A Survey on Resource Allocation Strategies in Cloud Computing

V.Vinothina, Sr.Lecturer,
Dept. of Computer Science,
Garden City College,
Bangalore, Karnataka, India.

Dr.R.Sridaran, Dean
Faculty of Computer Applications,
Marwadi Education Foundation's
Group of Institutions,
Rajkot, Gujarat, India.

Dr.PadmavathiGanapathi
Professor and Head,
Dept. of Computer Science,
Avinashilingam Institute of
Home Science and Higher
Education for Women,
Coimbatore, Tamil Nadu.
India.

***Abstract*— Cloud computing has become a new age technology that has got huge potentials in enterprises and markets. Clouds can make it possible to access applications and associated data from anywhere. Companies are able to rent resources from cloud for storage and other computational purposes so that their infrastructure cost can be reduced significantly. Further they can make use of company-wide access to applications, based on pay-as-you-go model. Hence there is no need for getting licenses for individual products. However one of the major pitfalls in cloud computing is related to optimizing the resources being allocated. Because of the uniqueness of the model, resource allocation is performed with the objective of minimizing the costs associated with it. The other challenges of resource allocation are meeting customer demands and application requirements. In this paper, various resource allocation strategies and their challenges are discussed in detail. It is believed that this paper would benefit both cloud users and researchers in overcoming the challenges faced.**

Keywords- Cloud Computing; Cloud Services; Resource Allocation; Infrastructure.

I. INTRODUCTION

Cloud computing emerges as a new computing paradigm which aims to provide reliable, customized and QoS (Quality of Service) guaranteed computing dynamic environments for end-users [22].Distributed processing, parallel processing and grid computing together emerged as cloud computing. The basic principle of cloud computing is that user data is not stored locally but is stored in the data center of internet. The companies which provide cloud computing service could manage and maintain the operation of these data centers. The users can access the stored data at any time by using Application Programming Interface (API) provided by cloud providers through any terminal equipment connected to the internet.

Not only are storage services provided but also hardware and software services are available to the general public and business markets. The services provided by service providers can be everything, from the infrastructure, platform or software resources. Each such service is respectively called Infrastructure as a Service (IaaS), Platform as a Service (PaaS) or Software as a Service (SaaS) [45].

There are numerous advantages of cloud computing, the most basic ones being lower costs, re-provisioning of resources and remote accessibility. Cloud computing lowers cost by avoiding the capital expenditure by the company in renting the physical infrastructure from a third party provider. Due to the flexible nature of cloud computing, we can quickly access more resources from cloud providers when we need to expand our business. The remote accessibility enables us to access the cloud services from anywhere at any time. To gain the maximum degree of the above mentioned benefits, the services offered in terms of resources should be allocated optimally to the applications running in the cloud. The following section discusses the significance of resource allocation.

A. Significance of Resource Allocation

In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed precisely. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module.

Resource Allocation Strategy (RAS) is all about integrating cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application. It requires the type and amount of resources needed by each application in order to complete a user job. The order and time of allocation of resources are also an input for an optimal RAS. An optimal RAS should avoid the following criteria as follows:

a) ***Resource contention*** *situation arises when two applications try to access the same resource at the same time.*

b) ***Scarcity of resources*** *arises when there are limited resources.*

c) ***Resource fragmentation*** *situation arises when the resources are isolated. [There will be enough resources but not able to allocate to the needed application.]*

d) ***Over-provisioning*** *of resources arises when the application gets surplus resources than the demanded one.*

*e) **Under-provisioning** of resources occurs when the application is assigned with fewer numbers of resources than the demand.*

Resource users' (cloud users) estimates of resource demands to complete a job before the estimated time may lead to an over-provisioning of resources. Resource providers' allocation of resources may lead to an under-provisioning of resources. To overcome the above mentioned discrepancies, inputs needed from both cloud providers and users for a RAS as shown in table I. From the cloud user's angle, the application requirement and Service Level Agreement (SLA) are major inputs to RAS. The offerings, resource status and available resources are the inputs required from the other side to manage and allocate resources to host applications [25] by RAS. The outcome of any optimal RAS must satisfy the parameters such as throughput, latency and response time. Even though cloud provides reliable resources, it also poses a crucial problem in allocating and managing resources dynamically across the applications.

TABLE I. INPUT PARAMETERS

Parameter	Provider	Customer
Provider Offerings	√	-
Resource Status	√	-
Available Resources	√	-
Application Requirements	-	√
Agreed Contract Between Customer and provider	√	√

From the perspective of a cloud provider, predicting the dynamic nature of users, user demands, and application demands are impractical. For the cloud users, the job should be completed on time with minimal cost. Hence due to limited resources, resource heterogeneity, locality restrictions, environmental necessities and dynamic nature of resource demand, we need an efficient resource allocation system that suits cloud environments.

Cloud resources consist of physical and virtual resources. The physical resources are shared across multiple compute requests through virtualization and provisioning [23]. The request for virtualized resources is described through a set of parameters detailing the processing, memory and disk needs which is depicted in Fig.1. Provisioning satisfies the request by mapping virtualized resources to physical ones. The hardware and software resources are allocated to the cloud applications on-demand basis. For scalable computing, Virtual Machines are rented.

The complexity of finding an optimum resource allocation is exponential in huge systems like big clusters, data centers or Grids. Since resource demand and supply can be dynamic and uncertain, various strategies for resource allocation are proposed. This paper puts forth various resource allocation strategies deployed in cloud environments.

The rest of the paper is organized as follows: In section II, a few work related to this topic is presented. Various resource allocation strategies and their impacts in cloud environments are discussed in section III. In section IV, some of the advantages and limitations of resource allocation in cloud are addressed. Finally the conclusion of the paper is given as section V.

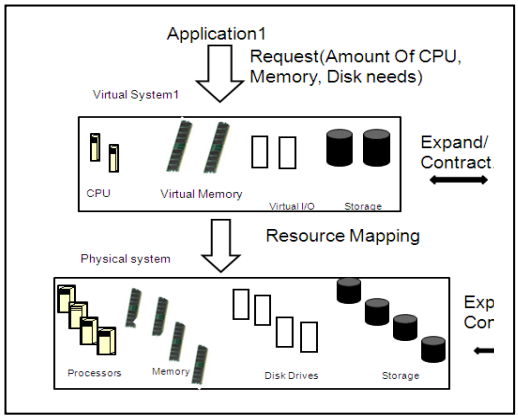

Figure1. Mapping of virtual to physical resources

II. RELATED WORK

Very little literature is available on this survey paper in cloud computing paradigm. Shikharesh et al. in paper [30] describes the resource allocation challenges in clouds from the fundamental point of resource management. The paper has not addressed any specific resource allocation strategy.

Patricia et al. [25], investigates the uncertainties that increase difficulty in scheduling and matchmaking by considering some examples of recent research.

It is evident that the paper which analyzes various resource allocation strategies is not available so far. The proposed literature focuses on resource allocation strategies and its impacts on cloud users and cloud providers. It is believed that this survey would greatly benefit the cloud users and researchers.

III. RESOURCE ALLOCATION STRATEGIES (RAS) AT A GLANCE

The input parameters to RAS and the way of resource allocation vary based on the services, infrastructure and the nature of applications which demand resources. The schematic diagram in Fig.2 depicts the classification of Resource Allocation Strategies (RAS) proposed in cloud paradigm. The following section discusses the RAS employed in cloud.

A. Execution Time

Different kinds of resource allocation mechanisms are proposed in cloud. In the work by Jiani at.al [15], actual task execution time and preemptable scheduling is considered for resource allocation. It overcomes the problem of resource contention and increases resource utilization by using different modes of renting computing capacities. But estimating the execution time for a job is a hard task for a user and errors are made very often [30]. But the VM model considered in [15] is heterogeneous and proposed for IaaS.

Using the above-mentioned strategy, a resource allocation strategy for distributed environment is proposed by Jose et al. [16]. Proposed matchmaking (assign a resource to a job) strategy in [16] is based on Any-Schedulability criteria for

assigning jobs to opaque resources in heterogeneous environment. This work does not use detailed knowledge of the scheduling policies used at resources and subjected to AR's (Advance Reservation).

B. Policy

Since centralized user and resource management lacks in scalable management of users, resources and organization-level security policy [6], Dongwan et al. [6] has proposed a decentralized user and virtualized resource management for IaaS by adding a new layer called domain in between the user and the virtualized resources. Based on role based access control (RBAC), virtualized resources are allocated to users through domain layer.

One of the resource allocation challenges of resource fragmentation in multi-cluster environment is controlled by the work given by Kuo-Chan et al. [20], which used the most-fit processor policy for resource allocation. The most-fit policy allocates a job to the cluster, which produces a leftover processor distribution, leading to the most number of immediate subsequent job allocations.

It requires a complex searching process, involving simulated allocation activities, to determine the target cluster. The clusters are assumed to be homogeneous and geographically distributed. The number of processors in each cluster is binary compatible. Job migration is required when load sharing activities occur.

Experimental results shows that the most-fit policy has higher time complexities but the time overheads are negligible compared to the system long time operation. This policy is practical to use in a real system.

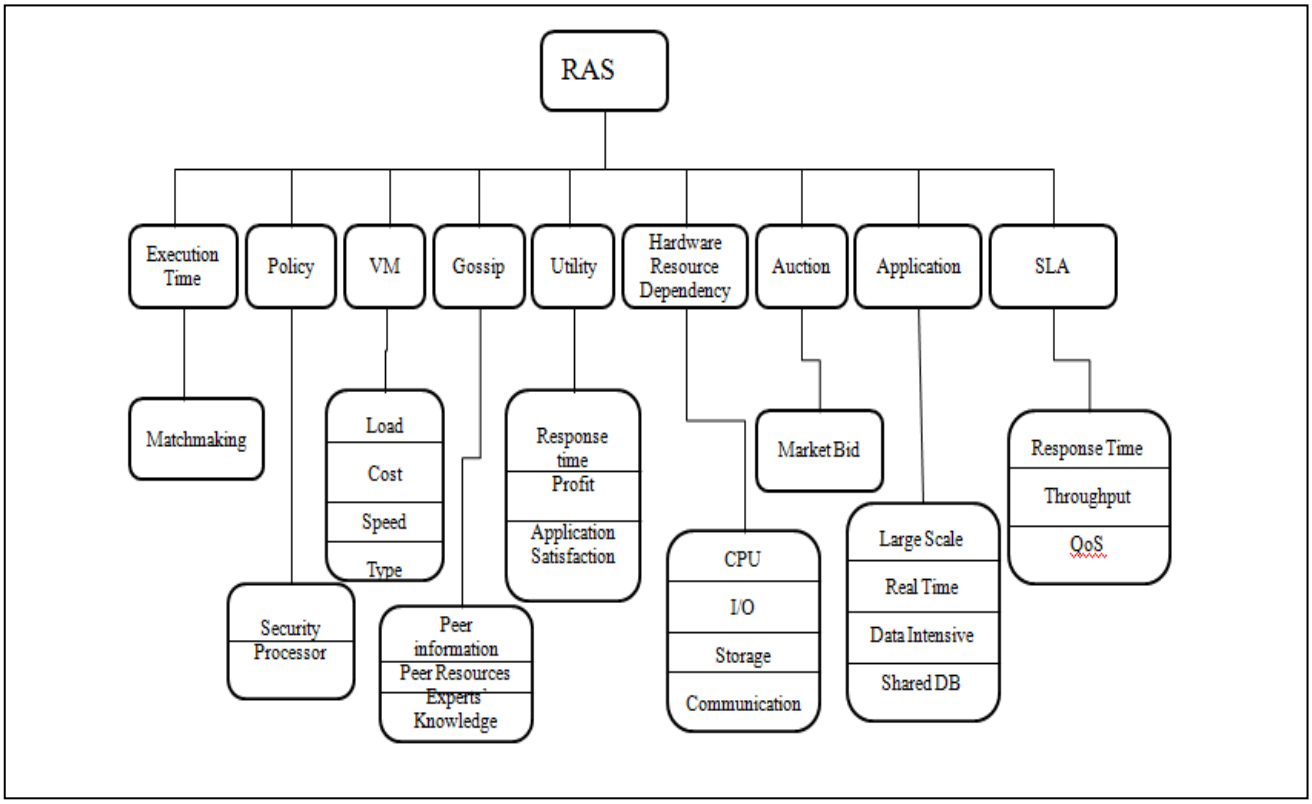

Figure2. Resource Allocation Strategies in Cloud Computing

C. Virtual Machine (VM)

A system which can automatically scale its infrastructure resources is designed in [24]. The system composed of a virtual network of virtual machines capable of live migration across multi- domain physical infrastructure. By using dynamic availability of infrastructure resources and dynamic application demand, a virtual computation environment is able to automatically relocate itself across the infrastructure and scale its resources. But the above work considers only the non-preemptable scheduling policy.

Several researchers have developed efficient resource allocations for real time tasks on multiprocessor system. But the studies, scheduled tasks on fixed number of processors. Hence it is lacks in scalability feature of cloud computing [18]. Recent studies on allocating cloud VMs for real time tasks [36], [31], [17] focus on different aspects like infrastructures to enable real-time tasks on VMs and selection of VMs for power management in the data center. But the work by Karthik et al. [18], have allocated the resources based on the speed and cost of different VMs in IaaS. It differs from other related works, by allowing the user to select VMs and reduces cost for the user.

Users can set up and boot the required resources and they have to pay only for the required resources [3]. It is implemented by enabling the users to dynamically add and/or delete one or more instances of the resources on the basis of VM load and the conditions specified by the user. The above mentioned RAS on IaaS differs from RAS on SaaS in cloud because SaaS delivers only the application to the cloud user over the internet.

Zhen Kong et al. have discussed mechanism design to allocate virtualized resources among selfish VMs in a non-cooperative cloud environment in [44]. By non-cooperative means, VMs care essentially about their own benefits without any consideration for others. They have utilized stochastic approximation approach to model and analyze QoS performance under various virtual resource allocations. The proposed stochastic resource allocation and management approaches enforced the VMs to report their types truthfully and the virtual resources can be allocated efficiently. The proposed method is very complex and it is not implemented in a practical virtualization cloud system with real workload.

D. Gossip

Cloud environment differs in terms of clusters, servers, nodes, their locality reference and capacity. The problem of resource management for a large-scale cloud environment (ranging to above 100,000 servers) is addressed in [28] and general Gossip protocol is proposed for fair allocation of CPU resources to clients.

A gossip-based protocol for resource allocation in large-scale cloud environments is proposed in [9]. It performs a key function within distributed middleware architecture for large clouds. In the thesis, the system is modeled as a dynamic set of nodes that represents the machines of cloud environment. Each node has a specific CPU capacity and memory capacity. The protocol 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 simulation results show that the protocol 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.

But in the work by Paul et al. [26] cloud resources are being allocated by obtaining resources from remote nodes when there is a change in user demand and has addressed three different policies to avoid over-provisioning and under-provisioning of resources. Recent research on sky computing focuses on bridging multiple cloud providers using the resources as a single entity which would allow elastic site for leveraging resources from multiple cloud providers [19]. Related work is proposed in [24] but it is considered only for preemptable tasks. Yang et al. [43] have proposed a profile based approach for scaling the applications automatically by capturing the experts' knowledge of scaling application servers as a profile. This approach greatly improves the system performance and resource utilization. Utility based RAS is also proposed for PaaS in [12].

In paper [8], Gossip based co-operative VM management with VM allocation and cost management is introduced. By this method, the organizations can cooperate to share the available resources to reduce the cost. Here the cloud environments of public and private clouds are considered. They have formulated an optimization model to obtain the optimal virtual machine allocation. Network game approach is adopted for the cooperative formation of organizations so that none of the organizations wants to deviate. This system does not consider the dynamic co-operative formation of organizations. Related work is discussed in [2] that use desktop cloud for better usage of computing resources due to the increase in average system utilization. The implication for a desktop cloud is that individual resource reallocation decisions using desktop consolidation and decision based on aggregate behavior of the system.

E. Utility Function

There are many proposals that dynamically manage VMs in IaaS by optimizing some objective function such as minimizing cost function, cost performance function and meeting QoS objectives. The objective function is defined as Utility property which is selected based on measures of response time, number of QoS, targets met and profit etc.

There are few works [4], [38] that dynamically allocate CPU resources to meet QoS objectives by first allocating requests to high priority applications. The authors of the papers do not try to maximize the objectives. Hence the authors' Dorian et al. proposed Utility (profit) based resource allocation for VMs which use live VM migration (one physical machine to other) as a resource allocation mechanism [7]. This controls the cost-performance trade-off by changing VM utilities or node costs. This work mainly focuses on scaling CPU resources in IaaS. A few works [1],[32] that use live migration as a resource provisioning mechanism but all of them use policy based heuristic algorithm to live migrate VM which is difficult in the presence of conflicting goals.

For multitier cloud computing systems (heterogeneous servers), resource allocation based on response time as a measure of utility function is proposed by considering CPU, memory and communication resources in [10]. HadiGoudarzi et al. characterized the servers based on their capacity of processing powers, memory usage and communication bandwidth.

For each tier, requests of the application are distributed among some of the available servers. Each available server is assigned to exactly one of these applications tiers i.e. server can only serve the requests on that specified server. Each client request is dispatched to the server using queuing theory and this system meets the requirement of SLA such as response time and utility function based on its response time. It follows the heuristics called force-directed resource management for resource consolidation. But this system is acceptable only as long as the client behaviors remain stationary.

But the work proposed in [13] considers the utility function as a measure of application satisfaction for specific resource allocation (CPU, RAM). The system of data center with single cluster is considered in [13] that support heterogeneous applications and workloads including both enterprise online applications and CPU-intensive applications. The utility goal is computed by Local Decision Module (LDM) by taking current work load of the system. The LDMs interact with Global Decision Module (GDM) and that is the decision making entity within the autonomic control loop. This system relies on a two-tier architecture and resource arbitration process that can be controlled through each application's weight and other factors.

F. Hardware Resource Dependency

In paper [35], to improve the hardware utilization, Multiple Job Optimization (MJO) scheduler is proposed. Jobs could be classified by hardware-resource dependency such as CPU-bound, Network I/O-bound, Disk I/O bound and memory bound. MJO scheduler can detect the type of jobs and parallel jobs of different categories. Based on the categories, resources are allocated. This system focuses only on CPU and I/O resource.

Eucalyptus, Open Nebula and Nimbus are typical open source frame works for resource virtualization management [39]. The common feature of these frameworks is to allocate virtual resources based on the available physical resources, expecting to form a virtualization resource pool decoupled with physical infrastructure. Because of the complexity of virtualization technology, all these frameworks cannot support all the application modes. The system called Vega LingCloud proposed in paper [39] supports both virtual and physical resources leasing from a single point to support heterogeneous application modes on shared infrastructure.

Cloud infrastructure refers to the physical and organizational structure needed for the operation of cloud. Many recent researches address the resource allocation strategies for different cloud environment. Xiaoying Wang et al. have discussed adaptive resource co-allocation approach based on CPU consumption amount in [25]. The stepwise resource co-allocation is done in three phases. The first phase determines the co-allocation scheme by considering the CPU consumption amount for each physical machine (PM). The second phase determines whether to put applications on PM or not by using simulated annealing algorithm which tries to perturb the configuration solution by randomly changing one element. During phase 3, the exact CPU share that each VM occupies is determined and it is optimized by the gradient climbing approach. This system mainly focuses on CPU and memory resources for co-allocation and does not considered the dynamic nature of resource request.

HadiGoudarzi et al. in paper [11] proposed a RAS by categorizing the cluster in the system based on the number and type of computing, data storage and communication resources that they control. All of these resources are allocated within each server. The disk resource is allocated based on the constant need of the clients and other kind of resources in the servers and clusters are allocated using Generalized Processor Sharing (GPS). This system performs distributed decision making to reduce the decision time by parallelizing the solution and used greedy algorithm to find 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.

G. Auction

Cloud resource allocation by auction mechanism is addressed by Wei-Yu Lin et al. in [37]. The proposed mechanism is based on sealed-bid auction. The cloud service provider collects all the users' bids and determines the price. The resource is distributed to the first k^{th} highest bidders under the price of the $(k+1)^{th}$ highest bid. This system simplifies the cloud service provider decision rule and the clear cut allocation rule by reducing the resource problem into ordering problem. But this mechanism does not ensure profit maximization due to its truth telling property under constraints.

The aim of resource allocation strategy is to maximize the profits of both the customer agent and the resource agent in a large datacenter by balancing the demand and supply in the market. It is achieved by using market based resource allocation strategy in which equilibrium theory is introduced (RSA-M) [41]. RSA-M determines the number of fractions used by one VM and can be adjusted dynamically according to the varied resource requirement of the workloads. One type of resource is delegated to publish the resource's price by resource agent and the resource delegated by the customer agent participates in the market system to obtain the maximum benefit for the consumer. Market Economy Mechanism is responsible for balancing the resource supply and demand in the market system.

H. Application

Resource Allocation strategies are proposed based on the nature of the applications in [33] [34]. In the work by Tram et al. [33], Virtual infrastructure allocation strategies are designed for workflow based applications where resources are allocated based on the workflow representation of the application. For work flow 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.

Real time application which collects and analyzes real time data from external service or applications has a deadline for completing the task. This kind of application has a light weight web interface and resource intensive back end [34]. To enable dynamic allocation of cloud resources for back-end mashups, a prototype system is implemented and evaluated for both static and adaptive allocation with a test bed cloud to allocate resources to the application. The system also accommodates new requests despite a-priori undefined resource utilization requirements. This prototype works by

monitoring the CPU usage of each virtual machine and adaptively invoking additional virtual machines as required by the system.

David Irwin et al. [5] have suggested the integration of high bandwidth radar sensor networks with computational and storage resources in the cloud to design end-to-end data intensive cloud systems. Their work provides a platform that supports a research on broad range of heterogeneous resources and overcomes the challenges of coordinated provisioning between sensors networks, network providers and cloud computing providers. Inclusion of nontraditional resources like Steerable sensors and cameras and stitching mechanisms to bind the resources are the requirement of this project. Resource allocation strategy plays significant role in this project.

Database replicas allocation strategy is designed in [27]. In that work, the resource allocation module divides the resource (CPU, Memory and DB replicas) allocation problem in two levels. The first level optimally splits the resources among the clients whereas the database replicas are expandable (dynamic) in the second level, based on the learned predictive model. It achieves optimal resource allocation in a dynamic and intelligent fashion.

I. SLA

In cloud, the works related to the SaaS providers considering SLA are still in their infancy. Therefore in order to achieve the SaaS providers' objective, various RAS specific to SaaS in cloud has been proposed. With the emergence of SaaS, applications have started moving away from pc based to web delivered-hosted services. Most of the RAS for SaaS focused towards customer benefits. Popovivi et al. [14] have mainly considered QoS parameters on the resource provider's side such as price and offered load.

Moreover Lee at.al [42] have addressed the problem of profit driven service request scheduling in cloud computing by considering the objectives of both parties such as service providers and consumers. But the author Linlin Wu et al. [21] have contributed to RAS by focusing on SLA driven user based QoS parameters to maximize the profit for SaaS providers. The mappings of customer requests in to infrastructure level parameters and policies that minimize the cost by optimizing the resource allocation within a VM are also proposed in [21].

Managing the computing resources for SaaS processes is challenging for SaaS providers [29]. Therefore a framework for resource management for SaaS providers to efficiently control the service levels of their users is contributed by Richard et al. [29]. It can also scale SaaS provider application under various dynamic user arrivals/departures. All the above mentioned mainly focus on SaaS providers' benefits and significantly reduce resource waste and SLO violations.

IV. ADVANTAGES AND LIMITATIONS

There are many benefits in resource allocation while using cloud computing irrespective of size of the organization and business markets. But there are some limitations as well, since it is an evolving technology. Let's have a comparative look at the advantages and limitations of resource allocation in cloud.

A. Advantages:

1) The biggest benefit of resource allocation is that user neither has to install software nor hardware to access the applications, to develop the application and to host the application over the internet.

2) The next major benefit is that there is no limitation of place and medium. We can reach our applications and data anywhere in the world, on any system.

3) The user does not need to expend on hardware and software systems.

4) Cloud providers can share their resources over the internet during resource scarcity.

B. Limitations

1) Since users rent resources from remote servers for their purpose, they don't have control over their resources.

2) Migration problem occurs, when the users wants to switch to some other provider for the better storage of their data. It's not easy to transfer huge data from one provider to the other.

3) In public cloud, the clients' data can be susceptible to hacking or phishing attacks. Since the servers on cloud are interconnected, it is easy for malware to spread.

4) Peripheral devices like printers or scanners might not work with cloud. Many of them require software to be installed locally. Networked peripherals have lesser problems.

5) More and deeper knowledge is required for allocating and managing resources in cloud, since all knowledge about the working of the cloud mainly depends upon the cloud service provider.

In Appendix A, various resource allocations strategies and their impact are listed.

V. CONCLUSION

Cloud computing technology is increasingly being used in enterprises and business markets. In cloud paradigm, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. This paper summarizes the classification of RAS and its impacts in cloud system. Some of the strategies discussed above mainly focus on CPU, memory resources but are lacking in some factors. Hence this survey paper will hopefully motivate future researchers to come up with smarter and secured optimal resource allocation algorithms and framework to strengthen the cloud computing paradigm.

REFERENCES

[1] A.Singh ,M.Korupolu and D.Mohapatra. Server-storage virtualization:Integration and Load balancing in data centers. In Proc.2008 ACM/IEEE conference on supercomputing (SC'08) pages 1-12, IEEE Press 2008.

[2] AndrzejKochut et al. : Desktop Workload Study with Implications for Desktop Cloud Resource Optimization,978-1-4244-6534-7/10 2010 IEEE.

[3] Atsuo Inomata, TaikiMorikawa, Minoru Ikebe, Sk.Md. MizanurRahman: Proposal and Evaluation of Dynamin Resource Allocation Method Based on the Load Of VMs on IaaS(IEEE,2010),978-1-4244-8704-2/11.
[4] D. Gmach, J.RoliaandL.cherkasova, Satisfying service level objectives in a self-managing resource pool. In Proc. Third IEEE international conference on self-adaptive and self organizing system.(SASO'09) pages 243-253.IEEE Press 2009 .
[5] David Irwin, PrashantShenoy, Emmanuel Cecchet and Michael Zink:Resource Management in Data-Intensive Clouds: Opportunities and Challenges .This work is supported in part by NSF under grant number CNS-0834243.
[6] Dongwan Shin and HakanAkkan :Domain- based virtualized resource management in cloud computing.
[7] Dorian Minarolli and Bernd Freisleben: Uitlity –based Resource Allocations for virtual machines in cloud computing(IEEE,2011),pp.410-417.
[8] DusitNiyato, Zhu Kun and Ping Wang : Cooperative Virtual Machine Management for Multi-Organization Cloud Computing Environment.
[9] FetahiWuhib and Rolf Stadler : Distributed monitoring and resource management for Large cloud environments(IEEE,2011),pp.970-975.
[10] HadiGoudaezi and MassoudPedram : Multidimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems IEEE 4th International conference on Cloud computing 2011,pp.324-331.
[11] HadiGoudarzi and MassoudPedram: Maximizing Profit in Cloud Computing System Via Resource Allocation: IEEE 31st International Conference on Distributed Computing Systems Workshops 2011: pp,1-6.
[12] Hien et al. ,'Automatic virtual resource management for service hosting platforms, cloud'09,pp 1-8.
[13] Hien Nguyen et al.: SLA-aware Virtual Resource Management for Cloud Infrastructures: IEEE Ninth International Conference on Computer and Information Technology 2009, pp.357-362.
[14] I.Popovici et al,"Profitable services in an uncertain world". In proceedings of the conference on supercomputing CSC2005.
[15] Jiyani et al.: Adaptive resource allocation for preemptable jobs in cloud systems (IEEE, 2010), pp.31-36.
[16] Jose Orlando Melendez &shikhareshMajumdar: Matchmaking with Limited knowledge of Resources on Clouds and Grids.
[17] K.H Kim et al. Power-aware provisioning of cloud resources for real time services. In international workshop on Middlleware for grids and clouds and e-science, pages 1-6, 2009.
[18] Karthik Kumar et al.: Resource Allocation for real time tasks using cloud computing (IEEE, 2011), pp.
[19] Keahey et al., "sky Computing",Intenet computing, IEEE,vol 13,no.5,pp43-51,sept-Oct2009.
[20] Kuo-Chan Huang &Kuan-Po Lai: Processor Allocation policies for Reducing Resource fragmentation in Multi cluster Grid and Cloud Environments (IEEE, 2010), pp.971-976.
[21] Linlin Wu, Saurabh Kumar Garg and Raj kumarBuyya: SLA –based Resource Allocation for SaaS Provides in Cloud Computing Environments (IEEE, 2011), pp.195-204 .
[22] Lizhewang,JieTao,Kunze M.,Castellanos,A.C,Kramer,D.,Karl,w, "High Performance Computing and Communications",IEEE International Conference HPCC,2008,pp.825-830.
[23] M.SuhailRehman,MajdF.Sakr : Initial Findings for provisioning Variation in Cloud Computing(IEEE,2010),pp.473-479 .
[24] P.Ruth,J.Rhee, D.Xu, R.Kennell and S.Goasguen, "Autonomic Adaptation of virtual computational environments in a multi-domain infrastructure", IEEE International conference on Autonomic Computing, 2006,pp.5-14.
[25] Patricia Takako Endo et al. :Resource allocation for distributed cloud :Concept and Research challenges(IEEE,2011),pp.42-46 .
[26] Paul Marshall, Kate Keahey& Tim Freeman: Elastic Site(IEEE,2010),pp.43-52.
[27] PenchengXiong,Yun Chi, Shenghuo Zhu, Hyun Jin Moon, CaltonPu&HakanHacigumus: Intelligent Management Of Virtualized Resources for Database Systems in Cloud Environment(IEEE,2011),pp.87-98.
[28] RerngvitYanggratoke, FetahiWuhib and Rolf Stadler: Gossip-based resource allocation for green computing in Large Clouds: 7th International conference on network and service management, Paris, France, 24-28 October, 2011.
[29] Richard T.B.Ma,Dah Ming Chiu and John C.S.Lui, Vishal Misra and Dan Rubenstein:On Resource Management for Cloud users :a Generalized Kelly Mechanism Approach.
[30] ShikhareshMajumdar: Resource Management on cloud : Handling uncertainties in Parameters and Policies (CSI communicatons,2011,edn)pp.16-19.
[31] Shuo Liu Gang Quan ShangpingRen On –Line scheduling of real time services for cloud computing. In world congress on services, pages 459-464, 2010.
[32] T.Wood et al. Black Box and gray box strategies for virtual machine migration. In Proc 4th USENIX Symposium on Networked Systems Design and Implementation (NSDI 07),pages 229-242.
[33] Tram Truong Huu& John Montagnat: Virtual Resource Allocations distribution on a cloud infrastructure (IEEE, 2010), pp.612-617.
[34] WaheedIqbal, Matthew N.Dailey, Imran Ali and Paul Janecek& David Carrera: Adaptive Resource Allocation for Back-end Mashup Applications on a heterogeneous private cloud.
[35] Weisong Hu et al. : Multiple Job Optimization in MapReduce for Heterogeneous Workloads : IEEE Sixth International Conference on Semantics, Knowledge and Grids 2010,pp.135-140.
[36] Wei-Tek Tsai Qihong Shao Xin Sun Elston, J. Service-oriented cloud computing. In world congress on services, pages 473-478, 2010.
[37] Wei-Yu Lin et al. : Dynamic Auction Mechanism for Cloud Resource Allocation: 2010 IEEE/ACM 10th International Conference on Cluster, Cloud and Grid Computing, pp.591-592.
[38] X.Zhu et al. Integrated capacity and workload management for the next generation data center. In proc.5th international conference on Automatic computing(ICAC'08),pages 172-181,IEEE Press 2008.
[39] XiaoyiLu,Jian Lin, Li Zha and ZhiweiXu: Vega Ling Cloud: A Resource Single Leasing Point System to Support Heterogenous Application Modes on Shared Infrastructure(IEEE,2011),pp.99-106.
[40] Xiaoying Wang et al. : Design and Implementation Of Adaptive Resource Co-allocation Approaches for Cloud Service Environments : IEEE 3rd International Conference on Advanced Computer Theory and Engineering 2010,V2,pp,484-488.
[41] Xindong YOU, Xianghua XU, Jian Wan, Dongjin YU:RAS-M :Resource Allocation Strategy based on Market Mechanism in Cloud Computing(IEEE,2009),pp.256-263.
[42] Y.C Lee et.al,"Project driven service request scheduling in clouds". In proceedings of the international symposium on cluster & Grid Computing.(CC Grid 2010),Melbourne, Australia.
[43] Yang wt.al A profile based approach to Just in time scalability for cloud applications, IEEE international conference on cloud computing ,2009,pp 9-16.
[44] Zhen Kong et.al : Mechanism Design for Stochastic Virtual Resource Allocation in Non-Cooperative Cloud Systems: 2011 IEEE 4th International Conference on Cloud Computing :pp,614-621.
[45] ZhixiongChen,JongP.Yoon,"International Conference on P2P, Parallel,Grid,Cloud and Internet Computing",2010 IEEE:pp 250-257

AUTHORS PROFILE

1. **Ms. Vinothina V., M.Sc. (Computer Science), M.B.A. (Systems), M.Phil.,**is presently a Senior Lecturer, Department of Computer Science, **Garden City College**, Bangalore. She has done M.Sc. Computer Science from Madurai Kamarai University, Tamil Nadu. She has completed M.B.A. with specialization in Systems from Alagappa University and M.Phil. inImage Processing from Bharathidasan University, Tamil Nadu. She has presented 4 papers in the National conference, attended many workshops and Webinars. Her areas of interest are networking and Web Technologies. She is currently pursuing Ph.D in Bharathiyar University. She may be reached at rvvino@yahoo.com.

2. Dr. Sridaran has done his post graduation in Computer Applications and Management. He has been awarded the Ph.D in Computer Applications in 2010. Having started his career as an Entrepreneur, he has offered his consultancy services to various service sectors. He has also designed and delivered various training programs in the areas of IT & Management. He has published 14 research papers in leading Journals and Conferences and presently guiding four research scholars. He has got 17 years of academic experience and served in leading educational institutions at different capacities. He is currently the Dean, Faculty of Computer Applications, Marwadi Education Foundation's Group of Institutions, Rajkot, Gujarat. He may be reached at sridaran.rajagopal@gmail.com

3. Dr. PadmavathiGanapathi is the Professor and Head of the Department ofComputer Science, Avinashilingam University. Her area of research
includes communication network, security and real time and multimediasystems. she has 170 publications at her credit.she is a life member of many professional organizations like, CSI, ISTE, ICSA, AACE, WSEAS,UWA etc. she has executed funded projects worth 2 crore from variousfunding agencies like AICTE, UGC, DRDO-NRB, DRDO-ARMREB.sheiscurrently guiding research scholars at M.Phil and Ph.D levels. She may be reached at ganapathi.padmavathi@gmail.com.

APPENDIX A

S.No	Resource Allocation Strategy	Impacts
1	Based on the estimated execution time of job .(Advanced Reservation, Best effort and immediate mode)	Estimation may not be accurate. If job could not finish its execution in estimated tine, it will affect the execution of other jobs.
2	Matchmaking strategy based on Any-Schedulability criteria.	Strategy mainly depends upon the user estimated job execution time of a job.
3	Based on role based security policy.	Follows decentralized resource allocation.
4	Most Fit Processor Policy.	Requires complex searching process and practical to use in real system.
5	Based on cost and speed of VM.	Allows the user to select VM.
6	Based on the load conditions specified by the user.	Instances of resources can be added or removed.
7	Based on gossip protocol (resources allocated by getting information for other local nodes)	It used decentralized algorithm to compute resource allocation and this prototype is not acceptable for heterogeneous cloud environment.
8	Utility function as a measure of profit based on live VM migration.	Focused on scaling CPU resources in IaaS.
9	Based on the utility function as a measure of price.	Allocate resources only in the lowest level of cloud computing and considered only CPU resource.
10.	Utility function as a measure of response time.	Lacks in handling dynamic client requests.
11	Based on utility function as a measure of application satisfaction.	Relies on two-tier architecture.
12	Based on the CPU usage of VM, active user requests are served. Adaptively new VM spawns, when the CPU usage reaches some critical point.(VR)	There is a limitation in the number of concurrent user monitor and the prototype is not capable of scaling down as the number of active user decreases.
13	Based on hardware resource dependency.	Considered only CPU and I/O resource.
14	Auction mechanism.	Not ensure profit maximization
15	Based on online resource demand predication.	Prediction may not be accurate and leads to over provisioning or under provisioning.
16	Based on workflow representation of the application.	The application logic can be interpreted and exploited to produce an execution schedule estimate. Again estimation may not be accurate.
17	Based on the machine learning technique to precisely make decisions on resources.	This prototype reduces the total SLA cost and allocate resources considering the both the request rates and also the weights.
18	Simulated annealing algorithm.	Lacks in handling dynamic resource request.
19	Based on constant needs of client and GPS.	Solution can be improved by changing the resource allocation and lacks in handling the large changes in parameters.
20	Stochastic approximation approach.	Very complex in nature.
21	Network game theory approach.	Lack in dynamic cooperative organization formation