Survey on Resource Management Technique in Cloud Computing

M. Anitha Mary
Post-Graduate Student
Department of Computer Science and Engineering,
Karunya University
India

Mrs. Jeno Lovesum
Assistant professor
Department of Computer Science and Engineering,
Karunya University
India

Abstract

Cloud computing is an on demand technology which provides dynamic and versatile resource allocation for user demands. Remote servers and web are used by this for taking care of the data and applications. The users can access and process the applications without install in their computers. This technology support many service providers and any users can make use of the variety of service providers. It provides more compactable services to users and provides more storage space an scalability. The resource allocation is the major task in cloud computing. Resource management techniques are used take care the resource allocation process. Here in this paper describe the various resource management techniques for efficient resource allocation to the users demand. Moreover, the merits and limitations of using Resource Allocation in Cloud computing systems are also discussed.

Keywords: **Cloud computing, Resource Management, Genetic Algorithm, Scheduling.**

I. Introduction

Cloud computing is the delivery of computing services over the internet. Cloud services can allow every individuals and businesses to use software and hardware that are managed by third parties at remote locations. The characteristics of cloud computing include rapid elasticity, on-demand self service, broad network access, resource pooling, and measured services.

Cloud computing providers offer their services according to several fundamental models: infrastructure as a service, platform as a service, and software as a service. In Software as a Service model, a pre-made application, along with the required software, hardware operating system and network are provided. In Platform as a Service, an operating system, hardware and network are provided, and the customers can installs or develops its own software and applications. The Infrastructure as a service model that provides just the hardware and network; the customer installs or develops its own operating systems, applications and software.

Private cloud, public cloud, community cloud, and hybrid cloud are the various types of cloud services. Public cloud is offered over the internet and are owned and operated by cloud provider. In private cloud that the cloud infrastructure is operated solely for a specific organization. And which is managed by the organization or a third party. In a community cloud, the services can shared by several organizations and made available only to those shared groups. A hybrid cloud is not just like before mentioned methods but which is the combination of different methods of resource pooling. Virtualization is the process of running multiple operating systems on a single physical system and shares the underlying hardware resources [21].

Resource management is one of the biggest challenges in cloud computing environment. Resource allocation is the technique of providing services and storage space to the particular task is given by users. The processing units in cloud environments are called as virtual machines (VMs). It should make sure that the tasks are fulfilled their motivations about focusing on the resources. Resource allocation can be done by two methods. One of that statically allocates the resources and dynamically allocates the resources. Two methods are having advantages and disadvantages according to its processing environments.

II. Survey Among Various Techniques

Abirami S.P. and Shalini Ramanathan [1], propose a linear scheduling algorithm for obtain the efficient scheduling resources according to the user task. Cloud computing provides various resources and services to different users requirements and task. Scheduling the resource allocation according to the users task is one of the major issue. The proposed method of Linear Scheduling for Tasks and Resources(LSTR) strategy used to scheduling the resources among the requestors and maximize resource utility. The scheduling must be done according to consider the users task and availability of virtual machine. In linear scheduling the arrival time does not consider for scheduling the resources. Shortest Job First Scheduling algorithm using for sort the requesting task. Dynamic allocation could be carried out by the scheduler for different type of task. Shortest time first resource allocation is much better than the first come first serve. Merits of this paper are to improve the resource utilization and response time and avoid starvation and deadlock. Demerits of this paper is not suitable for inter active real time applications.

AmitNathan, Sanjay chaudharya, gaurav soman [2], proposed the algorithm of dynamic planning and scheduling to achieve the effective resource allocation and maximize the utilization of resources. Four policies are used here such as immediate, best effort, advanced reservation and deadline sensitive. Proposed dynamic planning based scheduling algorithm is implemented in Haizea that can admit new leases and prepare the schedule whenever a new lease can be accommodated. It maximizes resource utilization and acceptance of leases. The advantage of this method is maximizing the rate of resource utilization. Disadvantages are requiring more preemption and it increase the overall overhead.

Bo Yin, Ying Wang, Luoming Meng, Xuesong Qiu [3], consider the issue of how to arrange large-scale jobs submitted to cloud in order to optimize resource allocation and reduce cost. To allocate method should decide which virtual machines should be assigned with a new set of jobs. The use of lightweight node to allocate resource is considered as performance metric in this paper. Authors proposed the Multi-Dimensional Resource allocation algorithm considers the multi-dimensional constraints in resource allocation process, which assure jobs can be processed in nodes selected by object function. Under the purpose of optimize utility of nodes, It can select problem as a binary integer programming problem, and object function can assure that using working nodes remain resources to process more jobs. The advantages of the method are increasing the resource utilization and reduce cost of data center. But performance bottleneck may occur and scheduling process of resource allocation is difficult.

Fetahi Wuhib,Rolf Stadler., [4], Propose the Dynamic resource management which declares the particular challenges in large-scale cloud environment. To obtain the efficient heuristic solution to the problem such as Minimize the adaption cost for resource allocation and resource utility using gossip protocol. Scalability and adaptability as taken as the parameters in this paper. The proposed Gossip protocol used to execute task as a middleware platform. The protocol continuously executes while its input changes and it ensure the three design goals as fairness, scalability and adaptability. Resource allocation refers to the allocation of cloud resources to the demands sent by the users. The most free memory virtual machine in the cloud will receive the process and allocate resources to that process. Adaptation is the process of resource allocation for modified process. Demand Sharing is based on heuristic algorithm and it refers

to sharing of memory demand of the process while demand exceeds the capacity of virtual machine. Merits of this paper are it can continuously adapt changes in cloud input and dynamically solve the problem. Demerits are computational complexity high and Overhead may occur.

G.Sireesha, L.Bharathi [5], Parallel data processing has emerged to be one of the major issue applications for Infrastructure-as-a-Service (IaaS) clouds. A new processing framework designed for exploiting the dynamic resource allocation offered by IaaS clouds for both, task scheduling and execution. The both can be received as a valid Job Graph from the users; And the Nephele's Job Manager can transforms it into a so-called Execution Graph. An Execution Graph is nothing but the Nephele's primary data structure for scheduling and monitoring the execution of a Nephele job. Parallelization and Scheduling Strategies are used to constructing an Execution Graph from a user's submitted Job Graph may leave different degrees of freedom to Nephele. The user can provides any job annotation that may contains more specific instructions we currently pursue simple default strategy: Each vertex of the Job Graph is transformed into one Execution Vertex. And the default channel types are network channels. In that each execution vertex is assigned to its own Execution Instance unless the user's annotations or other scheduling restrictions (e.g. the usage of in-memory channels) prohibit it. Parallel execution of instructions thereby saving processing time and cost, Secure mode of transmission, Reliable and efficient transmission, Only required data will be provided are the advantages of this method. But it has some demerits also such that bandwidth consumption and it consume more energy.

Gihun Jung and Kwang Mong Sim [6], the proposed system defined the performance degradation according to response and also consider the location of physical machine to allocate user request as a virtual machine. To design the hybrid resource management architecture to perform location aware VM placement and dynamic resource utilization management. Response time have to increased in this method. In this paper consider the utility function is used to find out which PM is appropriate for a new VM or migration, the provider evaluates each PM using a utility function.VM Placement Decision Making is the process, When a provider receives a new VM placement for a user, the providers should evaluate the network delays between the user and each data center. If the provider finds closest data center for the user VM, the provider now evaluate each and every utilization level of PMs based on utilization report, which is sent from each PM. After VM placement, each PMs can keep on monitoring its utilization level, and it can reports to the providers when the utilization is changed. When a PM's utilization level exceeds a given threshold, it reports to the provider for to decide whether VMs are needed to migrate to another PM. When a provider decides to migrate a problematic VM, the providers can instructs the hypervisor to migrate the VM to another appropriate PMs after migration; each PM keep on monitoring its utilization level and reports it. This method provides better performance and response time. The demerit is bandwidth consumption and not cost effective.

Paul C. K. Kwok, Minjie Zhang [7], propose a method for finding work load of data center for efficient resource management. In cloud computing, the resources are provided by the service providers based on virtualization to satisfy the demands of users. Every users can use different variety of resources and variance in using the resource also. For that the service providers have to offer different amount of virtualized resources per request. To provide good services, a provider may have data centers that are geographically distributed throughout the world. Like that, the user locations may vary in geographic location. Though cloud computing services delivered over the internet, there may be occur undesirable response latency between the users and the data centers. The proposed method of Finding work load of data center and distance between user and data center is used to avoid the response latency. It provide Better response time and resource utilization. But it more cost effective and may overhead occur.

Gunho Lee,Niraj Tolia., [8], Defined that the Current IaaS systems are usually unaware of the hosted application's requirements and therefore allocate resources independently of its needs. To estimate the performance of a given resource

allocation using prediction engine and a genetic algorithm to find an optimized solution in the large search space. The prediction engine maps resource allocation candidates to scores that measures their "fitness" with respect to a given objective function. The Topology Aware Resource Allocation (TARA) can compare and rank different candidates. Search algorithm efficiently used to identify an approximate solution; we chose a genetic algorithm (GA) to generate possible candidates for the prediction engine to evaluate. GA is a search technique inspired by evolutionary biology for finding solutions to optimization and search problems. It reduces the job completion time of applications. Here migration problems may occur and more expensive in cost.

Gunho Leey, Byung-Gon Chunz, Randy H. Katzy [9], In this paper,the resource allocation and job scheduling is the major concern problem of the data analytics cluster in the cloud. To improve performance and cost-effectiveness of a data analytics cluster in that the data analytics system must account for heterogeneity of the environment and workloads. Scheduling is the process of allocating the units of resources that are heterogeneous and shared among various jobs. Progress Share (PS) used to realize fair and effective job scheduling in shared and heterogeneous cluster. The computing rate (CR) is used to calculate the progress share of a job. The sum of the progress share for all jobs indicates the effectiveness of the resource assignment. The advantage of this paper is give good performance and fairness and more cost effective. But it decreases the bandwidth and increase the overhead.

Hadi Goudarzi and Massoud Pedram [10], Consider the resource allocation to the distributed system is the major issue. We consider problem depends on the users SLA with the service provider for resource allocation. To design a well optimized Force directed search algorithm for SLA based resource allocation problem of multi-tier applications in cloud computing. Considering two important things in SLA model average response time guaranteed SLA and SLA that has a price pre request based on the response time. In that model consider two types of SLA classes are Gold SLA class and Bronze SLA class. Gold SLA which specifies an average response time target, a utility value for each serviced request, a maximum arrival rate for the client's requests and a penalty if the average request response time is missed and the Bronze SLA class can specifies a maximum arrival rate and a utility function that specifies a profit per request based on the response time. The probability of distributed function used to determine the amount resources allocate to users. Force directed search algorithm used for SLA based resource allocation problem of multi-tier applications in cloud computing. Merits are Improve the total profit of the system. A demerit of this method is not homogeneous and also occur performance bottleneck.

Jiayin li,Meikang Qiu,Jain-Wei Niu,Yu chen,Zhong Ming [11], proposed an adaptive resource allocation algorithm for preempt able tasks for cloud system which algorithms adjust the resource allocation adaptively based on the updated of the task executions. Adaptive list scheduling (ALS) and adaptive min-min scheduling (AMMS) algorithms are use 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 reevaluation process, the schedulers are re-calculating the finish time of their respective submitted tasks, except the tasks that are assigned to that cloud. A merit of this method is increase the resource utilization but cost and time consumption is high.

Koti Reddy S, Ch. Subba Rao [12], In this paper, Particular tasks of a user's job can be assigned to different types of virtual machines which are automatically instantiated and terminated during the job execution. The proposed model using location based and task scheduling. If user creates a fake ID for access the services, but users location alone can discloser users identify. If users wants to access services they should sent their location to trusted server. Then it will allow you to access information if the users are correct. Task scheduling considers the location based services. It provides the privacy protection. Malicious can't know about your personal identification. Through in this method we can achieve quick response and user privacy to access the

services. Dynamic resource allocation and Parallelism are implemented here. It is designed to run data analysis jobs on a large amount of data, Many Task Computing (MTC) has been developed, less expensive are the merits. Demerits are Performance bottleneck and computation complexity.

Preeti Agrawal,Yogesh Rathore [13], Resource allocation is the issue in the cloud computing environment. In high performance computing providing the adequate resources to the user applications is difficult. For instance, the cloud center cannot handle the user applications which have shortest deadlines to due limited. For that users have to access many computing center to complete the application because of deadline. In this paper, the proposed strategy for the resource managed in cloud environment. To make applications available on flexible execution environments primarily located in the Internet. The proposed model which is efficiently reallocates the resource to get complete. Job scheduling system plays a very important role in how to meet Cloud computing users' job QoS requirements and use cloud resources efficiently in an economic way. from the Cloud computing resources users sides users always think which Cloud computing resource can meet their job QoS requirements for computing, how much money they must pay for the cloud computing resources. From the Cloud Computing service providers side, the CCSP always think about to gain the maximum profits by offering Cloud Computing resources, apart from the CCU's job quality of service requirements. To make these two ends meet, the job scheduling system must take efficient and economic strategies for CCU's differentiated service QoS requirements. It provides better throughput and resource utilization.

Radhika T V & K C Gouda [14], propose profit model algorithm to minimize the infrastructure cost and SLA violations. The developed resource algorithm is used to maintain the resource available in the service center. And the development priority algorithm is used for better resource allocation jobs in the cloud environment. After the efficient resource allocation of various jobs, The profit model has been developed to calculate maximum profit of cloud administrator by serving of each user request. The proposed priority algorithm helps cloud admin to decide priority among the users and allocate resources efficiently according to priority. Demerit of this method is to take proper decision for job scheduling, execution of job, managing the resources.

Rajkamal Kaur Grewal,Pushpendra Kumar Pateriya [15], In this paper, The rule based resource manager used for Hybrid environment, which increase the Performance of private cloud on-demand and also reduce the cost. And it can able to set the time for public cloud and private cloud to fulfill the request and provide the services in time. Depends on the resource utilization and cost we can evaluate the performance of Resource Manager in the hybrid cloud environment. Merits of this method is High priority resource serve by private cloud, less cost and time consumption and demerits of the method is efficiently managing the allocation of resource needs between two clouds.

Rashmi. K. S, Suma. V and Vaidehi. M [16], Defined that the cloud has inherited characteristic of distributed computing and virtualization there is a possibility of occurrence of deadlock. When there are more requests competing for the same resource, at the same time the available resources are insufficient to service the arrived requests. The objective of load balancing in the cloud computing environment is to provide on demand resources with high availability. In this paper using enhanced load balancing for Cloud Manager analyses the availability of the VMs at the time of job arrivals to update the data structure thereby having less overhead involved in maintenance of the data structure. The advantages of this paper are better response time, High performance, Avoids deadlock and No overheads. But it not much cost effective.

Rostand Costa, Francisco Brasileiro [17], Capacity planning for the provision of cloud data centers. To optimize resource usage and to reduce the number of idle resources, The perfect solution is to set time interval and alter resources as persistently keeping with workload changes. Consider the Just in Time Provider does not represent any public cloud provider, but it acts as a legitimate and fully autonomous provider that takes advantage of resources that would be irretrievably wasted without its intervention. Each JiT DC brings some amount of

resources with certain characteristics and capabilities, called JiT Resources. It more cost effective. The disadvantage can occur prediction error and it may use of recursive data structures.

Shaminder Kaur [18],Proposed for design the improved genetic algorithm developed by merging the scheduling algorithm to improve the cost efficiency and performance. The performance metrics of Resource utility, Task scheduling are considered in this paper. Modified Genetic Algorithm (MGA) Generate an initial population of individuals with output schedules of algorithms Longest Cloudlet to Fastest Processor (LCFP), Smallest Cloudlet to Fastest Processor (SCFP) and 8 Random Schedules. The Method can calculate the fitness of all individuals and perform crossover, mutation to obtain good evaluation. It much cost effective and provide better response time, efficient performance. But the computation complexity is high.

Shikharesh Mujumdar [19], Match making and scheduling is used in this paper to represent the resource allocation in cloud computing. Match making is the method of allocating jobs associated with user requests to resources designated from the resource pool. Scheduling is used to determining the order in which jobs mapped to a selected resource are to be executed. The advantage of this method is cost effective and less delay. Demerit of this paper are Uncertainties that are associated with such type of match making, Error Associated with Estimation of Job Execution Times, Lack of Knowledge regarding Local Resource Management Policies.

Shin-ichi Kuribayashi [20], resource allocation is the major issue at every time in cloud computing. The optimization of resource allocations are very important for provide economically feasible cloud services. The effective processing ability and bandwidth for each service request can be achieved by optimal resource allocation method. The optimal resource allocation methods are Best-Fit approach and Round Robin used to perform efficient resource allocation for every users task. The multiple tasks can perform and share the available resources at the same time. Processing ability and Bandwidth are the parameters consider in the paper. Best-Fit approach preferred to reserve as much as possible for future request which may require large size of ability or bandwidth and also reduce the possibility of deadlocks. In Round robin method can select the appropriate center in terms of pre-defined order. This method reduces the request loss probability, Reduce the total resource required for the process and Increase the bandwidth. Demerit of this paper is time consumption is high.

III. Metrics Associated With Resource Management Techniques

Various parameters are used for analyzing the efficiency of different resource allocation strategies.

Table 3.1: Evaluation of metrics

Resource management strategy	Resource utilization	Scalability	Performance	Throughput	Response time	Overhead associated	Cost efficiency	Service level agreement
Linear scheduling strategy				✓				

policy based resource allocation	✓							✓
Multi-Dimensional Resource allocation algorithm	✓							
A Gossip Protocol For Dynamic Resource Management		✓						
Exploiting Dynamic Resource Allocation							✓	
Location-Aware Dynamic Resource Allocation			✓		✓			
A TIME-DRIVEN Adaptive Mechanism	✓				✓			
Topology-Aware Resource Allocation		✓	✓					
Heterogeneity-Aware Resource Allocation		✓					✓	
Multi-Dimensional SLA-Based Resource Allocation	✓		✓					
Adaptive Resource Allocation	✓				✓			
Dynamic Resource Allocation		✓	✓					
An Approach for Effective Resource Management in Cloud							✓	
Resource Allocation								✓

Model for Efficient Management								
A Rule-based Approach	✓							
Enhanced Load Balancing approach					✓			
Just in Time Clouds			✓		✓			
Efficient Approach to Genetic Algorithm for Task Scheduling	✓		✓				✓	
Resource management on cloud: Handling uncertainities in parameters and policies					✓		✓	
Optimal Joint Multiple Resource Allocation		✓		✓				

IV. Conclusion

Cloud computing technology is arising technology being used in business marketing and enterprises. This survey paper described the various resource management techniques used in cloud computing. In cloud computing, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. Some of the techniques discussed above mainly focus on resources allocation for users demand.

V. References

[1] Abirami S.P. and Shalini Ramanathan.,2012. Linear Scheduling Strategy for Resource Allocation in Cloud Environment.

[2] Amit Nathan,Sanjay chaudharya,gaurav soman.,2012. Policy based resource allocation in Iaas cloud
[3] Bo Yin, Ying Wang, Luoming Meng, Xuesong Qiu.,2012.A Multi-Dimensional resource allocation Algorithm in cloud computing.
[4] Fetahi Wuhib,Rolf Stadler,Mike Spreitzer.,2012. Gossip protocol for Dynamic Resource Management in Large cloud environment.
[5] G.Sireesha, L.Bharathi.,2012.Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud.
[6] Gihun Jung and Kwang Mong Sim.,2012.Location-Aware Dynamic Resource Allocation Model for Cloud Computing Environment.
[7] GIHUN Jung, Kwang Mong Sim, Paul C. K. Kwok, Minjie Zhang.,2011.A TIME-DRIVEN Adaptive Mechanism For Cloud Resource Allocation.
[8] Gunho Lee,Niraj Tolia,Parthasarathy Ranganathan,Randy H. Katz.,2011. Topology-Aware Resource Allocation for Data-Intensive Workloads.
[9] Gunho Leey, Byung-Gon Chunz, Randy H. Katzy.,2011. Heterogeneity-Aware Resource Allocation and Scheduling in the Cloud.
[10] Hadi Goudarzi and Massoud Pedram.,2010. Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems.
[11] Jiayin li,Meikang Qiu,Jain-Wei Niu,Yu chen,Zhong Ming.,2011.Adaptive Resource Allocation for Pre-empt able Jobs in Cloud Systems.
[12] Koti Reddy S, Ch. Subba Rao.,2012.Dynamic Resource Allocation In The Cloud Computing Using Nephele's Architecture
[13] Preeti Agrawal,Yogesh Rathore.,2011. An Approach for Effective Resource Management in Cloud Computing.
[14] Radhika T V & K C Gouda .,2013.Resource Allocation Model for Efficient Management in Cloud Computing
[15] Rajkamal Kaur Grewal,Pushpendra Kumar Pateriya.,2012. A Rule-based Approach for Effective Resource Provisioning in Hybrid Cloud Environment.
[16] Rashmi. K. S, Suma. V and Vaidehi. M.,2012.Enhanced Load Balancing Approach to Avoid Deadlocks in Cloud
[17] Rostand Costa, Francisco Brasileiro,Guido Lemos de Souza Filho,Denio Mariz Sousa.,2010.Just in Time Clouds: Enabling Highly-Elastic Public Clouds over Low Scale Amortized Resources
[18] Shaminder Kaur.,2012.An Efficient Approach to Genetic Algorithm for Task Scheduling in Cloud Computing Environment
[19] Shikharesh Mujumdar.,2011.Resource management on cloud: Handling uncertainities in parameters and policies.
[20] Shin-ichi Kuribayashi .,2011.Optimal Joint Multiple Resource Allocation Method for Cloud Computing Environments.
[21] www.wikipedia.com