An Efficient Auction Based Resource Allocation Algorithm in Cloud Computing

Thesis submitted in partial fulfilment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By

Richa Goyal
(Roll No. 801332021)

Under the supervision of:

Dr. Anju Bala
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2015

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*An Efficient Auction Based Resource Allocation Algorithm in Cloud Computing*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Anju Bala* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Signature:

(Richa Goyal)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Anju Bala)
Assistant Professor,
Computer Science and Engineering Department

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. S. Bhatia)
Dean (Academic Affairs)
Thapar University
Patiala

ACKNOWLEDGEMENT

No volume of words is enough to express my gratitude towards my guide **Dr. Anju Bala,** Assistant Professor, Computer Science & Engineering Department, for providing her immense help, guidance, simulating suggestions and encouragement all the time. She has helped me to explore this vast topic in an organized manner and provided me all the ideas on how to work towards a research-oriented venture.

I am also thankful to **Dr. S. S. Bhatia**, Dean of Academic Affairs, **Dr. Deepak Garg**, Head of Computer Science & Engineering Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, for the motivation and inspiration that triggered me for the thesis work.

I would also like to thank the staff members and my colleagues who were always there at the need of hour and provided with all the help and facilities, which I required, for the completion of my thesis work.

Most importantly, I would like to thank my parents and the almighty for showing me the right direction out of the blue, to help me stay calm in the oddest of the times and keep moving even at times when there was no hope.

(Richa Goyal)
801332021

ABSTRACT

Cloud Computing provides services to users through internet. Services can be offered on demand by the Cloud providers through pay-as-you-go model. Three layers of Cloud Computing offer different type of services according to requirements. The main feature of Cloud is that it is scalable and can expand and contract anytime. Users need not to invest in infrastructure. Other advantages of Cloud are performance, multi-tenancy and unlimited storage space.

Currently Cloud Computing faces many challenges like fault tolerance, security and resource allocation. Resource allocation is one of the key issues in Cloud environment. Various authors have proposed many techniques for market based resource allocation which considers current market scenarios in which trading of resources take place. Buyers demand resources and seller sells resources. This research work focus on developing efficient auction based resource allocation technique taking into consideration market scenarios and developing dynamic pricing model for on demand scenarios of resource allocation.

This algorithm has been proposed for resource allocation which allocates resources based on auction model. It also calculates dynamic pricing which helps generating revenue to the provider. Users pay lower price as compared to fixed priced model. The algorithm is beneficial both for user and Cloud provider as there is double sided competition. It is evaluated using CloudSim.

Table of Contents

List of Figures and Tables

Tables:

Chapter 1
Introduction

This chapter describes about evolution of Cloud Computing, its characteristics, layers and deployment models. Further, it describes the research issues in Cloud and also organisation of thesis is presented at the end of the chapter.

1.1 Background

The term Cloud came from the telecommunication world when VPN (Virtual Private Network) service was used for data communication." Cloud Computing is a model for convenient, on-demand network access to shared resources that can be used and then released without much effort." The growth of Cloud Computing is modifying the organizations' purchase and manages their computing resources and provides a different model in which various IT activities including maintenance, backup, security, data storage are provided by Cloud providers. It deals with data storage, data access, computation and software where users do not know physical location of the system that is providing the services.

Cloud Computing is the latest trend in the IT which leads to movement of data and computational processes to large data centres from desktop and portable PCs. Now, applications can be delivered as a service over the network which decreases overall cost. The main motive of Cloud Computing is better utilization of distributed resources to achieve higher throughput and solves problem of large scale computation. End-user based on his requirements access different services without actually knowing where the services are hosted and how services are being delivered.

1.2 Cloud Computing Evolution

In sixties J.C.R. Licklider had presented the idea of delivering computing resources through the use of global network. The idea of global network came before the world in the form of internet in 1969 which was actually a research project at Advanced Research Projects Agency (ARPA) under the ministry of Defence, United State (MoD, US). The internet was used in the beginning mainly for scientific and military

applications. By the starting of the year 1988, internet became commercialised by providing services like email and telnet. All the services provided by Cloud Service Providers (CSP) are mainly provided through internet. Many experts believe that the idea of Cloud Computing was the vision of American computer scientist John McCarthy of MIT (Massachusetts Institute of Technology) that he presented in sixties. According to him, "computation can be delivered as a public utility" [1]. In the last sixty years, there have been many evolutions in the history of computers from very large size computers in 1970s to small PCs in 1999. Then came the period of distributed computing. In 2010 companies took an initiative to migrate from centralised computing to distributed computing in which they started using compact machines for computational purposes. This cost effective method of fulfilling computational needs was called Cloud Computing. Through the use of Cloud Computing, Computing-as-a-Service can be delivered by using internet [2][3].

Cloud Computing has evolved through various phases which involves Grid Computing, Utility computing and Software-as-a-Service (SaaS) as shown in Figure 1.1. Grid Computing is the accumulation of different computing resources which are distributed and heterogeneous resources collected from multiple sites with the purpose of accomplishing a common task.

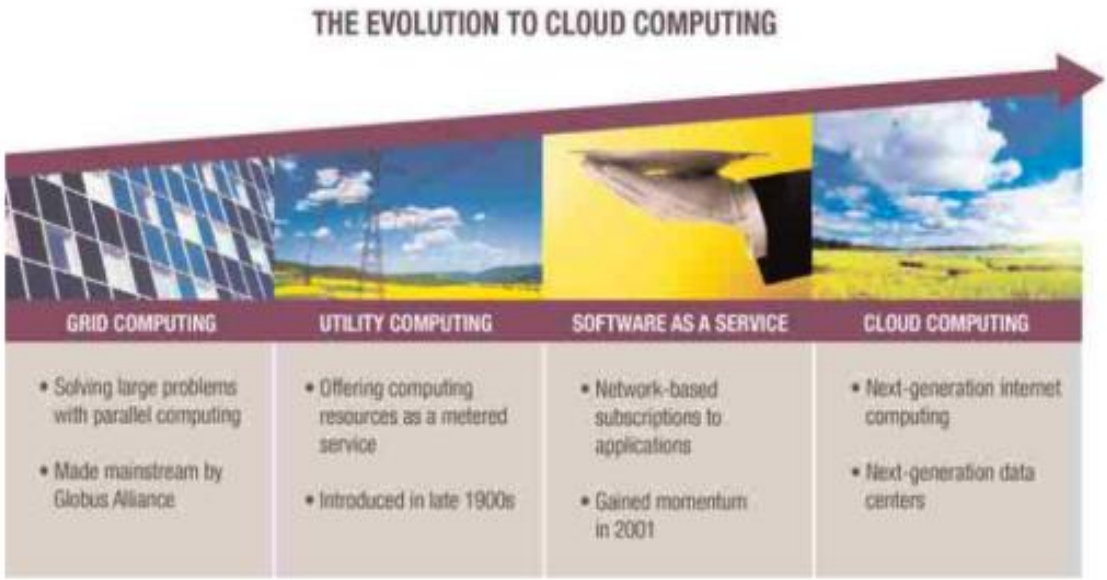


Figure 1.1: Evolution of Cloud Computing [4]

But the limitation of Grid Computing is its limited scope of research work and scientific applications [5]. Utility computing [6] is the providing metered services based on the consumption of the users liking traditional electricity consumption. So it can be said that Cloud has evolved from Utility and Grid Computing.

In 2010 third phase of the Cloud Computing "Software-as-a-Service" (SaaS) evolved in which both data and application are stored on the server and by using internet, users can connect themselves to the remote server. SaaS is only a part of Cloud services [10]. Other services provided by Cloud Computing are Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). There was no standard definition for Cloud Computing, so National Institute of Standards and Technology (NIST) has proposed a definition in 2011 which includes deployment models and essential characteristics of Cloud. Figure 1.2 demonstrates aspects of Cloud Computing as proposed by NIST.

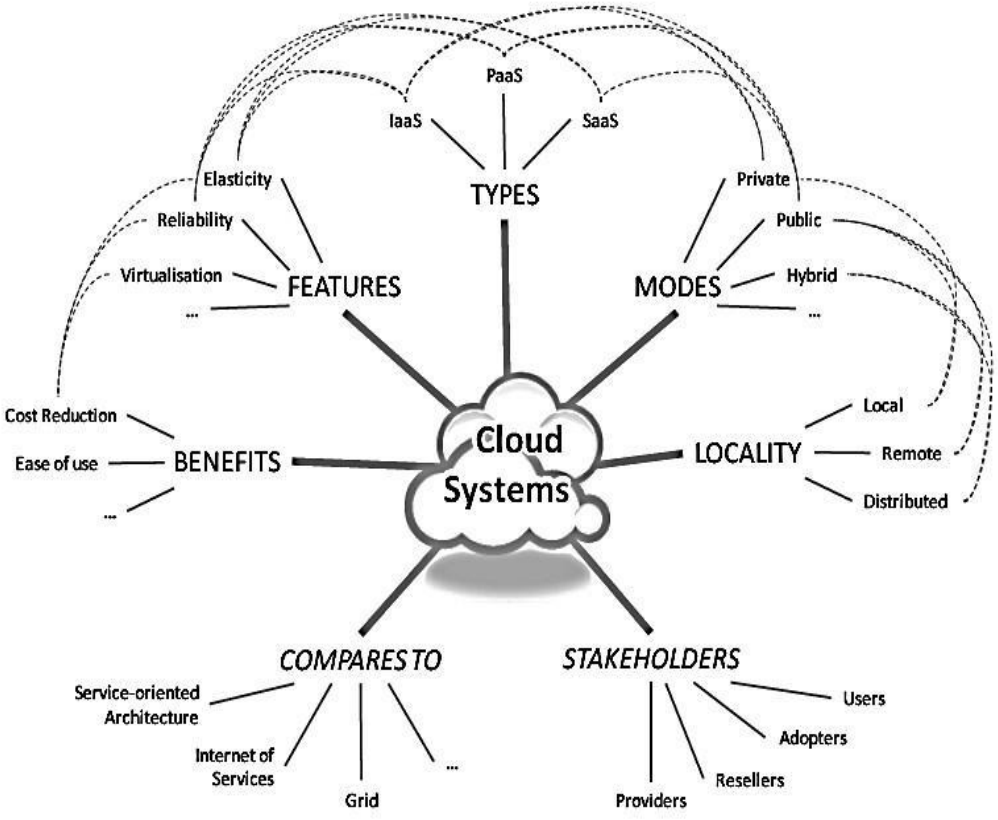


Figure 1.2 Aspects of Cloud Computing [13]

1.2.1 Characteristics of Cloud Computing

Some important characteristics of Cloud Computing [7] [8] [9] are discussed below:

- **Decreased Cost:** One of the most important features of Cloud Computing is pay-as-you go model. Users can pay according to their usage. Small business entrepreneur can easily scale up their business by renting infrastructure from Cloud providers without worrying about maintenance cost. They can get latest technology and they need not to worry about updating softwares. All they require is the internet connectivity to connect themselves to the remote server and access the services.

- **Large Storage:** In today's time, very large amount of data is generated and it is growing exponentially which has become a big challenge. Handling such a large amount of data has many aspects such as analysing, security, storage, maintenance of big data. Cloud Computing has offered its solution by offering large storage space for data as dynamic scaling up of Cloud is possible and it is even without affecting performance.

- **Flexibility:** Cloud provides flexibility in accessing data as it provides a feature of accessing data files anywhere anytime just by connecting with internet through various devices like mobile phones, PCs, tablets etc. User's data is stored centrally which helps them in sharing of data among family and friends and work with their colleagues in partnership. Storage space can be bought from Cloud providers according to the needs of the users and they can buy any more space if required, any time and can even relinquish space if not required anymore.

- **Reliability:** Cloud providers replicate the data of users so it offers reliability by proving undisrupted services. There is no data loss and it tries to prevent delay in services to the users as it replicates data on the servers specially used for replication purposes. So data is also available even if some problem occurs in any server.

- **Location Independence:** Data stored by users on Cloud in stored on unknown servers and their location is also unknown. But still data can be accessed by users anywhere anytime through internet. Users are incognisant about the

technical complexity of storing and retrieving data from the servers. Through these features of Cloud Computing, user's applications and data can be kept at optimum locations that help in reducing access time, quick recovery of data and greater level of security for the data.

1.2.2 Cloud Computing Services

Cloud Computing offers three level of services [7] as shown in Figure 1.3

- **Software-as-a-Service (SaaS):** These services can be accessed over internet by which user can run application as a web service by the use of web servers. These services are be accessed anytime when required. Users only interact with the simple interface provided by the Cloud provider through web browser and get access to information. They no longer require investing in servers or software license. Today in the market, there are many SaaS providers and Google, Amazon and Salesforce are a few big giants providing Software as a service. Google Docs is an example of these kinds of services.

- **Platform-as-a-Service (PaaS):** PaaS provides services which focus on development of applications and providing computational resources online which help developer in managing hardware or software. Users gain access to platform where they can develop or modify applications. Complete life cycle of development of applications is supported. Through PaaS design, deployment, testing of applications becomes easy and cost effective. It provides scalability and collaborative tools. Service providers provide storage, n/w, operating system etc. Examples of PaaS providers are Google App Engine and Amazon's AWS Elastic Beanstalk.

- **Infrastructure-as-a-Service (IaaS):** These services provide computer infrastructure. Softwares, connections, servers and other computational resources are provided as a service by Cloud providers. The main difference between IaaS and PaaS is level of the control on resources by the Cloud user. They get almost full control to manage the resources in IaaS. They can set all

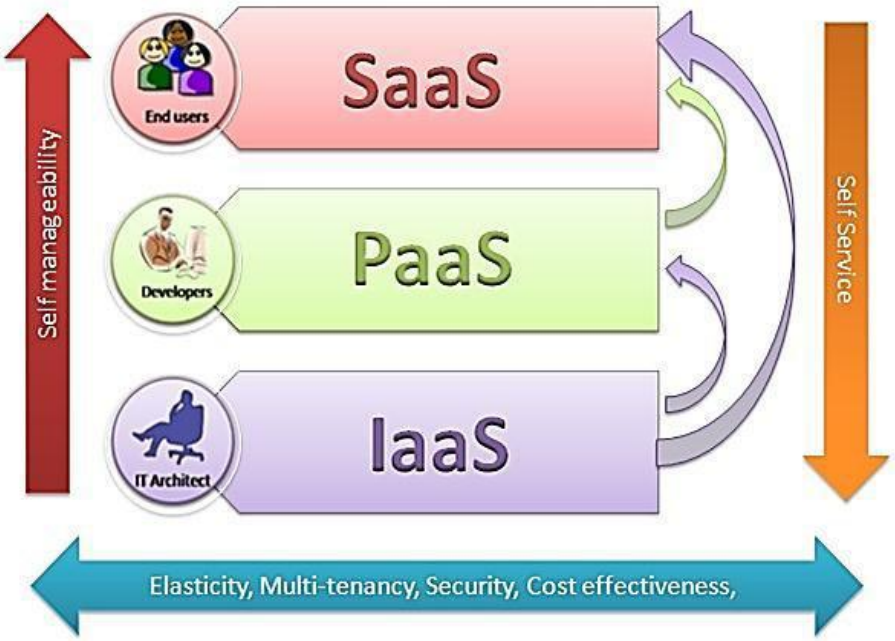


Figure 1.3 Service of Cloud Computing [14]

the parameters according to them like type of servers, configuration of operating system etc. This type of service is mainly suitable for those companies which have software packages available with them and now they want to run them, so they put these softwares on Cloud. Examples of IaaS are GoGrid, Amazon Elastic Compute Cloud (EC2), Rackspace.

1.2.3 Deployment Models of Cloud Computing

Cloud services can be deployed by four different ways as shown in the Figure 1.4 [11] [12].

- **Public Cloud:** It is also known as external Cloud and it is the traditional way of providing services and resources where everything is publically available which is provided by third party and can be shared by all Cloud users via internet. This deployment model mainly works on pay per use billing model. A large number of nodes are present in public Cloud so that millions of user's demand can be met. The biggest advantages of public Cloud are scalability and cost effectiveness. The availability for all users and resources and data

sharing among the Cloud users jeopardise the security of data present in Cloud. Google App Engine and Microsoft Azure are a few famous Public Clouds.

- **Private Cloud**: This Cloud is restricted within an organizational premise for hosting of private applications and computational resources for the use of only specific authorized users. This is for private use of the company simulating the Cloud. As the access is limited for the authorised users only, it is more secure Cloud. There is better control over the system resources like operating system and storage due to which it becomes easily customizable. The cost of infrastructure and maintenance is same as private Cloud but the difference lies in scalability and sharing of costs. A few famous private Cloud service providers are CloudBees, Amazon Virtual Private Clouds (Amazon VPC) and Rackspace Private.

- **Community Cloud:** Many organizations work together on some project or research work. For this they can use the services of community Cloud which can be shared among many organizations for specific interests. It is a collective effort which helps in sharing of infrastructure among specific community. It can be managed by third party or internal community and it is composite form of different Private Cloud.

- **Hybrid Cloud:** It is the accumulation of private Cloud platform with the public Cloud provider. It is usually designed for single organization. The independent infrastructures of public and private Cloud communicate with each other over encrypted connection. The sensitive data need not to be exposed to third party as it is stored on private Cloud and public Cloud is used to render computational resources. VMware's Vcare is an example of hybrid Cloud.

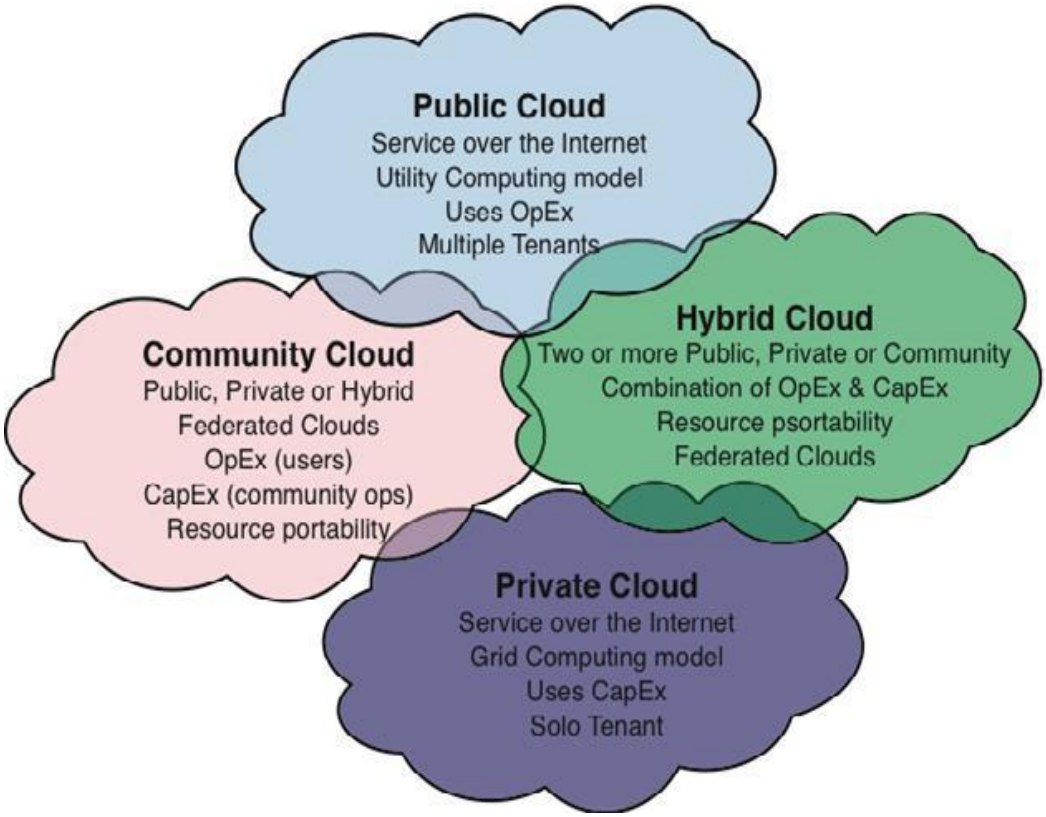


Figure 1.4: Deployment Models of Cloud Computing [12]

1.3 Research Issues in Cloud Computing

Cloud Computing is a service model which was actually introduced to render computational resources over the internet and it has been accepted and used by the industry. But still there are many open challenges [15][16][17][18][19][20] which need to addressed so as to be fully adopted by the industry and to make this model more secure, efficient, cost effective for user's requirements.

- **Virtual Machine Migration:** Virtualization is the technique to create virtual form of computing resources like physical servers, operating system, and storage and network devices so that many operating systems can run on single physical machine. It provides many advantages like resource sharing, fault tolerance, cost efficiency, split bandwidth of network by combining independent resources. Live migration is the migrating of a running machine from one physical server to another so as to the system more fault tolerant and to increase availability. If the load on a running machine is more, then Virtual

Machine Migration is used to balance the load of the server. The main research problem is how we can move complete running virtual machine from source server to destination along with current state of memory transparently, securely and without affecting availability of the server. Many Virtual Machine Migration techniques exist that try to efficiently migrate OS from one machine to another but none of the technique fully solves the purpose.

- **Energy Management:** In recent years, due to rising demands for computational power by web applications and scientific purposes, large scale data centres are created which are consuming tremendous amount of electrical power. A large amount of operation cost is spent on the power consumption. So Energy management has become a big challenge for researchers. There is a need for energy efficient techniques which could help in reduction of operational cost and saving energy. Energy can be saved by consolidation of Virtual Machines according to utilization of the resources, different virtual network topologies. Energy efficient hardware and power aware job scheduling are some of the few methods to attain energy efficient and green Cloud.

- **Fault Tolerance and Scalability:** The most fascinating characteristic of Cloud is fault tolerance and on demand scalability. Cloud helps in scaling IT infrastructure with a few minutes. So application should be designed considering scalability along with load and performance. In this environment it becomes necessary for applications to tolerate failure .A fault in hardware and software must be handled carefully without delaying which is very essential for providing reliable services. A survey by Gibson and Schroeder [31] discloses that Cloud suffers a lot due to hardware failure. Thus it becomes a big challenge for researchers to design highly scalable and fault tolerant applications which can be managed easily.

- **Interoperability:** To efficiently shift applications with the help of autonomous agent from one Cloud to another for acquiring required

performance and still maintaining transparency is an important research issue for Cloud researchers.

- **Data Security**: Users store their data in Cloud which is stored in the servers present at some unknown location, may be some another country, due to which data is affected both by laws of respective countries and Cloud Provider's policies [21]. Data needs to be protected from unauthorized access while maintain the integrity of user' data. So the challenge is developing secured and trustable system for maintaining data security and integrity.

- **Resource Allocation:** One of the key features of Cloud is infinite availability of computing resources that can be made available on demand. Users rely on infinite resource feeling as the Cloud attempts to deal with requirements of users by resource allocation system in a flexible way. This flexibility helps in statistical multiplexing of computing resources averting both over provisioning and under provisioning of IT infrastructure. Moreover there is necessity to deal with resource heterogeneity. The important feature of resource allocation is the assurance that the requirements of all Cloud users are satisfied. According to [57] resource allocation must restrict the degradation of performance up to certain range. Allocation techniques should be aware of status of every resource in distributed environment. And using this information efficiently applies algorithms so as to allocate computing resources in distributed environment. Users in Cloud can arrive and leave anytime without prior knowledge about their requirements. So developing efficient resource provisioning and allocation technique is one of the most challenging issues to meet heterogeneous requirements of Cloud users.

1.4 Research Motivation

The demand of the future generation research work is to solve the challenge of applications requiring high computational services. Cloud provides services to all types of applications including smaller execution time or longer execution time. The advantages of Cloud are no need to buy infrastructure as all resources are available on demand. Cloud providers have finite set of resources but in user's view there are

infinite resources. These resources should be properly allocated among users in order to earn more revenue. So, Resource Allocation is one of the most challenging issue in Cloud. Most of the existing Resource allocation techniques are static that try to allocate resources based on monitoring results and accordingly reservation of resources is done. But the demand of resources is dynamic in nature in which users enter or leave anytime they wish. These dynamic scenarios cannot be handled by static approach. Resources should be allocated according to demand and supply criteria. The prices of resources are also fixed or pay-as-you model is followed. But for proper generation of revenue in dynamic environment of demand and supply, dynamic model for pricing should be considered. These challenges need to be considered in detailed for proper adoption of Cloud by complete industry. The aim of the thesis is to present an auction based allocation algorithm that could allocate resource dynamically using auction according to real market based scenario.

1.5 Organization of Thesis

The rest of the thesis is organised as follows:
Chapter 2 – This chapter of the thesis presents the literature survey of resource allocation problem and detailed analysis of existing techniques in this area.

Chapter 3 – This chapter presents the gap analysis in the existing research work in the area of resource allocation in Cloud Computing, problem statement is discussed and objectives of thesis are presented.

Chapter 4 – In this chapter solution to the problem presented in Chapter 3 is presented by proposing a new algorithm for resource allocation with the help of sequence diagram and data flow diagram.

Chapter 5 – Implementation of the proposed algorithm with the description of CloudSim and Netbeans is discussed in this chapter. Experiments results and comparison with the existing techniques are described.

Chapter 6 – This chapter concludes the complete work of the thesis and presents the future research directions.

Chapter 2
Literature Review

In this chapter various research issues associated with resource allocation techniques in Grid and Cloud Computing have been discussed and detailed comparison among the existing techniques has also been performed.

2.1 Resource Management

Resource management consists of resource discovery, monitoring and allocation process as shown in Figure. 2.1. These processes controls physical resources like disk space, memory, CPU cores and network bandwidth. Virtual form of these resources is created and distributed among various virtual machines running heterogeneous workloads. The key element in resource management is discovery process. It consists of locating the appropriate resources that are requested by users and required by applications [22]. It is handled by Cloud service provider. The process is used by user broker to locate resources. Discovery process involves complete description of available resources. Discovery of resources provides mean for resource management system (RMS) to find resource's state handled by it and interoperated by other RMSs.

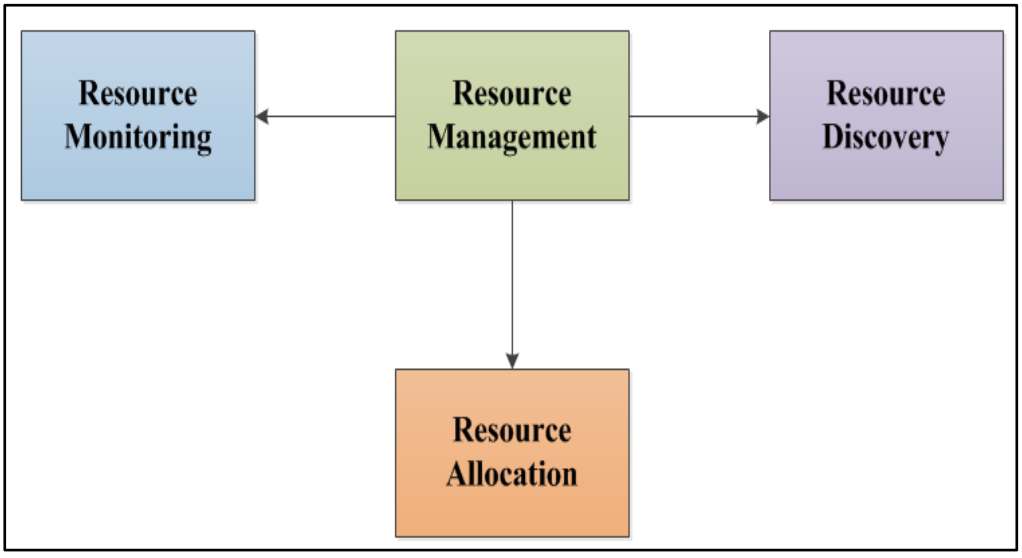


Figure 2.1 Elements of Resource Management.

The processing of allocating available resources to the required application is known as resource allocation [23]. Resources are allocated among users according to their request and pay-per-use model. When resources are assigned to respective user, dispatcher allocates the resources. According to [24], Resource monitoring is a tool for handling and controlling computing infrastructure. To help decision making process for resource allocation, resource monitoring renders information to both application and platform. Whenever there is an event failure in Cloud, it monitors the state of resources at all levels.

In the existing literature, there have been many techniques proposed for resource management. Castillo Claris et al [25] deal with the issue of dynamic submission of user's task and satisfying application time requirements. These jobs must be completed before deadline. Based on this, scheduling algorithm which is heterogeneity aware is presented and computational geometry to apply techniques to develop data structure is mentioned which helps in efficient management of resources in advance. Vanmechelen et al have introduced a Resource Management System (RMS) to trade CPU time for resource management by decentralized and centralized algorithm [44]. In centralized algorithm there is a broker which negotiates with resource provider as to satisfy user's requirements within QoS parameters mentioned by user. An auction is hosted by provider agent to sell his resources in decentralized algorithm. Both these algorithms support deadline requirements.

J. Espadas *et al.* [35] mentions that during peak hour loads can be predicted but during non peak hours computing resources are wasted as mostly resources are either under provision or over provision. And there is a loss of revenue for Cloud provider. Leased resources are controlled by users and provider only monitors resource usage. So, there must exist an advisory system that could bridge the physical resource monitoring and user's application monitoring to avoid over provisioning. V. Vinothina *et al.* [23] mentioned that over provisioning and resource contention must be avoided for optimal allocation of resources. There should be a solution which could satisfy all users' requirements without starving Cloud providers' capacity. The proposed technique may be fit for infrastructure level as all computing resources are managed by service provider. Kertesz and Kacsuk (2009) have introduced for grid resource management three different strategies to increase grid interoperability and presented their solution. Meta-brokering, the third approach presents the final solution

by using broker information [45]. When meta-broker is freed by serving grid users, final step will be considered for grid interoperability. R. Basmadjian *et al*. to save cost of energy, introduced a power consumption model in Cloud environment based on prediction [26]. A model to calculate power consumption of each component in data centre is presented using prediction. This prediction model can be enhanced further to provide benefit to user as well as inter Cloud provider.

2.2 Resource Allocation

Resource Allocation (RA) in Cloud environment is the process of allocating computing resources to the Cloud users on demand over the internet. If resources are not managed and allocated properly, there occurs the problem of resource starvation. It generally happens due to over provision and under provisioning of Cloud resources among users. The problem can be solved by resource provisioning which helps in managing Cloud providers' resources for every user module. Resource Allocation Strategy (RAS) is the way of integrating activities of Cloud provider for efficiently allocating scarce resources to satisfy the needs of Cloud users. It needs some prior information like amount of resources, type of resources and time for which resources are requested for efficiently allocating resources [23]. An optimal RA technique tries to avoid resource contention which is accessing same resource simultaneously by different applications, scarcity of resources when demand is more than supply, under provisioning, over provisioning and resource fragmentation which is availability of enough resources but at different location so cannot be allocated.

It is difficult for Cloud resource provider to predict the demands of users and applications in advance. There is a constraint to complete the execution of task within deadline and minimum cost. Thus there is a need of efficient RA technique in Cloud environment which could take into consideration resource heterogeneity and limited resources. Virtualization helps in sharing physical resources among many resource users. Virtual machines are created by giving them different computing power and they helps in making Cloud scalable .Scheduling algorithm is one of the ways to optimally allocate resources such as CPU and memory. A single process runs in uniprocessors. The task of the scheduler is to select which process to run among the different process running in various scheduling queues [29]. Scheduling decides

which task to be executed by the processor. Earlier researchers used first come first serve (FCFS) [28] scheduling technique for allocation process. With time various scheduling algorithms have been proposed considering different parameters in mind, like in [27] [29] [30] gang scheduling for resource allocation while in [31] energy aware algorithm has been discussed. In [27] by using virtual cluster a multi tier allocation scheme is proposed. [31] proved that it helped in reduction of energy consumption for Cloud environment by using technique of Dynamic Voltage Scaling. In [30] waiting time has been improved due to increase of waiting time. There are different conditions with lots of workload but it works only for single workload. M. Stillwell *et al.* [27] analyzed resource allocation by utilizing virtual clusters. These clusters run many tasks which are competing among themselves. They developed algorithm for allocation of tasks for different nodes. Users have the choice to choose algorithm to get best resources. For parallel task they equally allocated resources among each task considering sequential job algorithm. They used pragmatic formulation in dynamic workload for adaptation of resource allocation. The proposed solution is swift to adapt changes and resource allocation process.

The authors "Do *et al."* mentions that scheduling algorithms can be used for implementation of resource allocation strategies for Cloud infrastructure provider. In [32] there is a situation by using scheduling algorithm that a small task has wait for long time for its turn to get access to CPU. This will result in unnecessary waiting for shorter task and thus higher cost for the user. During prioritization of any task smaller task should be combined in single workload for better utilization of CPU. Task requiring bigger resources can be scheduled to another scheduler. Scheduler should have the capability to automatically allocate resources so that workload can be run faster. There are various ways to solve the resource allocation problem. In literature different techniques have been discussed for allocation of available resources considering different parameters and scenarios. These techniques are broadly discussed in the following section.

2.2.1 Resource Allocation Techniques

Lately many resource allocation strategies have discussed in the literature for Cloud and grid environment as now this technology has begun maturing. Various types of

resource allocation schemes have been proposed and implemented in Cloud and grid by the researchers. A few are discussed below.

- **Topology Aware Resource Allocation (TARA)**

Resources are allocated in IaaS without actually knowingly requirements of the tasks running on IaaS. So more resources can be allocated to applications having fewer requirements that affects performance of distributed data intensive applications. In this way resources can be wasted. To solve this problem [33] have proposed an architecture which is optimal for IaaS. This model considers 'what if' criteria to help IaaS in taking decision for resource allocation. The architecture consists of a prediction engine having lightweight simulator and a genetic algorithm for fast searching. Prediction engine is actually responsible for allocation resources at IaaS level optimally. It iterates through all possible resources to find optimal set of resources for any particular request. TARA uses genetic algorithm for searching. Figure 2.2 shows the architecture of the TARA. But due to the size of IaaS it is not feasibly possible to iterate among all the available resources even with lightweight prediction engine.

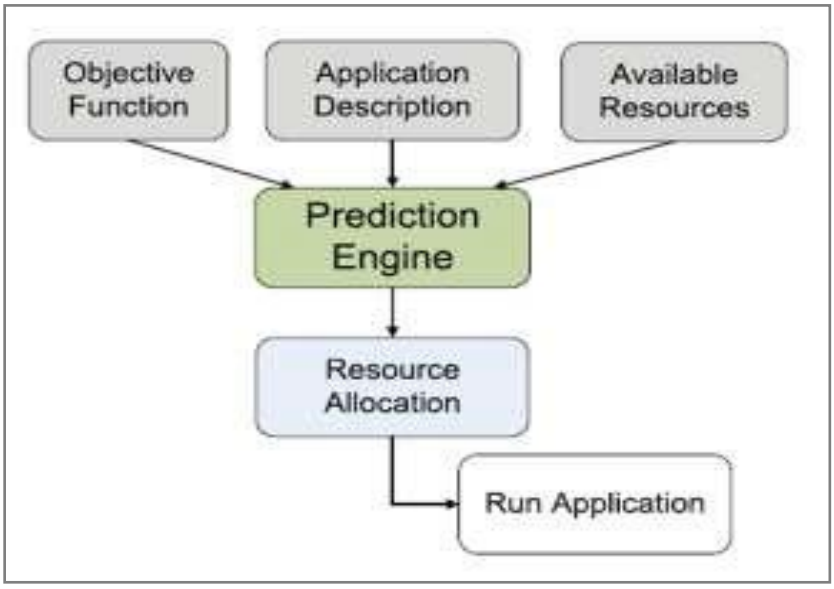


Figure 2.2 Architecture of TARA [33]

- **Dynamic Resource Allocation**

In [34] for parallel data processing, a dynamic resource allocation technique known as Nephele has been proposed which a framework is specially designed for Cloud. It is

first processing framework which helps in dynamic allocating and de-allocating of computing resources. Virtual Machines are assigned the task of processing job and these are instantiated and dismissed automatically while execution of job. Nephele's architecture as shown in Figure 2.3 uses master worker pattern. Before submitting a job, there must be a virtual machine (VM) started by the user which runs Job Manager (JM). The Job Manager schedules the users' job.

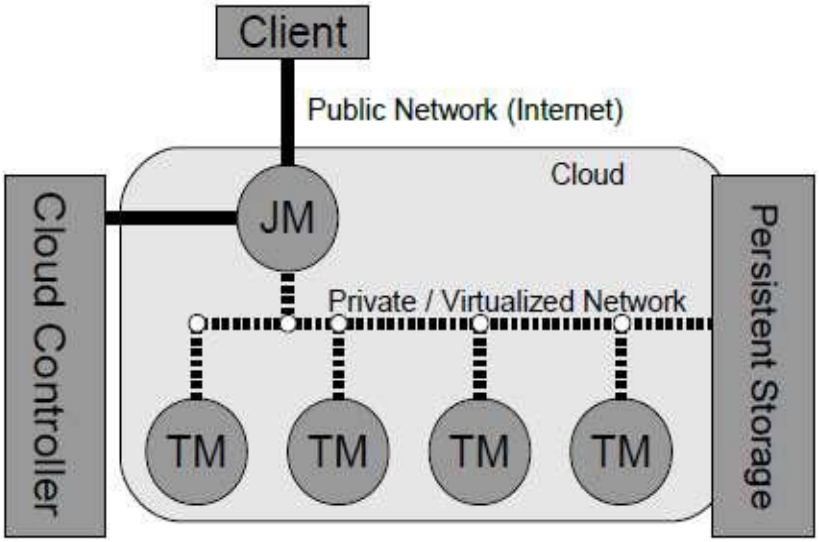


Figure 2.3: Architecture of Nepheles' Framework [34]

It communicates with an interface of Cloud provider for initiation of VM. This interface is known as Cloud controller. Job Manager helps in allocation and de-allocation of VMs. Task Manger after getting tasks from Job Manager executes the tasks and then informs the JM.

- **Market based Resource Allocation**

The main aim of resource management is to develop an agreement between resource provider and resource user so that provider makes available resources required for running the users task. Market based resource allocation presents pricing and money as the means for coordinating between resource provider and user. Users need resources for executing their task and resource providers need customers for selling their resource as to gain profit. Both will have to search for their client. One way to solve this problem is market based approach. In market based approach there is a trading of resources in which selling and buying of resources is performed. Various

economic models are used for trading of resources. So market based allocation can also be considered as economic based resource allocation.

Pourebrahimi et al (2006) introduced allocation of CPU time trough market based approach in central market of Grid. Whenever there is free node and some node waiting for CPU time they can approach each other. User and provider use their local knowledge to fulfil their requirements [46]. They have used continuous double auction for helping selfish agents in taking their decision. The mechanism is evaluated and was compared with other approaches like non market based. The results show that market based resource allocation performs better and provides good efficiency for resource utilization. Michal Feldman et al [47] use fixed budget allocation to check the performance of distributed market. Users decide how they can divide their budget for different machines according to their preferences. They try to maximize their utility. In their approach they studied performance of distributed market based strategy by utilizing both analytical and simulation method. Even high performance level can be attained in worst case at Nash equilibrium for efficiency. The system attains equilibrium for finite parallelism model with few exceptions as it uses best response algorithm and greedy adjustment when users are large. It used price anticipation method which works well for resource allocation.

Jacek Gomoluch *et al.* (2003), classified strategies as state based or model-based and pre-emptive and non-emptive for market based allocation in Grid. They tried to find out which strategy outperforms in round robin approach by proportional share protocol or by continuous double auction. Simulation results are presented for model for market, servers and clients. Various factors are studied like amount of load and heterogeneity of resources. In 2001, Rajkumar Buyya et al [37] introduced resource management for Grid environment as a complex process due to heterogeneous and distributed nature of the resources. These resources are controlled and managed by different owners, each having their own distinguished policies. They introduced a framework known as Grid Architecture for Computational Economy (GRACE) which by using middleware services offers new services for real world Grid. For trading of resources, economic model is used and experimental results, on cost based scheduling and deadline, are presented. Menasce and Casalicchio [49] introduced SLA based resource allocation technique for grid environment. Applications are dividend into

tasks and allocated assuring cost minimization and executing time based on agreed upon SLA.

Linlin Wu et al [36] presented the problem that SaaS provider in Cloud environment need resources to provide services to the users. They either get their own resources or rent from IaaS. So, they proposed algorithms that could minimise infrastructure cost and guarantee minimum level of service as specified in SLA. Various QoS parameters were taken into consideration for providing services. Three algorithms are presented, that schedules SaaS customers' request taken various QoS parameters into consideration. Algorithms try to minimize the cost for both SaaS provider and user and different algorithms are compared among themselves and results show that ProfminVMminAvaiSpace outperforms all in Cloud.

There are two ways to allocate resources in economic model among the different competing entities. One way to allocate resources is non pricing based economy model and another is pricing model. The non pricing models [38] are mainly cooperative mechanisms or game theory. Game theory is based on utilization and optimization of individual entities [39]. But cooperative models consider global optimization of all entities in distributed system. The distribution of economic model is depicted in Figure 2.4. Price based models are categorized into two parts Commodity market and Auction based models.

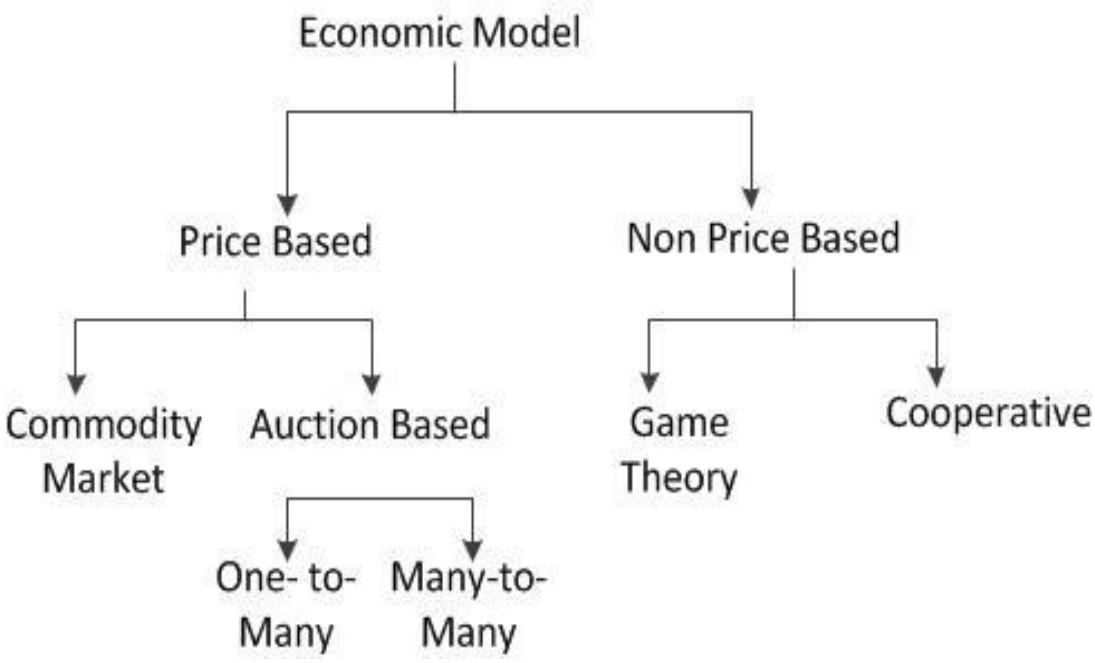


Figure 2.4 Distribution of Economic Model

Wolski et al [50] have considered the commodity based resource allocation approach in grid. Equilibrium price is used to allocate resources when demand and supply are same. Auction model can be either one-to-many or many-to-many models. Pricing based modes are further discussed in below section.

2.2.2 Pricing based Economic Models

In the literature of Cloud Computing, pricing of the Cloud resources has been discussed but most of the models use pay per use pricing schemes. But according to Harmon et al., pricing is an important issue to be considered by a firm for either new IT services or placing existing services [52]. Weinhardt et al. [53] state that success of Cloud services is only possible with appropriate pricing model. Paleologo (2004) describes that pricing models like cost plus pricing models are not appropriate for on demand services as there are changing requirements and other parameters like uncertain demand and shortened duration of contract [54]. Sainio et al [55] argue that earlier pricing approaches are quite operational but due to changing scenario and unpredictable environment of demand and supply of services strategic planning for pricing should be considered. Osterwalder (2004) argue that revenue models offers pricing approach and strategic planning to operational planning. There are basically three types of pricing approach for different Cloud services which are currently used by different Cloud providers [56].

Table 2.1 summarises different pricing approaches used in Cloud environment. In fixed pricing models prices do not vary according to different scenarios and do not consider real market scenarios. Differential pricing models consider pricing according to both customer and product features and other parameters like customer choice and volume based but do not consider real market scenarios. Market pricing models offer pricing according to real market scenarios.

Table 2.1 Pricing Models [56]

Category	Pricing Model	Description
Fixed Price	Pay per use	Customer pays according to his usage.
	Subscription	Customers pay flat price to access the services.
	Menu price	Fixed price mentioned in catalogues
Differential Pricing	Service based	Price is charged according to service features.
	Value based	Price is according to customer's value proportion.
	Volume based	Price is charged for volume of services consumed.
Market Pricing	Auction	Buyers bid in increasing price for pricing.
	Reverse Auction	Sellers bid in decreasing price value.
	Dynamic Market	Large number of buyers and sellers are present which influence pricing criteria.

Various Cloud providers have started using auction based pricing models for their Cloud resources. Auction model introduces more revenue for providers and cheaper prices for users. Prices vary according to increasing and decreasing demand. Like Amazon Spot Instances have been introduced by Amazon Web services for selling their unused capacity resources. As long as the customers' price is more than spot price, customer's instance is running. Spot price is set based on current utilization of data centres by Amazon (Amazon Web Services 2010).

2.2.3 Auction based Models

Continuous Double Auction (CDA) has been investigated by Pourebrahimi et al. [46] for resource allocation for grid computing. CDA is considered as one of the most famous auction mechanisms and it is applied in electronic stock market. Bids by participants can be submitted anytime when auction is going on. They argue that as compared to non market based mechanism, market based approach performs better for task utilization and resource allocation process. But the main drawback of this scheme is that it does not consider resource allocation for multiple items, only single item or resource allocation process has been considered. This approach was used for local grid environment for allocating CPU time. Another resource allocation technique for market based environment in grid is combinatorial double auction [51] was proposed which brought revenue maximization and economy efficiency. This proposed method had the advantage of monetary based approach and flexibility. A VCG based pricing scheme for resources was introduced in 2009 for dynamic markets [42]. This model was efficient in its approach as players in the auction had dominant strategy so that they declare true value of their resources. In 2010 a method of marginal bid was proposed to determine about the prices resources [42]. In this model to solve the problem of resource allocation double auction was used. This framework had the advantage of flexibility but the problem with the model was it did not thoroughly analyse the performance of the system.

On the basis of combinatorial auction two mechanism were proposed in 2010; Combinatorial Auction-Greedy (CA-GREEDY) and Combinatorial Auction-Linear Programming (CA-LP) [40]. For achieving higher utilization of resources and revenue CA-LP was a better option but its disadvantage is that it is limited to less number of users. CA-GREEDY gave better results when motive is to maximise social welfare. This approach restrained an auction to certain types of resources. In Cloud environment a reverse auction with batch matching technique was proposed [41]. Improvement of service quality and user satisfaction was the main objective of the presented model. This model taken into consideration user side only and it did not discuss details of resources sufficiently to apply the model to realistic Cloud environment. One more technique known as English combinatorial auction was discussed by Wang et al. [41] for resource allocation problem in Cloud environment. The price for resource trading was calculated by an English combinatorial auction

technique, which tried to minimize the execution time for finding the winner and maximize the profit of the resource provider. Paolo et al introduced procurement auction in which market conditions are discussed and different strategies used by Cloud provider for bidding are discussed. They used the model for selling unused resources. Overbooking is used for increasing their revenue.

All the models discussed so far tries to optimally allocate resources but only a few models consider real market scenarios and these models do not penalize providers if they fail to provide resources after committing the customer. These models do not provide QoS services according to real market scenarios in which multiple providers are competing among themselves for providing services to users and users want resources at a cost of as low as possible.

Chapter 3
Problem Analysis

In the previous chapter, various techniques, provided in the literature for solving the problem of resource allocation in Grid and Cloud have been discussed. This chapter deals with the various gaps found in the literature and these are discussed below. Objectives of the thesis are also presented in the end of the chapter.

3.1 Gap Analysis

Many market based models for resource allocation have been proposed in the literature so far. A few have been discussed in previous chapter. But there exists many gaps in literature which are discussed as below.

- Most of the techniques are biased towards provider. But none of the techniques favour provider and customer simultaneously in unregulated Cloud market.
- Current allocation of resources approach is static in nature. Resource monitoring is used for advance reservation of resources. But user demands are dynamic in nature which cannot be handled by static models.
- Most of the Cloud service providing companies use fixed price based model for pricing of their resources. But these models do not fit in every scenario of real market space. These models are static in nature but users' demands are dynamic and unpredictable. This way their resources are not utilised in an optimum way.
- Procurement of resources is done manually. As there are multiple Cloud providers, users have to go through all the vendors in order to find the most appropriate model which is very tedious work. Most of the work in literature considers single Cloud provider who satisfy the requirements of users. So there is need for automated and scalable technique for resource procurement.

- The work in market based allocation and dynamic pricing considers only single type of resources while using auction based technique. Customers often need resources in bundle of resources.
- Auction based techniques proposed in literature does not consider minimum restriction on Cloud provider so as to fulfil users' requirements
- In this way Cloud providers can cheat customers and earn more revenue by providing lower level of services.
- Nowhere in the literature, have penalties been imposed on Cloud provider for not satisfying committed services to the user.

3.2 Problem Statement

Optimal Resource allocation is one of the biggest challenges in Cloud environment. Cloud providers have resources which should be efficiently allocated otherwise there could occur the problem of under utilization or over utilization. Moreover there are multiple providers in Cloud market space who want to sell resources. Each provider wants to sell resources to users and earns maximum revenue. In this scenario, it becomes difficult for users to choose the best among the Cloud providers. It is also challenging for providers to choose the user who values the resources more. So the problem is to determine an appropriate Cloud provider for each user who could sell various resources in minimum possible amount according to his requirements and find suitable customer for Cloud provider who values his resources at the most. The model should satisfy all the requirements of users at a given point of time. It should also impose some minimum restrictions on the Cloud providers to fulfil requirements otherwise penalises the Cloud provider who violates the conditions. Thus, to summarize, the current thesis deals with designing of relevant market based resource allocation algorithm for the benefit of Cloud users as well as Cloud providers.

The aim of the thesis is to provide a solution for resource allocation problem in Cloud environment by using auction technique of the economic model in market based resource allocation scenario. The thesis satisfies the requirements of Cloud users who want resources in order to complete their task within the budget they are advertising. It also optimally allocates resources of Cloud provider's who are competing among themselves, as there are multiple providers in Cloud market space. Also providers

want maximum revenue for which this thesis designed a pricing technique and penalizes the Cloud providers violating specified conditions.

3.3 Objectives

The objectives of the thesis are discussed as:

i. To analyse the existing resource allocation techniques for optimum allocation of resources.
ii. To develop an algorithm that could efficiently allocate resources of providers in Cloud environment by using auction based technique of economic model.
iii. To evaluate the performance of proposed algorithm.

Chapter 4
Proposed Algorithm

This chapter proposes an algorithm as a solution for problem described in earlier chapter. The complete architecture for the proposed algorithm is presented including sequence diagram and data flow diagram.

4.1 Design of the Proposed Algorithm

The solution for the problem of resource allocation as mentioned in previous chapter is presented through the architecture of the proposed algorithm and data flow diagram. The following section presents the architecture.

4.1.1 Architecture

The proposed algorithm "Procurement Double Auction for Resource Allocation" (PDARA) in Cloud is depicted in Figure 4.1. In the proposed algorithm, the problem of resource allocation is solved using auction based resource allocation technique and by using dynamic pricing model for pricing of resources. Procurement auction which is also known as reverse auction is a type of auction in which role of seller and buyer is reversed. Buyer asks resources and seller usually bid for providing services to users. The auction winner provides services to users. The proposed architecture consists of multiple Cloud resource providers and multiple users who want resources to complete their task. The users will submit their requirements to the auctioneer. There are different types of resources available with Cloud resource providers. User can demand any amount of resources of each particular type of resource in the form of bundle. After receiving request, auctioneer will send messages to Cloud providers who will then start bidding. They can only bid if they have enough resources for bidding or if they have resources in overbooking. Overbooking means if Cloud provider knows that till the completion of auction process, his resources will become free from some another user who is currently using resources, he can go for current auction process. After bidding by providers, he will start mapping users and providers after calculating the lowest bid provider and highest resource demanding user. If lowest bid provider has enough resources he will be assigned first user. Otherwise

check for next resource provider. If the auction winner is unable to satisfy user requirement he will be penalized.

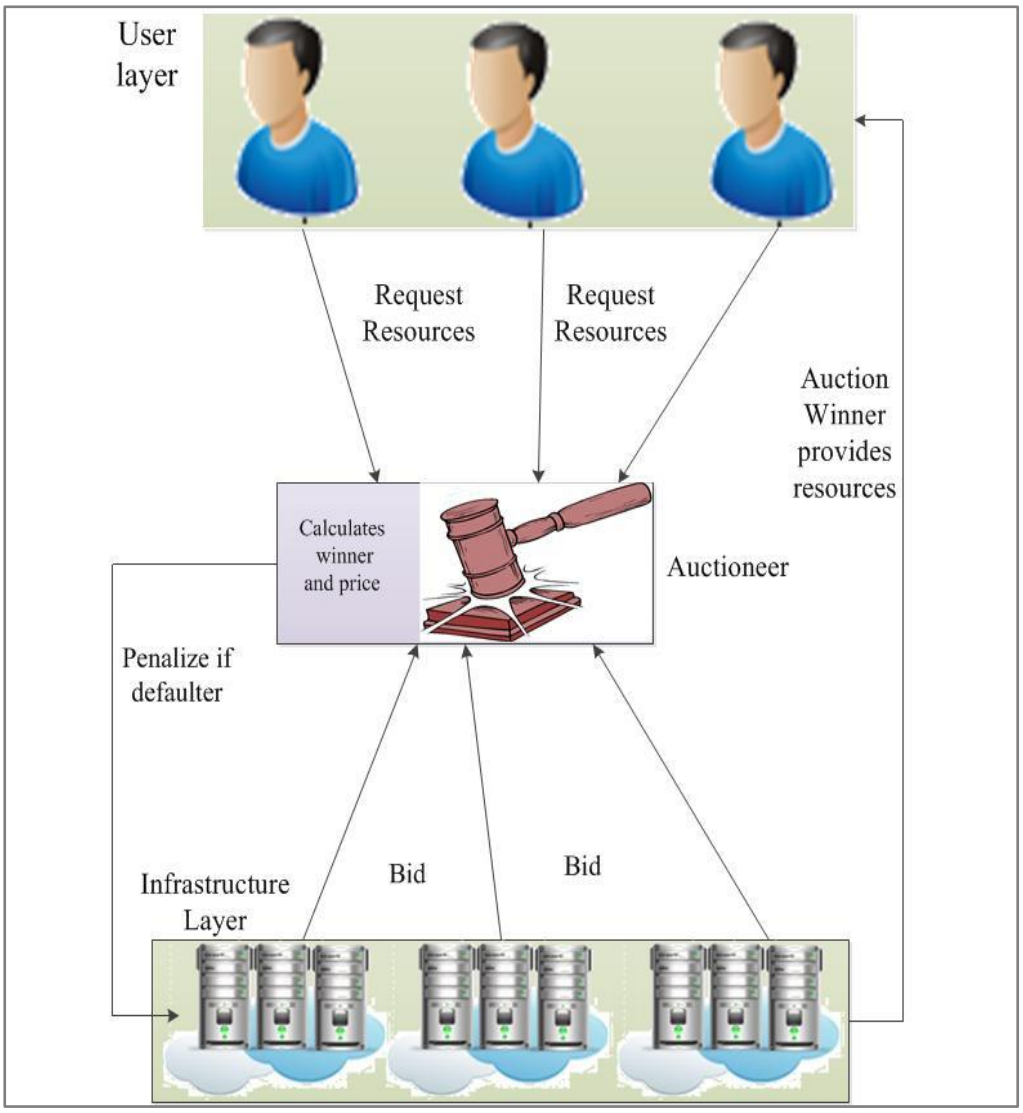


Figure 4.1 Architecture of Proposed Algorithm.

4.1.2 Participants of Proposed Algorithm

i. **User:** This is the entity who wants resources to fulfil its requirement as it wants to complete its task. It will go to auctioneer or broker for raising resource request in which it will mention number of resources required of each type, time for resource requirement and its budget. When user is assigned particular Cloud resource provider, he will be informed. When requirements are fulfilled, user will pay the amount calculated by the auctioneer.

ii. **Auctioneer or Broker:** This entity is the centre of this complete auction process. It is responsible for assigning particular resource provider to user. Auctioneer will finalize the price required to be paid by the user. User will raise requirement their request to auctioneer and it will send messages to Cloud resource providers for bidding. The complete auction process will be controlled and managed by him. Auctioneer is responsible for determining auction winners and informing them. If Cloud provider is unable to complete user's task after auction winning, he will be penalised by auctioneer. Auctioneer will calculate penalty for Cloud resource provider and then assign task to next provider.

iii. **Cloud resource provider:** This is the entity that has resources available with him. Memory, CPU, disk, storage space etc all resources are available with it. Provider wants to sell these resources as to gain maximum revenue. It searches for user as to provider services. For this provider will participate in auction process when auctioneer sends messages. If provider has enough resources it can participate and bid for resources. If it wins the auction, will provide service to users and will earn profit. But if after winning auction it is unable to provide service, it will be penalised.

Provider can participate in any number of auction processes according to his available resources as there is no limit. But the only condition is that if it commits it will have to provide service. Fig. 4.2 depicts the relationship between different participants of the algorithm.

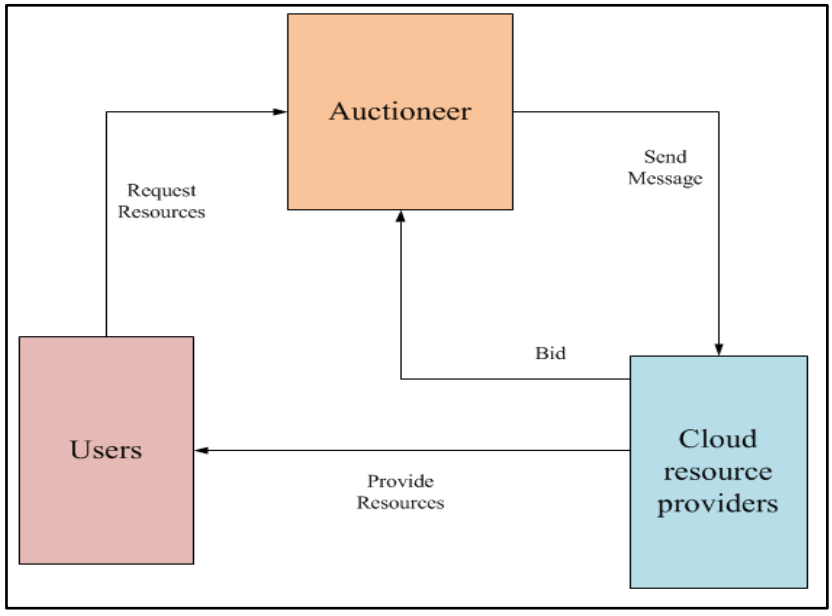


Figure 4.2 Relationship between entities

4.1.3 PDARA Parameters

- Let there be N number of users and n represents user number, $N = \{1, 2, 3 \ldots . n\}$.
- Consider M number of Cloud resource providers and m represents Cloud resource provider number. $M = \{1, 2, 3 \ldots . m\}$.
- Total number of resource types are K and $K = \{1, 2, 3 \ldots . k\}$.
- Each user asks for different amount of each resource type. And every Cloud provider has also different amount of each resource type.
- The user requirement vector can be represented as $A_n = \{ a_1^n , a_2^n, a_3^n \ldots . a_k^n, t_n, b_n \}$ where t_n represents the time for which user require resources on per hour basis and b_n is the budget of the user.
- When Cloud resource provider will bid, he will advertise resources available represented by vector $R_m = \{r_1^m, r_2^m , r_3^m \ldots . r_k^m\}$ and price for each resource type as $P_m = \{p_1^m, p_2^m, p_3^m \ldots . p_k^m\}$.
- Weight vector W of resources representing different weight for each resource can be represented as $W = \{w_1, w_2 \ldots \ldots . w_k\}$.

Table 4.1 shows the complete list of parameters used in the algorithm model. In the table various parameters used in algorithm model for calculating auction winner and price calculation are described.

Table 4.1 Parameters of the Algorithm

Parameter	Description
N	Number of users
n	User number
M	Number of Cloud Resource Providers
m	Provider Number
K	Total number of different resources
A	User requirement vector
R	Quantities available with provider vector
P	Provider bid price vector
t	Time for which user require resources

b	Budget of the user
Z	Per resource bid value of user
Zt	Time multiplied by per resource bid value
u	Total bid of provider
M	Per resource bid value of provider
W	Weighted vector of resources
w	Weight of each resource
wt	Weighted total of resources
Up	Price paid by user
Pt	Penalty
P_w	Price by ultimate winner provider
P_d	Price by defaulter provider

4.1.4 Winner Determination

The user will request resources from auctioneer who will then go to Cloud resource providers for bidding. After receiving bids from Cloud resource provider, auctioneer will calculate winners of the auction. Auctioneer will first calculate weighted total of resources requested by user which can be represented by wt as described in Eq. 1.

$$wt_n = \sum_{i=1}^{k} (a_i w_i) \qquad \text{Eq. (1)}$$

Then he will calculate per resource bid value of user as given by Eq. 2. Next per

$$Z_n = \frac{b_n}{\mathrm{w_t}} \qquad \text{Eq. (2)}$$

resource bid value will be multiplied by time for which user require resources as to give more preference to user who require resources for longer time which will ultimately generate profit for provider. Eq. 3 calculates it.

$$Z_{t_n} = Z_n \times t_n \qquad \text{Eq. (3)}$$

The users who require resources for more time will get higher value and the list will be sorted in descending order. Maximum value user will come at first place. In the same way, provider per resource bid density is calculated. And then list will be sorted in ascending order.

$$M_{\mathrm{m}} = \frac{\sum_{i=1}^{k}(p_i)}{\mathrm{k}} \qquad \text{Eq. (4)}$$

In the sorted list, first Cloud provider is the lowest bidder in the auction. If he has enough resources for first user, mapping will be done of first user and the corresponding Cloud provider. But if not, then check for next Cloud provider. Iterate through all providers until user's requirements are satisfied. Same process will be followed for all users. Price which user will have to pay is described in next section.

4.1.5 Price Calculation

The user is mapped to provider. Then user will have to pay price which is calculated in Eq. 5.

$$Up_n = \left(\frac{b_i}{(\sum_{i=1}^{k}(a_i) \times t_n)} + M_m \right) \times \frac{1}{2} \times t_n \times \sum_{i=1}^{k}(a_i) \qquad \text{Eq. (5)}$$

The total price calculate by auctioneer will be communicated to user which he will have to pay at the end of completion of task to the particular provider. But if Cloud provider after winning auction is unable to provider resources, he will have to pay penalty which is described in next section.

4.1.6 Penalty Calculation

When Cloud provider after winning auction is unable to provider resources, he will have to pay penalty which is proportional to every Cloud provider who is defaulter. Suppose first and second Cloud providers are defaulter, user's task will be assigned to third Cloud provider. The proportional of the difference of defaulter Cloud provider's price (P_d) and the price asked by third Cloud provider which is ultimate winner (P_w) will be divided proportionally between two Cloud providers as described in Eq. 6.

$$Pt_{\mathrm{m}} = (P_{\mathrm{w}} - P_{\mathrm{d}}) \,/\, \sum_{i=1}^{m-1}(P_{\mathrm{w}} - P_{\mathrm{d}}) \qquad \text{Eq. (6)}$$

4.1.7 Execution Flow of the Algorithm

- Execution flow the proposed algorithm is as follows:
- All Cloud providers initially register themselves with the broker or auctioneer so that in near future they can send bid in auction.
- Users at a given point of time send their requirements to the auctioneer. In the requirement message, they advertise their budget and the time for which require resources.
- The auctioneer will collect the list of all the users' requirements and send the messages to all registered providers for bidding.
- Auctioneer will set the timer. Providers who bids with in time limit will be considered for current auction process.
- Cloud provider after receiving message from auctioneer will check his resources. If he has enough resources or he used overbooking, he would bid for each type of resource he has and will advertise the quantity of resources available.
- Auctioneer after completing the time will stop the bidding process. After this if any provider bids, he will not be considered for auction process.
- Auctioneer will start calculating process. First he will calculate weighted total of requirements of each user according to Eq. 1. Then he will find per resource bidding value as mentioned in Eq. 2.
- To give the preference to users who require resources for longer duration, the above calculated value will be multiplied by time for which resources are required (Eq. 3).
- The users list prepared will be sorted in decreasing order. In this way user who require resources for longer duration will be placed in the starting.
- In the similar way, using Eq. 4, per resource value will be calculated for Cloud providers.
- The list will be sorted in increasing number as provider with minimum bidding will come at the starting position.
- Mapping of users and Cloud providers will be done and is calculated using Eq. 5. If first Cloud provider has enough available resources then first user will be mapped with him otherwise check next provider.

- After satisfying the needs of one user if he is left with enough resources, he will compete for next user otherwise will be removed from list.
- In this all users' requirements will be satisfied. Provider will initiate Virtual Machine to fulfil user's requirement.
- If after winning the auction he is unable to fulfil user's requirement, he will be penalised according to Eq. 6 and user's requirements will be fulfilled by next provider.
- When all the requirements of users are satisfied, auction process will be closed. Fig. 4.3 shows the sequence diagram of the execution flow of the PDARA algorithm.

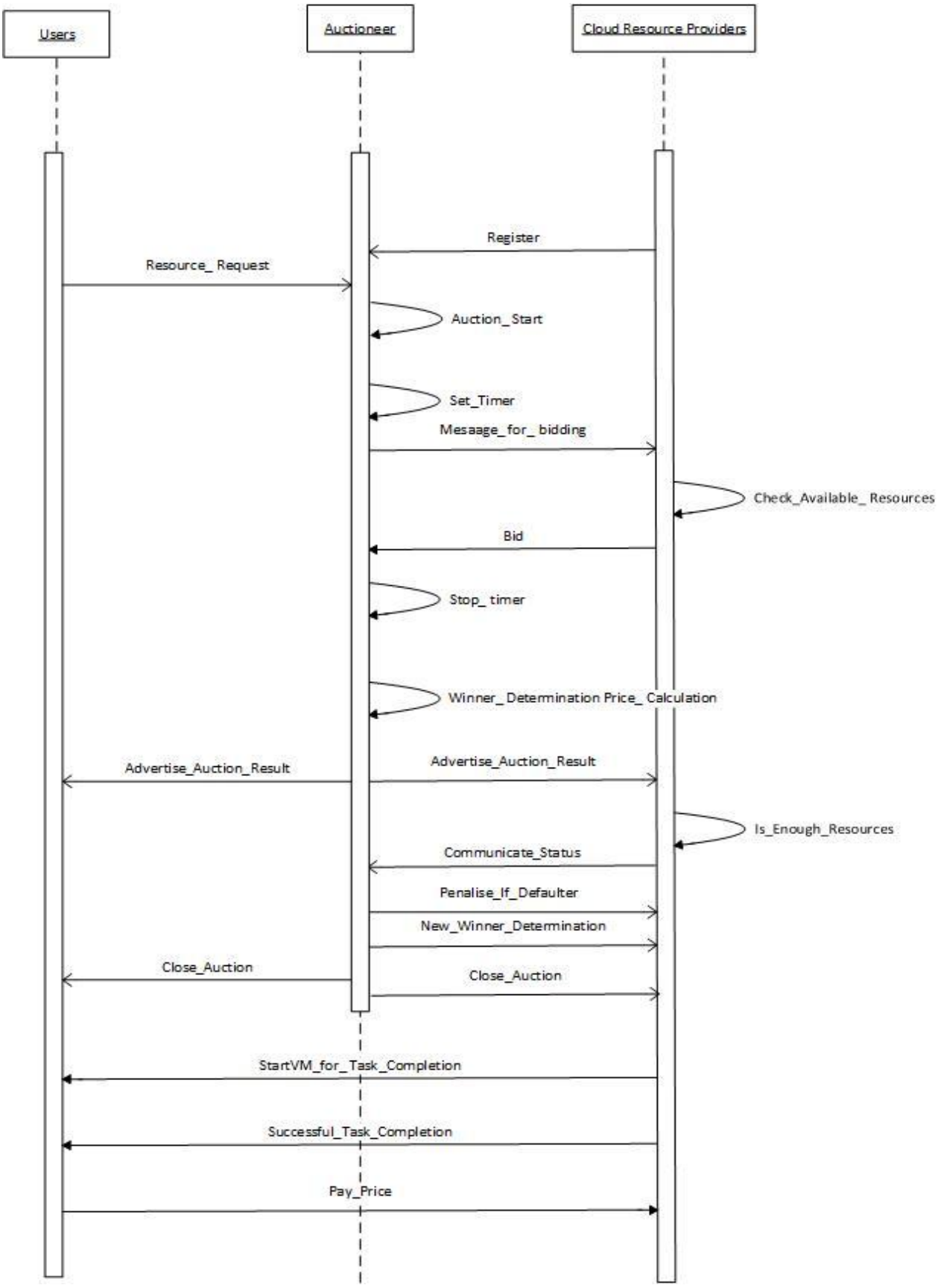
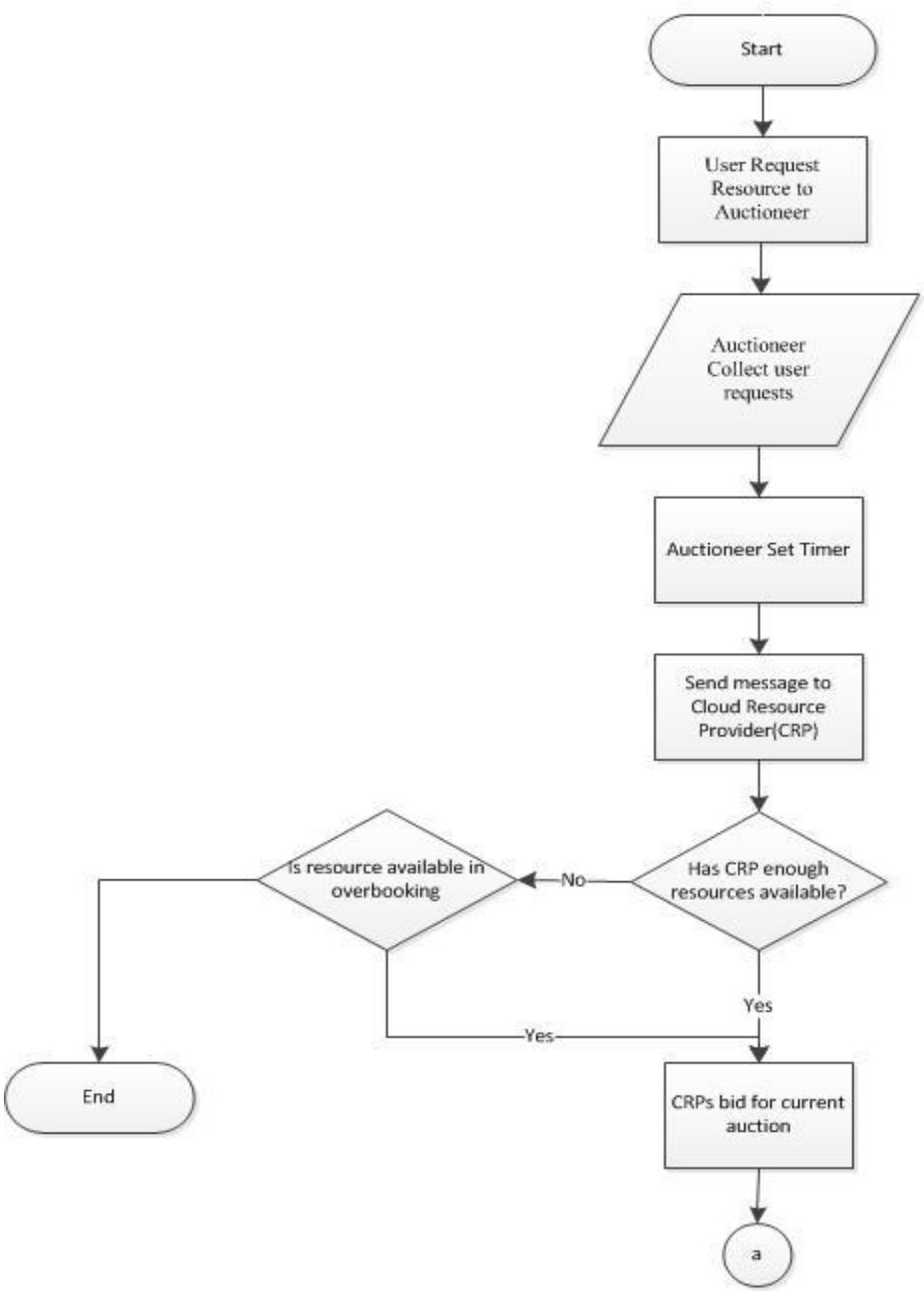
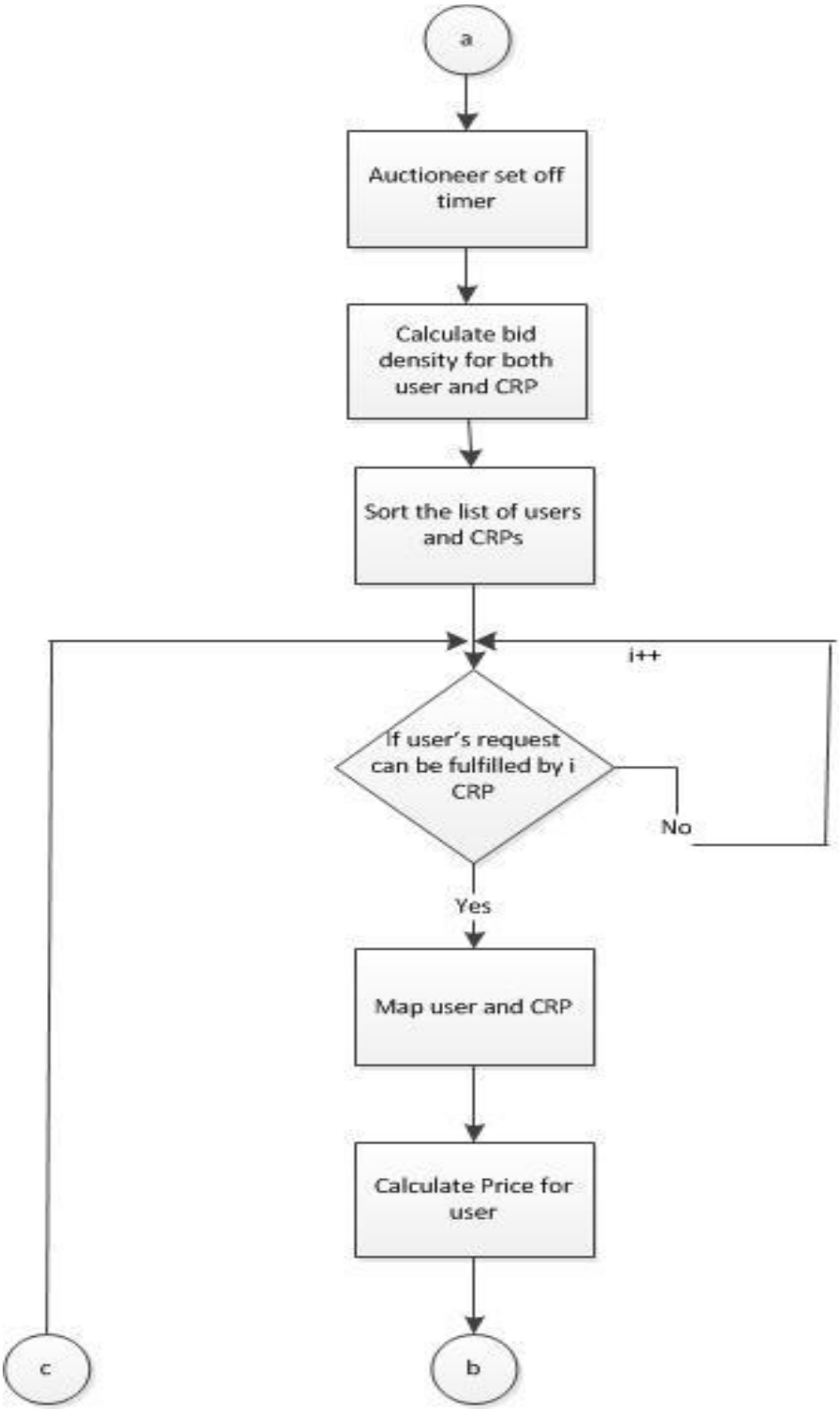


Figure 4.3 Sequence Diagram of the algorithm

4.1.8 Data Flow Diagram

The data flow diagram of the algorithm is shown in Figure 4.4. The algorithm starts with request of users for resources. Then auctioneer starts auction process. Cloud providers bid for resources. After winning the auction, they complete user task by starting VM. If provider is unable to complete the user task, even after winning the auction, he will be penalized.




a
Auctioneer set off timer
Calculate bid density for both user and CRP
Sort the list of users and CRPs
i++
If user's request can be fulfilled by i CRP
No
Yes
Map user and CRP
Calculate Price for user
c
b

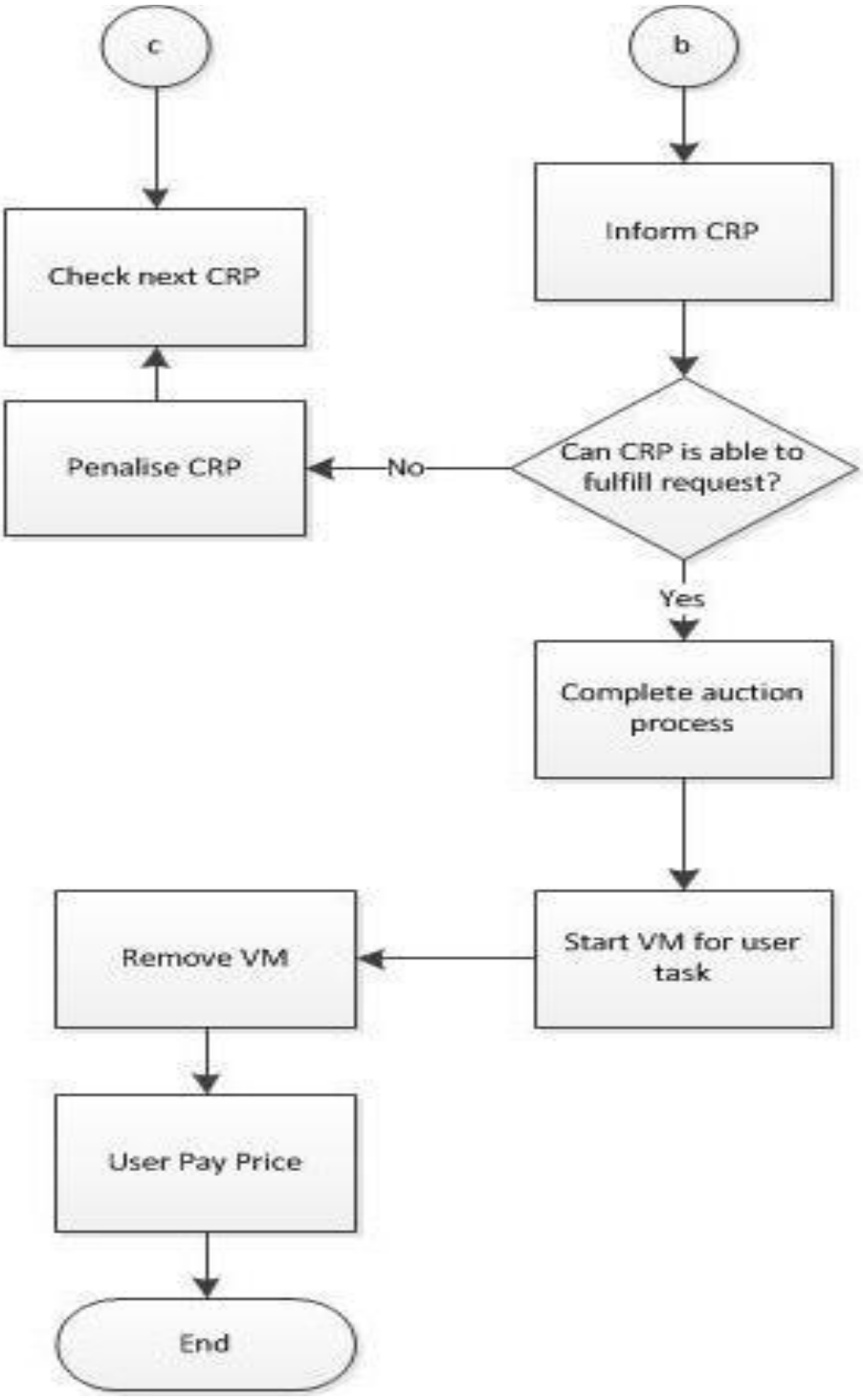


Figure 4.4 Data Flow Diagram of the Proposed Algorithm

4.2 Proposed Resource Allocation Algorithm

Procurement Double Auction Resource Allocation Algorithm (PDARA)

Input: RegisteredProvider_List

Output: Auction winners and price matrix

1. Set $i, j, temp \leftarrow 0$
2. Users request resources to auctioneer and inform their budget and time for which resources are required.
3. Auctioneer set timer and send message for bidding to Cloud resource providers.
4. Cloud resource providers will check their resources or check overbooking resources.
    a) If Enough resources available
        Begin
            Bid for current auction
            Advertise total available resources
        End
    Else
        Begin
            Do not bid for current auction
        End
5. Auctioneer sets off timer and close bidding process.
6. Find auction winners
    a) Calculate per resource bid value for user according to Eq. 1 and Eq. 4.
    b) Multiply time to per resource bid value for user according to Eq. 3.
    c) Sort the list of users in descending order and provider in ascending order and map users and providers.
    d) Set $n \leftarrow 1$
    e) for i = 1 to M
        for j $\leftarrow$ 1 to k
        if (R[m][j]) $< A[n][j]$)
            Begin

i++;

break;

End

End for

temp ← i;

break;

End for

f) Map n^{th} user with temp provider.

g) Decrement provider resources

m ← temp

for j ← 1 to k

R[m][j] = R[m][j] − A[n][j]

End for

h) Set n ← n + 1 and go to step e)

7. Calculate price according to Eq. 5.
8. Inform Cloud resource provider about auction results.
9. Cloud resource provider will start completing user task
    a) If resources not available

Begin

Inform auctioneer

Assign user task to next provider in the list and go to step 6(e)

Provider pay penalty according to Eq. 6

End

Else if

Begin

Start VM completion of task.

Complete user task and end VM.

User pay price.

End

10. Close auction process.

Chapter 5
Implementation and Experimental Results

This chapter presents various tools used for Cloud environment setup. The proposed algorithm has been implemented in the CloudSim simulator with the help of Netbeans IDE.

5.1 Tools for Setting up Cloud Environment

The applications of Cloud have various requirements for configuration and deployment. It is very difficult to check the performance of these applications on real Cloud. For this purpose simulation tools are used such as CloudSim which helps in evaluating different application without investing in the purchase of real time Cloud infrastructure. The simulation tools have many advantages like:

- To test the application there is no need buy infrastructure and applications can be tested repeatedly in controlled environment.
- Performance of the applications can be checked before actually deploying these applications on the Cloud.

Applications can be tested on simulation environment for different workloads or by using different resource leasing under provider's side which is difficult to study on real Cloud environment as it needs real services and resources for testing allocation algorithms. If there is no such simulator, then the only way to test the evaluation of algorithm is through theoretical study which does not presents real results. So these tools are helpful in saving money as it decreases the cost of experimental study of performance of the algorithms. Tools used for experimental setup are discussed as in next section.

5.1.1 CloudSim

CloudSim is a tool for simulation of Cloud environment [43]. It is an extension of GridSim as a layer that help in checking the performance of resource allocation algorithms by simulation of virtualized datacenters. It has many advantages which are as follows:

- It helps in simulation and modelling of Cloud Computing infrastructure which can be very large scale including datacenters on one physical machine.
- Platform is present in CloudSim for modelling allocation policies, datacenters.
- It is flexible enough to shift from time and space shared physical nodes to virtualized services.

5.1.2 Netbeans

Netbeans [58] is an software used for deployment of application in various languages. It is an multi language tool which can be used for languages like C/C++, HTML, Java, PHP etc. The Standard Document Kit (SDK) of Netbeans has tools for depeloping applications in Java. It provides plug-in support for other languages by extending the editor of Netbeans. There are various built in library classes for helping Java developers and they need to code strating. CloudSim Simulation toolkit is Java platform based so Netbeans SDK is helpful in creating simulated environment.

5.2 Implementation Details

The proposed auction based resource allocation algorithm has been implemented in CloudSim toolkit by extending the framework of toolkit. The starting of algorithm of auction process is depicted in Figure 5.1. Initially auction process asks for number of user participants and Cloud providers participants in the auction.

Figure 5.1 Main page of Auction Process

Four types of VM are created as to provide resources to users. Bundle of resources can be provided to user by initiating different type of VMs. The attributes are set according to CloudSim capability. Range of different parameters is set in CloudSim. Four parameters are considers i.e. processor speed, memory size, storage and bandwidth. The range of memory is [256, 512, 1024] MB, storage [2000, 10000] MB, bandwidth [120, 1000] b/s and processor range is [220, 1000]. In Figure 5.2 creation of VM is shown.

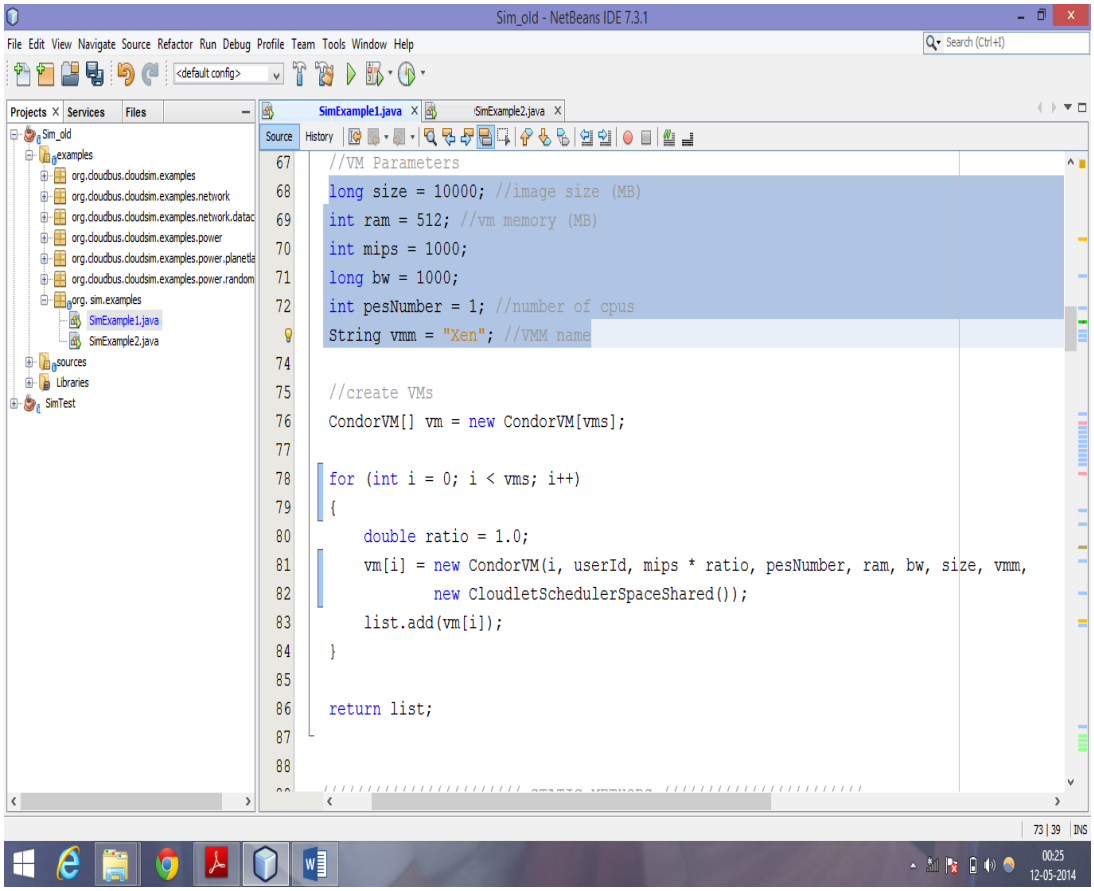


Figure 5.2 Creation of VM

5.2.1 Scenario

A scenario was created for testing of auction algorithm. In the scenario 25 users asked for resources from the auctioneer and 10 Cloud providers attended the auction means they bid for the current auction process. Figure 5.3 depicts the scenario for participants of auction process.

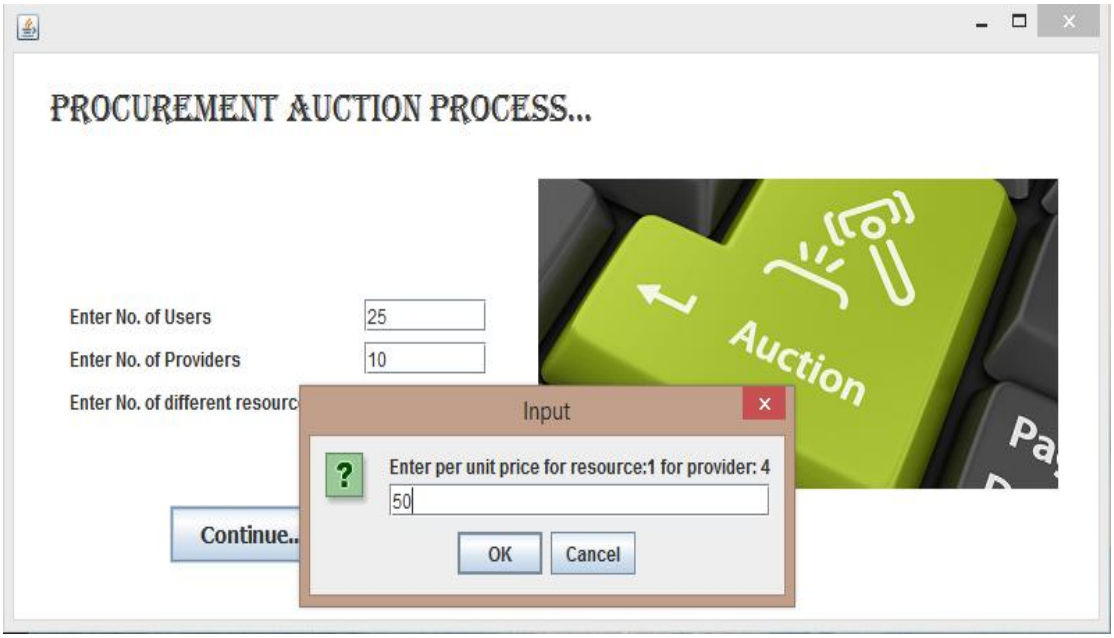


Figure 5.3 Auction Scenario

Different parameters of VM were setup so as user can ask resources in bundle form. Four parameters are considered as mentioned earlier. Range of resources has also been specified. Based on these VM parameters are set up as shown in Figure 5.4

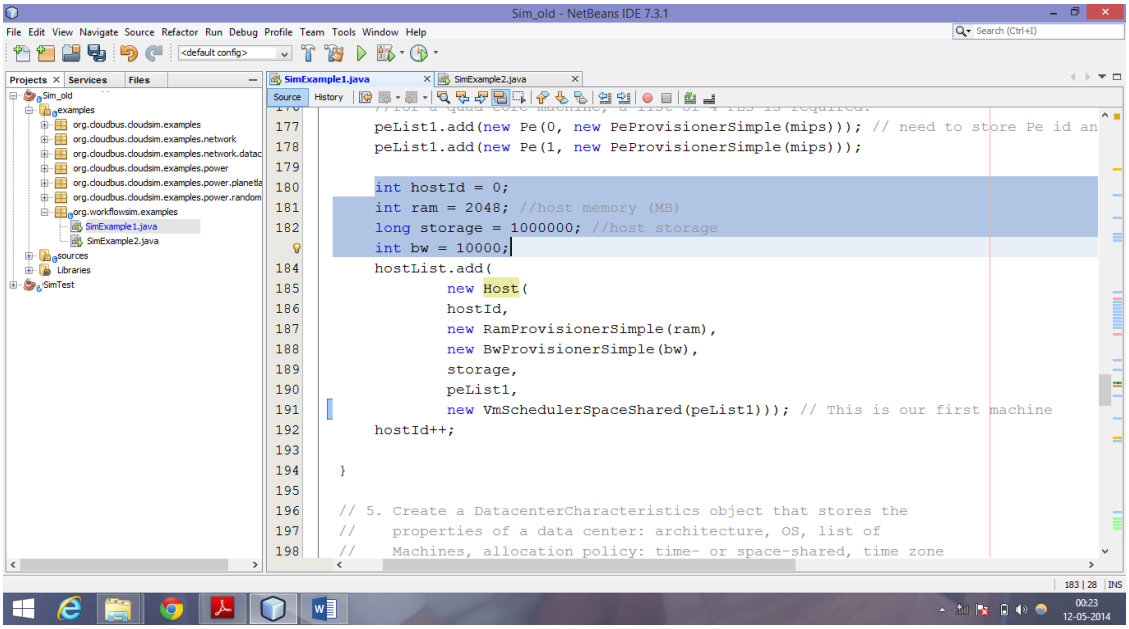


Figure 5.4 Setting VM parameters

After starting of auction process, bidding density for providers and users is calculated. Bid density is used to sort providers and users so that highest paying user and lowest bidder can be mapped.

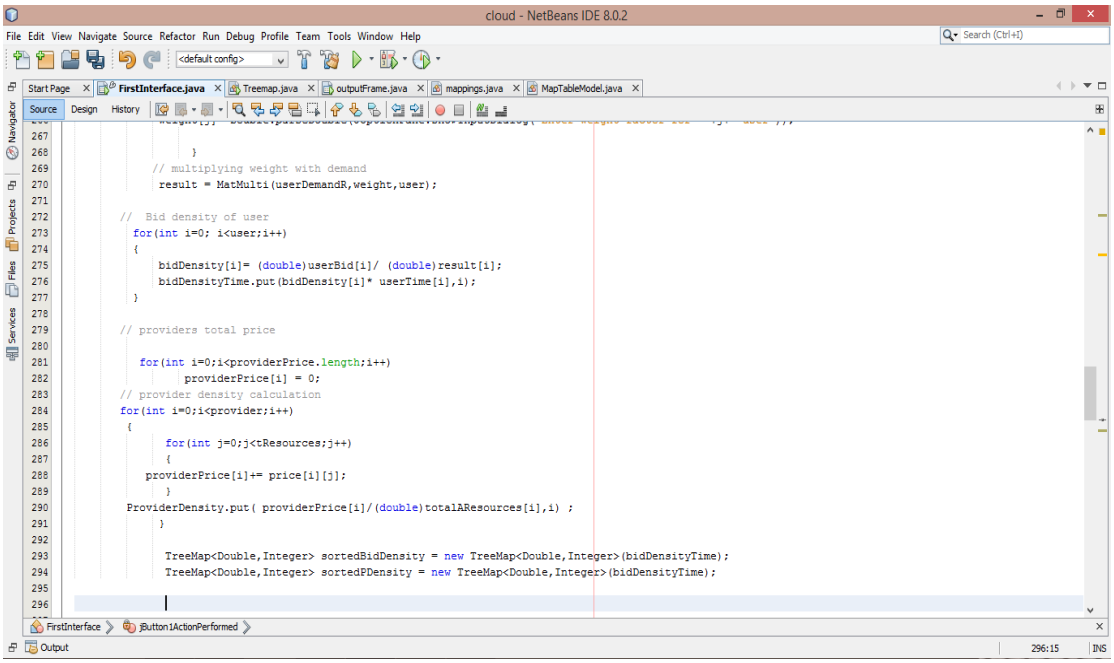


Figure 5.5 Bid density calculations

5.3 Experimental Results

Allocation of provider resources among 25 users is shown in Figure 5.6. Provider 1 is able to allocate resources among 4 users. Similarly others are able to allocate resources among multiple providers but provider 7 has failed in the auction so his resources are not allocated to any user. This can occur due to either very high bidding value or less number of resources available with him. The current auction process is successful because the auction has been able to satisfy users' requirements.

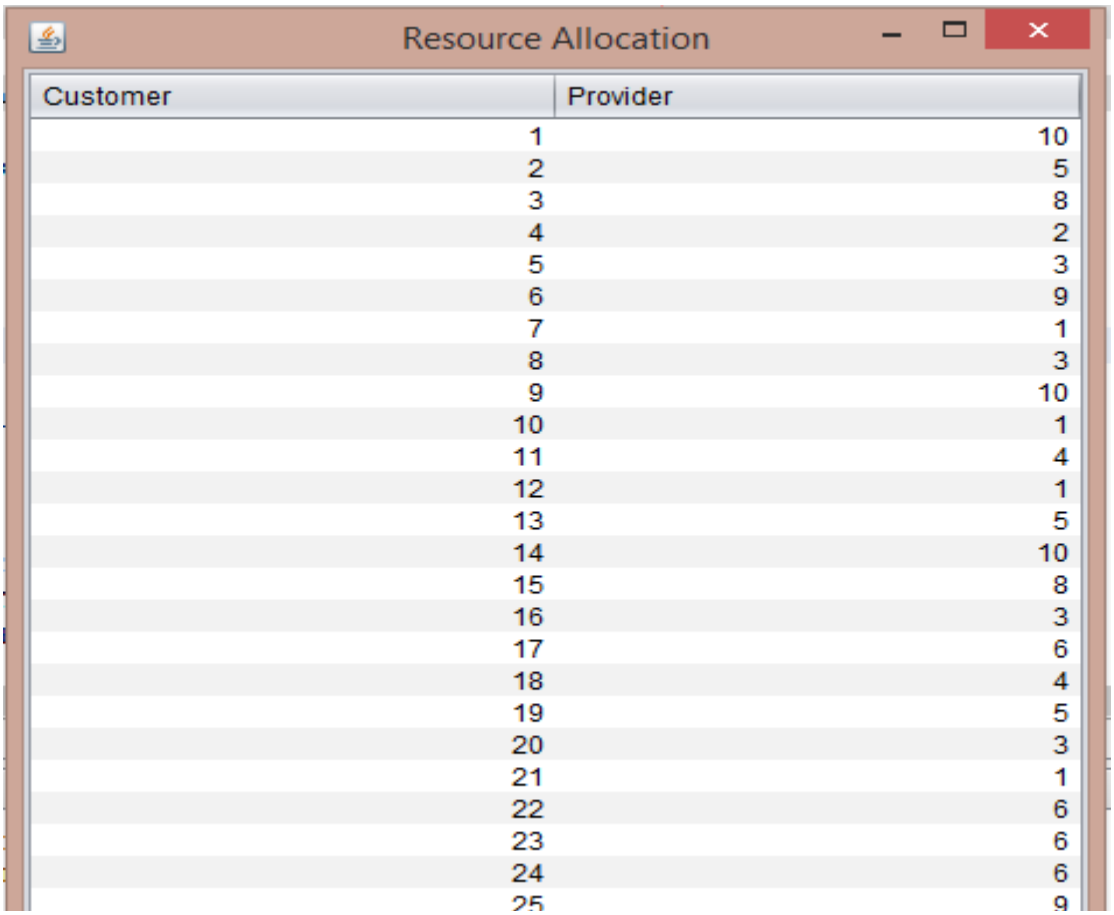


Figure 5.6 Allocation of Resources

The price for various scenarios is calculated by using both fixed price mode and dynamic price model using the proposed algorithm. And the results are depicted in the Figure 5.7. The graph shows that as compared to fixed price model, in most of the cases dynamic model provides benefit by lowering the total resource cost for the users. In few cases results of fixed price model are better. This graph shows that dynamic pricing model is beneficial as it is cost saving for user.

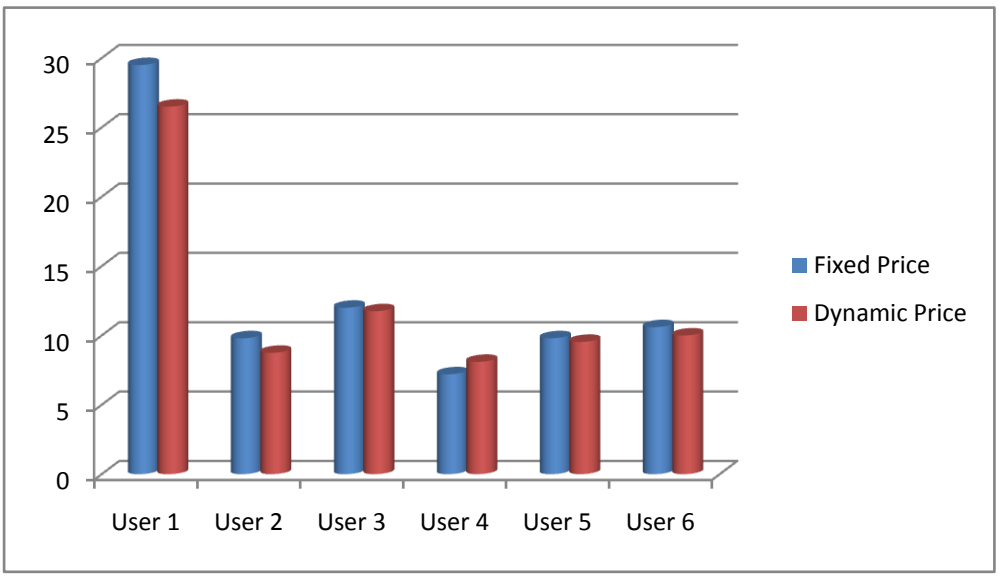


Figure 5.7 Comparison of Price Paid by User

Figure 5.8 depicts the price bid by 4 providers for each resource type and the trade price which is the price of actual selling of resource. Trade price is generally more than bidding value of the provider to due which they earn more revenue in dynamic price model.

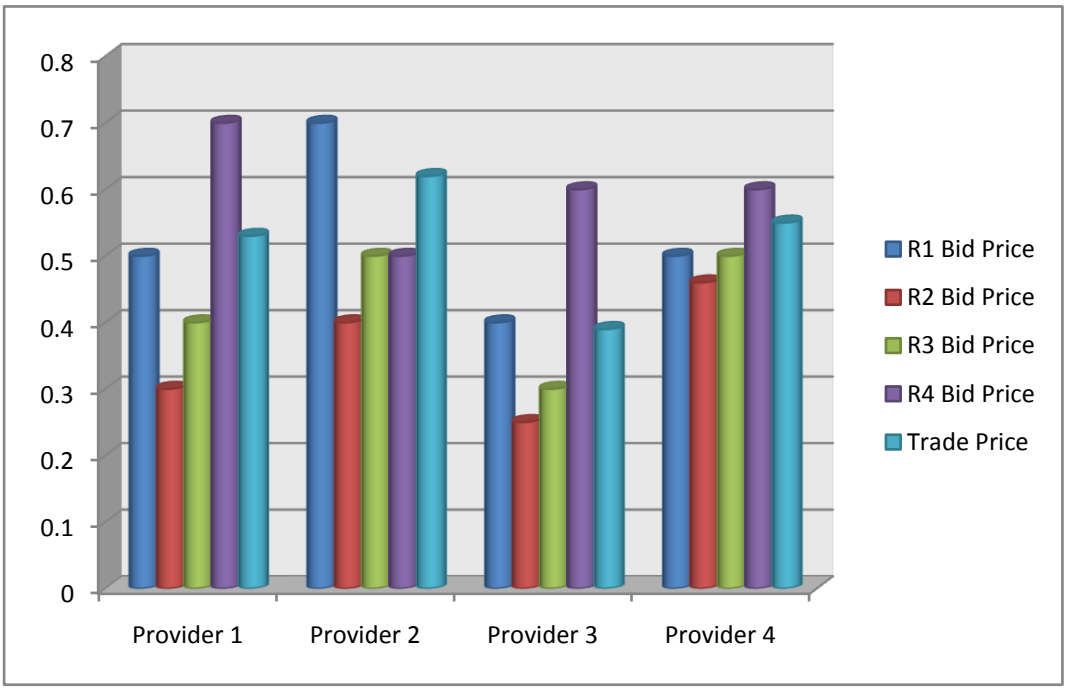


Figure 5.8 Comparisons of Bid Price and Trade Price

Chapter 6
Conclusion and Future Work

This chapter presents the conclusion of the work mentioned in this thesis and discusses the future work so as to extend this work.

6.1 Conclusion

Resource allocation is one of the most challenging issues in Cloud Computing. Cloud providers sell their resources through fixed price model by direct selling which are not much efficient. In the literature, many researchers have proposed approaches to solve the issue. But there exists many gaps in tradition algorithms. Even in unpredictable and dynamic demand scenario, Cloud providers use fixed price model or pay-as- you model which does not properly justifies the scenario. To solve resource allocation problem, auction based algorithms offer advantages for both user and Cloud resource provider. This algorithm helps to implement dynamic pricing model. Based on this economic model, a new auction based algorithm has been presented in this thesis. The algorithm efficiently allocates available resources of the Cloud provider and satisfies the requirements of users. The price has been calculated dynamically and the proposed algorithm is implemented in the CloudSim. The results of the proposed algorithm show that the algorithm is beneficial for Cloud providers as they earn more revenue and Cloud users for optimally satisfying their requirements and allocates resources of Cloud providers.

6.2 Thesis Contribution

a) In this thesis, previous resource allocation algorithms have been analysed and gaps in their work are discussed.
b) To solve the problem, a new Procurement Double Auction Resource Allocation (PDARA) has been proposed, which is based on economic auction model. This algorithm efficiently allocates resources of Cloud providers among the users who require resources to complete their task.
c) Dynamic pricing model is presented in the algorithm for pricing of resources.
d) The proposed algorithm is implemented in CloudSim simulator which proves the efficiency of the algorithm which increases revenue of the Cloud provider.

6.3 Future Scope

a) In the future business model for the auctioneer will be considered in order to provide benefit to all participants of auction based market model of resource allocation.

b) Discount based criteria for users demanding more resources or for longer time will be taken into consideration for providing benefits to users who want more resources and attract the customers for Cloud provider.

c) In the future experimental results of the algorithm can be compared with other existing auction based models.

References

[1] "Cloud Computing Evolution," [online] Available: www. computerweekly.com/feature/A-history-of-Cloud-computing.

[2] M. Creeger, "Cloud computing: an overview," *ACM Queue*, vol.7, no.5, pp. 2 2009.

[3] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems*, vol. 25, no. 6, pp. 599- 616, 2009.

[4] "Evolution of Cloud computing," [online] Available: www.tech.gaeatimes.com/index.php/archive/top-10-Cloud-computing serviceproviders-in-2010. [Feb. 20, 2014]

[5] I.Foster and C. Kesselman, "*The grid 2: blueprint for a future computing infrastructure*," Waltham: Morgan Kaufmann Publishers, 2004.

[6] M. A. Rappa, "The utility business model and the future of computing services," *IBM Systems Journal*, vol. 43, no. 1, pp. 32-42, 2004.

[7] V. R. Srinivasa, N. K. R. Nageswara and E. K. Kumari, "*Cloud* computing: an overview", *Journal of Theoretical and Applied Information Technology*, vol. 9, no. 1, pp. 71-76, 2009.

[8] M. Ahmed, A. Chowdhury and M. H. Rafee, "An advanced survey on cloud computing and state-of-the-art research issues", *International Journal of Computer Science Issues*, vol. 9, no. 1, pp. 201-2017, 2012.

[9] *Oracle JD Edwards Cloud Computing: choosing a deployment strategy that fits*, white paper, Oracle Corp., Oct. 2012.

[10] M. Turner, D. Budgen and P. Brereton, "*Turning software into a service*", Computer, vol. 36, no.10, pp. 38-44, 2003.

[11] D. C. Wyld, *"Moving to the cloud: An introduction to cloud computing in government,", IBM Center for the Business of Government*, 2009.

[12] "Cloud Computing Technology- the study of its model," [online] Available: www.websiteglobe.blogspot.in/2012/05/Cloud-computing-technologystudyof-its.html. [Mar. 5, 2014].

[13] K. Jeffery and B. Neidecker-Lutz, "*The future of cloud computing: opportunities for european cloud computing beyond 2010*," Expert Group Report, European Commission, Jan. 2010.

[14] "Cloud Computing: Acronyms (IaaS,PaaS,SaaS)," [online] Available: www. mahameeditpro.blogspot.in/2012/03/saas-software-as-service-essentially.html. [Mar. 5, 2014].

[15] Q. Zhang, L. Cheng and R. Boutaba, "Cloud computing: state-of-the-art andresearch challenges," *Journal of internet services and applications*, vol. 1, no. 1, pp. 7-18, 2010.

[16] R. P. Padhy, M. R.Patra and S. C. Satapathy, "Cloud computing: security issues and research challenges", *International Journal of Computer Science and Information Technology & Security*, vol. 1, no. 2, Dec. 2011.

[17] M. Nazir, "Cloud computing: overview & current research challenges, "*Journal of Computer Engineering*, IOSR-JCE, vol. 8, issue 1, pp. 14-22, Dec. 2012.

[18] M. A. Vouk, "Cloud computing–issues, research and implementations, *"Journal of Computing and Information Technology*, CIT, vol. 16, no. 4, pp. 235-246, 2008.

[19] V. K. Reddy, B. T. Rao and L. S. S. Reddy, "Research issues in cloud computing," *Global Journal of Computer Science and* Technology, vol. 11, no. 11, 2011.

[20] T. Dillon, C. Wu and E. Chang, "Cloud computing: issues and challenges," In the Proceedings of *24th IEEE International Conference on Advanced Information Networking and Applications*, AINA, pp. 27-33, Apr. 2010.

[21] C. Rong, S. T. Nguyen and M. G. Jaatun, "Beyond lightning: A survey on security challenges in cloud computing," *Computers and Electrical Engineering*, vol. 39, no. 1, pp. 47-54, 2013.

[22] R. Buyya and R. Ranjan, "Special section: Federated resource management in grid and cloud computing systems," *Future Generation Computer Systems*, vol. 26, no. 8, pp. 1189–1191, June, 2010

[23] V. Vinothina, R. Sridaran, and P. Ganapathi, "*A survey on resource allocate strategies in cloud computing*," International Journal of Advanced Computer Science and Applications, vol. 3, no. 6, pp. 97–104, 2012.

[24] G. Aceto, A. Botta, W. De. Donato, and A. Pescapè, "Cloud monitoring: A survey," *Computer Networks*, vol. 57, no. 9, pp. 2093– 2115, 2013.

[25] C. Castillo, G. N. Rouskas, K. Harfoush, "On the Design of Online Scheduling Algorithms for Advance Reservations and QoS in Grids," IEEE Transactions on parallel and distributed systems, pp. 1-10, 2007.

[26] R. Basmadjian, H. D. Meer, R. Lent, and G. Giuliani, "Cloud computing and its interest in saving energy: the use case of a private cloud," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 1, no. 1, pp. 5, 2012.

[27] M. Stillwell, D. Schanzenbach, F. Vivien, and H. Casanova, "Resource allocation using virtual clusters," In *Proceedings of 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, (Section III), pp. 260–267,2009.

[28] Y. Hu, J. Wong, G. Iszlai, and M. Litoiu, "Resource provisioning for cloud computing," In *Proceedings of the 2009 Conference of the Center for Advanced Studies on Collaborative Research*, pp. 101–111,2009.

[29] Types of Scheduling-Go4Expert. [Online]. Available: http://www.go4expert.com/articles/types-of-scheduling-t22307/

[30] I.A. Moschakis, and H. D. Karatza, "Evaluation of gang scheduling performance and cost in a cloud computing system," *The Journal of Supercomputing*, vol. 59, no. 2, pp. 975–992, 2010.

[31] Y. C. Lee and A. Y. Zomaya, "Energy conscious scheduling for distributed computing systems under different operating conditions," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 8, pp. 1374-1381, 2011.

[32] T. V. Do and C. Rotter, "Comparison of scheduling schemes for on demand IaaS requests," *Journal of Systems and Software*, vol. 85, no. 6, pp. 1400–1408, 2012.

[33] G. Lee, N. Tolia, P. Ranganathan, and R. H. Katz, "Topology aware resource allocation for data-intensive workloads*", In Proceedings of the first ACM asia-pacific workshop on Workshop on systems",* pp.1-6, Aug.2010

[34] D. Warneke and O. Kao, "Exploiting dynamic resource allocation for efficient parallel data processing in the cloud", *IEEE Transactions on Parallel and Distributed Systems*, vol.22, no.6, pp.985-997, 2011.

[35] Espadas, Javier, A. Molina, G. Jimenez, M. Molina, R. Ramirez and D. Concha "A tenant-based resource allocation model for scaling Software-as-a-Service applications over cloud computing infrastructures."*Future Generation Computer Systems, vol.* 29, no. 1, pp. 273-286, January 2013.

[36] W. Linlin, S. Garg, and R. Buyya. "SLA-based resource allocation for software as a service provider (SaaS) in cloud computing environments." *In Proceedings of 2011 11th International Symposium on Cluster, Cloud and Grid Computing,* pp. 195-204. IEEE, 2011

[37] R. Buyya, Rajkumar, D. Abramson, and J. Giddy. "Nimrod/G: architecture for a resource management and scheduling system in a global computational grid." In *Proceedings of The 4th International Conference on High Performance Computing*, vol. 1, pp. 283-289, 2000.

[38] G.M. Heal. "Planning without prices. Review of Economic Studies", vol. 36 no. 107, pp. 347-62, 1969.

[39] M. Shubik. "Game Theory in Social Science." The MITPress, Cambridge, Mass,1983

[40] S. Zaman and D. Grosu, "Combinatorial auction-based allocation of virtual machine instances in clouds", *In Proceedings of Second IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, vol. 73,no. 4, pp. 127–134,2010.

[41] X. Wang, J. Sun, M. Huang and C. Wu, "A resource auction based allocation mechanism in the cloud computing environment", In Proceedings of *26th International Parallel and Distributed Processing Symposium*, pp. 2111–2115, 2012.

[42] T. Y. Meng and M. Mihailescu, "A strategy-proof pricing scheme for multiple resource type allocations", *In Proceedings of International Conference on Parallel Processing (ICPP)*, pp. 172–179, 2009.

[43] R.N. Calheiros, R. Ranjan and A. Beloglazov, C.A.F. D. Rose, R. Buyya, "CloudSim a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms", Software: Practice Experience vol. 41, no. 1, pp. 23–50,2011.

[44] Vanmechelen, Kurt, W. Depoorter and J. Broeckhove, "Economic Grid Resource Management for CPU Bound Applications with Hard Deadlines", *In*

Proceedings of 8th IEEE International Symposium on Cluster Computing and the Grid, pp. 258-26, 2008

[45] Kertesz A. and Kacsuk P., "Grid Interoperability Solutions in Grid Resource Management", *IEEE Transaction on Systems Journal*, vol. 3, no 1, pp. 131-141,2009.

[46] B. Pourebrahimi , K. Bertels, G.M. Kandru and S. Vassiliadis,"Market-based Resource Allocation in Grids", *In Proceedings of the Second IEEE International Conference on e-Science and Grid Computing*, 2006.

[47] M. Feldman, K. Lai and L. Zhang,"A price-anticipating resource allocation mechanism for distributed shared clusters", *In Proceedings of the 6th ACM conference on Electronic commerce Electronic Commerce*, pp. 127-136, 2005.

[48] J. Gomoluch and M. Schroeder, "Market-Based Resource Allocation for Grid Computing: A Model and Simulation", *In Proceedings of the 1st International Workshop on Middleware for Grid Computing*, pp. 211-218, 2003.

[49] D.A Menasce. and E. Casalicchio, "A framework for resource allocation in grid computing", *In Proceedings of the 12th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems*, pp. 259-267,2004.

[50] R. Wolski, S.P. James, J. Brevik and T. Bryan, "Analyzing Market-based Resource Allocation Strategies for the Computational Grid", *The International Journal of High Performance Computing Applications*, vol. 15, no. 3, pp. 258-281,2001.

[51] L. Li, Y. A. Liu, K. M. Liu, X. L. Ma and M. Yang, "Pricing in combinatorial double auction based grid allocation model", *The Journals of China University. Posts and Telecommunication*, vol. 16, no. 3, pp. 59–65, 2009.

[52] Harmon, Robert, H. Demirkan, B. Hefley, and N. Auseklis. "Pricing strategies for information technology services: A value-based approach." *In Proceedings of 42nd Hawaii International Conference on System Sciences*, pp. 1-10, 2009.

[53] Weinhardt, C. Anandasivam, D. I. W. A. Blau, B. Borissov, D. I. N. Meinl, D. M. T. Michalk, D. I. W. W. Michalk & J. Stößer, "Cloud computing–a classification, business models, and research directions". *Business & Information Systems Engineering*, vol.1, no. 5, pp. 391-399, 2009.

[54] Paleologo and A. Giuseppe, "Price-at-Risk: A methodology for pricing utility computing services." *IBM Systems Journal, vol.* 43, no.1, pp. 20-31, 2004.

[55] Sainio, L. Maija, and E. Marjakoski. "The logic of revenue logic? Strategic and operational levels of pricing in the context of software business.", *Technovation, vol.* 29, no. 5, pp. 368-378,2009.

[56] A. Osterwalde, Y. Pigneur, and C. L. Tucci. "Clarifying business models: Origins, present, and future of the concept." *Communications of the association for Information Systems,* vol. 16, no. 1 pp. 1, 2005.

[57] Warneke, Daniel, and O. Kao, "Exploiting dynamic resource allocation for efficient parallel data processing in the cloud." *Parallel and Distributed Systems, IEEE Transactions on,* vol. 22 no., pp. 985-997,2011.

[58] "Netbeans," [online] Available: https://netbeans.org/features/index.html, [Accessed: 7th May, 2014].

List of Publications

[1] R Goyal and A. Bala, "*Auction based Resource Allocation Strategy for Infrastructure as a Service*," 1st International Conference on Next Generation Computing Technologies, UPES, Dehradun, India, September 4-5, 2015 **[Communicated].**

[2] R Goyal and A. Bala, "*Procurement Double Auction for Resource Allocation in Cloud*" 3rd International Conference on International Conference on Information System Design and Intelligent Applications, ANITS, Visakhapatnam, India, January 8-9, 2016 **[Communicated].**