when the application gets surplus resources than the demanded one.
e) **Under-provisioning** of resources occurs when the application is assigned with fewer numbers of resources than the demand.

4.1 QoS Parameters [10]

a) **Response time:** It is the minimum amount of time taken by load balancing algorithm to respond the execution of the request in system. It is one type of delay which many time considers as a latency in many algorithms.
b) **Resource Utilization:** It is the amount of resources utilized by system to serve user request. Good load balancing algorithm always have an optimal resource utilization.
c) **Performance:** It represent that how much improvement can be occur after load balancing algorithm successfully executed. If all the above parameters are satisfied in its optimal manner then performance will be improved.
d) **Throughput:** In system throughput means total numbers of tasks that completed execution in some fixed time constraint. Higher throughput means improvement of the performance.

V. LITERATURE STUDY

Jiayin Li et al. [1]propose an adaptive resource allocation algorithm for the cloud system with preempt able tasks in which algorithms adjust the resource allocation adaptively based on the updated of the actual task executions. Adaptive list scheduling (ALS) and adaptive min-min scheduling (AMMS) algorithms are used for task scheduling which includes static task scheduling, for static resource allocation, is generated offline. The online adaptive procedure is use for re-evaluating the remaining static resource allocation repeatedly with predefined frequency. In each 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.

Lin, Wang et al. [2]introduced a dynamic Virtual Machine-Varying Based resource allocation using a threshold. Using this threshold their algorithm decides that the current counts of virtual machines which are assigned to an application are sufficient or not, it is the same for over provisioning. They have defined two other parameters in threshold formulation; one is a rate called normal rate which demonstrates the average amount of workload that one individual virtual instance can tolerate without any over utilization and the other is a parameter that would be defined by system admin based on the work load; those two made the approach very parametric which seems to be a weakness.

Ray, Sarkar[3]this paper presents a load balancing scheme through the concept of resource allocation strategy and then describe the importance of resource allocation in distributed cloud environment. Here author presents the process of allocating the resources for particular job in this dynamic environment. Allocation is made on the basis of the requirement submitted by the consumers or clients. Provider stores the requirement in the repository in xml format. Then final selection of the resource is done based on the resource occupancy matrix, duration of the job and service charge and finally a service level agreement is made between cloud service provider and cloud consumer.

Pawar et al. [4] proposed priority based scheduling algorithm (PBSA) for preemptable jobs in cloud. In proposed approach it considers multiple SLA parameter such as memory, network bandwidth, and required CPU time and resource allocation by preemption mechanism for high priority task execution can improve the resource utilization in Cloud. In proposed work priority based scheduling algorithm is modified for executing highest priority task with advance reservation by preempting best-effort task. Experimental results show that in a situation where resource contention is fierce, algorithm provides better utilization of resources.

Li, Ge et al. [5]this paper proposed a comprehensive QDA modeling & scheduling algorithm for the instance intensive workflow task scheduling in cloud environment, which takes users' experiences into consideration. First, the workflow task was modeled by DAG graph. Task parameters and dependencies were determined, and user preference type value was added. Then, the QoS of cloud service resource was modeled to get QoS utility function with user preferences. Finally, combined with staggered sub-deadlines allocation criteria, cloud service resources were sorted according to the corresponding QoS utility function, and then the task scheduling was quickly completed. According to results QDA has much less execution time, better user satisfaction, and improved load balancing rate.

Zhe Gao [6]In this paper, improved ant colony optimization algorithm to compute the allocation of the cloud computing resource and to analyze the impact of bandwidth, the load of network and response time on the cloud resource proposed. In which factor of energy consumption is introduced to determine heuristic factor based on that maximum and minimum path cost is determined and according to that resources allocated.

Peng, Lin et al. [7] To address the problems in existing resource allocation schemes for cloud data centers, author modeled the resource allocation problem in a cloud data center as a constraint satisfaction problem (CSP). By solving this constraint satisfaction problem, a Choco-Based algorithm (CB) is designed to minimize the number of PMs in a virtualized cloud data center. Moreover an improved FFD (IFFD) and an improved BFD (IBFD) and conduct performance evaluation using Choco and Java is proposed. Performance studies show that the proposed algorithms are effective and outperform existing resource

allocation algorithms in virtualized cloud data centers. Network bandwidth constraints into the source assignments in order to help the VM of the data center in cloud computing to assign the source reasonably and promote the performance of the VM as well as the usage rate of the data center.

Chang, Ren et al. [8]focus on an important problem for such enterprise users is to understand how many and what kinds of virtual machines will be needed from clouds. Author formulate demand for computing power and other resources as a resource allocation problem with multiplicity, where computations that have to be performed concurrently are represented as tasks and a later task can reuse resources released by an earlier task. This paper presents an approximation algorithm with a proof of its approximation bound that can yield close to optimum solutions in polynomial time.

VI. OPEN CHALLENGES IN RESOURCE ALLOCATION

1. Users do not have control over their resources because they only rent resources from remote servers for their purpose.
2. Bring out the techniques for allocation of services to applications depending on energy efficiency and expenditure of service providers
3. Migration problem occurs, when the users wants to switch to some other provider for the better storage of their data. It's not easy to transfer huge data from one provider to the other.
4. Devise a mechanism that allows controlling the tradeoff between the costs of reconfiguration and maximizing the cloud utility
5. In public cloud, the clients' data can be susceptible to hacking or phishing attacks. Since the servers on cloud are interconnected, it is easy for malware to spread.
6. Design SLA-oriented resource allocation strategies that encompass customer-driven service management, computational risk management, and autonomic management of clouds in order to improve the system efficiency, minimize violation of SLAs, and improve profitability of service providers
7. Peripheral devices like printers or scanners might not work with cloud. Many of them require software to be installed locally. Networked peripherals have lesser problems.
8. More and deeper knowledge is required for allocating and managing resources in cloud, since all knowledge about the working of the cloud mainly depends upon the cloud service provider.
9. Move from one cloud to another cloud considering vendor lock-in issues.

VII. CONCLUSION

Cloud computing technology is increasingly being used in enterprises and business markets. Resource management is one of the most important job of cloud computing and it's mostly accomplished by resource allocation. In cloud paradigm, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. In this paper we analyzed resource allocation in detail and its various strategies are reviewed in detail. It isbelieved that this paper would benefit both cloud usersand researchers in understanding the concepts of resource allocation.

REFERENCES

[1] Jiayin Li, Meikang Qiu, Yu Chen., "Adaptive Resource Allocation for Preemptable Jobs in Cloud Systems", IEEE 10th International Conference on Intelligent Systems Design and Applications, 2010.
[2] Weiwei Lin, James Z. Wang, Chen Liang, Deyu Qi, "A Threshold-based Dynamic Resource Allocation Scheme for Cloud Computing", Procedia Engineering, Volume 23, Pages 695-703, ISSN 1877-7058, Elsevier, 2011.
[3] S. Ray and A. De Sarkar, "Resource Allocation Scheme in Cloud Infrastructure," 2013 IEEE Int. Conf. Cloud Ubiquitous Comput. Emerg. Technol., pp. 30–35, Nov. 2013.
[4] Pawar, C.S.; Wagh, R.B., "Priority Based Dynamic Resource Allocation in Cloud Computing," 2012 IEEE International Symposium on Cloud and Services Computing (ISCOS), vol., no., pp.1,6, 17-18 Dec. 2012.
[5] Huifang Li; Siyuan Ge; Lu Zhang, "A QoS-based scheduling algorithm for instance-intensive workflows in cloud environment," IEEE The 26th Chinese Control and Decision Conference (2014 CCDC), vol., no., pp.4094,4099, May 31 2014-June 2 2014.
[6] Zhe Gao, "The Allocation of Cloud Computing Resources Based on the Improved Ant Colony Algorithm," 2014 IEEE Sixth International Conference onIntelligent Human-Machine Systems and Cybernetics (IHMSC), vol.2, no., pp.334,337, 26-27 Aug. 2014.
[7] Weiwei Lin; Baoyun Peng; Chen Liang; Bo Liu, "Novel Resource Allocation Model and Algorithms for Cloud Computing," 2013 IEEE Fourth International Conference on Emerging Intelligent Data and Web Technologies (EIDWT), vol., no., pp.77,82, 9-11 Sept. 2013.

[8] F. Chang, J. Ren, and B. Labs, "Optimal Resource Allocation in Clouds," IEEE 3rd International Conference on Cloud Computing, 2010.
[9] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster. Virtual infrastructure management in private and hybrid clouds. IEEE Internet Computing, vol. 13, no. 5, pp. 14–22, 2009.
[10] M. Ajit and G. Vidya, "VM level load balancing in cloud environment," IEEE Fourth Int. Conf. Comput. Commun. Netw. Technol., pp. 1–5, Jul. 2013.
[11] M. E. Frincu and S. Genaud, "On the efficiency of several VM provisioning strategies for workflows with multi-threaded tasks on clouds," Springer, 2014.

