A SURVEY ON RESOURCE ALLOCATION MECHANISM IN CLOUD ENVIRONMENT

Kishan M. Patel¹, Asst. Prof. Mohammed Husain Bohara²

¹Student Computer Science & Engineering Department, Parul Institute of Engineering and Technology, kishan053@gmail.com
²Computer Science & Engineering Department, Parul Institute of Engineering and Technology, mohammed.jeeranwala@gmail.com

Abstract—Cloud computing are distributed and parallel computing system, which facilitate virtualization of resources based on demand. It is a new computing paradigm that goal, to provide reliable, customized and quality of services guaranteed computing environment for cloud user. 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 detail survey on various existing resource allocation mechanisms in state in favor to the consumer and the producer.

Keywords: Cloud Computing, Resource Allocation, Virtualization, Survey, Resource Provision.

I. INTRODUCTION

Cloud computing is defined as a new way of computing dynamically scalable and virtualized resources which are provided as a service over the internet. It is a model for enabling on-demand network access to a shared pool of resources like servers, storage, which provides the services that can be provisioned and released with minimal management effort [1]. Cloud computing represents a recent trend in IT that moves computing and data away from desktop into large data centers. It is an application delivered as services over the Internet [2]. The computing power in a cloud computing environments is supplied by a collection of data centers, in many different locations and interconnected by high speed networks [3]. In cloud computing, a cloud is a cluster of distributed computers which provides on-demand computational resources or services to the remote users over a network [4].

The resource management mechanism helps to coordinate IT resources in response to management actions performed by both cloud consumers and cloud providers. It is the allocation of resources from resource providers to resource consumers. Resource management allows to dynamically re-allocating resources, so that user can more efficiently use available capacity.

In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. IaaS cloud allocates resources to competing requests based on predefined resource allocation policies. If the allocation is not managed properly resource allocation starves services, this problem is solved by allowing the service providers to manage the resources for each individual module. Resource allocation is a part of resource management and it is used to assign the available resources in an economic way.

A. Cloud Architecture

The architecture for Cloud Computing can be divided into three layers: Resource, Platform and Application. The resource layer is the infrastructure layer which is composed of physical and virtualized computing, storage and networking resources [5]. Taking storage as an example, when a user uses the storage service of cloud computing, he just pays for the consuming part without buying any disks or even knowing anything about the location of the data he deals with. Sometimes the IaaS is also called Hardware-as-a-Service (HaaS) [6].

Platform layer also called Platform-as-a-Service generally abstracts the infrastructures and supports a set of application program interface to cloud applications. It is the middle bridge between hardware and application. Examples of platform-as-a-services are Google App Engine and Microsoft’s Azure Services Platform [6].
Application layer or Software-as-a-Service replaces the applications that are running on the computer. If you are using SaaS then there is no need to install and run the special software on your computer. Instead of buying the software with higher cost, you just follow the pay-per-use pattern which can reduce your total cost [5].

![Fig 1: Architecture for cloud computing](image)

**B. Deployment Models**

The Cloud model promotes four deployment models:

1) Private Cloud: The Cloud infrastructure is operated merely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise.

2) Community Cloud: The Cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise.

3) Public Cloud: The Cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling Cloud services.

4) Hybrid Cloud: The Cloud infrastructure is a composition of two or more Clouds (private, community, or public) that remain unique entities but are bound together by standardized technology that enables data and application portability [4].

**C. Virtualization**

Virtualization technology provides the technical basis for Cloud Computing. In general, virtualization deals with the creation of virtual resources, such as operating systems, servers, or storage devices. Different kinds of virtualization: System virtualization adds a hardware abstraction layer on top of the hardware, which is called hypervisor or virtual machine monitor. Virtual Machines do not have direct access to the hardware. The hypervisor runs virtual machines in a non-privileged environment. Using system virtualization, multiple virtual machines, which may run various operating systems, can be run on a single physical machine. System virtualization is the very important technology that is used to provide IaaS, PaaS and SaaS resources. A Virtual Machine Monitor (VMM), also called as hypervisor, is a software that securely partitions the resources of a computer system into one or more virtual machines. A guest operating system is an operating system that runs under the control of a VMM rather than directly on the hardware. The VMM runs in kernel mode, whereas a guest OS runs in user mode. Different hypervisors support different aspects of the cloud.

Hypervisors come in several types:

1) Native hypervisors which sit directly on the hardware platform are most likely used to gain better performance for individual users.
2) Embedded hypervisors are integrated into a processor on a separate chip. Using this type of hypervisor is how a service provider gains performance improvements.

3) Hosted hypervisors run as a distinct software layer above both the hardware and the OS. This type of hypervisor is useful both in private and public clouds to gain performance improvements [7].

II. SIGNIFICATION 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 produces allocates the resource to the consumer less than the demand made.

c. Resource Congestion comes to play when two are 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.

III. EXISTING RESOURCE ALLOCATION MECHANISM

In this paper[8] they decreased the most costly SLA violations, and improve performance and low energy consumption for autonomic allocation workload. They have hierarchically structured all possible reallocation actions, and designed, implemented, and evaluated two knowledge management techniques, Case Based Reasoning and a rule-based approach to achieve the aforementioned goal for one reallocation level, i.e., VM reconfiguration. After a comparison, they determined the rule-based approach to outperform CBR with respect to violations and utilization, but also to time performance. Furthermore, they applied the rule-based approach to a real world use case evaluating a scientific workflow from the area of bioinformatics. They showed by simulation that the rule-based approach can effectively guarantee the execution of a workload with unpredictably large resource consumptions.

[9] In this matter, a tenant-based model is presented to tackle over and underutilization when SaaS platforms are deployed over cloud computing infrastructures. This model contains three complementary approaches: (1) tenant-based isolation which encapsulates the execution of each tenant, (2) tenant-based load balancing which distributes requests according to the tenant information, and (3) a tenant-based VM instance allocation which determines the number of VM instances needed for certain workload, based on VM capacity and tenant context weight. After running all tests and simulations, the results were gathered and averages were calculated. In general, over and underutilization averages were reduced but only averages for underutilization were statistically improved.
In this paper[10], they present a resource optimization mechanism for pre-emptible applications in federated heterogeneous cloud systems. They also propose two novel online dynamic scheduling algorithms, DCLS and DCMMS, for this resource allocation mechanism. Experimental results show that the DCMMS outperforms DCLS and FCFS. And the dynamic procedure with updated information provides significant improvement in the fierce resource contention situation. The energy-aware local mapping in our dynamic scheduling algorithms can significantly reduce the energy consumption in the federated cloud system.

[11] They considered the problem of QoS-based resource provisioning in a hybrid Cloud computing system where the private Cloud is failure-prone. Their specific contributions in this work were as follows:

• They developed a flexible and scalable hybrid Cloud architecture to solve the problem of resource provisioning for users’ requests. The proposed architecture utilizes the InterGrid concepts which are based on the virtualization technology and adopt a gateway (IGG) to interconnect different resource providers.

• They proposed brokering strategies in the hybrid Cloud system where an organization that operates its private Cloud aims to improve the QoS for the users’ requests by utilizing the public Cloud resources. Various failure-aware brokering strategies which adopt the workload model and take into account the failure correlations are presented. The proposed policies take advantage of the knowledge-free approach, so they do not need any statistical information about the failure model of the local resources in the private Cloud.

• They evaluated the proposed policies and consider different performance metrics such as deadline violation rate and job slowdown. Experimental results under realistic workload and failure events, reveal that we are able to adopt the user estimates in the brokering strategy while using the workload model provides the flexibility to choose the suitable strategy based on the desired level of QoS, needed performance, and available budget.

[12] Their algorithms with the incorporation of RS and MCER greatly contribute to reducing energy consumption. In essence, the energy saving of ECS and ECS+idle is enabled by the exploitation of the DVS technique—a recent advance in processor design. Their study provides promising results showing the significance and potential of DVS in the reduction of energy consumption. They have evaluated ECS and ECS+idle with an extensive set of simulations. They were also compared with two previous algorithms. The experimental results from our comparative evaluation study confirm the superior performance of ECS and ECS+idle over the other two, particularly in energy saving.

The paper[13] proposes an architectural framework for on-demand infrastructure services provisioning that comprises of the three main components: Compassable Services Architecture (CSA) that intends to provide a conceptual and methodological framework for developing dynamically configurable virtualized infrastructure services; Infrastructure Services Modeling Framework (ISMF) that provides a basis for the infrastructure resources virtualization and management, including description, discovery, modeling, composition and monitoring; Service Delivery Framework (SDF), which provides a basis for defining the whole compassable services life cycle management and supporting infrastructure services.

This paper[14] has studied the advantages of using a hybrid infrastructure composed of Grid and Cloud resources. These two technologies can work together providing the scientific community with an environment in which the researchers can execute computationally intensive scientific applications. The proposed prototype is able to efficiently execute HTC scientific applications on a hybrid infrastructure. A mixed infrastructure composed of Globus Toolkit resources, for the Grid, and Virtual Machines deployed through OpenNebula, for the Cloud, has been evaluated. The scheduling approach enables to outsource job executions to the Cloud when no spare Grid resources are available. In addition, other models of hybrid Grid/Cloud execution models have been covered, pointing out the benefits of the Cloud in terms of elasticity and configurability. The usage of hybrid infrastructure enables to access a larger pool of computational resources which reduces the execution time of HTC application when compared to single infrastructures.

In this paper[15], they considered the problem of dynamic resource allocation and power management in virtualized data centers. Prior work in this area uses prediction based approaches for resource provisioning. In this work, they have used an alternate approach that makes use of the queuing information available in the system to make online control decisions. This approach is adaptive to unpredictable changes in workload and does not require estimation and prediction of its statistics. Their approach uses the recently developed technique of Lyapunov Optimization that allows us to derive analytical performance guarantees of the algorithm.
They have developed an efficient and effective algorithm to determine the allocation strategy that results in the smallest number of servers required. They have also developed a novel scheduling discipline, called probability dependent priority, which is superior to FCFS and head-of-the-line priority in terms of requiring the smallest number of servers.

IV. CONCLUSION

Cloud computing is the new era of computing for delivering computing as a resource. The success and beauty behind cloud computing is due to the cloud services provided with the cloud. Due to the availability of finite resources, it is very important for cloud providers to manage and assign all the resources in time to cloud consumers as their requirements are changing dynamically. So in this paper various resource allocation techniques in cloud computing environments have been considered.

Many authors have proposed algorithms and methods for dynamic resource allocation in cloud computing. In summary, an efficient Resource Allocation Technique should meet the following criteria: Quality of Service (QoS) aware utilization of resources, cost reduction and power reduction/energy reduction. Some of the authors have focused on IaaS-based resource allocation with VM scheduling. The ultimate goal of resource allocation in cloud computing is to maximize the profit for cloud providers and to minimize the cost for cloud consumers.

REFERENCES