Resource Allocation Strategies in Cloud Computing: Overview

Noha Hamdy
Ph.D. Student of Computer Science
Faculty of Computers and Information, Helwan University, Cairo, Egypt

Amal Elsayed Aboutabl
Associate Professor of Computer Science
Faculty of Computers and Information, Helwan University, Cairo, Egypt

Nahla ElHaggar
Assistant Professor of Information Technology
Faculty of Computers and Information, Helwan University, Cairo, Egypt

Mostafa-Sami M. Mostafa
Professor of computer Science
Faculty of Computers and Information, Helwan University, Cairo, Egypt

ABSTRACT
Cloud computing is an attractive processing model, it allows clients to use the internet and central remote servers to manipulate data, applications and access their personal files at any computer without installation of extra software. This technology allows more efficient computing by centralizing storage, memory, processing and bandwidth.
Optimizing resources in the cloud is a main benefit, minimizing cost and satisfying client requests are the goal. In this paper, many resource allocation strategies and their challenges are presented. It is believed that this paper would help both cloud users and researchers to be aware with many applied resource allocation strategies.

Keywords
Cloud Computing, Resource Allocation Strategies (RAS), Virtual Machine, Service Level Agreement (SLA), CloudSim.

1. INTRODUCTION
Cloud Computing is a technology that utilizes central remote servers to maintain data and applications. According to NIST definition "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources e.g., networks, servers, storage, applications, and any other available resources that can be rapidly provisioned with minimal management effort or extra provider interaction."[1]
Cloud computing term is defined in 2006 or 2007, it is a technological progress that concentrates on the way of designing computing systems, develop applications, and existing services for building software. It is based on the concept of dynamic provisioning, which is applied not only to services, but also to compute capability, storage, networking, and Information Technology (IT) infrastructure in general. Resources are made available through the internet and offered on a pay-per-use basis from cloud computing service providers. Today, anyone can pay to get a cloud services, deploy and configure servers for an application in a few hours [21]. Cloud computing customers do not own the physical infrastructure, rather they rent the utilization from a third party cloud provider. They consume resources as an accommodation and pay only for resources that they utilize [3].
This paper provides a brief overview of the Cloud computing phenomenon.
The paper is organized as following, section 2 presents cloud architecture including cloud service models and cloud development models, section 3 presents different applied resource allocation strategies finally section 4 concludes the paper contribution.

2. CLOUD COMPUTING ARCHITECTURE
The goal of cloud computing is controlling, arranging, and reaching the hardware and software resources remotely. It offers online data storage, infrastructure, and applications. It presents an independent platform as the software is not required to be setup on the PC. Consequently, cloud computing helps making business applications portable and cooperative. Virtualization is the magic key in cloud, it represents a technology platform used for the creation of virtual instances of IT resources. A layer of virtualization software allows physical IT resources to provide multiple virtual images of themselves so that their underlying processing capabilities can be shared by multiple users [3]. National Institute of Standard and Technology (NIST) describes cloud computing with five characteristics, three service models and four deployment models as in figure 1.
The first layer includes five characteristics of cloud computing: on demand self-service, broad network access, resource pooling, rapid elasticity and measured service. On-demand self-service provides automatic computing capability management to systems, without requiring human interaction. Broad network access allows heterogeneous clients, such as mobile phones, laptops, to connect to cloud systems over the network. Resource pooling in cloud systems is available as pooling resources for multiple consumers which is able to dynamically assign and reassign according to consumer demand. Rapid elasticity offers rapid and elastic provision of capabilities. It can quickly scale out and dramatically release to quickly scale in automatically in order to support consumer's systems. Measure service the last characteristic provides monitoring, controlling and reporting of resource usage [4].

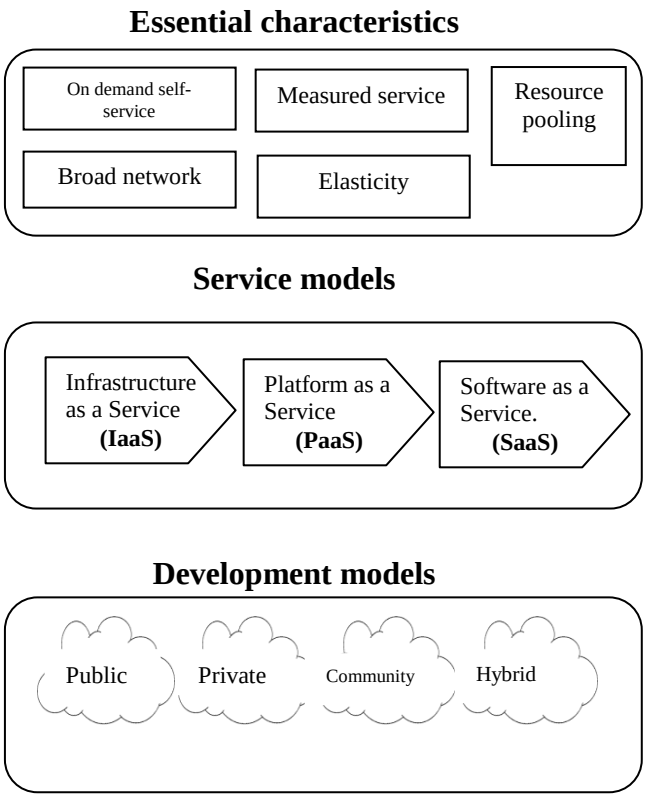

Figure 1. NIST visual model of cloud computing definition

The second layer of a cloud includes cloud computing service models that are categorized into three basic levels:

- Infrastructure-as–a-Service (IaaS).
- Platform-as-a-Service (PaaS).
- Software-as-a-Service (SaaS).

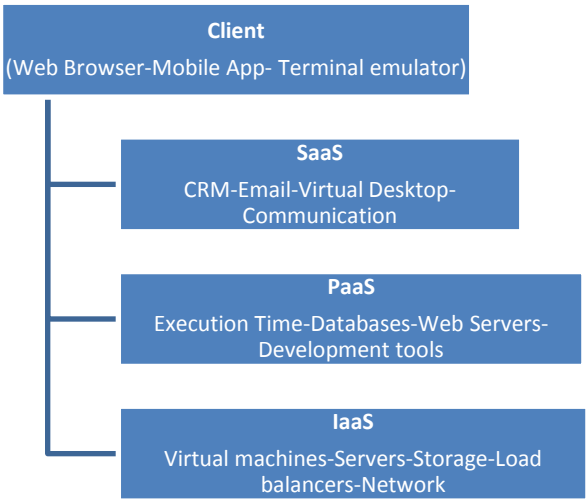

Figure 2 Cloud computing service models

Infrastructure-as Service (IaaS) provides the capability to the consumer to provision processing, storage, networks, and other fundamental computing resources. A consumer is able to deploy and run software, which can include operating systems through administrative access to virtual machines and applications but can't manage or control the underlying cloud infrastructure [3].

Platform as-a-Service (PaaS) enables the lone developers to deploy web-based applications without buying actual servers and setting them up as in figure 2.
Software as-a-Service (SaaS) enables a software distribution model in which a third party provider hosts applications and makes them available to customers over the internet. It can reduce the total cost of hardware and software development, maintenance, and operations. Now SaaS is offered by organizations, for example, Google, Salesforce, Microsoft and Zoho [3].

Development models is the last layer in cloud computing which define the communication between the customer and cloud provider in private, public, hybrid and community cloud.
In public cloud, cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services such as Microsoft, Amazon and Google. There are many benefits of deploying, the public cloud model as cost effectiveness, reliability, flexibility and high scalability however low security is one of the main disadvantages of public cloud [19].
In private, the cloud infrastructure is deployed, maintained and operated for a specific organization. The operation may be in-house or with a third party on the premises. High security and privacy and full control are the main benefits of private cloud but high cost and limited scalability are considered a limitation.
In community cloud, infrastructure is shared by a few associations and constructs a particular group that has shared concerns. It might be managed by the associations or an outsider party [3].

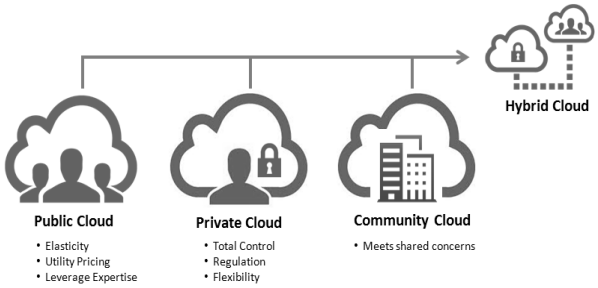

Figure 3 Cloud computing development models [5]

In hybrid cloud, infrastructure consists of a number of clouds of any type, but the clouds have the ability through their interfaces to allow data and/or applications to be moved from one cloud to another. This can be a combination of private and public clouds that support the requirement to retain some data in an organization, and also the need to offer services in the cloud. (e.g., cloud blasting for load balancing between clouds) [19].

3. RESOURCE ALLOCATION STRATEGIES

A Resource Allocation Strategy (RAS) in cloud computing can be defined as any mechanism that aims to guarantee that physical and/or virtual resources are assigned correctly to cloud users. This leads to minimizing resource struggle, lack of resources, resource fragmentation, over-provisioning and under-provisioning. Various parameters affect the applied resource allocation strategy which are overviewed in the following section.

3.1 Linear Scheduling Methods

The FIFO or LIFO scheduling methods are the golden keys in linear scheduling. Abirami and Ramanathan [9] propose a scheduling algorithm called Linear Scheduling for Tasks and Resources (LSTR), which applies scheduling for processing in tasks and resources respectively. It combines Nimbus [17] and Cumulus [18] services to a server node to establish the IaaS cloud environment and virtualization method is KVM/Xen along with LSTR scheduling to allocate resources. The dynamic allocation could be carried out by the scheduler dynamically on requests for additional resources with continuous evaluation of the threshold value. The resource requests are collected and are sorted in different queues based on a threshold value. Then, the requests are satisfied by the VM's.

3.2 Virtual Machine

Virtualization means to create a virtual image of a physical element such as a storage device, operating system, or any processing element. The cloud divides the resource into one or more execution environments [3]. The structure made out of a virtual system is ready for development as a physical element. VM migration offers incredible advantages, for example, load adjusting, server solidification, online maintenance and proactive adaptation to non-critical failure.

A model that proposes a virtualization technology to allocate available resources dynamically with optimization of number of servers in use and follow the application demands and support green computing [12]. The authors introduce the concept of "skewness" to get the difference in the multi-dimensional resource utilization of a server. They try to decrease the skewness value to combine different types of workloads and improve the overall utilization of server resources. A set of heuristics have been developed that prevent overload in the system while saving energy used.

Mofolo and Suchithra [13] propose an algorithm to minimize migration time and the number of migrations, it consider the VM placement problem as a bin packing problem where the physical servers are symbolized by bins, the VMs to be allocated are symbolized by items, and size of the bins is the available CPU capacities of those nodes. The VM is implemented through the CloudSim [16].

Virtual Machine Manager (VMM) [14] has been created as a scheduler of VMs, It asks for resources from the resource provider by sending the task needs, resource provider checks the availability of resources with resource owner, if the resources are available, the resource owner grants the access permission to use the resources to resource provider. Resource provider further provides access of the resources for creation of virtual machines. The execution of the task with considering performance factor reduces execution time and saves cost. The paper extended the CloudSim [16] by adding provider policies and new resource allocation (ERA) algorithm in VMM allocation policy as a decision support for resource provider.

3.3 Nature Inspired Optimization Methods

Biologically inspired methods are based on modeling animals' natural behavior to reach a solution for optimization problems. It includes ant colony, bee colony, and firefly and eagle strategy. A study based on ant colony optimization has been proposed [2]. Authors present a task classification based on QoS with network bandwidth, service completion time, the system reliability and costs as a QoS parameters. In this experiment they set the number of task from 20 to 100, the number of node calculation of 8, In order to show distinction, they designed the QoS attribute of node set up large gap, mainly including the CPU, memory and network bandwidth. Application of ant colony optimization and random distribution algorithm respectively carry out 10 times they realized with the increase of the quantity task, through the ant colony optimization algorithm performs all the tasks, it takes the time less than general algorithm. Due to the ant colony algorithm which chooses target path through the pheromone strength, so when the task amount is less (such as 20), this algorithm implementation effect is not obvious, But when the task quantity achieved 80, two algorithms' execution time nearly one seconds.

Hussain and Mishra present ABC Bee's Algorithm [10], a scheduler finds the job with lowest memory, input-output, and processor requirements. This job is represented as a scout bee which is required to get the suitable site. The scout job is sent to the location at which the task requires the resource at present. The scout job finds the location by using a fitness function. This fitness function runs the task in a specific instance and evaluates that the task is memory dependent or processor dependent. Fitness is the progress of the specific job with assigned resources. After identifying the resources and location, scout job returns back to the scheduler and performs a waggle function. A Waggle function characterize the tasks being in scheduled on the basis of the information provided by scout job such as cost, processor and memory requirements.

3.4 Priority Based Methods

Gouda, Radhika and Akshatha present a resource allocation model that decides priority among different user requests [5]. Each request consists of different tasks. For each task different parameters are considered such as time, processor request, importance and price. Time refers to computation time needed to complete a particular task. Processor request refers to number of processors needed to run the task. The more the number of processor, the faster will be the completion of task. Importance refers to how important the user is to a cloud administrator that is whether the user is old customer to cloud or new customer. Finally price parameter refers to cost charged by cloud admin to cloud users. Based on bin packing algorithm and all the parameters considered above, priority algorithm decides priority among different task submitted by different users.

Another priority method based on application nature in the work by Truong Huu and Montagnat [6]. Virtual infrastructure allocation strategies are prepared for workflow based applications where resources are allocated based on the workflow representation of the application. For workflow based applications, the application logic can be interpreted and exploited to produce an execution schedule estimate. This helps the user to estimate the exact amount of resources that will be consumed for each run of the application. Four strategies such as Naive, FIFO, Optimized and services group optimization are designed to allocate resources and schedule computing tasks.

3.5 Gossip Protocol Based Methods

A gossip based model for resource allocation in large cloud environments is proposed by Wuhib and Stadler [20]. The proposed model is represented as a dynamic group of nodes that constructs the machines (physical nodes) of cloud environment. Each node has a specific CPU capacity and memory capacity. The model implements a distributed scheme that allocates cloud resources to a set of applications that have time dependent memory demands and it

dynamically maximizes a global cloud utility function. The experimental simulated results show that the model produces optimal allocation when memory demand is smaller than the available memory in the cloud and the quality of the allocation does not change with the number of applications and the number of machines. But this work requires additional functionalities to make resource allocation scheme is robust to machine failure which spans several clusters and datacenters. Yanggratoke, Wuhib and Stadler [15] propose a decentralized design by the components of the middleware layer run on every server of the cloud environment. They develop an instantiation of the generic protocol of gossip method which aims to decreasing power consumption through server integration, while satisfying a changing load pattern. This protocol, called GRMP-Q, provides an efficient solution that performs in a good manner in many cases. Under overload, the protocol gives a fair allocation of CPU resources to clients.

3.6 Service Level Agreement (SLA)

A service-level agreement (SLA) is a contract between a service provider and its customer on what services the provider will furnish. SLAs originated with network service provider, but are now widely used by telecommunication service providers and cloud computing service providers [8]. A proposed resource allocation algorithms for SaaS providers who want to minimize infrastructure cost and SLA violations. That ensure that SaaS providers are able to manage the dynamic change of customers, mapping customer requests to infrastructure level parameters and handling heterogeneity of Virtual Machines. CloudSim [16] is used to simulate the cloud computing environment that utilizes the proposed algorithms for resource allocation. Performance is measured from both customers and SaaS providers' point of view. From customers' perspective, observed how many SLAs are violated. From SaaS providers' perspective, observed how much cost reduced and how many VMs are initiated.

3.7 Auction Mechanism

Cloud resource allocation by auction mechanism is presented by Fujiwara thesis [4] that proposes a combinatorial auction based marketplace mechanism for cloud computing services, which allows users to reserve arbitrary combination of services at requested timeslots, prices and quality of service. The participated agents are a seller agents stand for a provider of cloud computing services and a buyer agents stand for a user of cloud computing services, a dedicated protocol named CombiSVMP (stands for Combinatorial Simple Virtual Market Protocol), has been designed to exchange information between the marketplace server and the participant agents. Three experiments have been carried out to evaluate the marketplace designs. The Results showed that the forward /combinatorial design brings the best completion rate and cost performance for the users and the highest global utilization.
Wei-Yu Lin, Guan-Yu Lin and Hung-Yu Wei [7] present a dynamic auction mechanism for cloud resource allocation by construct a real time model consisting of two periods with n cloud users and a cloud service provider (CSP). The CSP has two tasks, performing time-insensitive background computing and distributing resource to the cloud users in the dynamic process. If the total input into the background task excess the threshold, the CSP will gain a fixed amount of value. The CSP will also sell its residual resources to the cloud users after deciding how much resource shall be distributed to the background task.

3.8 Other Methods

This section includes rarely applied methods. Such as a hardware resource dependency [11], this method proposed a procedure that partitions clusters in the cloud based on the number and type of computing , data storage and communication resources that they administrate. All of these resources are allocated within each server. The disk resource is allocated based on the fixed usage of each client and other kind of resources in the servers and are clustered and allocated using Generalized Processor Sharing (GPS). This procedure performs distributed decision making to minimize the decision time by paralleling the solution and used greedy algorithm to reach the best initial solution. The solution could be improved by changing resource allocation. But this system cannot handle large changes in the parameters which are used for finding the solution. Also Shin and Akkan [22] have implemented a decentralized user and virtualized resource management by including another layer called domain in the middle of the customer and the virtualized resources. In light of role based access control (RBAC), virtualized assets are assigned to clients through this layer.
Finally figure 4 summarizes commonly applied methods in resource allocation process in cloud computing.

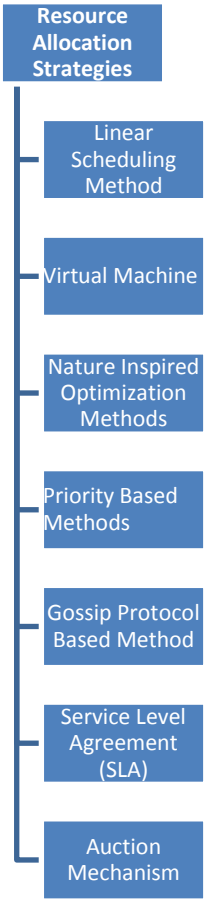

Figure 4. Applied RAS summary

4. CONCLUSION

This overview shows many RAS, focuses on the vital role of it in the cloud. The cloud provider selection of RAS will

affect the throughput, utilization, response time and latency of resources in the cloud. And our main goal is minimizing response time, avoid over provisioning and under provisioning, avoid resource fragmentation,

5. REFERENCES

[1] National Institute of Standards and Technology, "*The NIST Definition of Cloud Computing*", Computer Security Division, Information Technology Laboratory (2011).

[2] Linan Zhu, Qingshui Li, and Lingna He, "*Study on Cloud Computing Resource Scheduling Strategy Based on the Ant Colony Optimization Algorithm*", I International Journal of Computer Science, September 2012.

[3] Bhaskar Prasad, Eunmi Choi and Ian Lumb, "*A Taxonomy and Survey of Cloud Computing Systems*", Fifth International Joint Conference on INC, IMS and IDC, 2009

[4] Ikki Fujiwara," *Study on Combinatorial Auction Mechanism for Resource Allocation in Cloud Computing Environment*", Ph.D. thesis 2012.

[5] K C Gouda, Radhika T V and Akshatha M," *Priority based resource allocation model for cloud computing*", International Journal of Science, Engineering and Technology Research, January 2013.

[6] Tram Truong Huu and John Montagnat, "*Virtual Resource Allocations distribution on a cloud infrastructure*", IEEE, 2010.

[7] Wei-Yu Lin, Guan-Yu Lin and Hung-Yu Wei," *Dynamic Auction Mechanism for Cloud Resource Allocation*", IEEE/ACM 10th International Conference on Cluster, Cloud and Grid Computing.

[8] Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya, "*SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments*", IEEE Computer Society 2011.

[9] Abirami S.P. and Shalini Ramanathan, "*Linear Scheduling Strategy for Resource Allocation in Cloud Environment*", International Journal on Cloud Computing: Services and Architecture, February 2012.

[10] Javed Hussain and Durgesh Kumar Mishra," *An Efficient Resource Scheduling In Cloud Using Abc Algorithm*", International Journal of Computer Engineering and Applications, June 2015

[11] HadiGoudarzi and MassoudPedram, "*Maximizing Profit in Cloud Computing System via Resource Allocation*", 2012

[12] Seematai S. Patil and Koganti Bhavani , "*Dynamic Resource Allocation using Virtual Machines for Cloud Computing Environment*", International Journal of Engineering and Advanced Technology, August 2014

[13] Ts`epoMofolo, R Suchithra," *Heuristic Based Resource Allocation Using Virtual Machine Migration: A Cloud Computing Perspective*", International Refereed Journal of Engineering and Science, May 2013.

[14] Gaurav Raj, Ankit Nischal," *Efficient Resource Allocation in Resource provisioning policies over Resource Cloud Communication Paradigm*", International Journal on Cloud Computing: Services and Architecture, June 2012.

[15] Rerngvit Yanggratoke, Fetahi Wuhib and Rolf Stadler," *Gossip-based Resource Allocation for Green Computing in Large Clouds*", 7th International Conference on Network and Service Management, Paris, France, October, 2011.

[16] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov1 and C´esar A. F. De Rose," *CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms*", 24 August 2010 in Wiley Online Library.

[17] www.nimbusproject.org

[18] www.cumulus-project.eu

[19] Thomas Erl, Zaigham Mahmood, Ricardo Puttini, "*Understanding Cloud Computing*" in "*Cloud Computing Concepts, Technology and Architecture"* Second Edition September 2013, pp 27-49.

[20] Fetahi Wuhib and Rolf Stadler, "Distributed monitoring and resource management for Large cloud environments" IEEE, 2011.

[21] Rajkumar Buyya, Christian Vecchiola, S.Thamarai Selvi, "Mastering Cloud Computing Foundation and Application Programming, 1st Ed.,India: McGraw Hill Education Private Limited, September 2013.

[22] Dongwan Shin and Hakan Akkan, "Domain- based virtualized resource management in cloud computing", Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2010 6th International Conference on USA.