International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012

http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-489 PAGE NO: 3340-3356

AN OVERVIEW OF RESOURCE ALLOCATION STRATEGIES IN CLOUD COMPUTING

D.LOGASAKTHI
M.Sc Student,
Department of Computer Science,
Rajeswari College of Arts and Science for Women,
Bommayapalayam, Tamil Nadu, India.
Email ID: logasakthi.d@gmail.com

Dr.A.SARANYA
Assistant Professor,
Department of Computer Application,
Rajeswari College of Arts and Science for Women,
Bommayapalayam, Tamil Nadu, India.
Email ID: drsaranyarcw@gmail.com

Abstract

Cloud computing has emerged as a cutting-edge technology with enormous commercial and enterprise potential. Apps and related data can be accessed from any location thanks to clouds. Companies can drastically lower their infrastructure costs by renting resources from the cloud for storage and other processing needs. They can also utilize pay-as-you-go applications that are accessible to the entire firm. Therefore, obtaining licenses for specific products is not necessary. Optimizing the resources that are being allocated is one of the main challenges associated with cloud computing. Resource allocation is carried out with the goal of reducing the expenses related to the model because of its uniqueness. Meeting customer requests and application requirements presents additional hurdles for resource allocation. A detailed discussion of the issues associated with different resource distribution systems is provided in this work. Both cloud users and scholars are expected to find this study useful in overcoming obstacles.

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

Introduction

A new computer paradigm called cloud computing promises to give end users dynamic, dependable, and QoS (quality of service)-guaranteed computing environments. Cloud computing is an amalgam of distributed processing, parallel processing, and grid computing. The fundamental idea behind cloud computing is that user data is kept online in data centers rather than locally. These data centers could be operated and maintained by the cloud computing service providers. Customers have unlimited access to the stored data through the Application Programming Interface (API) provided by cloud providers and any internet-connected terminal device.

In addition to storage services, commercial and general public customers can also access hardware and software services. Any kind of infrastructure, platform, or software resource can be included in the services offered by providers. Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) are the names given to each of these services appropriately.

Cloud computing offers many benefits, the most fundamental being reduced expenses, resource redistribution, and remote accessibility. By saving the business money on capital expenditures associated with leasing physical infrastructure from a third-party supplier, cloud computing reduces costs. When we need to grow our company, we can easily obtain additional resources from cloud providers because of the adaptable nature of cloud computing. We can use the cloud services at any time and from any location thanks to their remote connectivity. In order to fully profit from the aforementioned advantages, the cloud applications need have optimal resource allocation from the services provided. The importance of resource allocation is covered in the section that follows.

A. Significance of Resource Allocation

Allocating available resources to required cloud applications over the internet is known as resource allocation, or RA, in the context of cloud computing. If resources are not allocated and managed properly, services will suffer. By enabling the service providers to oversee the resources for each distinct module, resource provisioning addresses that issue.

In order to meet the demands of the cloud application, Resource Allocation Strategy (RAS) integrates cloud provider activities for the purpose of using and assigning scarce resources within the constraints of the cloud environment. It needs the kind and quantity of resources that each program needs to finish a task for the user. For an ideal RAS, the sequence and timing of resource allocation are also inputs. The ensuing conditions should be avoided by an ideal RAS:

- ✓ When two apps attempt to access the same resource at the same time, resource contention conditions rise.
- ✓ When there are insufficient resources, a shortage of resources occurs.
- ✓ The condition of resource fragmentation occurs when resources are not easily accessible. [There won't be enough intelligence to devote to the intended application, but there will be sufficient resources.]
- ✓ When allocated tasks receive more spare resources than needed, this is known as over-provisioning of resources.
- ✓ When tasks assigned to users have less resources than required, this is known as under-provisioning of resources.

In order to finish a task ahead of schedule, resource users (cloud users) may anticipate resource demands, which could result in overprovisioning of resources. A shortage of resources could result from the distribution of resources by resource providers. To resolve these differences, data from cloud vendors and customers for a RAS are needed, as table I shows.

From the perspective of the cloud user, the Service Level Agreement (SLA) and application requirement are the two main factors into RAS. The other side's inputs—offers, resource status, and available resources—are needed for RAS to handle and distribute resources to host applications [22]. Any ideal RAS's output must meet requirements for response time, latency, and flow. While cloud computing offers dependable resources, it also presents a significant challenge for dynamic resource allocation and management among apps.

Table I. Input Variables

Factors	Supplier	Consumer
Vendor Products and Services	✓	-
Condition of Resources	✓	-
Accessible Resources	✓	-
Application Requirements	-	✓
Agreement Between the Provider and the Client	✓	✓

Forecasting user behavior, application demands, and user dynamics in advance is unfeasible from the cloud provider's point of view. The task ought to be finished quickly and affordably for cloud users. So, we require an effective resource allocation system that works in cloud environments because of resource scarcity, heterogeneity, locality constraints, environmental requirements, and the dynamic nature of resource demand.

Virtual and physical resources make up cloud resources. Through virtualization and provisioning, the physical resources are shared among several computing requests. A series of parameters that specify the processor, memory, and disk requirements are used to define the request for virtualized resources; this is shown in Fig. 1. By connecting virtualized resources to physical ones, provisioning fulfills the request. Hardware and software resources are allotted to the cloud apps according to demand. Distributed computing is achieved via renting virtual machines.

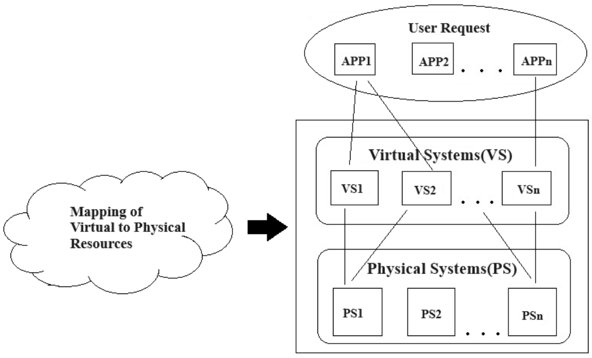


Figure 1. Virtual to Physical Resources

Large systems such as massive clusters, data centers, or grids have exponentially more complexity in determining the best resource allocation. Different solutions are recommended for resource distribution since the supply and demand of resources might be unpredictable and dynamic. Several resource allocation techniques used in cloud systems are presented in this study.

The following is how the remainder of the paper is structured: A few works on this subject are included in Part II. Section III discusses several resource allocation techniques and how they affect cloud systems. A few advantages and disadvantages of cloud resource allocation are covered in Section IV. In section V, the paper's conclusion is finally provided.

II. Related Work

In the cloud computing paradigm, there is a dearth of available material on this survey paper. From a basic resource management perspective, Shikharesh et al.'s study [27] explains the difficulties associated with resource allocation in clouds. None of the resource allocation strategies have been covered in this work.

Patricia et al. [22] examine several examples of current research to investigate the inconsistencies that contribute to the difficulty of planning and arranging matches.

It is clear that there is currently no publication available that assesses different strategies for allocating resources. The recommended readings focus on resource distribution strategies and their effects on clients and cloud service providers. This poll is anticipated to be very helpful to researchers and cloud users.

III. Resource Allocation Strategies (RAS):

A variety of RAS input parameters and resource distribution algorithms are applicable, contingent upon the services, infrastructure, and application types requiring resources. The categorization of Resource Allocation Strategies (RAS) suggested in the cloud paradigm is shown schematically in Fig. 2. The RAS used in the cloud is covered in the section that follows.

A. Execution Time

In the cloud, many resource allocation strategies are offered. Real task execution time and preemptable scheduling are taken into account for resource allocation in the study by Jiani et al. [14]. By employing various methods of renting computing capacities, it eliminates the issue of resource contention and improves resource use. However, users frequently make mistakes when predicting a job's execution time [27]. However, the VM paradigm suggested for IaaS in [14] is heterogeneous.

Jose et al. [15] suggest a resource allocation strategy for dispersed environments based on the previously discussed approach. Any-Schedulability criteria is the foundation of the proposed matching (assign a resource to a job) technique in [15] for allocating jobs to opaque resources in a heterogeneous environment. This job does not require in-depth familiarity with the resource scheduling policies that are subject to advance reservations, or ARs.

B. Policy

For Infrastructure as a Service (IaaS), Dongwan et al. [5] have proposed a distributed user and virtualized resource management by introducing an extra layer called a domain between the client and the virtualized resources. This is because centralized user and resource management lacks flexible management of users, resources, and organization-level security policy [5]. Role-based access control, or RBAC, governs how users have access to virtualized resources through the domain layer.

According to Kuo-Chan et al.'s study [19], the most-fit processor method is employed to reduce the resource allocation challenges caused by resource segmentation in a multi-cluster setting. The cluster receives a job from the most-fit policy, which results in a leftover processor distribution and the greatest number of following job allocations.

The target cluster is found by a laborious searching procedure that includes simulated allocation activities. It is assumed that the clusters are spread geographically and equal. Each cluster's processor count is binary compatible. When load sharing activities take place, job migration is necessary.

The findings of the experiment indicate that while the most-fit policy has greater time complexities, the time overheads are insignificant when compared to the system's extended operation. Using this policy in a real system makes sense.

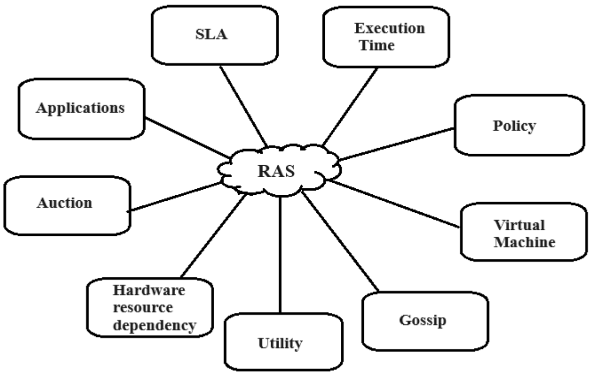


Figure 2: Resource Allocation Strategies (RAS)

C. Virtual Machine (VM)

It is intended in [21] to have an infrastructure resource scalability mechanism. With the ability to migrate live across physical infrastructure across many domains, the system is comprised of a virtual network of virtual machines. Virtual compute environments can automatically migrate themselves across the infrastructure and scale their resources by leveraging dynamic application demand and dynamic infrastructure resource availability. Only the non-preemptable scheduling policy is taken into account in the work mentioned above.

Efficient resource allocations for real-time workloads on multiprocessor systems have been created by several researchers. However, the research used fixed-number processors for scheduled tasks. Thus, cloud computing's scalability aspect is lacking [17].

Research on cloud virtual machine (VM) allocation for real-time applications is now focused on a number of factors, including infrastructures to support VM real-time jobs and VM selection for data center power consumption [33], [28], and [16]. Yet, Karthik et al.'s research [17] divided up the resources according to the cost and speed of various VMs under IaaS. Because it lets the user choose which virtual machines to use and saves money, it varies from other similar works.

The resources that are needed can be installed and booted by users, and they only need to pay for those. To implement it, users can add, remove, or modify instances of resources dynamically based on the load of the virtual machine and user-specified parameters Since SaaS only offers the application to cloud customers via the internet, the RAS on SaaS in cloud computing differs from the RAS on IaaS as previously discussed.

In a non-cooperative cloud setting, Zhen Kong et al. have explored the creation of mechanisms to distribute virtualized resources across self-serving virtual machines [40]. By using non-cooperative tactics, VMs only really care about their personal gain, disregarding the needs of others. They have used the random approximation technique to simulate and analyze QoS performance under different virtual resource allocations. To ensure the best possible distribution of virtual resources, the recommended randomized resource allocation and management strategies required the virtual machines (VMs) to provide proper type reports. Since it is not implemented in a genuine workload virtualization cloud system, the suggested solution is extremely difficult.

D. Gossip

Clusters, servers, nodes, their location reference, and capacity are all different in a cloud context In [25].A universal protocol called Gossip is proposed to guarantee a fair sharing of CPU resources among clients, solving the problem of resource management for massive cloud platforms (>100,000 servers).In large-scale cloud systems, a gossip-based resource allocation protocol is presented in [8]. In big cloud distributed middleware architecture, it serves an important role. The system that depicts the computers in a cloud environment is depicted in the thesis as a dynamic set of nodes. There is a fixed CPU and memory capacity for every node.

The simulation results demonstrate 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; however, additional functionalities are required to make the resource allocation scheme robust to machine failure which spans several clusters and datacenters. Through the implementation of a distributed strategy that allocates cloud resources to a group of applications with time-dependent memory demands, the protocol continuously enhances a global cloud utility function.

To prevent over- and under-provisioning of resources, Paul et al.'s work [23] addresses three distinct policies and allocates cloud resources by getting resources from remote nodes in response to changes in user demand. Elastic sites would be able to leverage resources from different cloud providers by linking several cloud providers employing resources as a single entity, according to recent research on sky computing [18]. Although related work is suggested in [21], it is only taken into account for tasks that can be completed beforehand. A profile-based method for automatically scaling applications has been developed by Yang et al. [39]. It involves encoding the application server scaling expertise of experts into a profile. The system's performance and resource usage are significantly enhanced by this method. Utility-based RAS for PaaS is also proposed in [11].

In the work [7], gossip-based virtual machine allocation and cost maintenance is presented as a unique method of cooperative virtual machine management. Through this approach, the entities can collaborate to distribute the accessible resources in order to save expenses. Here, both public and private cloud infrastructures are taken into account. To determine the ideal virtual machine allocation, they have created an optimization model. To ensure that no organization desires to stray, a cooperative organization building process using the network game technique is used. The flexible cooperative structure of organizations is not taken into account by this system. As a result of an increase in average system utilization, related work that uses desktop clouds for better computing resource use is covered in [2]. Applying desktop consolidation and making decisions based on the overall system behavior, individual resource reallocation decisions for a desktop cloud are implied.

E. Utility Function

Numerous suggestions aim to optimize specific target functions, such as cost performance, QoS, and cost minimization, in order to dynamically manage virtual machines in infrastructure as a service. Utility property is the definition of the objective function, and it is chosen based on metrics like profit, number of QoS, targets reached, and reaction time.

Few works (e.g., [3], [35]) assign CPU resources dynamically to prioritize high priority applications before meeting QoS targets. The authors of the articles don't make an effort to optimize the goals. As a result, Dorian et al. proposed a utility-based (profit-based) approach to resource allocation for virtual machines (VMs) that makes use of live VM transfer (from one actual computer to another) [6]. This balances the trade-off between performance and cost by changing node fees or virtual machine utilities. The main objective of this work is to scale CPU resources in IaaS. A few articles [1], [29] employ live migration as a resource provisioning technique, but they all use policy-based heuristic algorithms to live transfer virtual machines (VMs), which is difficult to do when competing goals are involved.

CPU, memory, and transmission resources are taken into account in a proposed resource allocation technique for multitier cloud computing systems (heterogeneous servers), that emphasizes on utility function measurement using response time [9]. The computing capability, memory utilization, and communication bandwidth of the servers were assessed by HadiGoudarzi et al.

A portion of the available servers are assigned to handle the application's requests for each tier. There is only one of these application tiers assigned to each available server, meaning that the server can only fulfill requests made to that particular server. By utilizing queuing theory to assign a distinct server request to every client and by offering a utility function that is reliant on response time, the system complies with SLA standards. When consolidating resources, it employs algorithms known as force-directed resource management. However, the client actions must remain consistent for this method to be considered acceptable.

However, for particular resource allocation (CPU, RAM, the methodology offered in [12] considers the utility function as a program acceptance measurement. One data center system that supports heterogeneous workloads and applications, including both CPU-intensive and enterprise web applications, is the single cluster system that is examined in [12]. By taking into account the system's present workload, the Local Decision Module (LDM) calculates the utility goal. As the decision-making body in the autonomic control loop, the Global Decision Module (GDM) communicates with the LDMs. The weight of each application and other variables can be used to regulate the resource arbitration procedure and two-tier architecture of this system.

F. Hardware Resource Dependency

The Multiple Job Optimization (MJO) scheduler is suggested in paper [32] as a way to increase hardware utilization. Jobs that depend on hardware resources, such as those that are memory-bound, CPU-bound, network-bound, or disk-bound, can be categorized. The type of jobs and continuous jobs of various types can be detected by the MJO scheduler. The classifications determine how resources are allocated. This machine is only equipped for managing CPU and I/O resources.

Common free-source frameworks for resource virtualization management include Nimbus, Open Nebula, and Eucalyptus [36]. In order to create a virtualization resource pool that is independent of physical infrastructure, these frameworks all share the characteristic of allocating virtual resources depending on the physical resources that are currently available. Since virtualization technology is so complicated, not all application modes can be supported by all of these frameworks. According to a report [36], a system named Vega LingCloud facilitates the leasing of both virtual and physical resources from a single point of contact, enabling varied application modes on shared infrastructure.

The term "cloud infrastructure" describes the organizational and physical framework required to run a cloud. Numerous recent studies discuss resource allocation techniques for various cloud environments. Xiaoying Wang et al. have addressed a flexible resource co-allocation technique based on the CPU utilization quantity in [37].

There are three stages involved in the stepwise resource co-allocation. The initial stage establishes the co-allocation plan by taking into account the CPU consumption quantity for every physical machine (PM). The second stage uses a simulated annealing technique that intends to simulate the configuration solution by altering one element at random to decide whether or not to place apps on PM. Phase 3 involves determining the precise CPU share that each virtual machine (VM) uses, and the gradient climbing technique is used to improve it. This approach ignores the dynamic nature of resource requests in favor of co-allocating CPU and memory resources.

According to the quantity and kind of processing, data storage, and communication resources that each cluster in the system controls, HadiGoudarzi et al. classified the cluster in their work [10] to suggest a RAS. Within every server, certain resources are allotted. Generalized Processor Sharing (GPS) is used to allocate additional resources among the servers and clusters, while the disk resource is distributed according to the ongoing needs of the clients. By parallelizing the solution and using a greedy approach to obtain the best initial solution, this system reduces decision time through distributed decision making. By allocating resources differently, the solution could be improved. However, significant changes to the parameters that are employed to identify the solution are beyond the capabilities of this system.

G. Auction

Wei-Yu Lin et al. address the topic of cloud resource allocation using auction mechanism in [34]. A sealed-bid auction serves as the foundation for the suggested method. After compiling every bid from every user, the cloud service provider sets the price. The resource is awarded to the top k bidders at the (k+1)th highest bid price. By transforming the resource problem into an ordering problem, this solution streamlines the decision-making process for cloud service providers and the clear-cut allocation rule. However, because of its truth-telling ability when operating under limitations, this system does not guarantee profit maximization.

By balancing supply and demand in the market, the distribution of resources approach aims to optimize profitability for both resource agents and customer agents in a big datacenter. It is done by implementing a market-based resource allocation approach that introduces equilibrium theory (RSA-M). The amount of fractions needed by a single virtual machine (VM) is determined by RSA-M, which can be dynamically modified based on the different resource requirements of the workloads. A resource is assigned by a resource agent to publish its pricing, and a customer agent assigns a resource to engage in the market system in order to maximize benefits for the consumer. The market economy mechanism balances the supply and demand for resources in the marketplace system.

H. Application

Based on the types of applications, resource allocation solutions are suggested in [30] [31]. Tram et al.'s work establishes virtual infrastructure allocation algorithms for workflow-based applications, in which resources are allotted according to the application's workflow representation. [30]. An estimate of the execution schedule can be generated for work flow-based applications by interpreting and using the application logic. This facilitates the user's evaluation of the exact resource usage for each application run. Resources are allocated and computing jobs are scheduled using four strategies: Naive, FIFO, Optimized, and Services Group Optimization.

There is a deadline for finishing a real-time application that gathers and evaluates data in real-time from external services or apps. This type of program features a back end that requires a lot of resources and a lightweight web interface [31]. A feasible model is created and assessed for both fixed and versatile allocation using a test bed cloud to assign resources to the application in order to enable dynamic cloud resource allocation for back-end mashups. Despite a-priori undefinable resource consumption requirements, the system also supports new requests. This prototype operates by keeping track of how much CPU time each virtual machine uses and dynamically calling in more virtual machines when the system calls for them.

In order to construct end-to-end data demanding cloud systems, High bandwidth radar sensor networks coupled with cloud-based computing and storage power have been proposed by David Irwin et al. [4]. Their work overcomes the difficulties of coordinated provisioning across sensor networks, network providers, and cloud computing providers, and offers a platform that supports research on a wide range of heterogeneous resources. This project requires the use of non-traditional resources, such as stitching techniques to link the resources and steerable sensors and cameras. In this project, the approach for allocating resources is important.

[24] is the design of the database replicas allocation mechanism. The resource allocation component divides the resource allocation problem (CPU, Memory, and DB replicas) into two phases in that job. While the second level's database copies are expandable (dynamic) depending on the learnt predictive model, the first level divides resources among the clients in an optimal manner. Through dynamic and intelligent means, it achieves optimal resource allocation.

I. SLA

SaaS providers' consideration of SLA in the cloud is still in its early stages of development. Consequently, many SaaS-specific RAS have been proposed for cloud SaaS in order to meet the goal of the SaaS providers. Web-hosted services have replaced PC-based ones in applications since the advent of Software as a Service (SaaS). Customer benefits were the primary emphasis of the majority of SaaS RAS. Popovivi et al. [13] primarily examined QoS characteristics related to the resource supplier, including proposed load and price.

Furthermore, Lee et al. [38] have tackled the issue of profit-driven scheduling of service requests in cloud computing by taking into account the goals of both consumers and service providers. However, by concentrating on SLA-driven, user-based QoS parameters in order to maximize profitability for SaaS providers, the author Linlin Wu et al. [20] have contributed to RAS. In [20], it is also suggested to tie client demands to infrastructure-level parameters and policies that optimize resource allocation within virtual machines (VMs) in order to decrease costs.

For SaaS providers, managing the computer resources for SaaS procedures is difficult [26]. Richard et al. [26] thus give a paradigm for resource management that enables SaaS providers to effectively regulate the service levels of their users. Furthermore, SaaS provider applications can be expanded under various dynamic user arrivals and departures While focusing particularly on the benefits of SaaS providers, all of the prior solutions significantly reduce resource waste and SLO breaches.

IV. Advantages and Limitations

Regardless of an organization's size or its target market, there are numerous advantages to resource allocation when utilizing cloud computing. Since this is a developing technology, there are certain restrictions as well. Let's examine side by side the benefits and drawbacks of cloud resource allocation.

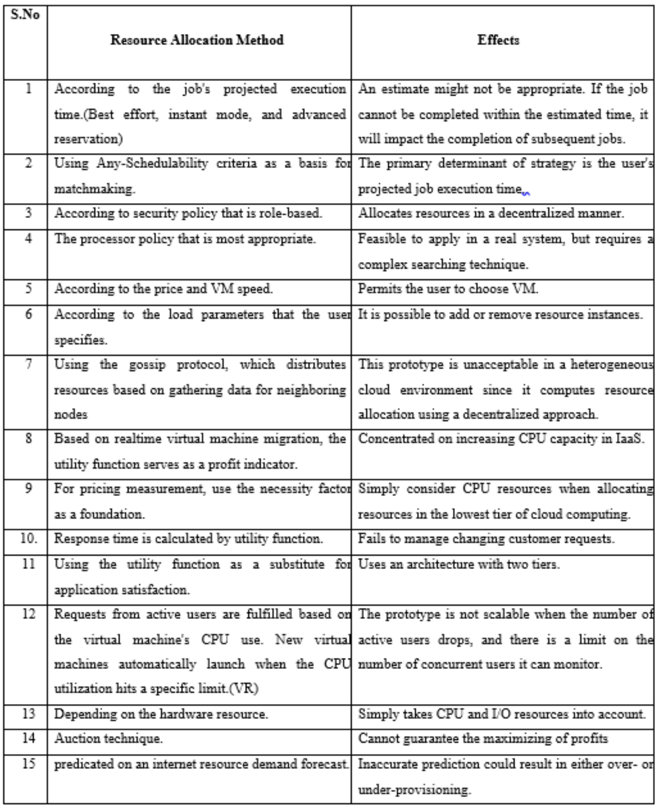
A. Advantages:

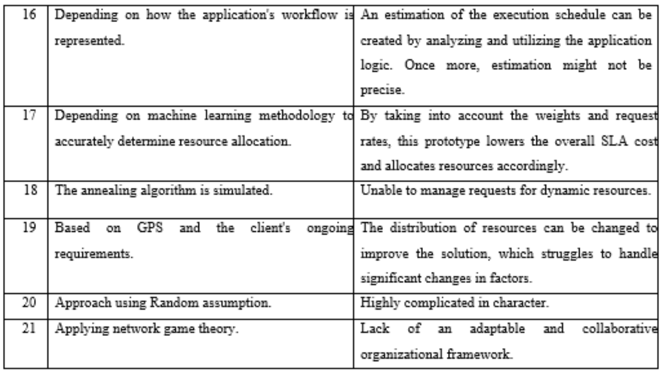
1. The primary advantage of resource allocation is that users can develop, host, and access applications over the internet without the need to install any hardware or software.
2. The lack of restrictions on location or media is the next significant advantage. On any machine, we can access our apps and data from anywhere in the globe.
3. Hardware and software systems do not require the user to spend money.
4. When there is a shortage of resources, cloud providers might share those resources online.

B. Limitations:

1. Users do not own the resources they utilize because they are rented from distant servers for their intended usage.
2. When a user wants to move to a different provider for better data storage, a migration problem arises. Large-scale data transfers between providers are difficult.
3. Data from clients may be vulnerable to phishing or hackers in public clouds. Malware may spread easily on cloud servers due of their interconnection.
4. It's possible that peripherals like printers and scanners won't function with the cloud. For many of them, local software installation is necessary. Peripherals connected to a network face fewer issues.
5. Since all information regarding how the cloud operates mostly depends on the cloud service provider, more and deeper expertise is needed for resource allocation and management.

Various ways for allocating resources and their effects are listed in below:





S.No	Resource Allocation Method	Effects
1	According to the job's projected execution time.(Best effort, instant mode, and advanced reservation)	An estimate might not be appropriate. If the job cannot be completed within the estimated time, it will impact the completion of subsequent jobs.
2	Using Any-Schedulability criteria as a basis for matchmaking.	The primary determinant of strategy is the user's projected job execution time.
3	According to security policy that is role-based.	Allocates resources in a decentralized manner.
4	The processor policy that is most appropriate.	Feasible to apply in a real system, but requires a complex searching technique.
5	According to the price and VM speed.	Permits the user to choose VM.
6	According to the load parameters that the user specifies.	It is possible to add or remove resource instances.
7	Using the gossip protocol, which distributes resources based on gathering data for neighboring nodes	This prototype is unacceptable in a heterogeneous cloud environment since it computes resource allocation using a decentralized approach.
8	Based on realtime virtual machine migration, the utility function serves as a profit indicator.	Concentrated on increasing CPU capacity in IaaS.
9	For pricing measurement, use the necessity factor as a foundation.	Simply consider CPU resources when allocating resources in the lowest tier of cloud computing.
10.	Response time is calculated by utility function.	Fails to manage changing customer requests.
11	Using the utility function as a substitute for application satisfaction.	Uses an architecture with two tiers.
12	Requests from active users are fulfilled based on the virtual machine's CPU use. New virtual machines automatically launch when the CPU utilization hits a specific limit.(VR)	The prototype is not scalable when the number of active users drops, and there is a limit on the number of concurrent users it can monitor.
13	Depending on the hardware resource.	Simply takes CPU and I/O resources into account.
14	Auction technique.	Cannot guarantee the maximizing of profits
15	predicated on an internet resource demand forecast.	Inaccurate prediction could result in either over- or under-provisioning.
16	Depending on how the application's workflow is represented.	An estimation of the execution schedule can be created by analyzing and utilizing the application logic. Once more, estimation might not be precise.
17	Depending on machine learning methodology to accurately determine resource allocation.	By taking into account the weights and request rates, this prototype lowers the overall SLA cost and allocates resources accordingly.
18	The annealing algorithm is simulated.	Unable to manage requests for dynamic resources.
19	Based on GPS and the client's ongoing requirements.	The distribution of resources can be changed to improve the solution, which struggles to handle significant changes in factors.
20	Approach using Random assumption.	Highly complicated in character.
21	Applying network game theory.	Lack of an adaptable and collaborative organizational framework.

V. Conclusion

Companies and commercial markets are using cloud computing technologies more and more. In the cloud paradigm, optimizing user satisfaction and cloud service providers' profits requires an efficient resource allocation plan. The classification of RAS and its effects on cloud systems are outlined in this research. While several of the above-discussed solutions concentrate primarily on CPU and memory resources, they are deficient in certain areas. In order to strengthen the cloud computing paradigm, it is hoped that this survey article will inspire future academics to develop more intelligent and secure optimal resource allocation algorithms and frameworks.

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. 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.

4. 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.
5. Dongwan Shin and HakanAkkan: Domain- based virtualized resource management in cloud computing.
6. Dorian Minarolli and Bernd Freisleben: Uitlity –based Resource Allocations for virtual machines in cloud computing (IEEE, 2011), pp.410-417.
7. DusitNiyato, Zhu Kun and Ping Wang: Cooperative Virtual Machine Management for Multi-Organization Cloud Computing Environment.
8. FetahiWuhib and Rolf Stadler: Distributed monitoring and resource management for Large cloud environments (IEEE, 2011), pp.970-975.
9. 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.
10. 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.
11. Hien et al. ,'Automatic virtual resource management for service hosting platforms, cloud'09,pp 1-8.
12. 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.
13. I.Popovici et al,"Profitable services in an uncertain world". In proceedings of the conference on supercomputing CSC2005.
14. Jiyani et al.: Adaptive resource allocation for preemptable jobs in cloud systems (IEEE, 2010), pp.31-36.
15. Jose Orlando Melendez &shikhareshMajumdar: Matchmaking with Limited knowledge of Resources on Clouds and Grids.
16. 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.
17. Karthik Kumar et al.: Resource Allocation for real time tasks using cloud computing (IEEE, 2011), pp.

18. Keahey et al.,"sky Computing",Intenet computing, IEEE,vol 13,no.5,pp43-51,sept-Oct2009.
19. 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.
20. Linlin Wu, Saurabh Kumar Garg and Raj kumarBuyya: SLA –based Resource Allocation for SaaS Provides in Cloud Computing Environments (IEEE, 2011), pp.195-204.
21. 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.
22. Patricia Takako Endo et al.:Resource allocation for distributed cloud Concept and Research challenges (IEEE,2011), pp.42-46.
23. Paul Marshall, Kate Keahey& Tim Freeman: Elastic Site (IEEE, 2010), pp.43-52.
24. PenchengXiong,Yun Chi, Shenghuo Zhu, Hyun Jin Moon, CaltonPu & Hakan Hacigumus: Intelligent Management Of Virtualized Resources for Database Systems in Cloud Environment (IEEE,2011), pp.87-98.
25. 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.
26. 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.
27. ShikhareshMajumdar: Resource Management on cloud: Handling uncertainties in Parameters and Policies (CSI communicatons, 2011,edn)pp.16-19.
28. Shuo Liu Gang Quan ShangpingRen On –Line scheduling of real time services for cloud computing. In world congress on services, pages 459- 464, 2010.
29. 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.
30. Tram Truong Huu& John Montagnat: Virtual Resource Allocations distribution on a cloud infrastructure (IEEE, 2010), pp.612-617.
31. WaheedIqbal, Matthew N.Dailey, Imran Ali and Paul Janecek& David Carrera: Adaptive Resource Allocation for Back-

end Mashup Applications on a heterogeneous private cloud.

32. 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.
33. Wei-Tek Tsai Qihong Shao Xin Sun Elston, J. Service-oriented cloud computing. In world congress on services, pages 473-478, 2010.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.