Resource Provisioning Methodologies: An Approach of Producer and Consumer Favorable In Cloud Environment

V.Vivek¹, Srinivasan R², Elijah Blessing Rajsingh³

¹Research Scholar (Part-time External), ²Research Supervisor (External), ³Professor and Director, School of Computer Science and Engineering, Karunya University, Coimbatore, Tamil Nadu, India.

¹vivek544@hotmail.com
²srini0402@gmail.com
³elijahblessing@gmail.com

Abstract— Cloud computing is an emerging technology, having ability to renovate a large part of computing environments by providing resources as services. Cloud environment is composed of a set of resource providers and consumer. Cloud offers two ways of resources provision to the consumers. Firstly resource on-demand and secondly resource on-reservation. Various case studies have proven that resource on-demand has increased cost than resource on-reservation. This paper explores a detailed survey on various existing resource provision mechanisms in state in favour to the consumer and the producer. We will use survey results to identify areas that require further research.

Keywords— Cloud Computing, Resource Distribution, Producer, Consumer.

I. INTRODUCTION

Cloud computing groups the unutilized/ideal inter connected or virtualized computing resource into different kinds of clusters and provides high-performance computing resources to various clients (called consumers) who are in need. This era of processing big data through the utilized resources has brought revolutionary change in the global IT infrastructure. As a result, clients are able to increases their computing power with low cost. Extreme growth in the data quantity, increase in the data speed and increases in types of data are some of the factors why clients adopt cloud environment. Many organizations such as Amazon, Google have already started offering cloud services as “pay as you use” basis [1].

Cloud environment offers four service models: 1) firstly ‘Infrastructure as a Service (IaaS)’ - this provides physical or virtual infrastructure, 2) secondly ‘Platform as a Service (PaaS)’ – provides operating system, programming languages executing environments, database and web servers, 3) thirdly ‘Software as a Service (SaaS)’ – provides access to the application software and database and lastly ‘Network as a Service (NaaS)’ – provides transport connectivity services and inter cloud connectivity services. These models have led to rapid change in the service market among resource producer, mediator and consumer. Here we present various resource provisioning techniques between producer and consumer.

However cloud consumer always have provision to utilize resource either by on-demand the resource instantly or reserve the resource in advance. In general, utilization of resource provisioned by on-demand costs higher than that provisioned by reservation [2]. For example the reservation plan offered by the Amazon EC2 can reduced the total provisioning cost up to 49 percent when the resource is fully utilized [2],[3].

A. Significance of Resource Allocation

Rapid changes in computation paradigm which provides trusted computing environment and growth in digitalization emerges cloud computing environment. This kind of environment resource allocation exists between producer and consumer with set of SLA’s (Service Level Agreements).
These SLA’s comprises of data storage, utilization of available bandwidth and security issues etc. Many cloud producers always end up with provisioning over resources in order to satisfy their consumers (called clients). Hence, this kind of over provisioning leads to needless utilization of resources, which will also lead to unavailability of resources for new consumers. In such cases resource provisioning algorithms helps in proper allocation, favoring producer and consumer. An efficient resource allocation should avoid the following criteria.

a. Over Enthusiasm comes to play when the producer allocates the resource to the consumer additionally than the demand made.

b. Less Enthusiasm comes to play when the producer allocates the resource to the consumer less than the demand made.

c. Resource Congestion comes to play when two or more consumers is trying to access same resource at a particular instance.

d. Resource overload comes to play when a set of resources are loaded heavily and at the same time few resources are not utilized.

e. Resource utilization comes to play when there is a demand from the consumer and the resource is left ideal. This situation will arise when there is no proper allocation.

Mapping resource between cloud consumer and resources available is a big task for the cloud producer. In general producer allocates the resource to the consumer with the minimal cost, but estimating the demand by the consumer is impartial as the request from the consumer are dynamic. At this point it should not lead to either resource over provisioning from the producer perspective and resource under provisioning from the consumer perspective. Minimizing both over provisioning and under provisioning is key highlight in this paper. A detail survey has been presented to reduce the total cost for provisioning resource over a period of time. We have considered both producer and consumer perspectives, requirements, outcomes and risks to compare the various resource allocation techniques.

Remaining sections are as follows: In section II, we have presented few terminologies along with related work for resource allocation in cloud computing. In section III we have focused on comparison of available resource allocation.

Section IV covers, restrictions and challenges of resource allocation followed by section V the conclusion. This survey has not addressed any new resource allocation methodologies.

II. RELATED WORK

We have seen a dramatic growth in cloud environments in the last few years [4]. Currently there are many resource provisioning methodologies available in the market, example Amazon EC2 [3]. System model for cloud environment is comprised of cloud producer, virtual machine repository, cloud brokers and cloud consumers as shown in Fig. 1.

Services are offered to users on rental basis to run their applications and pay-by-the-time basis for creating instances. Since these services are publicly available we often refer to them as public clouds. Parallel to this are the private clouds which are managed for solitary purpose. These clouds are dedicated to either inter organization or single consumer services. A quantity of hybrid clouds i.e. combination of private and public clouds are also available for consumers. Large number of cloud services are also available to share infrastructure between several organizations from specific community with common concerns for example security, compliance and jurisdiction. This can be managed internally or by a third party and hosted internally or externally [5]. All these models are shown in Fig. 2.

![Fig. 1. Cloud Environment System Model.](image-url)
In general, consumer makes a request to the producer for a specific resource. On receiving consumer’s request the producer will perform a search in his list of available resources. If the resource is available then the producer allocates resource based on the priority of the request for that particular resource. In case if the resource is not available, consumer has to send a request to another resource provider. For such cases the producer will encounter matchmaking problem i.e. for every request made by different consumers, the producer has to initiate the search mechanism. Consumer on the other end will send the request to multiple producers and will opt for the fastest available resource.

At the same time producer has to initiate the search for every request made by the consumer without any guarantee that the consumer will opt for the resource provided by him. We also call consumer waiting time for resource as user oriented waiting time and producer waiting time i.e. time taken to search for an available resource as system oriented waiting time. Lesser system oriented waiting time makes the consumer to opt for the resource provided by the particular producer. The logic has been illustrated in Fig. 4. On other hand this situation make i.e. to reduce the system oriented waiting time for searching a resource the producers to go for a cloud broker and virtual resource repository who will maintain the details about available resource database of various cloud producers, illustrated in Fig. 1.

III. RESOURCE PROVISIONING METHODOLOGIES

Moving on to the various resource provisioning methodologies, pairing/matching the resource according to the consumer requirements is a phase where consumer and producer has to be vigilant.

At this junction consumer waiting i.e. user oriented waiting time plays a significant role. Consumer opts for the providers who can execute their services with high speed performance. Taking this ahead, Jiayin Li [6] has proposed a resource provisioning algorithm by preempting the task. This kind of preemption will increase the speed of execution time for a set of consumers who is in need of executing their services at high priority. There is another methodology for improving the execution time by opting for advance reservation [3]. Keeping the execution time in mind, many cloud providers have started using the services of cloud virtual repositories (VR’s).

Keeping the execution time in mind, many cloud providers have started using the services of cloud virtual repositories (VR’s). These virtual repositories maintains their infrastructure in such a way that one or more cloud providers can register their available services under them and these VR’s will be under various cloud brokers also called as cloud mediators control. Cloud providers will have SLA’s with these VR’s i.e indirectly with the cloud brokers. One cloud broker may have control over one or more VR. Whenever consumer needs a resource or service he has provision to approach these cloud brokers with SLA’s. This is one way consumers are able to get their job/task done in efficient manner.

SLA’s covers, sharing amount and usage cost between producer and VR’s according to the work done by Kyong Hoon Kim and Rajkumar Buyya [7]. Amit Nathani, Sanjay Chaudhary, and Gaurav Somani proposed Immediate and Best Effort policies where, Immediate refers to resource will be allocate if available otherwise it will be rejected and Best Effort refers to requested resource will be allocated if available otherwise the request is placed in FIFO queue. Due to finite request it may not be possible to satisfy all requests at a time. To solve this Haizea is proposed which adopts lease mechanism for resource allocation [8]. T.P. Baker proposed stack base policy for resource allocation which permits to process different priorities to share a single run time stack [9].

In fact managing all these is a critical task. When cloud offers virtualization there are many security concerns. These concerns vary from private cloud to public cloud. A cloud infrastructure is a kind of another computer network. So it is obvious that you can face all kinds of attacks which are applicable in normal networks and also in cloud environment.
Even though cloud infrastructure and management capabilities are powerful and reliable, consumers still face both internal and external security including privacy threats, including media failures, software bugs, malware, administrator errors and malicious insiders. Apple’s iPad subscriber privacy leak (http://techcrunch.com/2010/06/15/ipad-breach-personal-data/), Amazon S3’s recent downtime (http://status.aws.amazon.com/s3-20080720.html), and Gmail’s mass email deletions (www.techcrunch.com/2006/12/28/gmail-disaster-reports-of-mass-email-deletions) are examples of such threats. [23]. Dimitrios Zissis proposed trusted third party software which assures specific security characteristics within a cloud environment [24].

On other hand when it comes to resource provisioning, many producers use static resource allocation [10]. Static resource allocations have become basic form of resource allocation. In case of resource migration dynamic resource allocation is more flexible than static. Mochizuki. K in his paper pronounced that it is important for a resource not to be static as changing network dynamics and user requirements may lead to seemingly optimal allocation to an under provisioned or over provisioned at later stages [11]. Rajkumar Buyya proves that the dynamic reallocation of VMs according to the current CPU utilization can bring higher energy savings compared to static resource allocation policies [12].
Nephele is the first data processing framework who accurately accomplished the dynamic resource allocation offered by today’s IaaS clouds for both, task scheduling and execution [13]. Dynamic in nature the systems with high loading are automatically migrated to another low loading physical machine without service interrupting [14].

As authors [14] said, migration takes place when system loading too high. But such type of migration should only happen where both producer and consumer are beneficial. Migration of job is not an easy step; producer has to search another resource which can complete the job without effecting the consumer’s operations. K. Ye in his paper said that live migration of virtual machines can be used to implement energy saving and load balancing in cloud data center [15]. Christopher Clark highlighted various implications and requirements that need to be taken into consideration for migration including timeline and the system infrastructure [16]. He also concludes in his paper that, migration reduces IT costs and improves flexibility with server consolidation and decrease downtime with improved reliability etc. Rajkumar Buyya [17] in his paper clearly states that live migration of VMs often allows workload movement with a short service downtime. However, service levels of running applications are likely to be negatively affected during a live VM migration. Based on the performance evaluation, Rajkumar Buyya concludes that migration overhead is acceptable in many situations but it cannot be ignored, especially in systems where service availability and responsiveness are governed by strict Service Level Agreements (SLAs).

In cloud environment resource dependency parameters are CPU, I/O, memory and communication. Cloud computing becomes as high performance computing only when it satisfy these dependency parameters to their cloud users. In fact allocation of resource to virtual machines has to match their workload [18]. Marjan Gusev in his works states that the consumer cost depends upon the resource lease time and provider cost mostly depends on the CPU utilization of the active leased resources. Zhen Xiao has develop a resource allocation system that can avoid overload in the system effectively while minimizing the number of servers used and also introduced the concept of “skewness” to measure the uneven utilization of a server [19]. Gaurav Somani proposed a novel, VM Placement algorithm which takes into account the application specific resource usage made by the VM. The resources are classified into three categories, CPU, network, and disk I/O.

Resource allocation will be done based on the above mentioned requirements or based on the technology. Another kind of resource allocation is auction model where consumers bid for the best resource within their organizational budget. Since there are multiple cloud providers, same resource will be available in multiple places. But the efficiency differs based on the flexibility, budget and success rate. Improper allocation leads to high damage. To avoid such kind of damage every cloud provider’s resource will have a rank. This rank will be set to the resource based on the success rate and the capacity i.e. CPU speed, available I/O devices, and memory. One are more consumers have an option to bid for single resource at particular instance. These auctions take place in two ways i.e. open forum and closed forum [20].
In auction model, each provider and consumer makes a decision dynamically and agree privately on the selling price. On the other hand in commodity model, resource provider specifies the price and the consumer are charged accordingly [21]. Xingwei Wang in his paper ‘A Resource Auction Based Allocation Mechanism in the Cloud Computing Environment’ proposed a reverse auction model, known as reverse batch matching model.

In real-time the resource should be allocated based on the nature of the application. Since resource allocation is non deterministic problem, allocation methodology has to be consider depending on the nature of the application. All available resource provision methodologies are proposed based on the workflow. Few applications may not be suitable for this kind of resource provisioning. Real time applications need to collect the data dynamically at the time of executions. Kuo-Chan Huang proposed service deployment strategy based on the structural property of application service flows and a dynamic resource provisioning mechanism utilizing the estimation of future service requests [22].

IV. CONCLUSION AND FURTHER RESEARCH SCOPE

Cloud computing is an emerging technology which brought rapid changes in many organizations in terms of infrastructure, services and cost savings. In this paper we have classified resource provisioning methodologies under eight categories in favour to the consumer and producer. They are Execution Time/Matchmaking, Guiding Principles, SLA’s, Virtual Machines (VM’s), Information Exchange, Resource Dependency, Auction and Application. We have not proposed any specific RPM in this paper. For each classification we have given justification along with at least one reference paper. For further work we have open the form in the following areas.

1. Since resource reservation is not always possible an efficient advanced resource allocation mechanism can be proposed.
2. Resource matchmaking algorithms can be proposed based on the consumer requirements and producer success rate.
3. Migration from one producer to another producer is not a simple task to the consumer especially moving large amount of data.
4. More security mechanisms can be proposed for big data public clouds.
5. Dynamic resource allocation and migration algorithms can be proposed by focusing on the drawbacks of static resource allocation.
6. A nature inspired resource allocations can be proposed.

Not limiting to above, hope this paper will be helpful to the researchers to come up with new resource allocation methodologies in cloud environment.

REFERENCES


[3] Amazon EC2 Reserved Instances, [ ONLINE]


