International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 3 Issue: 1 15-Jan-2014,ISSN_NO: 2321-3337

Analysis of Resource Allocation Approaches in Cloud Computing

Dr. N. Nagadeepa M.Sc., MCA, M.Phil., Ph.D., H.O.D,
Department of Computer Science and Computer Applications,
Jairams Arts and Sciemce College,
Karur, Tamil Nadu, India

Abstract*--Cloud computing is a technology emerging at a rapid speed. The data and various applications can be accessed from anywhere, at any time with the help of clouds. Cloud helps enterprises to cut down their infrastructure cost to a large extent by renting resources from cloud for computational purpose and storage. The applications can be used on a pay – as – you - use basis by the entire company. Hence, getting license for each product is not required. But, the major problem in cloud computing is the cost optimization of the allocated resources. Also, satisfying customer needs as well as application requirements is a great challenge while allocating resources.*

INTRODUCTION

Cloud computing is arising as a new computing model that intends to provide consistent and quality of service guaranteed lively computing environment for the users. Cloud computing is a combination of 3 technologies namely, grid computing, parallel processing and distributed computing. Storing user data in the data center of internet is the core idea behind cloud computing. The stored data can be accessed by the users at any time and they are maintained by the companies that offer cloud service.

Other than storage services, there are hardware and software services by the providers. The names of the services are namely Platform as a Service (PaaS), Infrastructure as a Service (IaaS), Software as a Service (SaaS). All the services are offered for the public as well as for business.

The major benefits of cloud computing are reduced costs and remote access of resources. Reasons of how the above stated benefits are attained are discussed. By the help of cloud computing, the company can reduce the investment of large sums in the physical infrastructures. More and more resources can be accessed from cloud providers when a need arises to expand the business. Next, the cloud services can be used at anytime from anywhere.

I. HIGH PERFORMANCE CLOUD CAPABILITIES

There are several features required for a private cloud to have high performance.

Rapid Elasticity

Based on the demand, the resources can be added or deleted to the system whenever required.

Measured Service

The usage can be measured and it can be controlled, monitored.

Broad Network Access

The resources can be used by anyone in network from anywhere.

On Demand Self Service

The computing resources can be changed automatically by the users depending on the need without the help of human. The changes could be done by them with the help of interactive portal.

II. Resource Allocation's Importance

The task of assigning the resources available to the needed applications over the internet is called Resource Allocation (RA). If the resources are allocated properly, then problem arises. This problem is solved by resource provisioning. For each module, the service providers manage resources with the help of resource provisioning.

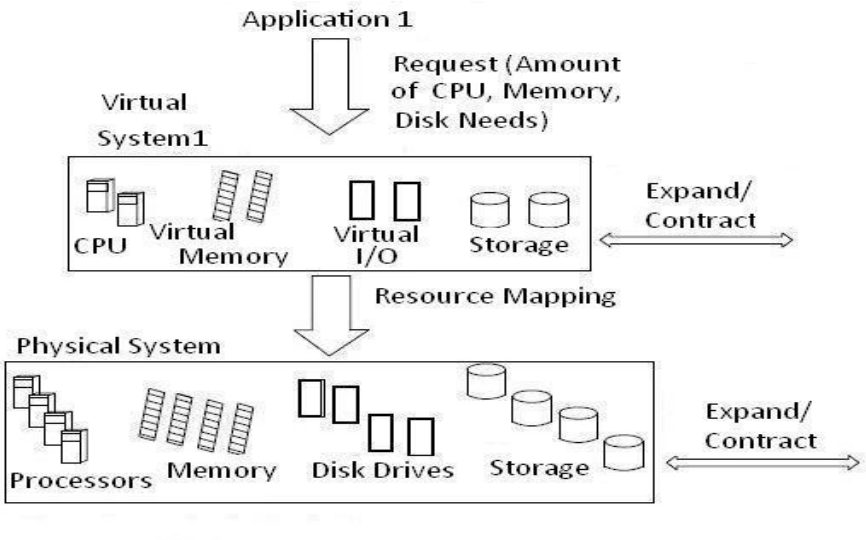

Fig. 1 Mapping of Virtual to physical resources

The process of integrating all the activities required for using and assigning limited resources within the cloud environment to satisfy the needs of cloud application is called Resource Allocation Strategy (RAS). The type and amount of resources required to complete a user job is needed by RAS. Along with it, the time and order of allocation is also supplied as input for optimal RAS. The conditions to be avoided by optimal RAS are:

a) Resource Fragmentation

When there are sufficient resources but they could not be allocated to the needed ones.

b) Resource Contention

When a single resource is tried to be accessed simultaneously by two or more applications, then resource contention occurs.

c) Under - Provisioning

This situation occurs when resources are allocated lesser than required for the application.

d) Over - Provisioning

This situation occurs when the application is allocated resources more than it needs.

The estimation of the resources required to complete a job may sometimes be under - provisioned or over – provisioned. This problem could be avoided by getting the inputs from user and cloud providers for RAS as shown in table1. The important inputs for RAS from the point of cloud providers and users are shown in the table. Optimal RAS results must satisfy various parameters. The cloud provides reliable resources as well as poses an issue in allocating and handling resources.

TABLE I. INPUT PARAMETERS

Parameter	**Provider**	**Customer**
Provider Offerings	√	-
Resource Status	√	-
Available Resources	√	-
Application Requirements	-	√

Agreed Contract Between Customer and Provider	√	√

Predicting the changing nature of users, demands of application and users are impractical from the perspective of cloud providers. The users expect the job to be done with minimum cost and on time. Therefore, due to the inadequate resources, locality restrictions, etc., there is a need for effective resource allocation system.

There are 2 types of cloud resources – physical resources and virtual resources. The physical resources are shared by various computing request through virtualization and provisioning. The virtualized resources request is described by a list of constraints that pertains to processing, etc. To satisfy the request, provisioning maps virtual resources to physical ones. In an on-demand basis, the resources are allocated to the cloud applications.

As the demand and supply of resources can be unpredictable and changing, different techniques are found for resource allocation. Finding an optimum resource allocation is quite complex.

III. Resource Allocation Strategies (RAS)

The input parameters and the way of resource allocation vary depending on the services, nature, etc. of the application that demands the resource. Fig. 2 describes the classification of proposed Resource Allocation Strategies.

A. Execution Time

In cloud, various resource allocation strategies are used. When actual task execution time and preemptable scheduling is used for resource allocation, the issue of resource contention is avoided. Moreover, the resource utilization can be increased when different styles of renting computing capacities are used. But estimation of execution time for a job is a difficult task.

B. Policy

The most-fit processor resource allocation policy is used to control the fragmentation problem in resource allocation. In this method, a job is assigned to a cluster that creates the remaining processor distribution. This leads to assignment of maximum number of jobs.

To identify the target cluster, a difficult search has to be carried out. The nature of the clusters is presumed to be similar and distributed. The quantity of processors used or present in each cluster is binary compatible.

Various experiments have proved that the most-fit policy has larger time complexities but time overhead is lesser than the system's operational time. Hence, in real time it can be used.

C. Virtual Machine

Design of an automatically infrastructure resources scalable system composing of virtual network of virtual machines is presented. It can perform live migration across various domain physical infrastructure. With the help of dynamic availability of resources and demand of application, a virtual computation environment can change its location across the infrastructure and configure the resources.

Effective resource allocations for multiprocessor system's real time task have been done by various researchers. Various aspects like choosing virtual memory for power management, etc. in the data center are focused in recent studies on allocation of cloud's virtual memory for real time task.

The needed resources are set up and booted by the resources. Also the costs for the used resources are only paid. By permitting the users to dynamically include or remove any number of instances of the resources based on the constraints specified by the user and load of VM enables this concept. This strategy differs from IaaS to SaaS as over the internet only applications are delivered to the cloud user.

D. Gossip

Based on the servers, nodes, clusters, etc. cloud environment differs based on cloud environment. For efficient management of resources in a large-scale cloud environment a gossip protocol is used. This protocol is used in large clouds to perform key function within distributed middleware architecture. The basic idea behind this is that the machines of cloud environment are demonstrated as a dynamic set of nodes. Every node has an exact CPU and memory capacity. Cloud resources are allocated to a set of application by the protocol. It dynamically increases a global cloud utility function and has time dependent memory demands. Based on the results of simulation, it is proved that when there are smaller memory demands than cloud's available memory, there is no change in the allocation quality based on the number of machines and applications.

E. *Utility Function*

By optimizing certain objective function like cost performance, etc., the virtual memories in IaaS can be managed dynamically. The selection of objective function is done based on targets achieved, profit, etc. Allocating requests to higher priority applications first by allocating resources dynamically the CPU resources to meet quality of service objectives are described only in few works.

Resource allocation for heterogeneous systems is done based on response time. The application requests are distributed among various available servers which are allocated to each of application tiers exactly. Queuing theory is used to send the client's request to the selected server. Force directed resource management is used for resource consolidation.

F. Hardware Resource Dependency

Multiple Job Optimization (MJO) scheduler is used to improve the utilization of hardware. Based on the hardware resource dependency jobs are classified into various types like Network I/O bound, Memory bound, etc. The type of job and different category parallel job can be detected using MJO scheduler. The resources are allocated to the system that focuses on CPU and I/O based on categories.

For management of resource virtualization open source frameworks like Eucalyptus, Open Nebula is used. Allocating virtual resources on the basis of available physical resources is the common feature of these frameworks. All the frameworks cannot support all the application modes due to the complexity of virtualization technology. Leasing resources from a single point for both the virtual and physical resources are supported by Vega Ling Cloud.

The physical and organizational arrangement needed to carry out cloud operations is known as Cloud Infrastructure. There are 3 phases in the step by step resource co-allocation. Co-allocation scheme is determined in first phase. It is done by taking into account the consumption of CPU for each physical machine. In the next phase, by using simulated annealing algorithm it is decided whether to put an application on physical machine or not. In the third phase, the CPU share that is occupied by

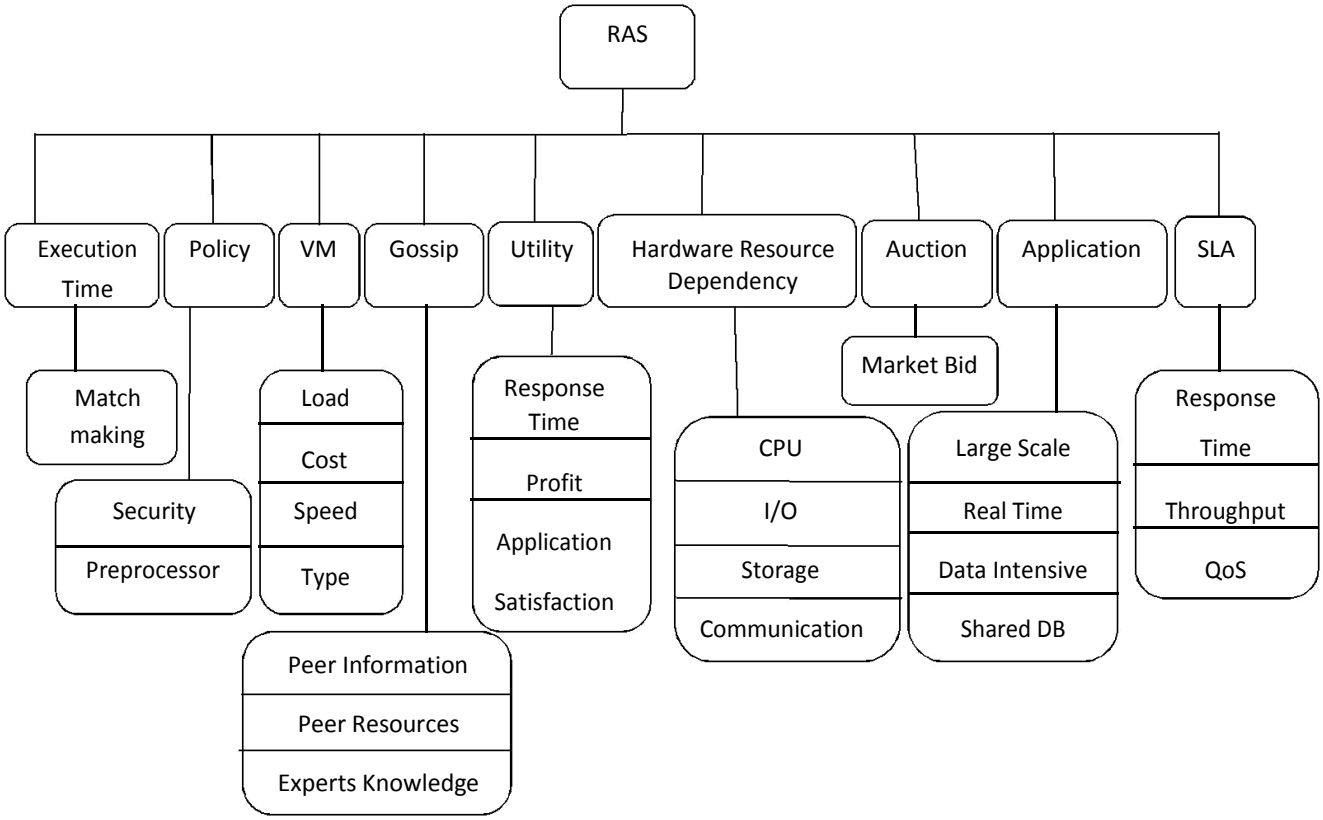

Fig. 2 Cloud Computing Resource Allocation Strategies

each virtual machine is determined. CPU and memory resources for co – allocation are mainly focused in this system.

In a paper, a RAS is proposed in which the clusters are categorized based on type and number of computing, communication resources and data storage. Within each server the resources are allocated. Based on the constant need of the clients, the disk resource is allocated. Based on Generalized Processor Sharing, the other kind of resources in servers and clusters are allocated. In

order to decrease the time taken to make a decision distributed decision making is done by the system. Also to find the optimal initial solution greedy algorithm is used. By varying the resource allocation the solution can be improved.

G. Auction

The allocation of resources using auction technique is based on sealed-bid auction. Initially, each user's bid is collected by the provider. Then, the price is determined. Based on the price of $(k+1)^{th}$ highest bid, the resources are distributed to the first

K^{th} highest bidders. By converting the resource problem into ordering problem, the system simplified the cloud service provider's decision rule and the allocation rule.

The gain of both the customer and resource agent in a large data center are maximized by the resource allocation strategy. This s done by balancing the demand and supply in the market by using market based resource allocation strategy (RSA-M). In this, equilibrium theory is introduced. The amount of fraction used by one virtual memory is determined by RSA-M. Based on the varying requirements of workloads, it can be adjusted dynamically.

H. Application

On the basis of application's nature, resource allocation strategies are proposed. To allocate resource for workflow based application, virtual infrastructure strategies are used. For the allocation of resources strategies like FIFO, naïve, etc. are used. There exists a limit to complete a task for real time application that gathers and analyses the real time data. The interfaces and resources for these kinds of applications vary from rest.

In Database Replica allocation strategy, the resource allocation problem is divided into 2 levels by resource allocation module. The resources are distributed among the clients in the first level. Based on the learning predictive model in the next level, the database replicas are expandable.

I. SLA

Service Level Agreement (SLA) is the agreement between the cloud providers and customers. The agreement specifies the terms of service that a provider must follow. SLA's specify the amount of time the service would be provided, number of simultaneous users, performance levels, etc.

The SLA's are still not properly considered by SaaS providers as they are still in the beginning stage. Several resource allocation strategies are proposed that are specific to SaaS in cloud environment. The movement of computer based applications to web delivered or web based application is begun due to the emergence of SaaS.

V. ADVANTAGES AND DISADVANTAGES

In cloud computing, there are several advantages as well as certain limitations in allocation of resources.

A. Advantages

1) The software or hardware used with a cloud need not be installed. It can be used with the help of internet.

2) The data and applications in the cloud can be used from anywhere.

3) The resources over the internet can be shared by cloud providers when there are only limited resources available

B. Disadvantages

1) When a user wishes to change the provider, then mitigation problem arises as a large amount of data needs to be transferred.

2) To use peripheral devices like printers on cloud, the respective software needs to be installed in that respective system.

VI. CONCLUSION

Cloud computing is one of the most fast emerging technology currently. It is used in various fields for achieving various goals. For the fulfillment of customer's needs, the resources have to be allocated efficiently. This could be achieved only if best resource allocation strategies are used that increases the gain of service providers and also gives better satisfaction for the user. This paper mainly focuses on the resource allocation strategies and issues. Also it discusses a few points about cloud computing.

REFERENCES

[1] Jiyani et al.: Adaptive resource allocation for premptable jobs in cloud systems *(IEEE, 2010)*

[2] Patricia Takako Endo et al.: Resource allocation for distributed cloud: Concept and Research Challenges *(IEEE, 2011)*

[3] Shikharesh majumdar: Resource Management on Cloud: Handling Uncertainities in Parameters and Policies *(CSI communications, 2011, edn.)*